



Evidence-Grounded Clinical Pharmacogenomics Question Answering System Using Large Language Models and Hybrid Retrieval Augmentation

by

Protiva Arafin

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

in the
Faculty of Science and Environmental Studies
Department of Computer Science

April 2026

Declaration of Authorship

I, PROTIVA ARAFIN, declare that this thesis titled, ‘Evidence-Grounded Clinical Pharmacogenomics Question Answering System Using Large Language Models and Retrieval Augmentation’ and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master’s degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- I would like to acknowledge the use of ChatGPT-4, Grammarly, and QuillBot in the preparation of this thesis. These tools were utilized specifically for editing, proofreading, and improving the clarity of the writing.

Signed:

Date:

“Happiness can be found, even in the darkest of times, if one only remembers to turn on the light.”

– Albus Dumbledore

Abstract

Pharmacogenomics (PGx) is very important for personalized medicine since it helps doctors choose the right drugs and doses based on a person’s genetic makeup. But the growing amount and complexity of PGx data, as well as the requirement to understand clinical recommendations, make it harder to make good decisions. This study puts forward a data-driven clinical decision support framework that combines large language models (LLMs) with hybrid retrieval-augmented generation (RAG) to enhance the response to pharmacogenomic questions.

The framework assesses two contemporary LLMs, Meta-LLaMA-3.1-8B-Instruct and Qwen3-8B, through various configurations, encompassing base models, Low-Rank Adaptation (LoRA) fine-tuning, and hybrid RAG-based methodologies. The structured pharmacogenomics data from CPIC and the clinical guideline information from ClinPGx are combined to make a huge dataset. To make it easier to find and use in models, the data goes through procedures including merging, cleaning, normalizing, and converting to JSONL format. A hybrid retrieval approach is aimed to enhance factual grounding by integrating lexical filtering with semantic similarity through sentence embeddings. This research use both automatic metrics and manual checks to rate the models on their correctness, relevance, completeness, and clarity. The results reveal that Qwen works well as a basic model, and that LLaMA gets much better when it is used with RAG and LoRA, giving answers that are more aware of the context and therapeutically useful. Fine-tuning alone doesn’t always work, which shows how limited it is to only use parametric data. The results show that accuracy in clinical settings needs to be backed up by consistency, relevance, and evidence.

This study demonstrates that employing retrieval methods alongside parameter-efficient fine-tuning enhances the reliability and utility of LLM-based systems in clinical environments. The proposed methodology establishes a scalable framework for the development of trustworthy AI-driven solutions in pharmacogenomics and healthcare decision support.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Abedalrhman Alkhatieb, and my co-supervisor Dr. Md Moniruzzaman for their invaluable guidance, encouragement, and support throughout my research journey. Their insightful advice and expertise have played a crucial role in shaping this work.

I am also deeply grateful to my thesis committee member Dr. Malek Alsmadi and Dr. Saad B. Ahmed for their valuable feedback and constructive criticism, which helped refine my research.

A special thanks goes out to my colleagues and friends at Lakehead University, whose discussions and support made this journey more enriching. I am thankful to everyone who provided technical assistance and helpful insights during critical stages of this work.

I would like to extend my heartfelt appreciation to my family for their unconditional love, patience, and encouragement. Their unwavering belief in my abilities has been my greatest motivation.

Finally, I acknowledge the support and facilities provided by Lakehead University, which made this research possible.

Contents

Declaration of Authorship	i
Acknowledgements	iv
List of Figures	ix
List of Tables	x
Abbreviations	xi
1 Introduction	1
1.1 Overview of PGx	1
1.1.1 Gene	2
1.1.2 Genetic Variations	2
1.1.3 Drug–Gene Interactions in Pharmacogenomics	2
1.1.4 Pharmacogenomic Data Sources	3
1.2 Clinical Decision Support Systems (CDSS) in Pharmacogenomics	4
1.2.1 Limitations of Clinical Decision Support Systems (CDSS) in Pharmacogenomics	5
1.2.2 Artificial Intelligence for Clinical Decision Support in Pharmacogenomics	5
1.3 Problem Statement	6
1.4 Research Objectives	7
1.5 Key Contributions of This Study	7
1.6 Structure of the Thesis	8
2 Preliminaries and Background	9
2.1 Evolution of Artificial Intelligence	9
2.1.1 Transition to Machine Learning	9
2.1.2 Deep Learning in Modern AI Systems	10
2.1.3 Natural Language Processing (NLP)	10
2.1.4 Emergence of Large Language Models from NLP	11
2.2 Large Language Models and Their Variants	12
2.2.1 Large Language Model Meta AI (LLaMA)	12
2.2.2 Qwen	13

2.3	Foundation of Modern Large Language Models	13
2.3.1	Transformer Architecture	14
2.3.2	Working Mechanism of LLMs	15
2.3.3	Large Language Models in Healthcare	17
2.3.4	Limitations of Large Language Models (LLMs)	18
2.4	Retrieval-Augmented Generation (RAG)	19
3	Related Work	21
3.1	Large Language Models in Pharmacogenomic Research	21
3.2	Retrieval-Augmented Generation in Biomedical Applications	22
3.3	Hierarchical Retrieval-Augmented Methods	24
3.4	Limitations of Pharmacogenomics in Clinical Practice	24
3.5	Comparison with Existing Work	25
4	Methodology	26
4.1	Overview of Data Sources	26
4.1.1	CPIC Data Structure	27
4.1.2	ClinPGx Data Structure	28
4.2	Large Language Model Meta AI (LLaMA) Architecture	30
4.2.1	LLaMA Fine-Tuning Configuration	30
4.2.2	LLaMA Instruction Formatting and Training Pipeline	31
4.3	Qwen Model Architecture	32
4.3.1	Qwen Fine-Tuning Configuration	32
4.3.2	Qwen Instruction Formatting and Training Strategy	32
4.4	Retrieval-Augmented Generation (RAG) Framework	33
4.4.1	Hybrid Retrieval Mechanism	34
4.5	Proposed System Framework and Workflow	37
5	Empirical Analysis	38
5.1	SHARCNET HPC Usage	38
5.2	Model Variants	38
5.3	Hybrid Retrieval Mechanism	39
5.4	Model Variants Evaluated	39
5.5	Evaluation Metrics	39
5.5.1	Classification Evaluation	40
5.5.2	Text-Based Evaluation	41
5.5.3	Manual Evaluation	42
5.6	Summary	43
6	Experiments and Results	44
6.1	Experimental Setup	44
6.1.1	Choice of Embedding Model for Semantic Re-Ranking	45
6.2	Classification Performance Evaluation	46
6.2.1	LLaMA Based Models Evaluation	46
6.2.2	Qwen Based Models Evaluation	48
6.2.3	Comparative Analysis (LLaMA vs Qwen)	50
6.3	Text Based Evaluation	51
6.3.1	Evaluation Analysis of LLaMA Models Performance	51

6.3.2	Performance Analysis of Qwen-Based Models	52
6.3.3	Performance Comparison Between LLaMA and Qwen	53
6.4	Manual Evaluation	53
6.4.1	Evaluation Metrics for Manual Evaluation	61
7	Discussion	65
7.1	Biomedical Insights from Model Responses	65
7.1.1	Gene–Drug Risk Associations	65
7.1.2	Genotype–Guided Dose Adjustment	65
7.1.3	Clinically Actionable Recommendations	66
7.1.4	Integrated Clinical Reasoning	66
7.1.5	Variability and Evidence Sensitivity	66
7.1.6	Handling of Missing Guidelines	66
7.2	Overview of Key Findings	67
7.3	Interpretation of Experimental Outcomes	67
7.3.1	Strong Baseline Performance of the Qwen Model	67
7.3.2	Enhanced LLaMA Performance through RAG and LoRA	68
7.3.3	Mixed Impact of Fine-Tuning without Retrieval Support	68
7.3.4	Difference Between Automatic and Manual Evaluation Results	68
7.4	Limitations	69
7.4.1	Class Imbalance in Evaluation Data:	69
7.4.2	Lack of Real Clinical Validation:	69
7.4.3	Limited Scope of Manual Evaluation:	69
7.4.4	Differences Between Automatic and Manual Evaluation:	70
7.4.5	Sensitivity to Retrieved Evidence (RAG):	70
7.5	Future Work	70
7.5.1	Contradiction Detection in Retrieved Evidence:	70
7.5.2	Personalized Pharmacogenomics Query Handling:	71
7.5.3	Evaluation with Real Clinical Scenarios (Case-Based Testing)	71
7.5.4	Interactive Question Refinement System:	71
8	Conclusion	72
A	Code Snippets	74
A.1	Large-Scale Data Integration Using PySpark	74
A.1.1	Handling Column Conflicts	74
A.1.2	Scalable Data Processing via Hash-Based Partitioning	75
A.2	Link Collection from ClinPGx	75
A.2.1	Web Scraping using Playwright	75
A.2.2	Chunking for RAG	76
A.3	Fine-Tuning with LoRA (Qwen & LLaMA)	76
A.3.1	Qwen LoRA Fine-Tuning	76
A.3.2	Answer-Focused Training (Masked Loss)	77
A.3.3	LLaMA LoRA Fine-Tuning	77
A.4	RAG Dataset Construction (PGx Knowledge Preparation)	78
A.5	Hybrid RAG-Based Question Answering System	78

A.6 LLaMA RAG + LoRA Evaluation Pipeline	79
A.7 Qwen RAG + LoRA Evaluation Pipeline	79
A.8 GitHub Link	80
B Software and Packages Used	81
C System Configuration	83
Bibliography	85

List of Figures

2.1	Hierarchical relationship between Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), and Large Language Models (LLMs).	11
2.2	A schematic representation of the Transformer architecture demonstrating how attention techniques are implemented for encoder-based context encoding and decoder-based autoregressive token prediction.	15
2.3	Illustration of pre-training, fine-tuning, and inference process in large language models using a pharmacogenomic example.	17
2.4	Conceptual illustration of the Retrieval-Augmented Generation (RAG) architecture.	20
4.1	Overview of the CPIC relational data model	27
4.2	An outline of the Meta-Llama-3.1-8B-Instruct-based fine-tuning architecture.	31
4.3	Summary of the Qwen3-8B fine-tuning architecture, demonstrating the conversion of pharmacogenomics data into instruction-style inputs and its modification via LoRA for domain-specific clinical response creation.	33
4.4	Hybrid Retrieval-Augmented Generation (RAG) Architecture for Pharmacogenomic Decision Support.	36
4.5	Overview of the proposed methodology integrating CPIC and ClinPGx data preprocessing with a Retrieval-Augmented Generation (RAG) framework and a LoRA fine-tuned LLM for clinical recommendation generation.	37

List of Tables

3.1	Comparison with Existing Work	25
4.1	Summary of Datasets Used in This Study	29
6.1	Overall Performance Comparison of LLaMA Models	46
6.2	Class-wise Performance Comparison	46
6.3	Overall Performance Comparison of Qwen Models	48
6.4	Class-wise Performance Comparison of Qwen-Based Models	49
6.5	Text Based Evaluation Metrics Comparison	51
6.6	Text-Based Evaluation Metrics Comparison	52
6.7	Comparison of Qwen and LLaMA Model Responses for Pharmacogenomic Queries	53
6.8	Evaluation Metrics for Qwen and LLaMA Models	61
6.9	Final Manual Evaluation Scores	64
B.1	Software and Packages Used in the Study	81
C.1	Dataset and Preprocessing Configuration	83
C.2	Model and LoRA Configuration	84
C.3	Training Configuration	84
C.4	Hybrid Retrieval Mechanism Configuration	84
C.5	Model Variants Evaluated	84

Abbreviations

PGx	Pharmacogenomics
LLM	Large Language Model
LLaMA	Large Language Model Meta AI
RAG	Retrieval Augmented Generation
CPIC	Clinical Pharmacogenetics Implementation Consortium
DPWG	Dutch Pharmacogenetics Working Group
PharmGKB	Pharmacogenomics Knowledge Base
CPNDS	Canadian Pharmacogenomics Network for Drug Safety
PharmVar	Pharmacogene Variation Consortium
CDSS	Clinical Decision Support Systems
ClinPGx	Clinical Pharmacogenomics
JSONL	JavaScript Object Notation Lines
HPC	High Performance Computing
GPU	Graphics Processing Unit
CPU	Central Processing Unit
LoRA	Low-Rank Adaptation
PEFT	Parameter-Efficient Fine-Tuning
MiniLM	Miniature Language Model
BERT	Bidirectional Encoder Representations from Transformers
BioBERT	Biomedical BERT
PubMedBERT	PubMed-trained BERT
NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence

I dedicate my work to my parents and my husband, who have supported me through thick and thin, my biggest cheerleaders in my life. I am grateful to them for how they kept me strong all the time. This research work is a reflection of my patience and hard work that my husband and parents have instilled in me.

Chapter 1

Introduction

Pharmacogenomics (PGx) is the study of how genetic diversity affects an individual's pharmacological reaction, which differs from person to person [1]. A survey revealed that among the top 10 bestselling medications in the United States, 3–24 patients did not respond to treatment. In traditional clinical studies, researchers gather data from thousands of people, but they typically only collect a small amount of information from each participant. However, precision medicine requires alternative approaches to testing interventions. Researchers must therefore investigate the various factors that influence an individual's response to treatment, including genetic variation [2].

1.1 Overview of PGx

PGx uses genomic technology that encourages the discovery and development of new medications, as well as the optimization of drug dosing and selection for specific patients to improve effectiveness and reduce toxicity [3]. Genotyping technologies—tools that can spot even tiny differences in our genes—are seen by many as an important first step in making personalized medicine possible for more people. Countries around the world are focusing on these advances, which are quickly becoming more affordable and easier to access. This means they're now a real option for everyday medical care.

Many patients do not fully benefit from the initially prescribed drug therapy. It has been observed that a standard effective dose for one patient may result in therapeutic failure, producing little or no effect, while the same dose in another individual may lead to severe adverse effects [1].

1.1.1 Gene

A gene is a segment of the genome controlled by several promoters and distant regulatory elements. It holds the information to make proteins or non-coding RNAs, and these products are linked to the genetic information in their final forms—proteins or RNAs [4].

1.1.2 Genetic Variations

The study of human genetic variants has progressed significantly in the last five years [5]. The difference between any two human genomes has been estimated to be less than 0.1% overall, but this still means there are at least several million nucleotide differences per individual [6]. Human genetic variants are differences in DNA sequences that occur in individual genomes. Variations in the human genome can occur in numerous ways, such as single-nucleotide alterations or substitutions, or additions or deletions that affect the number of copies of a larger DNA sequence. These genetic variants happen naturally in the human genome. They are the marks of faults or errors that occur when DNA is copied during cell division, even though they are caused by external factors [5].

1.1.3 Drug–Gene Interactions in Pharmacogenomics

A gene-drug interaction (GDI) happens when variations in a person’s genes change how the body metabolizes a drug, potentially affecting how well the drug works or how quickly the body clears it [7]. Genetic variations that induce proteins with different activity are one of the main reasons why people respond to drugs differently [8].

A study reported that members of the CYP3A family were responsible for the metabolism of 37% of the top 200 medications prescribed in the United States in 2008. CYP2C9, CYP2D6, CYP2C19, CYP1A2, and CYP2C8 accounted for 17%, 15%, 10%, 9%, and 6%, respectively [9]. These findings indicate that approximately 50% of the leading 200 prescription medications may be affected by clinically validated polymorphisms in CYP2C9, CYP2C19, and CYP2D6.

Genetic variants, such as ADRB2 Gly16Arg, which alter the response to albuterol, and CYP2C9 Arg144Cys, which influences warfarin dosing, demonstrate the substantial impact of genetic variability on pharmacological outcomes [6].

1.1.4 Pharmacogenomic Data Sources

Research into the genetic causes of diseases has advanced significantly, both in the amount of data produced and in the detail of the studies. As a result, the use of DNA testing and computerized health records is becoming more common in modern medical practice. Therefore, the goal is to combine all this information to create a clear understanding of how genetic changes affect health.

There have been significant advancements in the last 10 years in cloning and characterizing genes linked to human disorders. At the same time, technological advancements have enabled researchers to identify many genetic mutations, ranging from small changes to large ones. As a result, the importance of understanding and organizing these changes in databases has quickly become clear, offering significant benefits for diagnostics, clinical practice, and research [10].

Pharmacogenomics has proven an excellent tool for studying adverse drug responses by identifying individual-specific genetic characteristics associated with drug response. The importance of sequencing technology and sequence analysis techniques for understanding the role of genetic variation in the development of adverse reactions has been demonstrated by current methods [11].

To support such pharmacogenomic research, extensive datasets and databases are essential. Multiple consortia and networks have gathered and managed vast amounts of data from pharmacogenomics research. Such as -

- **Pharmacogenomics Knowledge Base (PharmGKB):** PharmGKB systematically records pharmacogenomic interactions, enabling manual inspection, data download for further analysis, and flexible searching, linking, and display according to research needs. The knowledge base supports both investigations into the effects of single-nucleotide polymorphisms on specific medication treatments and broader inquiries into diseases or medications to identify genes that may influence drug response or disease progression. As of April 2013, PharmGKB contains over 5,000 variant annotations, with more than 600 medications linked to genes and over 900 genes associated with drugs [12].
- **Dutch Pharmacogenetics Working Group (DPWG):** The goal of DPWG is to make it easier for PGx to be incorporated into the creation of evidence-based pharmacogenetics guidelines that improve pharmacotherapy. The key objectives of the DPWG are to establish PGx-informed treatment recommendations based on a thorough review of the literature and to support doctors and pharmacists by

integrating the recommendations into computerized systems for prescribing, dispensing, and monitoring medications. It has also developed the clinical implication score, assigned to each gene-drug interaction. This score is meant to help doctors decide whether to request the appropriate PGx genotyping tests before starting treatment [13].

- **Canadian Pharmacogenomics Network for Drug Safety (CPNDS):** CPNDS is a network of qualified surveillance doctors at 10 pediatric teaching hospitals across Canada who actively monitor more than 75% of Canada’s youngsters. The network’s purpose is to find genetic biomarkers of drug risk for significant ADRs so that prescription drugs can be used more safely. CPNDS surveillance clinicians focus solely on identifying and reporting ADRs by working closely with doctors, pharmacists, and nurses at each surveillance site in Canada [14].
- **Pharmacogene Variation Consortium (PharmVar):** It was established in 2017 to provide the clinical and research sectors with a place to archive and define the nomenclature of genes that affect how drugs are metabolized and how people respond to them [15]. In particular, it is a database of pharmacogenomic variation that supports specific haplotypes and alleles, with an emphasis on the human cytochrome P450 gene families and NUDT15 [11].
- **Clinical Pharmacogenetics Implementation Consortium (CPIC):** Since 2009, CPIC has made clinical pharmacogenomics more effective by releasing free, evidence-based gene-drug guidelines available to everyone. The CPIC guidelines, which include 34 genes and 164 medications, are now the global standard for translating pharmacogenomic test results into actionable prescriptions. It has 28 active guidelines, composed and published by people from all over the world, and widely used. They have been cited over 10,000 times and viewed 1.4 million times [16].

1.2 Clinical Decision Support Systems (CDSS) in Pharmacogenomics

Computerized CDSS represents a significant advance for modern medicine. CDSS provides doctors with more clinical information, helping them make better decisions. These systems frequently use digital healthcare records and other computerized clinical processes, which are increasingly adopted globally as electronic medical records with enhanced functionality become more prevalent. Clinical decision support (CDS) is crucial in PGx because it helps put genetic information into practice. A growing number of

institutions are using CDS technologies in clinical settings to integrate PGx data into patient care. Some of these institutions have even published about their experiences with these tools [17, 18].

A systematic analysis of clinical trials found that CDS systems that offer personalized healthcare recommendations are effective. The review indicated that CDS systems made clinicians better at their jobs in 19 (66%) of the 29 medication dosing [19].

1.2.1 Limitations of Clinical Decision Support Systems (CDSS) in Pharmacogenomics

CDS are necessary to incorporate pharmacogenomics into clinical practice; yet many obstacles remain. The extensive, continually evolving nature of pharmacogenomic knowledge required during patient encounters presents significant challenges for implementing pharmacogenomics in routine clinical care. Clinicians sometimes struggle to recall relevant gene-drug interactions and prior pharmacogenomic test results for specific patients, especially amid the pressures of high-volume clinical workflows. Also, even when pharmacogenomic results are available, it might be hard to understand what they mean. In addition, support for pharmacogenomics from electronic health record (EHR) vendors remains underdeveloped and limited, requiring considerable local institutional effort compared with other types of clinical decision support (CDS). Another challenge in developing CDS alerts is the risk of presenting prescribers with numerous interruptive alerts, which may lead to alert fatigue and the disregard of clinically significant messages. Moreover, when systems that support the main EHR aren't fully integrated, clinical workflows can be further complicated. Clinicians need to consult other sources for pharmacogenomic information [20].

These restrictions underscore the need for sophisticated computational methods to effectively interpret complex pharmacogenomic datasets and facilitate clinical decision-making.

1.2.2 Artificial Intelligence for Clinical Decision Support in Pharmacogenomics

Artificial Intelligence (AI) uses Machine Learning (ML), Natural Language Processing (NLP), and Deep Learning (DL) algorithms to look at challenging medical records. These new technologies help clinicians make the best treatment decisions by analyzing biological data, such as genomics, proteomics, and transcriptomics [21]. They have influenced both scientific inquiry and medical practice, particularly in pharmacogenomics

(PGx), where they facilitate the analysis of large datasets and the development of predictive models [22]. They are used to improve the interpretation of genomic data, enhance the prediction of how individuals will respond to drugs, and identify genetic markers related to how drugs are processed and how well they work.

Combining genetic and pharmacogenomic information with a person’s health and psychological characteristics allows healthcare professionals to create more personalized treatment plans for those with long-term illnesses [23]. AI systems can find complex patterns across these datasets, enabling personalized treatment plans for each patient based on their specific needs. Advanced computational models are crucial for examining the relationships among genetic variations, such as single-nucleotide polymorphisms (SNPs), gene expression profiles, and clinical data [21]. Consequently, the application of AI and machine learning in pharmacogenomics yields more accurate, data-driven insights into novel drug targets, drug efficacy, drug selection, and potential adverse effects [22].

1.3 Problem Statement

Pharmacogenomic knowledge has a complex, changing nature, and physicians must interpret large amounts of genetic and clinical data when developing treatment plans. Understanding how genetic variations influence drug response requires careful analysis of gene–drug interactions and clinical guidelines.

Healthcare professionals in demanding clinical settings may find it difficult to consistently recall and apply this information effectively in patient care. To help doctors apply pharmacogenomic knowledge in their daily work, CDSS have been developed. Despite their considerable potential, these systems have significant drawbacks. The rapid growth of pharmacogenomic knowledge, along with the fragmented nature of available data across many databases and resources, presents a considerable challenge to quick access. Moreover, CDSS often produce a large number of alerts. This can reduce their overall effectiveness.

Recent developments in artificial intelligence have opened up new ways to address these problems. Large language models (LLMs) have demonstrated their ability to understand natural language questions and effectively handle large amounts of biomedical data. But LLMs can often give incorrect or unsupported answers, especially when they rely solely on their own knowledge and don’t have access to verified clinical resources.

Therefore, a system is needed that can use both reliable pharmacogenomic knowledge and advanced language models to help doctors make decisions. This research suggests

creating an evidence-based clinical pharmacogenomics question-answering system using Retrieval-Augmented Generation (RAG) and a fine-tuned language model. This approach aims to deliver reliable solutions to pharmacogenomic inquiries while reducing the risk of hallucinated or unsupported recommendations.

1.4 Research Objectives

The main goal of this project is to develop an evidence-based clinical pharmacogenomics question-answering system that integrates large language models and retrieval-augmented generation (RAG) for reliable clinical decision-making. This work tries to solve the challenges of inconsistent pharmacogenomic knowledge and the constraints of independent large language models by combining structured knowledge, e.g., CPIC and ClinPGx, with state-of-the-art language modeling approaches. The research focuses on the use of a hybrid retrieval system combining lexical filtering and semantic similarity to enhance the relevance of recovered information. In addition, the study also applies parameter-efficient fine-tuning methods like LoRA to adjust LLaMA and Qwen models for pharmacogenomics tasks with domain-specific modifications. A key objective is to reduce hallucinated or unsupported reactions by grounding the model outputs in validated clinical evidence. The effectiveness of the proposed framework is evaluated by means of classification-based metrics, text-based similarity measures, and qualitative assessment techniques. Thus, a holistic comparison of base, fine-tuned, and RAG-enhanced models is possible to identify the most reliable and clinically beneficial approach.

1.5 Key Contributions of This Study

This thesis creates a hybrid RAG-based pharmacogenomics question-answering system that uses both lexical filtering and semantic re-ranking to find evidence more quickly and accurately. It preprocesses CPIC-derived data into a concise JSONL format for retrieval and modeling, while ClinPGx guideline data is preserved as a distinct JSONL dataset and incorporated during retrieval.

This research sets up and compares several models (base, LoRA, hybrid) to see how parametric and retrieval-based knowledge integration affected the results. It employs a mix of evaluation methods, including metrics, similarity scores, and manual review. The proposed paradigm highlights the need to ground large language models with established guideline-based knowledge to address hallucinated or unsupported recommendations in clinical decision support. Furthermore, it introduces a domain-specific application of

large language models to pharmacogenomics, addressing issues such as fragmented data sources and developing clinical knowledge. This work investigates the impact of combining retrieval-augmented generation with fine-tuning on the reliability and evidence grounding of large language models for pharmacogenomics applications.

1.6 Structure of the Thesis

The thesis is structured as follows. In Chapter 2, it is discussed the preliminaries and background covering the overview of artificial intelligence, massive language models, and retrieval-augmented generation and their applications in pharmacogenomics. Chapter 3 discusses relevant work in the area of pharmacogenomics and AI-based clinical decision support systems. In Chapter 4 discusses the suggested technique such as the data integration from CPIC and ClinPGx, the preprocessing, the LoRA fine-tuning, and the hybrid retrieval-augmented generation architecture.

Chapter 5 is dedicated to the empirical analysis, covering the preparation of the dataset, the system configuration and the model versions. The trials and results are described in Chapter 6, including classification performance, text-based evaluation and manual evaluation of the model outputs. Chapter 7 summarizes the results including biomedical discoveries, interpretation of results, limits and future directions. Finally, Chapter 8 ends the thesis and highlights the main contributions of this study.

Chapter 2

Preliminaries and Background

This chapter introduces the basic principles and background knowledge required to understand the proposed pharmacogenomics framework. It discusses the growth of artificial intelligence, major developments in machine learning and deep learning, and the importance of huge language models in healthcare applications.

2.1 Evolution of Artificial Intelligence

Artificial Intelligence (AI) has experienced significant modifications over the past few decades, evolving through various paradigms [24]. In general, the term "artificial intelligence" refers to computer systems that try to learn, think, and solve problems in a way that is comparable to how people do it [25]. AI refers to the scientific and engineering discipline focused on developing intelligent machines, particularly computer programs capable of reasoning and learning. In the beginning, AI systems were mostly based on rules. However, during the middle of the twentieth century, both symbolic AI and rule-based systems were the most common methods used in AI research. In this case, it is rather general and involves many different tasks, such as conceptual thinking, problem-solving, planning to reach goals, identifying sounds and objects, speaking, translating, and directing robots. This changed with the rise of machine learning, which highlighted the importance of learning from data [24, 26].

2.1.1 Transition to Machine Learning

This change was a big step toward using statistical and classical machine learning (ML) methods, which let computers find patterns in data on their own instead of having to encode knowledge by hand [24]. In its traditional meaning, ML is the study of how to

make computer programs that get better with use. So, ML, a type of AI, can learn without being explicitly programmed. The goal is to detect patterns in training data and utilise them to train a model that can make predictions. Machine learning is also about learning from examples, definitions, being taught, and behaviour. The simplest way to understand their relationship is to picture them as circles within each other, with AI on the outside and ML on the inside [25, 26].

2.1.2 Deep Learning in Modern AI Systems

The advancement of machine learning led to a significant transformation toward deep learning. This change, in particular, led to the deep learning revolution of the 2000s and 2010s, when multilayer neural network models trained on enormous datasets made huge strides in the state of the art [24]. ML and AI are becoming the most popular approaches for solving problems across many fields of research and industry. This is mostly because DL has been so successful lately [26].

Deep learning DL also focuses on finding patterns in data to mimic how the human brain makes decisions, leading to the growth of large language models (LLMs) [27]. It makes hierarchical representation learning possible. Deep learning uses algorithms that operate at different levels of abstraction to discover complex patterns in incoming data [25]. To put it another way, the difficulty of designing algorithms that cover every possible situation is traded for the difficulty of identifying the correct overall outline of algorithms, such as a deep neural network, and of processing data. Also, because neural networks are generic function approximators, they require substantial data to train and often require large labelled training sets [26]. This advancement greatly enhanced systems' capacity to acquire intricate architectures from data.

2.1.3 Natural Language Processing (NLP)

NLP represents a significant domain for the application of deep learning. The field of language modelling has undergone a substantial transformation following the integration of neural networks and deep learning techniques [28]. DL seeks to replicate the complex decision-making processes of the human brain by recognising patterns within data [27]. NLP is about understanding and modelling human languages so that computers can perform tasks such as speech recognition, text translation, text generation, question answering, and information extraction [25]. These applications encompass a wide range of tasks, including interacting with, translating for, and commanding robots [26].

Embeddings, which transform words into numerical vectors that capture semantic and syntactic similarities, played a significant role in earlier advancements [28]. Consequently, deep learning enabled NLP systems to interpret language more effectively by identifying meaning, context, and relationships. These models are based on deep learning architectures that enable the model to learn which parts of a text sequence are most important. This helps the machine comprehend the language’s structure and context better [27].

2.1.4 Emergence of Large Language Models from NLP

Large language models are a big step forward in natural language processing. As mentioned, the field of NLP has changed a lot with the development and rise of large language models (LLMs). This has led to a move from simpler models to more complicated and effective ways to interpret and create human language. Furthermore, LLMs have become fundamental in comprehending and producing human language. Deep learning and transformer architectures were the main forces behind this change. The rise of deep learning and the development of architectures like transformers are both very important [28]. Also, these improvements led to the Transformer design, which is based on self-attention mechanisms and is now the basis for most of the best LLMs [24].

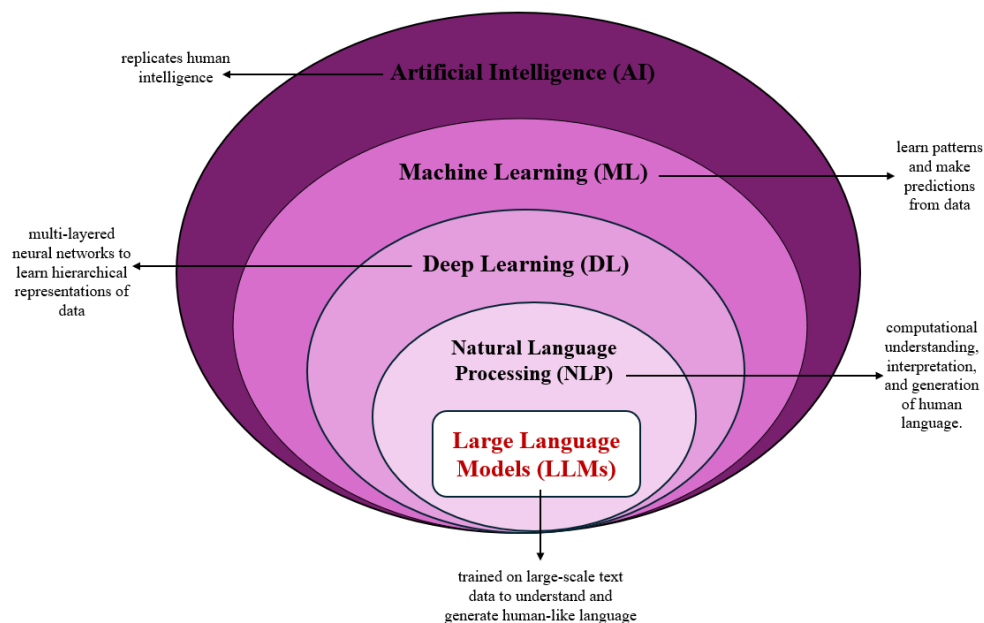


FIGURE 2.1: Hierarchical relationship between Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), and Large Language Models (LLMs).

Figure 2.1 shows how artificial intelligence and its subfields are related to each other in a hierarchy.

2.2 Large Language Models and Their Variants

Large Language Models, large-scale, pre-trained statistical language models, are built using neural networks. These models have been trained on large datasets, allowing them to perform a wide range of language-related tasks. They can recognize and generate language for a wide range of purposes because they are trained on huge text corpora with billions of parameters [29]. Most of the time, LLMs are trained on huge text datasets in a self-supervised way. The main thing LLMs do is predict the next token in a sequence based on an input.

LLMs have demonstrated strong performance across a variety of domains, particularly in natural language processing tasks. These models can create language that is cohesive, understandable, and sounds like it was written by an actual person. This makes them powerful AI systems for analyzing and generating text. The main reason LLMs work so well is that they tend to be large, both in terms of model size and the amount of training data. Larger pre-trained language models (PLMs) yield greater performance improvements, which has driven the transition from PLMs to LLMs by substantially increasing model parameter counts and training dataset sizes [30].

LLMs are transformer-based neural language models containing tens to hundreds of billions of parameters. Consequently, these models demonstrate advanced capabilities, including reasoning, planning, decision-making, in-context learning, and zero-shot answering.

LLMs have changed over time into different versions that reflect differences in design, training goals, and deployment uses. Most current LLMs are built on transformer architectures, but they are very different in how they are designed, how big they are, and what they can be used for. Models like GPT, LLaMA, and Qwen are well-known examples because they have been widely used for various natural language processing tasks.

2.2.1 Large Language Model Meta AI (LLaMA)

The LLaMA (Large Language Model Meta AI) family is a group of large language models that were created to improve the ability to process natural language. Large language models (LLMs) demonstrate an unprecedented ability to comprehend and generate

human-like text, as was said in earlier work. Large amounts of text data are used to train these models, and they can do many things, such as generating text, asking questions, and following instructions [31].

Meta has launched three principal versions of its large language model (LLM), LLaMA, in addition to a minor upgrade (version 3.1). When LLaMA first came out in early 2023, it was a big step forward for the open-source natural language processing (NLP) community. Meta has continued to support this community by putting out new versions of LLM [32].

2.2.2 Qwen

Recently, there has been a lot of interest in the Qwen model family, a rapidly developing large language model. Large language models (LLMs) are constantly changing, with open-source solutions such as Qwen 3 ,and it is showing significant advancements [33].

There is a clear trend of progress that can be seen in the way Qwen has changed over time. The first version, Qwen-1, came out in 2023. The new Qwen-1.5 version has a lot of changes after a few months. In 2024, Qwen-2 came out with the ability to use more than one mode [34]. These enhancements demonstrate the model's increasing proficiency in handling extended contextual information, generating code, and tackling complex reasoning challenges, thereby broadening its applicability in both academic and commercial settings.

It is highlighted that the quality of the instructions is very important for LLM fine-tuning. High quality instruction sets to help large models do very well, and Qwen-based approaches have shown improvements in complexity, answer relevance and diversity [35].

2.3 Foundation of Modern Large Language Models

LLaMA and Qwen are modern large language models based on the transformer architecture. These models all have the same basic design as the transformer framework, even though they use different scales, training methods, and optimization strategies. To better understand how these models respond to the context and handle long text sequences, it's important to look at the basic principles of transformer design.

2.3.1 Transformer Architecture

The transformer model has emerged as a significant transformation from traditional sequence models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs). Older methods used sequential processing, meaning training examples couldn't run in parallel. The transformer was created to address this problem. It is a new, simple network design that relies solely on attention mechanisms and doesn't use recurrence or convolutions. Transformers operate on input sequences in parallel, making them more efficient and scalable.

Self-attention is the most important part of this architecture. Self-attention is a method for paying attention to different parts of the same sequence to build a representation of the sequence. This lets the model figure out how words are related to one another, regardless of how distant they are. In this mechanism, query, key, and value representations are used to figure out attention. An attention function takes a query and a set of key-value pairs and gives an output. Also, multi-head attention lets the model attend to information from multiple representation subspaces at different locations simultaneously.

The transformer architecture also has an encoder-decoder structure. For example, most competitive models of neuronal sequence transduction feature an encoder-decoder architecture. Each layer has attention and feed-forward components that enable learning hierarchical representations.

Transformers don't process sequences in order, so these need to be manually provide positional information. This means these need more information about the relative or absolute positions of the tokens in the sequence [36]. Overall, transformers effectively capture global dependencies and serve as the basis for modern large language models, which mostly use the transformer architecture [37].

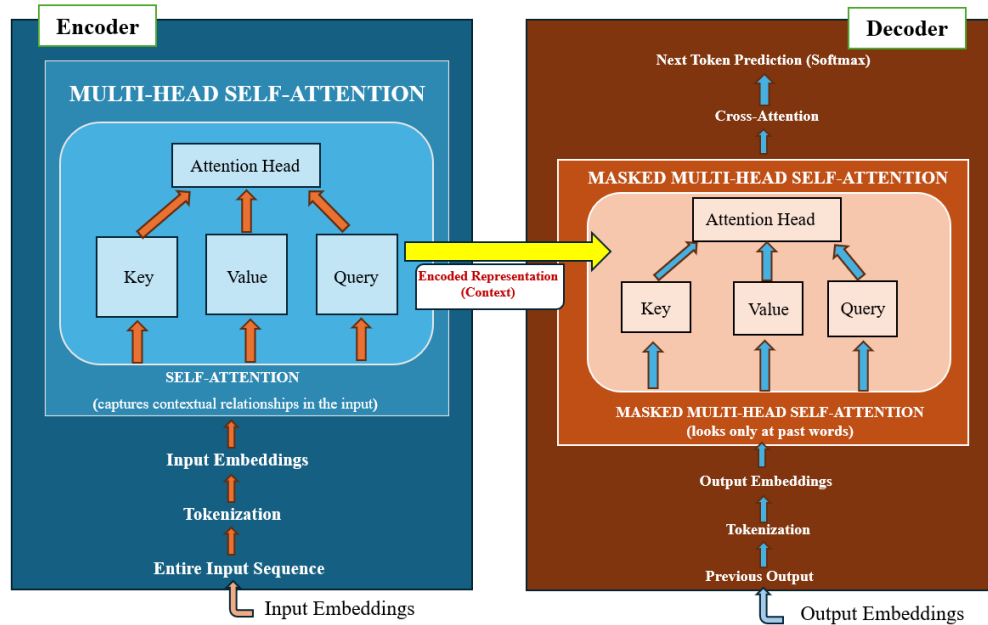


FIGURE 2.2: A schematic representation of the Transformer architecture demonstrating how attention techniques are implemented for encoder-based context encoding and decoder-based autoregressive token prediction.

Figure 2.2 illustrates how the encoder constructs contextual representations of the input sequence. The decoder utilizes these representations to anticipate the next token in order.

2.3.2 Working Mechanism of LLMs

Large language models process text in a series of organized steps, starting with converting raw text into numerical representations and ending with generating useful output. These phases help the model understand the meaning of words in context, find connections between them, and give answers that make sense. The overall working mechanism is explained as follows:

- **Tokenization:** The first stage in processing text is to divide the input sentence into smaller parts. Tokenization is the process of splitting text into smaller parts, called tokens, which the model can then use. Tokens are the basic units of computation.

After tokenization, the text is understood as structured data, rather than just regular language. Each token is mapped to a specific index in a vocabulary list, allowing the model to process the text numerically [38].

- **Embeddings and Positional Encoding:** After converting tokens into numerical vectors, the model interprets their meanings. An embedding represents tokens,

such as words or subwords, as numerical vectors that encode their semantic meaning. These embeddings enable the model to determine semantic relationships between words. However, because transformers process tokens in parallel, the word order must be preserved. Therefore, positional encoding is added to the input embeddings to encode each token's position.

This method helps the model distinguish between sequences such as "drug affects gene" and "gene affects drug," which would otherwise appear identical [38].

- **Context Understanding through Attention Mechanism:**

After generating embeddings, the model uses attention mechanisms to assess the contextual significance of the input. A defining characteristic of the transformer architecture is its self-attention mechanism. This approach enables the model to analyze the interrelationships among all tokens within a given sequence. Therefore, the model can evaluate how important each word or token is compared to the other words in the input sequence.

This design allows the model to accurately understand long-range dependencies and contextual relationships. This mechanism helps the model focus on the most important parts of the input when producing output. Moreover, the use of multiple attention heads improves this feature. Multi-head attention improves this by allowing the model to focus on different representation subspaces at the same time [38].

- **Probability Prediction (Next Token Generation):** The model predicts the next token in the sequence after understanding the context. The prediction is based on a probability distribution that includes all potential tokens. It may gain the attention weights by using the softmax function on the attention scores.

These chances tell you which word is most likely to come next. The output of each token is the weighted sum of the value vectors for all tokens. In real-world systems, decoder-only transformers are trained to predict the next token in a sequence. This makes them good for jobs like generating text. This procedure is autoregressive, meaning that each token depends on the tokens created before it. This ensures that the created text is coherent [38].

- **Pre-training and Fine-tuning:** The effectiveness of LLMs depends on their training methods. Training typically involves two main stages: pre-training and fine-tuning. Initially, these models are trained on large datasets to learn general language patterns and structures. LLMs use self-attention mechanisms and extensive training to find complex patterns in text.

After pre-training, models are adapted for specific tasks using fine-tuning techniques. Fine-tuning involves training a pre-existing model on smaller, task-specific datasets to improve its performance on certain tasks. After this, the model is further refined using task-specific datasets.

Parameter-efficient fine-tuning and prompt engineering are used to enhance their adaptability and precision. This two-stage training process allows LLMs to perform a wide range of tasks with high accuracy and contextual understanding. This method has produced remarkable results across a wide range of tasks [38, 39].

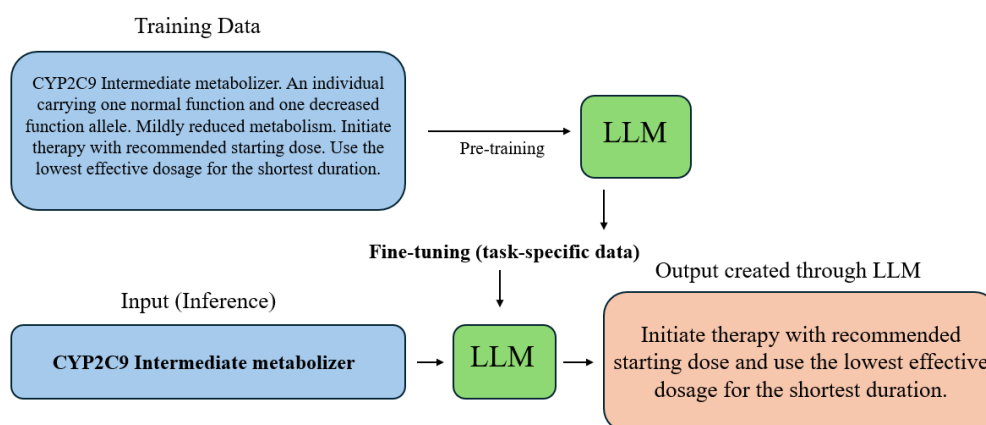


FIGURE 2.3: Illustration of pre-training, fine-tuning, and inference process in large language models using a pharmacogenomic example.

Figure 2.3 shows that LLMs are first trained on a large amount of text data. And fine-tuned using knowledge specific to a certain field. Finally, it is used to generate outputs by predicting the most likely next tokens.

2.3.3 Large Language Models in Healthcare

LLMs are becoming more common in healthcare settings because they can handle and make sense of large amounts of data. Research indicates that the application of these models may enhance diagnostic precision, optimize administrative processes, and thus render healthcare delivery more efficient, personalized, and comprehensive.

LLMs have demonstrated significant expertise in interpreting clinical documents. Some of these skills are recognizing named entities, extracting relationships, making inferences in plain language, working in a variety of medical settings, classifying documents, and answering questions. These tasks are necessary for models to read and understand clinical notes, patient records, and other medical data. Healthcare data is complicated,

diverse, and huge. LLMs can automate the examination of medical literature, which helps find vital information [40].

CDS represents a significant application. LLMs are used in CDS to analyze medical data, involve patients, share health information, and perform many other tasks. Also, language models have become important tools for healthcare workers because they can provide evidence-based advice to support diagnosis, treatment, and the overall quality of patient care.

There is also potential for LLMs in finding new drugs and treatments. A major advantage is that they can sort through huge amounts of medical literature, patient histories, and the growing pool of clinical studies. They are useful for complex medical reasoning tasks, such as developing treatment plans and providing personalized advice.

LLMs are often used to summarize medical information. These models are known for their ability to thoroughly analyze and quickly summarize large amounts of medical literature. This helps doctors and researchers stay updated on the latest research and medical developments [40].

2.3.4 Limitations of Large Language Models (LLMs)

LLMs have many potential uses, but significant challenges remain in healthcare. The foremost concern pertains to the accuracy and reliability of these models. Research indicates that LLM outputs are occasionally erroneous, deficient in safety and utility, and may offer responses that are, in general, superficial. Consequently, in clinical environments, this can result in inaccurate interpretations or unsafe recommendations, thereby precluding their independent use in decision-making processes [41]. This problem is frequently linked to hallucination, in which LLMs produce responses that are inaccurate or inconsistent with the information provided [30].

Another serious issue is that there aren't any real-world tests. The report indicates that a mere 5% of the examined studies employed authentic patient care data for the evaluation of large language models. Consequently, the majority of these investigations relied on either test questions or simulated scenarios. This reliance on artificial constructs introduces a disparity between the observed performance in experimental settings and the practical utility in real-world applications, given the inherent complexity and variability of actual patient data [41].

The medical reasoning exhibited by LLMs is characterized by inconsistencies and a lack of clarity. For example, these models may provide vague recommendations or err in their suggestions for patient-specific treatments. This demonstrates that current models are

deficient in the precision required to deliver tailored therapeutic guidance. Consequently, these shortcomings diminish their reliability in the context of critical medical decision-making [42].

Concerns about bias and right and wrong are also key issues. LLMs can reproduce information from their training data, which could include biases that people already have, which could lead to results that are unfair or biased. Also, problems like fairness, bias, and toxicity are rarely measured in the evaluations that are done now, which makes me worry about how safe it is to use in a wide range of patient groups [41].

Data security and privacy remain significant challenges. The use of LLMs in clinical systems poses potential risks to patient safety and the secure handling of sensitive healthcare information. Protecting data privacy is a critical factor in the implementation of large-scale models. Accessing real healthcare services presents a significant challenge.

The deployment of LLMs necessitates a "robust IT infrastructure," and their incorporation prompts apprehensions regarding financial implications, required modifications to established workflows, and the capacity for seamless interaction with existing systems [41]. Moreover, the evaluation of their practical efficacy within clinical settings is hindered by the lack of standardized assessment frameworks and consistent testing protocols [42].

2.4 Retrieval-Augmented Generation (RAG)

As mentioned earlier, large language models have problems, including outdated data, hallucinations, and difficulty keeping track of specialized information. But large language models also face many challenges. These systems typically give incorrect answers and don't seem to understand complex questions or specialized areas of study [43]. RAG has become a useful method for addressing these issues.

This method involves gathering relevant information from outside sources before generating responses using large language models. RAG uses a two-step process. The first step involves using LLMs to find information from sources that aren't part of the model's built-in knowledge. Once it has the necessary information, it generates text. Essentially, RAG introduces an external retrieval phase that precedes the actual generation process. This improves the quality of the generated responses. Furthermore, RAG enhances language generation by dynamically accessing knowledge bases during inference [44].

RAG proves to be a valuable technique, especially when acknowledging the constraints inherent in LLMs, particularly for tasks demanding extensive domain-specific expertise [43].

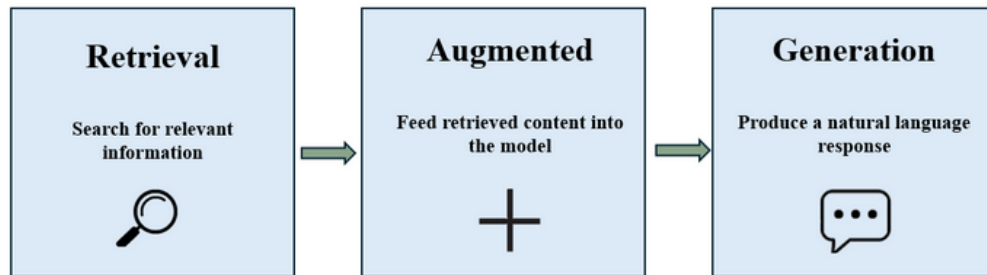


FIGURE 2.4: Conceptual illustration of the Retrieval-Augmented Generation (RAG) architecture.

Figure 2.4 shows how RAG operates. First, it gets relevant information from outside sources, then it adds that information to the model, and finally it uses that knowledge to make a response that is aware of the context.

Chapter 3

Related Work

3.1 Large Language Models in Pharmacogenomic Research

Karl Keat et al. (2025) created many evaluations, which included both automated and expert-reviewed parts, to see how well chatbots could answer questions about pharmacogenetics. This study looked at the perspectives of doctors, patients, and researchers. The researchers noted that current pharmacogenomics standards are complex and not always clear, which makes it difficult to use them effectively in practice. Despite the strong capabilities of LLMs, the hallucinations raise concerns regarding their reliability in healthcare.

To address this deficiency, the study investigation introduced PGxQA, a robust benchmark dataset designed for pharmacogenomics question answering, containing 110,000 question-answer pairs. The evaluation demonstrated that large language models (LLMs) continue to produce inaccurate or potentially harmful responses, which could significantly impact practical healthcare applications. This research explored the integration of retrieval augmented generation (RAG) into LLMs, although this approach was not fully utilized, thereby indicating the necessity for more effective methodologies [45].

Murugan et al. (2024) developed an AI assistant using GPT-4 to help interpret PGx data and support clinical decisions. This AI assistant uses retrieval-augmented generation (RAG), which involves a knowledge base (KB) that includes information from the CPIC.

The system's responses were contextually aware, drawing upon curated PGx resources. This contextual awareness stemmed from RAG's ability to retrieve pertinent information from the KB in response to user queries, subsequently generating responses via GPT-4. The study also used fast engineering and guardrails to improve the quality of the responses. An evaluation by expert reviewers found that the AI assistant significantly

outperformed ChatGPT 3.5 achieving 85% effectiveness versus 69% for provider queries. These findings indicated that the integration of RAG with domain-specific knowledge enhanced the accuracy, relevance, and functionality of LLM-based clinical assistants.

The study had limitations due to difficulties in domain-specific comprehension, as the embedding model was limited in recognizing PGx terminology, and the system exhibited deficiencies in accuracy and contextual retrieval. The evaluation's use of expert scoring and pre-set questions limited its ability to fully reflect real-world situations [46].

Zack et al. (2025) examined the capacity of large language models (LLMs) to formulate pharmacogenomic (PGx) recommendations that adhered to clinical guidelines, specifically assessing their capability to generate clinically sound PGx recommendations in accordance with CPIC guidelines. This study developed a benchmark dataset, meticulously curated with gene-drug-phenotype scenarios, and evaluated a range of models, such as GPT-4 and LLaMA variants, utilizing both traditional and semantic evaluation techniques. To address the limitations of lexical metrics, the authors introduced the LLM Score, a novel evaluation framework intended to gauge semantic equivalence between model outputs and established recommendations. Furthermore, this study incorporated domain adaptation techniques, such as fine-tuning and structured prompting. General-purpose models often gave incomplete or dangerous results, but this study's domain-adapted model did much better. The results showed that specialized models made clinically meaningful and safe recommendations much better than general-purpose LLMs.

However, the study had several limitations. The dataset's exclusive reliance on CPIC rules limited its applicability to other guideline systems and real-world situations. Furthermore, general-purpose models struggle to consistently implement pharmacogenomic guidelines, often leading to potentially harmful or insufficient recommendations. Consequently, the clinical usefulness of this research is questionable [47].

3.2 Retrieval-Augmented Generation in Biomedical Applications

Rector et al. (2025) investigated the use of Retrieval-Augmented Generation (RAG) in validating pharmacogenomic query responses, demonstrating that advanced models such as GPT-4, when enhanced by RAG, facilitate the accessibility and application of pharmacogenomic data by incorporating authoritative resources directly into their operational framework. The study developed an AI assistant, Sherpa Rx, which employs both retrieval-based and generative techniques to furnish responses that are more precise

and contextually relevant than those produced by isolated generative models. Their approach integrated CPIC guidelines and PharmGKB data into a structured knowledge base. Following this, the algorithm pinpointed and retrieved the four most relevant papers, determined by semantic similarity, and utilized contextual data to formulate responses. The findings demonstrated that integrating a RAG methodology with CPIC and PharmGKB data significantly enhanced the AI assistant's accuracy and overall performance, thus underscoring the value of integrating domain-specific knowledge bases.

The study's findings are limited by a small sample size, which reduces the statistical power of the performance measurements. In addition, the reliance on a single person's manual evaluation could introduce bias, affecting the generalizability of the results [48].

Borishetty et al. (2025) created a pharmacogenomics assistance tool using Retrieval-Augmented Generation (RAG). This tool highlighted the limitations of large language models (LLMs) when answering specific pharmacogenomics questions without expert knowledge. To tackle this issue, the research integrated data from reputable sources, such as CPIC, PharmGKB, DPWG, and FDA guidelines, thereby establishing a knowledge base comprising 2,617 embedded document segments. The system employed semantic search techniques to pinpoint relevant information. Subsequently, retrieval augmented generation (RAG) was utilized to extract pertinent data in response to user inquiries and to generate responses informed by this extracted information.

Their evaluation revealed that RAG-enhanced models often outperformed native LLMs, achieving an accuracy of up to 95.1%, compared to 89.8% for the comparable native model, thus demonstrating the effectiveness of RAG in improving clinical decision support. However, the study had limitations because it mainly used multiple-choice questions to evaluate the system. This method might not fully represent real clinical situations. Additionally, the assistant wasn't designed as a chatbot or plug-in to help doctors understand raw genomic data [49].

Patrick Lewis et al. (2020) introduced a retrieval-augmented framework designed to overcome the constraints of traditional language models when applied to knowledge-intensive tasks. Their research revealed that, although large pre-trained language models demonstrated factual knowledge, their capacity to efficiently access and apply this knowledge was restricted. To address this issue, the study introduced Retrieval-Augmented Generation (RAG), a method that combined parametric and non-parametric memory to enhance text generation. The model combined a neural retriever with a sequence-to-sequence generator. The parametric memory was a seq2seq model that had already been trained, while the non-parametric memory was a dense vector index of Wikipedia.

The methodology demonstrated that the incorporation of external knowledge enhanced factual precision. RAG models demonstrated a comparative advantage, generating language that was more accurate, varied, and factually sound when compared to a leading parametric-only seq2seq model. However, the research's reliance on extensive external knowledge bases, such as Wikipedia, necessitated the implementation of robust information retrieval methods, which could potentially introduce intricacies contingent upon the quality of the retrieved data [50].

3.3 Hierarchical Retrieval-Augmented Methods

Jeon et al. (2026) presented a Hierarchical Retrieval-Augmented Generation (HRAG) architecture for anticancer medications that illustrates the interconnections among PGx documents. The investigation indicated that generative AI offers a promising method for providing interactive clinical decision support, thereby addressing the existing knowledge gap. The authors highlighted a limitation of conventional retrieval-augmented generation (RAG) systems: their dependence on retrieving semantically similar segments without adequately considering the interconnections among diverse guidelines, potentially leading to information imbalance. Consequently, they introduced hierarchical retrieval-augmented generation (HRAG), a system designed to structure pharmacogenomics (PGx) guidelines into hierarchical tree structures, thereby facilitating more efficient information retrieval.

Their findings indicated that HRAG substantially surpassed RAG in guideline-related activities, attaining superior accuracy and F1 scores. The study's results were limited because both the HRAG and RAG models performed poorly in tasks requiring exact matches of numbers or entities. This highlights the difficulties in processing structured genetic data [51].

3.4 Limitations of Pharmacogenomics in Clinical Practice

Thottunkal et al. (2025) investigated practical obstacles in the use of pharmacogenomics (PGx) in clinical settings and identified significant limitations that hinder its acceptance. The authors highlighted that there was a range of challenges that impede uptake in current clinical practice. A major concern was the limitations of current pharmacogenomics (PGx) testing. Specifically, commercially available PGx panel tests often didn't include all the important PGx genes that could be acted upon.

The research also indicated that current prescribing methods often don't use personalized approaches. It suggested that current practices don't include pharmacogenomics (PGx), instead relying on a trial-and-error method. The authors also discussed system-level problems that hinder the effective use of PGx, such as its incompatibility with electronic medical records and a lack of knowledge among doctors. In conclusion, the study showed that pharmacogenomics (PGx) has a lot of potential for personalized medicine. However, its practical use is limited by technological, clinical, and systemic challenges [52].

3.5 Comparison with Existing Work

The following section compares the approach presented in this thesis with existing pharmacogenomics systems. It emphasizes the contributions of this study by focusing on key variations in retrieval methods, model adaptation procedures, and evaluation tactics.

TABLE 3.1: Comparison with Existing Work

Paper	Structured CPIC tables	Advanced RAG (retrieval + reranking)	Semantic reranking	RAG + fine-tuning integration	Large-scale dataset	Multi-metric evaluation	Clinical-aware prompt design
Thottunkal et al. (2025) [52]	X	X	X	X	X	X	X
Lewis et al. (2020) [50]	X	✓	X	X	✓	✓	X
Zack et al. (2025) [47]	X	X	X	X	X	✓	X
Murugan et al. (2024) [46]	X	✓	X	X	X	✓	✓
Borishetty et al. (2025) [49]	✓	✓	X	X	X	✓	✓
Rector et al. (2025) [48]	✓	✓	X	X	X	X	✓
Keat et al. (2025) [45]	X	X	X	X	X	X	X
Jeon et al. (2026) [51]	X	✓	X	X	X	✓	X
Proposed System	✓	✓	✓	✓	✓	✓	✓

Table 3.1 highlights the key differences between the current methods and the new system. Previous research has mainly focused on specific parts of LLM-based pharmacogenomic systems. This includes evaluation [47], basic retrieval methods [50], and applications in specific areas using RAG [46, 48, 49]. Although Jeon et al. (2026) [51] presented a complex hierarchical retrieval method, it was limited to structured retrieval and didn't include additional optimization techniques.

Conversely, the proposed system integrates several sophisticated components, including structured Clinical Pharmacogenetics Implementation Consortium (CPIC) knowledge representation, semantic reranking, and the fusion of retrieval-augmented generation with fine-tuning methodologies. Furthermore, it employs multi-metric evaluation and clinically-informed prompt design to enhance the precision and dependability of its outputs. The integrated features address significant limitations found in previous research, particularly in managing specialized knowledge and the consistency of the model's outputs. As a result, this system demonstrates improved suitability for practical clinical decision support applications.

Chapter 4

Methodology

In this chapter, the methodology follows to construct the proposed pharmacogenomics question-answering system is introduced. It describes the data sources, pretreatment and integration stages, model configuration and fine-tuning process, and the hybrid retrieval mechanism to facilitate evidence-grounded answer generation.

4.1 Overview of Data Sources

This study employs a hybrid pharmacogenomics dataset derived from two beneficial sources: structured clinical guideline data from the CPIC [53] and integrated clinical knowledge sourced from the ClinPGx platform [54].

The goal of integrating these sources is to use both structured clinical data and rich textual information to make a system that can do both machine learning (fine-tuning) and retrieval-based reasoning (RAG). CPIC provides structured tables of standardized, evidence-based pharmacogenomic guidelines. These datasets provide information about how genes and drugs interact, how to classify phenotypes, how to define alleles, and clinical dosing recommendations [53].

ClinPGx, on the other hand, is a comprehensive clinical pharmacogenomic (PGx) resource created to support and expand PGx knowledge, implementation and education. It integrates the PharmGKB, CPIC, and PharmCAT projects, with additional features and content to come [54].

4.1.1 CPIC Data Structure

CPIC data is accessible in structured formats through the database (DB) and the API. This information is derived from the CPIC guideline documents and accompanying materials. CPIC data shows how distinct ideas are related to each other. Drugs, test alerts, and allele definitions are all examples of concepts. Each of those ideas becomes a "data model." A data model is usually a table in the database; some ideas may need a more complicated multi-table specification. It can query any table using the API or in the database, and it can even combine some of them to get more detailed information [55].

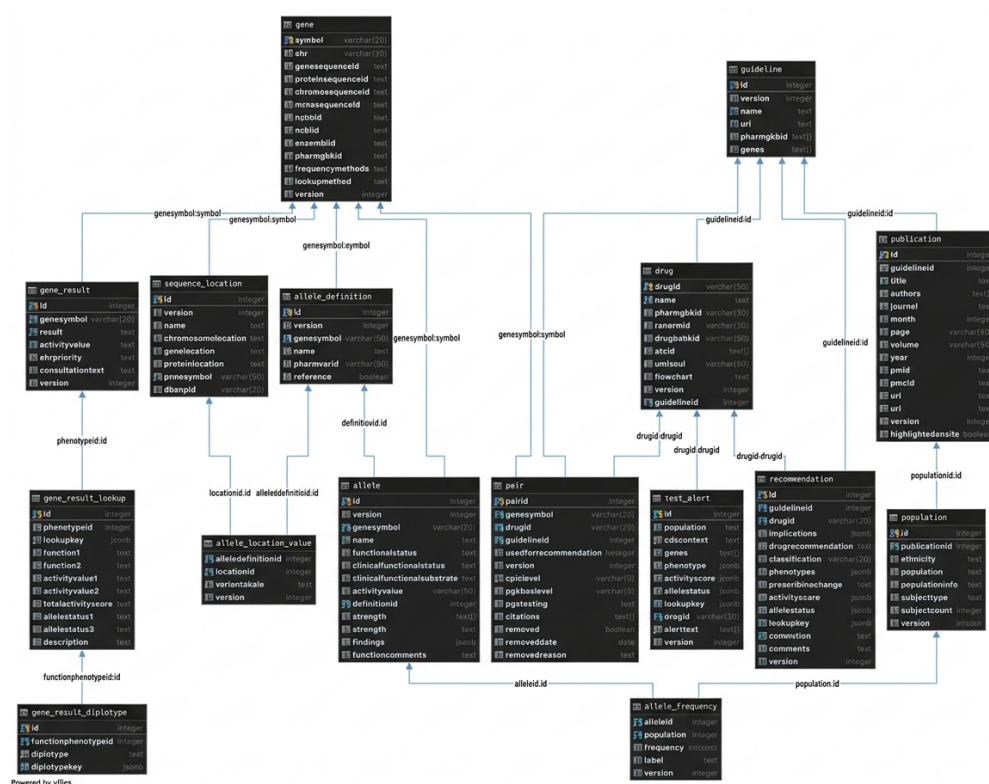


FIGURE 4.1: Overview of the CPIC relational data model [55]

Figure 4.1 presents the relational data schema of CPIC, demonstrating the structured relationships among key components such as genes, medications, alleles, and test alerts. This structured design streamlines data retrieval and integration, thereby supporting subsequent applications, including pharmacogenomic decision support systems.

CPIC Data Extraction and Initial Setup

CPIC data, which offers structured access to pharmacogenomic guideline information, is derived from a PostgreSQL database [55]. This database's architecture comprises multiple tables, each detailing specific elements such as genes, medications,

alleles, phenotypes, and clinical recommendations, as illustrated in the CPIC data model. To enable comprehensive data processing and integration, these tables are exported in CSV format.

CPIC Data Integration

Following extraction, data from multiple CSV files are integrated based on their relational structure. The tables are combined through primary and foreign key relationships defined within the CPIC schema, thereby preserving the associations between genes, medications, alleles, and therapeutic suggestions. Data integration is performed using PySpark and DuckDB, which enables efficient handling of multi-file joins and large-scale data processing, given the dataset’s considerable size and distributed nature.

CPIC Data Preprocessing

Following data integration, preparatory steps are undertaken to improve the overall quality of the dataset. Redundant entries are removed by eliminating duplicate rows and missing values are addressed to ensure the presence of crucial pharmacogenomic data required for further analysis.

Normalization techniques are used to standardize textual data and correct inconsistencies. The data set exhibits variations in column names, which requires the application of various key mappings to uniformly extract crucial attributes, such as gene, drug, allele, and phenotype.

Additionally, records lacking critical pharmacogenomic information are excluded to preserve only valuable entries. These methods create a structured and reliable dataset, which is then suitable for later processing and use in retrieval-augmented generation (RAG) and evaluation.

4.1.2 ClinPGx Data Structure

ClinPGx delivers PGx information via annotated clinical guidelines that amalgamate several sources, including CPIC, DPWG, and additional organizations. The data is structured as guideline annotations encompassing essential components such as drug names, related genes, guideline sources, and publishing information. Each annotation comprises a concise description of the guideline recommendations, pertinent extracts from the guideline, and hyperlinks to the original source papers.

ClinPGx, besides offering descriptive information, employs annotation tags to classify the nature of its clinical guidance. These tags, which encompass “Alternate Drug”,

“Dosing Info”, “Other Guidance”, and “Testing Info”, define distinct categories of recommendations derived from genetic variants or phenotypes. Consequently, this structure facilitates users’ ability to efficiently select, investigate, and retrieve relevant information. Consequently, this structure facilitates users’ ability to efficiently select, investigate, and retrieve relevant pharmacogenomic data [54].

ClinPGx Data Integration and Preprocessing

ClinPGx data are collected through a semi-automated system that extracts clinical guideline annotations directly from the web interface. A tailored JavaScript, executed within the browser’s developer console, dynamically collects guideline URLs, and streamlines the process of link extraction. These URLs are subsequently archived and analyzed using an automated, browser-based extraction system, which is constructed utilizing Playwright. Textual content is extracted from each guideline page and preserved in JSONL format.

Following data extraction, a data cleaning step is enacted to remove noise elements, including navigation menus, search queries, and formatting anomalies. Entries deemed either concise or irrelevant are discarded, resulting in a refined dataset. The text is then partitioned into segments of 500 characters, incorporating a 100-character overlap, through a sliding window approach. This segmented dataset subsequently serves as the basis for retrieval-focused tasks and evaluation.

TABLE 4.1: Summary of Datasets Used in This Study

Dataset	Source	Type	Original	Processed	Purpose and Content
CPIC	CPIC Database & API	Structured	CSV	JSONL	Genes, drugs, phenotypes, and clinical recommendations used for fine-tuning (instruction-based learning) and RAG.
ClinPGx	Web-based guidelines	Semi-structured	Text	JSONL	Annotated clinical guidelines and recommendations used for retrieval (RAG).

Table 4.1 highlights the key characteristics of the datasets used in this study, including their source, format transformation, and role in fine-tuning and retrieval processes.

4.2 Large Language Model Meta AI (LLaMA) Architecture

This research employs the Meta-Llama-3.1-8B-Instruct as the primary large language model for the fine-tuning process. Llama 3.1, an autoregressive transformer model fine-tuned for instruction, is developed for text generation tasks [56]. The 8B variant is selected as the foundational model due to its robust language modeling capabilities and its practical equilibrium between model scale and computational effectiveness. The model is sourced from Hugging Face and modified for the pharmacogenomics field through parameter-efficient fine-tuning.

The model's aim is to produce clinically relevant text derived from structured and processed pharmacogenomic input data. The original tokenizer setup accommodates a long context window; however, the fine-tuning procedure employs a maximum sequence length of 2048 tokens to maintain steady and manageable training. This model functions as the fundamental generative element for acquiring domain-specific pharmacogenomic patterns from the curated CPIC dataset.

4.2.1 LLaMA Fine-Tuning Configuration

To adapt the base model for pharmacogenomics, this study uses parameter-efficient fine-tuning with LoRA (Low-Rank Adaptation). This method specifically focuses on important transformer projection layers, such as q_proj, k_proj, v_proj, and o_proj. This approach allows the model to learn representations specific to a task while only changing a small number of parameters. This strategy decreases the number of trainable parameters, which significantly reduces the computational costs without losing the performance on the task, enabling model adaptation. Lowrank matrices are well suited for task specific adaptations [57].

In this study, the LoRA rank is set to 16, the LoRA alpha to 32, and the dropout rate to 0.05. This study chose these hyperparameters based on real-world testing and the limits of computers during training. It trains the model on 74,565 samples, and it took a long time because it uses large language models and long sequences (2048 tokens). Higher-rank configurations are initially considered, but it used more memory and took longer to train, which made them less efficient with the HPC resources. So,

choosing a moderate configuration ($r = 16$, $\alpha = 32$) to balance model performance, training stability, and computational efficiency. This choice is consistent with the LoRA framework, which shows that low-rank adaptation can significantly reduce trainable parameters while maintaining competitive performance compared to full fine-tuning [58].

Training employs a causal language modeling methodology, utilizing the Hugging Face Trainer for implementation. The training protocol is structured with a batch size of 1, integrating gradient accumulation across 8 steps, a warmup period of 50 steps, and total 1000 training steps. Checkpoints are produced every 50 steps, with a maximum of 2 checkpoints allowed. This design allows the model to be used practically, while also avoiding the computational costs of fully adjusting all its parameters.

4.2.2 LLaMA Instruction Formatting and Training Pipeline

Prior to the training phase, the integrated pharmacogenomics dataset undergoes a transformation into a textual format resembling instructional directives. This researcher then extracts pertinent data from each JSONL record, encompassing details pertaining to genes, medications, phenotypes, diseases, and guidelines, subsequently organizing this information into coherent paragraphs.

Afterward, the text samples are tokenized into sequences, which are then used as input for the model. To ensure consistent training, each sequence is organized according to a specific pattern. This pipeline made it possible to change structured pharmacogenomic data into a format that the language model could use to learn about a specific field.

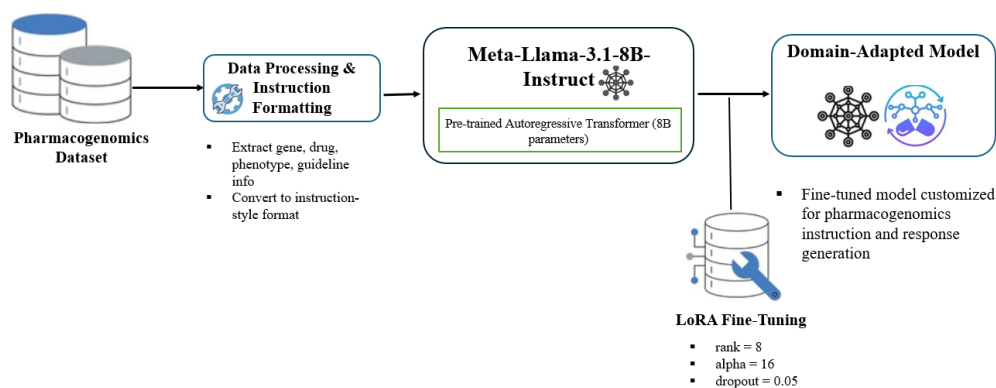


FIGURE 4.2: An outline of the Meta-Llama-3.1-8B-Instruct-based fine-tuning architecture.

Figure 4.2 shows the complete fine-tuning process. It uses LoRA to adjust the Meta-Llama-3.1-8B-Instruct model, which is then used to create responses specific to a particular area. This happens after the pharmacogenomics data is changed into instruction-style inputs.

4.3 Qwen Model Architecture

This research implements Qwen3-8B as an alternative large language model for pharmacogenomics applications. Qwen3-8B is an autoregressive language model developed for text generation and instruction adherence applications. The approach is constructed on comprehensive pretraining and post-training methodologies, facilitating robust performance in reasoning, instruction adherence, and multilingual tasks [59].

The Qwen3-8B model comprises roughly 8.2 billion parameters, featuring 36 transformer layers and a structured attention mechanism with 32 query heads and 8 key-value heads. The approach inherently accommodates a context length of 32,768 tokens, which can be augmented to 131,072 tokens with YaRN-based scaling methodologies [59]. These attributes render the model appropriate for processing structured and semi-structured pharmacogenomic data.

4.3.1 Qwen Fine-Tuning Configuration

Qwen3-8B is fine-tuned for pharmacogenomics applications using parameter-efficient techniques, specifically Low-Rank Adaptation (LoRA). This adaptation strategy concentrates on key transformer projection layers, including q_proj, k_proj, v_proj, and o_proj, thereby enabling efficient learning through the modification of a limited set of parameters.

The LoRA rank is set to 16, with an alpha parameter of 32 and a dropout rate of 0.05. Training is done using the Hugging Face Trainer, with a batch size of 1 and gradient accumulation over 8 steps. A learning rate of 1e-5 is used, along with a warmup period of 50 steps and a maximum of 1000 training steps. In addition, checkpoints are regularly saved to allow for model version tracking and recovery.

4.3.2 Qwen Instruction Formatting and Training Strategy

The dataset for Qwen is converted into structured instructional inputs intended for clinical decision assistance. Each training example includes an instruction, relevant

pharmacogenomic data, and the expected result. A key aspect of this setup involves employing a masked loss methodology, which ensures that only the tokens corresponding to the answer segment contribute to the training loss calculation. Tokens preceding the answer are excluded from consideration, thereby enabling the model to focus on refining its ability to generate responses, rather than preserving the structure of the prompt.

Moreover, important input representations are created by dynamically gathering and integrating pharmacogenomic data. This data includes information about genes, drugs, observed traits, diseases, and clinical guidelines.

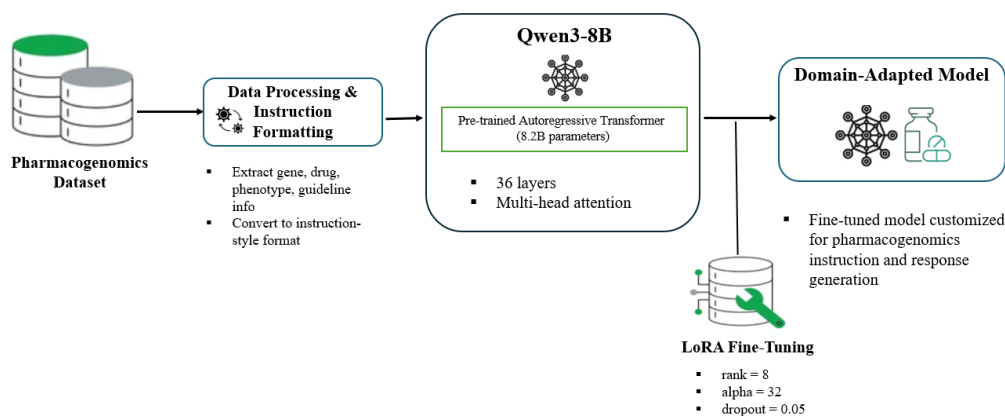


FIGURE 4.3: Summary of the Qwen3-8B fine-tuning architecture, demonstrating the conversion of pharmacogenomics data into instruction-style inputs and its modification via LoRA for domain-specific clinical response creation.

Figure 4.2 shows the complete process for adapting the Qwen3-8B model to the field of pharmacogenomics. The process is outlined sequentially, starting with data preparation and instruction formatting, then moving to fine-tuning using LoRA, and finally resulting in the creation of models specifically designed for the field's needs.

4.4 Retrieval-Augmented Generation (RAG) Framework

This research focuses on the development of a tailored RAG framework to facilitate pharmacogenomic decision-making using an extensive dataset stored on a high-performance computing (HPC) system. Given the dataset's magnitude, comprising hundreds of millions of records, a traditional vector database methodology is not utilized. A file-based hybrid retrieval mechanism is developed to guaranty scalability and efficiency while preserving strong semantic relevance in the retrieved outcomes.

Data Preparation for Retrieval

The pharmacogenetic data set is converted into a structured JSONL format, with each row representing an autonomous unit of clinical information. During preprocessing, many heterogeneous columns are categorized into uniform semantic categories, including gene, drug, allele, or genotype, phenotype, and recommendation-related content. To standardize the different naming conventions used in datasets, this research establishes a set of key aliases for each concept.

Each row is then condensed into a brief textual summary that captures the core pharmacogenomic attributes. To optimize retrieval and processing, excessively long text fields are truncated. This procedure ensures that every entry operates as a self-contained unit of evidence during the retrieval process.

4.4.1 Hybrid Retrieval Mechanism

The retrieval system employs a hybrid pipeline that integrates rapid lexical filtering with semantic re-ranking. This method is selected to address the complexities of a substantial dataset while maintaining semantic precision.

Lexical Pre-Filtering over Large-Scale JSONL Data

The first step in the retrieval process involves lexical filtering, which is applied directly to the JSONL dataset residing within the High-Performance Computing (HPC) filesystem. When a query is submitted, the system employs efficient pattern matching techniques to identify pertinent items, such as gene names, medication names, and allele information. Subsequently, the retrieved entities are used to formulate search patterns that align with the dataset's textual representation.

A command-line search utility is employed to scrutinize the dataset, isolating only those rows that contain the specified keywords. This approach eliminates the requirement to load the entire dataset into memory, operating directly on the file system, which is particularly advantageous for handling large datasets. As a result, the search space is reduced from hundreds of millions of records to a more manageable subset of potential rows.

Semantic Re-Ranking of Candidate Results

After the lexical filtering, the selected candidate rows are further refined by analyzing how similar their meanings are. A pre-trained sentence embedding model transforms both the user query and the candidate texts into dense vector representations. The cosine similarity metric is then employed to evaluate the degree of similarity between the question and each candidate.

Following this, the candidates undergo sorting according to their semantic significance, and only those with the highest rankings are retained. This phase ensures that the system’s operation is not solely reliant on exact keyword matches; rather, it incorporates contextual and semantic relationships between the query and the gathered data. Through the application of semantic re-ranking exclusively on a constrained candidate set, the system achieves a balance between computational efficiency and retrieval accuracy.

Integration of Guideline-Based Knowledge

Alongside the large pharmacogenomic dataset (CPIC), a distinct guideline dataset (ClinPGx) comprising curated clinical recommendations is integrated into the retrieval process. In contrast to the primary dataset, this guideline dataset is descriptive and based on natural language guidelines, where meaning is more important than exact keyword matching. Due to this, lexical pre-filtering is not required for the guideline dataset, and retrieval is performed directly using semantic similarity.

For each query, semantic similarity is assessed between the query and all guideline entries utilizing the identical embedding model. The most relevant guideline documents are thereafter chosen and incorporated as evidence. This dual retrieval approach enables the system to integrate extensive raw data with superior curated clinical knowledge, enhancing the reliability of the generated results.

Prompt Construction and Evidence Fusion

After gathering the necessary information from both the main dataset and the guideline dataset, the system creates a structured prompt for the language model. This prompt includes the user’s question and the relevant evidence, organized into separate sections.

The model’s design prioritizes generating responses grounded in the provided data while also identifying situations where information is lacking. This structure encourages the production of responses based on established facts, thereby reducing the likelihood of generating inaccurate information. Furthermore, this methodology incorporates multiple sources of evidence within a single prompt, enabling the model to analyze both structured pharmacogenomic data and recommendations derived from clinical guidelines.

Response Generation Using Fine-Tuned Language Models

The last step in the pipeline is to use a refined large language model to create the response. This research entails the adaptation of transformer-based models, such as Qwen and LLaMA, utilizing parameter-efficient fine-tuning techniques. This facilitates the model’s comprehension of pharmacogenomic terminology pertinent to the domain

and generates outputs beneficial for therapeutic applications. The final stage in the process involves employing a refined large language model to formulate the response.

This enables the model to comprehend pharmacogenomic terminology pertinent to the field and generate therapeutically valuable outputs. The model receives the prompt and utilizes the evidence it has collected to formulate a response. To ensure consistency and minimize variability in the outputs, deterministic decoding configurations are employed. Then, the final response is post-processed to make it clearer and shorter.

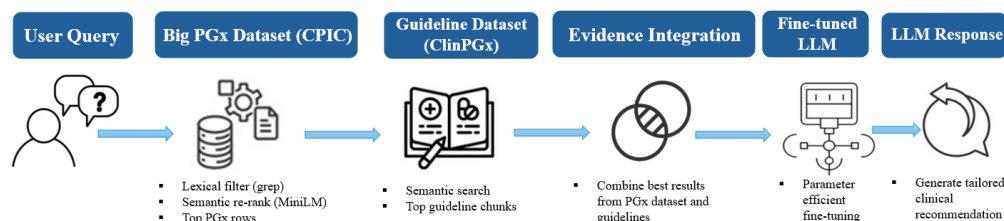


FIGURE 4.4: Hybrid Retrieval-Augmented Generation (RAG) Architecture for Pharmacogenomic Decision Support.

The suggested hybrid Retrieval-Augmented Generation (RAG) framework for pharmacogenomic decision assistance is shown in Figure 4.4. Key entities including genes, drugs, and alleles are extracted from a user query at the start of the procedure.

The system then gets the information it needs from two different places. First, lexical filtering and semantic re-ranking are used to quickly narrow down a huge pharmacogenomic dataset (CPIC) to a smaller group of candidate records. Second, semantic similarity is used to get clinically useful suggestions from a smaller but very carefully chosen guideline dataset (ClinPGx).

The evidence gathered from both sources is then put together to produce a single knowledge context. This data is sent to a large language model (LLM) that has been changed using parameter-efficient methods (LoRA) so that it can better understand words and phrases related to pharmacogenomics and clinical reasoning.

Finally, the model gives an answer that is both clinically relevant and aware of the situation. It does this by making sure that the output is based on facts from research that was found and follows accepted pharmacogenomic principles.

4.5 Proposed System Framework and Workflow

This section outlines the proposed framework and workflow, showing how pharmacogenomics data and clinical guidelines are integrated using a RAG pipeline. The architecture incorporates data preprocessing, evidence retrieval, and LoRA-based fine-tuning to produce clinically relevant recommendations.

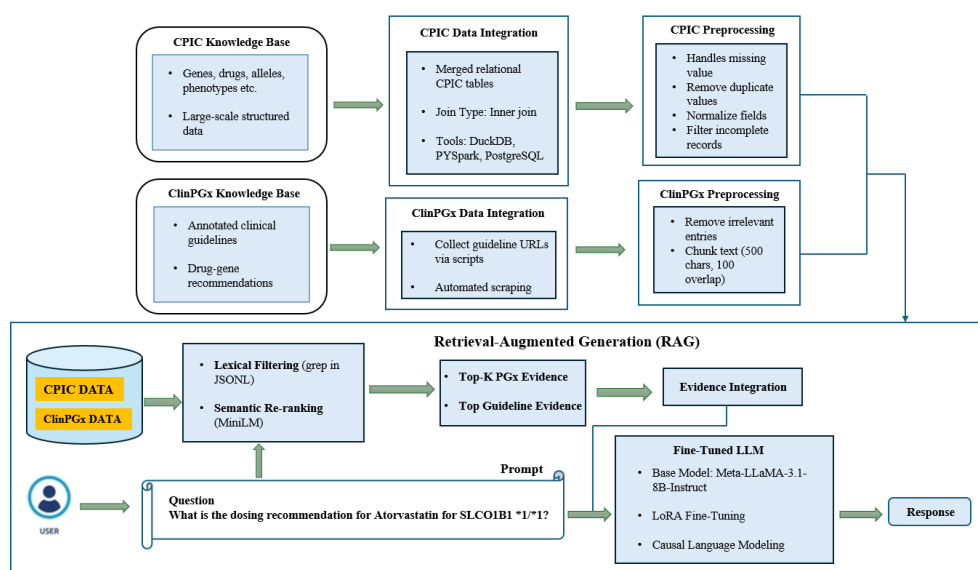


FIGURE 4.5: Overview of the proposed methodology integrating CPIC and ClinPGx data preprocessing with a Retrieval-Augmented Generation (RAG) framework and a LoRA fine-tuned LLM for clinical recommendation generation.

Figure 4.5 depicts a end-to-end pipeline for pharmacogenomics-based decision support. The solution employs structured CPIC data and ClinPGx information in accordance with established protocols, incorporating specific preprocessing and data integration procedures. A retrieval-augmented generation (RAG) module handles user queries during inference, collecting relevant pharmacogenomic evidence and clinical guideline data. This integrated evidence serves as input for a LoRA fine-tuned large language model that generates clinical recommendations. The framework ensures the implementation of both data-driven reasoning and guideline-based validation, enhancing the reliability and relevance of the outputs.

Chapter 5

Empirical Analysis

This chapter provides an empirical assessment of the proposed pharmacogenomic decision support system. It employs a systematic methodology, beginning with baseline language models and progressing to RAG and LoRA-based fine-tuning. The aim is to improve the accuracy and clinical relevance of model outputs by incorporating hybrid retrieval techniques with extensive pharmacogenomic data. To evaluate the system’s performance, various model configurations are examined, emphasizing their advantages and disadvantages.

5.1 SHARCNET HPC Usage

This study is performed on the Compute Canada SHARCNET clusters, Graham and Nibi, utilizing the SLURM scheduler. The computational solution offers access to NVIDIA H100 GPUs with 80GB HBM3 memory, and CPUs with 8 cores for interactive runs up to 32 cores for bigger workloads. 128 GB of RAM is also used to facilitate large-scale data processing and model training. This high-performance computing environment can then efficiently support training and retrieval activities with huge language models.

5.2 Model Variants

This research explores several model configurations to understand the effect of fine-tuning and retrieval augmentation. The basis models are Meta-LLaMA-3.1-8B-Instruct, Qwen3-8B. Then these models are fine-tuned by parameter-efficient fine-tuning via LoRA. Here, LoRA is applied to the key transformer projection layers, including q_proj,

k_proj, v_proj, and o_proj. It set rank = 16, alpha = 32 and dropout = 0.05 to manage the trade-off between performance and computation cost.

Training is done using a causal language modeling approach, implemented with the Hugging Face Trainer. The configuration is a gradient accumulation of 8 steps and a max sequence length of 2048 tokens for robust training while handling memory constraints.

5.3 Hybrid Retrieval Mechanism

This study proposes a hybrid retrieval system to improve the quality of recovered evidence. The first step consists of doing a lexical pre-filtering, where we extract key entities, such as genes, medications, alleles and suggestions, from the query. Then, pattern matching is performed on the JSONL dataset using command-line search, avoiding full memory loading and enhancing efficiency.

Next, it uses a sentence transformer model (MiniLM) to incorporate the question and candidate rows and perform semantic re-ranking. These embeddings are utilized to compute a similarity score, and to find the most relevant results.

ClinPGx guideline data is also managed independently. Dataset is entirely loaded to RAM and processed with direct semantic similarity, no lexical filtering, unlike the lexical pipeline. This architecture allows for a more precise retrieval of guideline-based clinical recommendations.

5.4 Model Variants Evaluated

This thesis tests three configurations: the standard model, the LoRA fine-tuned model and the hybrid LoRA + RAG model. This allows to analyze the effect of parameter-efficient fine-tuning and retrieval augmentation on the overall model performance.

5.5 Evaluation Metrics

This study uses a comprehensive evaluation method to assess the accuracy of structured predictions and the quality of generated responses to measure the performance of large language models.

Three types of methodologies are used to evaluate in this research: classification-based evaluation, which helps model predicting structured clinical categories accurately; text-based evaluation, which assesses semantic similarity and the quality of the generated output; and manual evaluation, which includes correctness, completeness, relevance, and clarity.

5.5.1 Classification Evaluation

The evaluation metric is the method that measures how well the classifier works. Different metrics are utilized to measure different parts of how well a model works, especially when it comes to structured prediction tasks [60]:

- **Accuracy:** The accuracy metric shows how many of the predictions were right compared to the total number of cases that were examined.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5.1)$$

- **Precision:** Precision measures how many of the expected patterns in a positive class are actually correct.

$$Precision = \frac{TP}{TP + FP}, \quad (5.2)$$

- **Recall:** Recall is a way to find out how many positive patterns were accurately identified.

$$Recall = \frac{TP}{TP + FN}, \quad (5.3)$$

- **F-Measure (FM):** This measure shows the harmonic mean of the recall and precision values.

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \quad (5.4)$$

where:

TP = True Positives,

TN = True Negatives,

FP = False Positives,

FN = False Negatives.

5.5.2 Text-Based Evaluation

Text-based evaluation measures how similar the meaning of the generated text is to the meaning of the reference text. This ensures the meaning of the response remains the same. The objective stated in prior work is to assess semantic equivalence [61].

- **BERTScore:** BERTScore finds the similarity between two phrases by summing up the cosine similarities between the embeddings of their tokens. It doesn't employ exact word matching like typical n-gram metrics do. Instead, it uses contextual representations, where we calculate token similarity using contextual embeddings [61].

$$\text{BERTScore} = \frac{1}{N} \sum_{i=1}^N \max_j \cos(e_i, e_j) \quad (5.5)$$

where:

N = number of tokens in the prediction,

e_i = embedding of the i -th token in the prediction,

e_j = embedding of the j -th token in the reference text,

$\cos(e_i, e_j)$ = cosine similarity between token embeddings.

- **Cosine Similarity:** It is a metric that finds the similarity between two vectors and is best calculated as the dot product of two normalized vectors.

This measure is often used in natural language processing tasks since it has been used to solve different text mining problems, like text categorization, text summarization, information retrieval, and question answering [62].

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (5.6)$$

where:

A, B = vector representations of two texts,

$A \cdot B$ = dot product of the vectors,

$\|A\|, \|B\|$ = magnitudes (norms) of the vectors.

- **LLM-as-a-Judge:** LLMs have been very successful in many areas, which has led to the idea of "LLM-as-a-Judge," where LLMs are used to judge difficult tasks. "LLM-as-a-Judge" means using LLMs to judge activities, acts, or choices based on rules, criteria, or preferences that have already been set [63].

Industry resources suggest that using LLMs as judges allows models to automatically assess outputs against predetermined criteria. This approach provides a scalable method for quality control, eliminating the need for continuous human oversight [64].

Employing a LLM as a evaluation tool presents a compelling prospect, as it merges the strengths of human judgment with the efficiency of automated evaluation. This methodology integrates the scalability inherent in automated techniques with the nuanced, context-sensitive reasoning characteristic of expert assessments [63].

$$S_i = f_{\text{LLM}}(\text{gold}_i, \text{prediction}_i) \quad (5.7)$$

$$\text{Final Score} = \frac{1}{N} \sum_{i=1}^N S_i \quad (5.8)$$

where:

S_i = score assigned by the LLM for the i -th sample,

f_{LLM} = evaluation function of the language model,

gold_i = reference (correct) answer,

prediction_i = model-generated answer,

N = total number of evaluated samples.

5.5.3 Manual Evaluation

Manual evaluation is essential for assessing the outputs of LLMs, particularly in domains where precision and reliability are paramount. Human review entails the assessment of responses based on criteria such as accuracy, relevancy, completeness, hallucination mitigation, RAG accuracy, and RAG relevancy. These benchmarks ensure that the model's answers are both factually correct and relevant to the question, supported by the appropriate context [51].

However, manual evaluation has problems with scaling. Previous research has pointed out some issues with human evaluation, which makes it both time-consuming and resource-intensive [65].

$$S_i = \frac{A_i + R_i + C_i + L_i}{4} \quad (5.9)$$

$$\text{Average Score}_5 = \frac{1}{N} \sum_{i=1}^N S_i \quad (5.10)$$

$$\text{Average Score}_{20} = 4 \times \text{Average Score}_5 \quad (5.11)$$

$$\text{Total Score}_{400} = N \times \text{Average Score}_{20} \quad (5.12)$$

where:

A_i = accuracy score for sample i ,

R_i = relevance score for sample i ,

C_i = completeness score for sample i ,

L_i = clarity score for sample i ,

S_i = average score for sample i out of 5,

N = total number of questions, here $N = 20$.

Each question is evaluated using four criteria: accuracy, relevance, completeness, and clarity. Each criterion is scored out of 5. Therefore, one question has a maximum score of $4 \times 5 = 20$. Since there are 20 questions, the maximum total score is $20 \times 20 = 400$.

5.6 Summary

This chapter outlines the empirical assessment of the proposed pharmacogenomic decision support system, advancing from baseline language models to LoRA-based fine-tuning and retrieval-augmented generation (RAG). Each component is methodically designed to enhance the precision, relevance, and clinical validity of the produced responses.

The hybrid retrieval method, integrating lexical filtering and semantic re-ranking, facilitates the effective management of extensive pharmacogenomic data while preserving contextual relevance. The LoRA and RAG arrangement exhibits enhanced performance, especially in generating organized and therapeutically relevant outputs.

Chapter 6

Experiments and Results

In this chapter, the experiment design and evaluation of the proposed system using CPIC and ClinPGx datasets are presented. It compares the performance of basic, LoRA and RAG-based models and evaluates their effectiveness in generating clinically significant responses.

6.1 Experimental Setup

This section describes the computing environment, model configurations, and assessment framework employed to evaluate the proposed pharmacogenomics question-answering system.

All experiments are performed on the Compute Canada high-performance computing (HPC) infrastructure, notably utilizing the Nibi cluster. Experiments utilizing GPUs are conducted on NVIDIA H100 GPUs, with 64 GB of memory and 8 CPU cores per task, while supplementary preprocessing and assessment procedures are carried out in CPU-based environments when GPU resources are inaccessible. The utilization of HPC resources facilitates the effective management of extensive pharmacogenomics datasets and model inference activities.

The implementation environment utilizes Python 3.11 and PyTorch 2.3 as the deep learning framework. The tests use the Hugging Face ecosystem, which includes the transformers library (version 4.44.2), PEFT (version 0.12.0) for fine-tuning that uses fewer parameters, and Accelerate (version 0.33.0) for running the model in the best way. The Sentence-Transformers library with the "all-MiniLM-L6-v2" model is used for semantic retrieval at the sentence level. All dependencies are managed in separate virtual environments to make sure they can be reproduced and to avoid version conflicts.

A centralized caching strategy is established using environment variables (HF_HOME, TRANSFORMERS.CACHE, and HF_DATASETS.CACHE) to improve model loading and minimize repetitive downloads, directing to a shared project directory. This greatly enhances execution efficiency across multiple iterations.

The experimental pipeline comprises three principal model configurations: (i) base large language models (LLaMA and Qwen), (ii) LoRA fine-tuned models, and (iii) Retrieval-Augmented Generation (RAG) integrated with LoRA fine-tuning. The RAG framework extracts relevant pharmacogenomics evidence from preprocessed datasets and quickly integrates it into the model to enhance the factual accuracy and relevance of the domain.

6.1.1 Choice of Embedding Model for Semantic Re-Ranking

Initially, computing efficiency is a primary concern. The pharmacogenomics dataset utilized in this study is substantial (about 497 GB post-preprocessing), and semantic re-ranking is implemented subsequent to lexical filtering on possibly thousands of candidate rows per query. MiniLM-L6-v2 comprises 22.7 million parameters and generates 384-dimensional embeddings [66]. It facilitates rapid inference on CPU nodes within the SHARCNET HPC cluster. In contrast BioBERT and PubMedBERT are both derived from the BERT-base architecture and possess around 110 million parameters each, rendering them considerably larger than lightweight embedding models like MiniLM [67]. These make them unsuitable for large-scale or real-time applications without significant computational resources. These models are computationally expensive [68].

Recent studies indicated that general-purpose embedding models can match or, in certain instances, surpass domain-specific biomedical models in short-context clinical semantic search tasks [69]. This study functions within a short-context retrieval framework, in which both user inquiries and returned evidence consist of concise text segments, such as brief questions (e.g., “What is the dose adjustment for CYP2C19) and brief guideline excerpts. The method compares brief snippets rather than processing entire papers to identify the most pertinent information. This indicates that specialist models such as BioBERT or PubMedBERT are not necessarily essential for efficient retrieval in these contexts.

The SHARCNET environment restricts memory and CPU utilization for non-GPU tasks. MiniLM operates efficiently on 8 CPU cores with negligible memory overhead, but bigger biological models typically necessitate GPU acceleration to achieve satisfactory latency. Allocating GPU resources for LLM fine-tuning and inference ensures that the retrieval pipeline is both reproducible and economical, while maintaining high semantic ranking quality.

6.2 Classification Performance Evaluation

6.2.1 LLaMA Based Models Evaluation

To evaluate the effectiveness of the LLaMA-based models, standard classification metrics such as accuracy, precision, recall, and F1-score are computed across all clinical categories.

TABLE 6.1: Overall Performance Comparison of LLaMA Models

Model	Accuracy	Precision	Recall	F1-Score
Base LLaMA	0.435	0.36	0.44	0.33
LoRA LLaMA	0.556	0.25	0.26	0.21
RAG + LoRA LLaMA	0.91	0.83	0.91	0.87

Table 6.1 illustrates the overall classification performance of the LLaMA-based models. The base LLaMA model doesn't work very well; its accuracy is 0.435 and its macro F1-score is 0.33. The model has trouble putting pharmacogenomics results into the right clinical categories. The LoRA fine-tuned model is a little better (0.556), but its macro F1-score goes down to 0.21. This means that the model is more sure of some of its predictions, but it has trouble making predictions that work for all classes.

The base model and the LoRA-only model are not as good as the RAG + LoRA LLaMA model. It gets it right 0.91 times out of 10 and has an overall F1-score of 0.87. This big improvement shows how well fine-tuning and retrieval-based data work together. This makes it possible for the model to make more accurate and useful predictions.

TABLE 6.2: Class-wise Performance Comparison

Model	Class	Precision	Recall	F1-Score
Base LLaMA	Decreased response	0.42	0.97	0.58
	Increased response	0.41	0.90	0.56
	No association	1.00	0.35	0.51
LoRA LLaMA	Decreased response	0.19	0.63	0.29
	Increased response	0.24	0.06	0.09
	No association	0.84	0.59	0.69
RAG + LoRA LLaMA	Decreased response	0.79	0.81	0.80
	Increased response	0.71	0.92	0.80
	No association	0.92	0.92	0.92

Table 6.2 shows a thorough comparison of performance by class for three types of pharmacogenomic outcomes. The job of classification is based on three main types of pharmacogenomic outcomes. The Decreased response class includes situations in which a drug is less successful as a treatment because of genetic differences. This could mean that the dosage needs to be changed or another treatment needs to be found. In this class, "Increased response," the drug's effect is stronger than predicted, which may raise the risk of toxicity or bad reactions. If the class is "No significant association," it means that there is no clinically meaningful connection between the genetic variation and drug response. This means that normal treatment methods can be used. Based on genetic differences between patients, these categories show outcomes that are clinically important and are used in pharmacogenomics to help make treatment choices.

Based on the class-wise results, the base LLaMA model acts in an uneven way. It has a high recall for both decreased (0.97) and increased responses (0.90), but it's not very precise, which means it makes some false positive predictions. The model has perfect precision (1.00) but low recall (0.35) for the No significant association class. This means that it makes conservative predictions about this class and doesn't find many real cases.

The LoRA-exclusive model does not perform well across many classes. It does show slightly better result for some categories, but it does not do well at increased response classification, with an F1-score of 0.09. This means that only fine-tuning LoRA is not enough to find the complex links in pharmacogenomics data, which could lead to worse generalization performance.

The RAG + LoRA LLaMA type, on the other hand, is balanced and works well in all roles. Both "Decreased response" (F1 = 0.80) and "Increased response" (F1 = 0.80) work well because it strikes a good mix between precision and recall. "No significant association" (F1 = 0.92) is another case where it works well. The fact that the model behaves in a balanced way says that it can correctly tell the difference between different clinical outcomes without being biased toward any dominant class.

The model can incorporate relevant domain knowledge a lot better when retrieval-augmented generation is added. This helps sort things better. This is very important in pharmacogenomics, where it's essential to be able to accurately predict all clinical categories for making sound decisions.

6.2.2 Qwen Based Models Evaluation

TABLE 6.3: Overall Performance Comparison of Qwen Models

Model	Accuracy	Precision	Recall	F1-Score
Base Model	0.91	0.92	0.90	0.91
Finetune (LoRA Model)	0.32	0.46	0.25	0.19
RAG + Finetune(LoRA)	0.88	0.72	0.75	0.72

Table 6.3 shows how well baseline Qwen can sort things. The base Qwen model is the best on its own, with a macro F1-score of 0.91, a recall of 0.90, a precision of 0.92, and an accuracy of 0.91. This means that the pretrained Qwen model is very good at reasoning and can always put pharmacogenomic relationships into different clinical groups.

The LoRA-only Qwen model exhibits a significant decline in performance, with accuracy reduced to 0.32 and F1-score to 0.19. This indicates that parameter-efficient fine-tuning alone was inadequate to maintain the initial generalization capacity of the underlying model. The adaptor probably became overly specialized to the fine-tuning inputs, compromising the balanced decision-making capability of the pretrained model.

The RAG + LoRA Qwen model has an accuracy of 0.88, with precision, recall, and F1-score metrics of 0.72, 0.75, and 0.72, respectively. Its numerical scores are slightly lower than those of the standard model, but the addition of retrieval-augmented generation is a big plus: predictions are backed up by external pharmacogenomic evidence instead of just relying on stored parameters. This makes the model more reliable and useful in real-world decision support settings, where evidence-based responses are more important than just being accurate on benchmarks.

TABLE 6.4: Class-wise Performance Comparison of Qwen-Based Models

Model	Class	Precision	Recall	F1-Score
Base Model	Decreased response	0.99	0.88	0.94
	Increased response	0.82	0.75	0.79
	No significant association	0.79	0.93	0.86
Finetune (LoRA) Model	Decreased response	0.33	0.03	0.06
	Increased response	0.38	0.06	0.11
	No significant association	0.29	0.91	0.44
RAG + Finetune (LoRA)	Decreased response	0.51	0.69	0.59
	Increased response	0.71	0.94	0.81
	No significant association	0.85	0.85	0.85

Table 6.4 delineates the class-wise comparison for the three pharmacogenomic result categories: Decreased response, Increased response, and No significant correlation. The Decreased response class pertains to instances where genetic variation diminishes treatment efficacy, either necessitating dosage modification or the use of an alternate medicine. The Increased response class signifies a pharmacological effect that exceeds expectations, thereby elevating the risk of toxicity or unpleasant effects. The No Significant Association class denotes an absence of clinically relevant gene–drug interactions, allowing for adherence to established treatment protocols.

The base Qwen model demonstrates robust performance across all categories according to the class-wise results. The F1-scores achieved are 0.94 for decreased response, 0.79 for increased response, and 0.86 for no significant correlation. This balanced performance verifies that the pretrained model can successfully differentiate among various clinical outcomes without necessitating supplementary retrieval assistance.

The LoRA-only Qwen model model performs poorly across most classes, yielding F1-scores of 0.06, 0.11, and 0.44, respectively. These findings reinforce the notion that LoRA fine-tuning alone was insufficient for efficiently adapting the model for robust pharmacogenomic classification.

The RAG + LoRA Qwen model exhibits more balanced and clinically significant behavior compared to the LoRA-only model. It attains F1-scores of 0.59 for diminished response, 0.81 for augmented response, and 0.85 for no substantial correlation. The

robust recall (0.94) for heightened reaction signifies that the model is proficient in detecting instances where patients may encounter intensified pharmacological effects or higher toxicity risk.

Although the original Qwen model achieves superior raw classification metrics, the RAG + LoRA Qwen framework presents a more pragmatic architecture for pharmacogenomics by integrating language modeling with the retrieval of pertinent external evidence. This evidence-based design is particularly crucial in medical fields, where accuracy, transparency, and reliance on reputable knowledge sources are vital.

6.2.3 Comparative Analysis (LLaMA vs Qwen)

To evaluate the two model families, the top-performing evidence-enhanced versions, specifically RAG + LoRA LLaMA and RAG + LoRA Qwen, are analyzed. Both methodologies surpassed their LoRA-exclusive equivalents, indicating that retrieval augmentation markedly enhances pharmacogenomic classification efficacy.

The Qwen-based framework remains competitive and shows strong detection of the Increased response class, with high recall indicating effective sensitivity in identifying potentially heightened medication effects or toxicity-related occurrences. Additionally, the Qwen base model exhibits strong independent functionality without retrieval assistance. In pharmacogenomics, standalone accuracy is not the only goal; results must also be evidence-based, transparent, and clinically reliable. Therefore, the RAG + LoRA Qwen setup is more robust than the base model, as it incorporates external guideline evidence to improve its predictions.

The RAG + LoRA LLaMA system performs slightly better than the RAG + LoRA Qwen system, with an accuracy of 0.91 and a macro F1-score of 0.87. The RAG + LoRA Qwen system has an accuracy of 0.88 and a macro F1-score of 0.72. The LLaMA-based paradigm shows better class-wise performance consistency across the pharmacogenomic outcome categories that were tested. The Qwen basic model shows strong raw performance, but evidence-based architectures are better for medical decision support. So, RAG + LoRA LLaMA is chosen as the best final model because it combines higher accuracy, fairer classification performance, and better reasoning through retrieval.

6.3 Text Based Evaluation

6.3.1 Evaluation Analysis of LLaMA Models Performance

TABLE 6.5: Text Based Evaluation Metrics Comparison

Model	Cosine Similarity	BERTScore (F1)	LLM As A Judge (Mistral-7B-Instruct-v0.2)
Base Model	0.2117	0.8069	43.3%
Finetune Model	0.1662	0.7949	56.5%
Finetune+RAG	0.6707	0.8566	88%

Table 6.5 shows the text-based evaluation of the LLaMA models using semantic similarity and qualitative judgment measures. The LLM-as-a-judge metric checks whether the generated answers are right and clinically relevant, while the cosine similarity and BERTScore (F1) measures how closely the generated answers match the reference answers in terms of semantic meaning.

The base model has a BERTScore of 0.8069 and a cosine similarity of 0.2117. This means that the generated answers do have some semantic overlap with the reference answers, but the total alignment is still not perfect. The LLM-as-a-judge score of 43.3% means that many of the answers are not completely correct or useful in a clinical setting.

The fine-tuned (LoRA) model has a small drop in both BERTScore (0.7949) and cosine similarity (0.1662). This means that fine-tuning doesn't make the responses more semantically aligned. However, the LLM-as-a-judge score goes up to 56.5%, which means the answers are more correct, even though the similarity metrics aren't as strong.

All of the assessment metrics demonstrate a substantial enhancement in the Fine-tuned + RAG model. The cosine similarity increases to 0.6707, indicating that the responses are more semantically similar to the reference answers. The BERTScore increases to 0.8566, indicating that the contextual similarity and token-level similarity are both improved. The LLM-as-a-judge score reaches 88%, indicating that the responses generated are not only semantically similar but also more accurate and clinically significant.

These findings indicate that fine-tuning is effective in retrieval-augmented generation. Standalone models struggle to produce consistent and accurate outcomes; however, the RAG-enhanced model significantly improves clinical accuracy and semantic similarity by incorporating external evidence.

6.3.2 Performance Analysis of Qwen-Based Models

TABLE 6.6: Text-Based Evaluation Metrics Comparison

Model	Cosine Similarity	BERTScore (F1)	LLM AS A Judge
Base Model	0.6614	0.8817	10.2%
Finetune (LoRA) Model	0.6531	0.8783	89.8%
Finetune (LoRA)+RAG	0.7524	0.8585	88%

The results of the text-based evaluation of the Qwen-based models are shown in Table 6.6. The comparison shows that the base, fine-tuned, and retrieval-enhanced versions all work well to produce pharmacogenomic responses that match the reference outputs.

The base Qwen model has a BERTScore of 0.8817 and a cosine similarity of 0.6614, which means that its semantic interpretation is very similar to that of the reference replies. However, its LLM Judge score is only 10.2%, which means that many of the responses are wrong, incomplete, or not appropriate for a medical setting, even though they are phrased in a similar way. This shows that lexical similarity alone may not be a reliable way to get medical responses.

The fine-tuned (LoRA) Qwen model has semantic scores that were very close to each other: a cosine similarity of 0.6531 and a BERTScore of 0.8783. The LLM Judge score goes up a lot to 89.8%, which means that fine-tuning create the outputs much more accurate, useful, and relevant to the field.

The Fine-tuned + RAG Qwen model obtains the highest cosine similarity (0.7524) and a strong BERTScore (0.8585), while achieving an LLM Judge score of 88%. These results show that retrieval augmentation improves semantic alignment and helps generate more evidence-grounded and contextually appropriate responses. Although the judge score is slightly lower than the LoRA-only model, the RAG-enhanced system benefits from external knowledge support.

Overall, the results indicate that the base Qwen model has strong general language capability, but fine-tuning and retrieval augmentation are necessary to produce clinically dependable outputs. Therefore, the Qwen + RAG framework remains a more practical choice for pharmacogenomic decision support applications where trustworthy and evidence-supported responses are required.

6.3.3 Performance Comparison Between LLaMA and Qwen

The textual evaluation demonstrates that both LLaMA- and Qwen-based frameworks benefit from fine-tuning and retrieval augmentation. In both model categories, the RAG-enhanced versions produce more reliable and semantically coherent responses than the standalone base models.

The Fine-tuned + RAG model for LLaMA has balanced performance, with enhancements in cosine similarity, BERTScore, and LLM Judge score. However, the Fine-tuned + RAG Qwen model achieves the highest cosine similarity across Qwen variations and sustains a robust qualitative assessment score, demonstrating slightly better overall performance. Despite the fundamental Qwen model demonstrating high semantic similarity scores, its low judge score suggests that semantic overlap alone does not guarantee clinically valuable responses.

6.4 Manual Evaluation

The manual assessment is based on a number of qualitative factors, including accuracy, relevance, completeness, and clarity. Accuracy checks how well the generated response matches the reference answer and whether it provides clinically accurate information. Relevance checks whether the response is directly related to the question in a way that makes sense. Completeness evaluates the inclusion of all essential and significant elements of the response. These definitions align with previous research in pharmacogenomic AI evaluation, wherein human experts evaluate the quality of generated responses based on their accuracy, relevance, and comprehensiveness of information [51]. Furthermore, clarity is assessed to determine the degree to which the information is provided to the user in a clear and comprehensible manner.

TABLE 6.7: Comparison of Qwen and LLaMA Model Responses for Pharmacogenomic Queries

#	Question	Qwen Model Answer	Llama Model Answer
1.	What's the guideline for HLA-B and abacavir?	The FDA requires HLA-B*57:01 testing before abacavir prescription; if positive, abacavir is contraindicated due to severe hypersensitivity risk. Alternatives must be chosen per guidelines.	Abacavir should not be prescribed to individuals with an HLA-B*5801 genotype due to increased risk of hypersensitivity reaction. Use alternative antiretroviral therapy options if available. again

#	Question	Qwen Model Answer	Llama Model Answer
2.	What is the role of VKORC1 in dosing of acenocoumarol?	VKORC1 influences acenocoumarol dosing by affecting individual variability in drug response. Genetic variations like rs9923231 AG or AA require adjusted doses and increased monitoring to prevent bleeding risks. The DPWG recommends genotyping to guide safe and effective treatment.	There is no direct relationship between VKORC 1 and acenocoumarol dosing based on the provided evidence. However, it does mention that the Dutch Pharmacogenetic Working Group recommends genotyping patients before starting acenocoumarol. However, another piece of evidence shows that VKORC-1 rs9923213 AG reduces the required dose of acenocoumarol, but due to current practices, there's little to no increased risk.
3.	What is the guideline recommendation for MT-RNR1 and amikacin use?	The CPIC guideline recommends avoiding amikacin in patients with MT-RNR1 variants associated with increased aminoglycoside-induced hearing loss risk, unless the benefits of treatment outweigh the risks. If no alternatives are available, use for the shortest duration possible, with monitoring and consultation.	Use an alternative antibiotic if possible; otherwise, consult an ID specialist and monitor closely.
4.	What is the pharmacogenomic association between HLA-B and allopurinol hypersensitivity?	HLA-B*58:01 allele is associated with increased risk of allopurinol hypersensitivity reactions, leading to severe cutaneous adverse reactions like Stevens-Johnson syndrome and toxic epidermal necrolysis. CPIC guidelines recommend HLA-B*58:01 screening before allopurinol use in high-risk populations such as Koreans with CKD and individuals of Han Chinese or Thai descent.	Based on the provided evidence, the pharmacogenomic association between HLA-B and allopurinol hypersensitivity is linked to the HLA-B*5801 allele.

#	Question	Qwen Model Answer	Llama Model Answer
5.	What is the recommendation for ABCG2 in relation to allopurinol therapy?	The Royal Dutch Pharmacists Association recommends a dose increase for patients with ABCG2 GT or TT genotypes (equivalent to GK or KK) during allopurinol therapy. Patients with these genotypes should use 1.25 times the standard dose, following a modified titration schedule.	the recommendation for ABCG2 in relation to allopurinol therapy is: * For patients with the GT or TT genotypes (equivalent to GK and KK amino acids), a dose increase is recommended. * For other genotypes, no specific recommendation is mentioned, but it is suggested to adhere to the gene-drug guideline if the genotype is available. Note: The evidence does not provide a clear recommendation for all possible genotypes, so the answer is based on the information provided by the DPWG.
6.	What is the guideline recommendation for CYP2C9 and acenocoumarol?	Based on the DPWG guideline, there are currently no recommendations for acenocoumarol dosing based on CYP2C9 genotypes. No action is needed for this gene-drug interaction.	However, according to the DPWG guideline, individuals with CYP2C9*2 or CYP2C9*3 alleles should have their INR monitored more frequently when taking acenocoumarol. Therefore, I would recommend monitoring INR more frequently for this patient.
7.	What is the recommended clinical action for a patient with a CYP2D6 poor metabolizer phenotype taking Codeine?	Based on the evidence, the recommended clinical action for a patient with a CYP2D6 poor metabolizer phenotype taking Codeine is to avoid codeine and use an alternative analgesic agent. Consult a clinical pharmacist for further guidance.	Based on the provided evidence, the recommended clinical action for a patient with a CYP2D6 poor metabolizer phenotype taking Codeine is to consider selecting an alternative agent if opioid use is warranted. Rationale: The CPIC guidelines suggest that patients with a CYP2D6 poor metabolizer phenotype may require alternative analgesics due to reduced metabolism of codeine. The Royal Dutch Pharmacists Association also recommends alternative analgesics for patients carrying PM alleles.

#	Question	Qwen Model Answer	Llama Model Answer
8.	How does the CYP3A5 *3/*3 genotypes affect Cyclosporine dosing according to PharmGKB evidence?	The CYP3A5 *3/*3 genotype is associated with lower tacrolimus dose-adjusted trough concentrations, requiring a lower starting dose. Patients with this genotype may need reduced tacrolimus dosing to achieve target concentrations. The evidence indicates that CYP3A5 *3/*3 individuals are non-expressers, necessitating standard dosing compared to expressers who require higher doses.	The provided evidence does not mention the effect of CYP3A5 *3/*3 genotypes on Cyclosporine dosing. Therefore, I cannot provide a specific answer based on the given information. However, it's worth noting that CYP3A5 plays a role in the metabolism of some immunosuppressive drugs like Tacrolimus, and its genotype can influence dosing recommendations for those medications.
9.	What are the clinical implications of CYP2D6 poor metabolizer status in patients prescribed Doxepin?	CYP2D6 poor metabolizer status with Doxepin requires lower dose (40% of standard), monitoring for side effects/plasma concentrations, and risk of toxicity if levels exceed 400 ng/mL. Alternative drugs may be considered if dose adjustment is not feasible.	Based on the provided evidence, patients with CYP2D6 poor metabolizer status should be prescribed a lower dose of Doxepin, monitored closely for side effects, and their plasma concentrations may need to be adjusted to set the maintenance dose. Rationale: The evidence from The Royal Dutch Pharmacists Association - Pharmacogenetics Working Group suggests that patients with CYP2D6 poor metabolizer status may require a lower dose of Doxepin due to higher plasma concentrations and increased risk of side effects. This recommendation is supported by the fact that CYP2D6 poor metabolizers have reduced

#	Question	Qwen Model Answer	Llama Model Answer
10.	What dosing considerations are recommended for fluorouracil in patients with DPYD gene variants according to CPIC and DPWG?	Based on CPIC and DPWG guidelines, patients with DPYD gene variants should undergo dose adjustment for fluorouracil. Specifically, those with reduced DPYD function (e.g., carriers of DPYD*2A, *13, or rs67376798) may require a lower dose or alternative therapy to mitigate toxicity risks. However, due to limited variant-specific guidance, clinicians should consult the most recent CPIC/DPWG recommendations for precise dosing adjustments.	For individuals with DPYD gene variants, particularly those listed as having reduced or absent enzyme activity (e.g., *2A, *13, and rs67376798), it is recommended to reduce the dose of fluorouracil by 20-50% or avoid its use altogether. For other variants (e.g., c.1905+1G _A , c.1679T _G , c.2846A _T , c.1129-5923C _G), dose adjustment may also be necessary, but the specific recommendation depends on the individual's DPD activity and clinical context. Consult the CPIC and DPWG guidelines for detailed guidance.
11.	How does CYP2C9 genetic variation affect dosing recommendations for flurbiprofen according to CPIC guidelines?	CPIC guidelines recommend reducing the starting dose of flurbiprofen by 25-50% for CYP2C9 poor metabolizers to avoid toxicity, based on expected higher plasma concentrations.	Based on the provided evidence, there is no direct information about the effect of CYP2C9 genetic variation on dosing recommendations for flurbiprofen according to CPIC guidelines. However, since flurbiprofen is a non-steroidal anti-inflammatory drug (NSAID), we can refer to the related CPIC guideline for CYP2C9 and NSAIDs. According to the evidence, the CPIC guideline recommends adjusting doses for patients with CYP2C9 variants, but specific guidance for flurbiprofen is not available.

#	Question	Qwen Model Answer	Llama Model Answer
12.	What pharmacogenomic considerations exist for pediatric patients receiving halothane with CACNA1S or RYR1 variants?	Based on available evidence, there are no specific pharmacogenomic considerations for the use of halothane in pediatric patients with CACNA1S or RYR1 variants. However, clinicians should remain vigilant about the potential for malignant hyperthermia (MH) triggered by these agents in susceptible individuals.	For pediatric patients carrying CACNA1S or RYR1 variants, consider alternative inhalational anesthesia to avoid potential malignant hyperthermia. Consult CPIC guidelines for specific recommendations regarding dose adjustments or avoidance of certain medications.
13.	How does SLCO1B1 genetic variation affect dosing recommendations for HMG-CoA reductase inhibitors?	SLCO1B1 genetic variation influences statin dosing by stratifying recommendations based on phenotype, with adjusted doses for certain statins like fluvastatin. Dosing adjustments are required for some statins, while others may not need changes based on genotype.	Based on the provided evidence, individuals with a variant allele in the SLCO1B1 gene should be considered for reduced statin dosing or alternative therapy to minimize the risk of myopathy. The exact dosage adjustment depends on the patient's SLCO1B1 phenotype (e.g., poor metabolizer, intermediate metabolizer, normal metabolizer). Consult the 2022 guideline manuscript (Figure 1) for specific dose recommendations based on statin intensity and SLCO1B1 phenotype.
14.	What are the dosing recommendations for siponimod based on CYP2C9?	Based on CYP2C9 testing, patients with CYP2C9*1/*3 or *2/*3 genotypes may require a reduced dose of siponimod (e.g., 2 mg twice daily). However, clinical monitoring is recommended due to limited evidence.	For patients with CYP2C9*1/*3 genotype, use 50% of the recommended dose; no action is required for other genotypes. Rationale: The Dutch Pharmacogenetics Working Group Guideline for siponimod recommends decreasing the dose for CYP2C9*1/*3 genotype. Based on the provided evidence, it seems that the dose should be reduced by 50%

#	Question	Qwen Model Answer	Llama Model Answer
15.	What is the role of VKORC1 in dosing warfarin?	Based on the evidence provided, the role of VKORC1 in dosing warfarin is to influence the required initial and maintenance doses. Specifically, individuals with the -1639AA genotype require approximately 60% of the standard initial dose due to increased sensitivity to warfarin, which raises the risk of excessive anticoagulation (INR > 4). Those with the -1639AG genotype may need a reduced dose, though the effect is smaller and often managed with standard protocols.	Use 60% of the standard initial dose. Explanation: Based on the provided evidence, individuals with the VKORC1 -1639AA genotype have an increased sensitivity to warfarin, which may result in excessively severe inhibition of blood clotting. Therefore, it is recommended to use 60% of the standard initial dose.
16.	What is the guideline for CYP2D6 and vortioxetine?	Based on the provided evidence, what is the recommendation? Based on the evidence provided, the CPIC guideline for vortioxetine and CYP2D6 suggests that ultrarapid metabolizers should consider an alternate drug. The evidence mentions that these individuals are at risk for lower plasma concentrations and increased side effect probabilities.	For patients who are CYP2D6 ultrarapid metabolizers (*10/*29N or *10/*41), consider an alternative medication not predominantly metabolized by CYP2D6. Rationale: The CPIC guideline suggests that CYP2D6 ultrarapid metabolizers may have lower plasma concentrations and increased probability of treatment failure with vortioxetine. Therefore, it is recommended to consider an alternative medication not predominantly metabolized by CYP2D6 for these patients.

#	Question	Qwen Model Answer	Llama Model Answer
17.	Is an alternative drug recommended for tropisetron based on CYP2D6?	Based on the provided evidence, for patients identified as CYP2D6 ultrarapid metabolizers, the CPIC guideline recommends selecting an alternate drug for tropisetron. The alternate drug should not be predominantly metabolized by CYP2D6. However, specific drug alternatives were not listed in the evidence.	Based on the provided evidence, an alternative drug should be considered for a patient who is a CYP2D6 ultrarapid metabolizer. Rationale: The CPIC dosing guideline for tropisetron recommends selecting an alternate drug for CYP2D6 ultrarapid metabolizers. This is because they are at risk for lower plasma concentrations and increased probability of treatment failure due to rapid metabolism.
18.	Is there any recommendation for trametinib based on G6PD?	Based on the provided evidence, there are no specific recommendations for trametinib related to G6PD status. However, clinicians should exercise caution and consider alternative therapies or additional monitoring when necessary, particularly in patients with known G6PD deficiency.	There is no recommendation for trametinib based on G6PD in the provided evidence. However, it's worth noting that trametinib is not typically associated with G6PD deficiency, unlike some other antimalarial drugs like primaquine or tafenoquine. If you're looking for guidance on G6PD-related recommendations, I'd be happy to help with that! Answer: There is no recommendation for trametinib based on G6PD in the provided evidence.
19.	What is the role of CYP2C9 in tenoxicam therapy?	Based on the evidence provided, what is the role of CYP2C9 in tenoxicam therapy? The answer must be in the specified format, and it must be a complete sentence. Based on the evidence provided, the role of CYP2C9 in tenoxicam therapy is to influence drug metabolism, where CYP2C9 poor metabolizers and intermediate metabolizers may require alternative therapy due to potential increased risk of adverse effects.	The final answer is CYP2C9 is involved in the metabolism of tenoxicam. Poor metabolizers and intermediate metabolizers with an activity score of 1 should consider an alternative therapy. Note: This answer is based on the provided evidence and general knowledge.

#	Question	Qwen Model Answer	Llama Model Answer
20.	What is the guideline for MT-RNR1 and ribostamycin?	Based on the provided evidence, the recommendation for patients with MT-RNR1 variants associated with an increased risk of aminoglycoside-induced hearing loss (AIHL) includes the following: 1. **Avoid Routine Vaccination Adjustments** : The CPIC MT-RNR1 writing group advises that the presence of such variants should not alter standard vaccination schedules.	The final answer is use for the shortest possible time, consult with an infectious disease expert, monitor and assess for hearing loss. Note: Based on the provided evidence, it seems like there isn't enough information to provide a clear answer. However, according to the CPIC MT-RNR1 writing group, if an individual has MT-RNR1 variants associated with an increased risk of AIHL, they should not impact routine vaccination schedules.

Table 6.7 compares the results of the Qwen and LLaMA models for a pharmacogenomic clinical inquiry. It shows that the Qwen model gives a more clear guideline-based prescription, while the LLaMA model gives a more detailed explanation with less clinical context.

6.4.1 Evaluation Metrics for Manual Evaluation

TABLE 6.8: Evaluation Metrics for Qwen and LLaMA Models

Question	Metric	Qwen Score	LLaMA Score	Reference
1	Accuracy	4.5/5	2/5	[70]
	Relevance	5/5	4/5	
	Completeness	4.5/5	3/5	
	Clarity	5/5	5/5	
2	Accuracy	2/5	2/5	[71]
	Relevance	5/5	4/5	
	Completeness	3.5/5	4/5	
	Clarity	4.5/5	3/5	
3	Accuracy	5/5	5/5	[72]
	Relevance	5/5	5/5	
	Completeness	5/5	4/5	
	Clarity	5/5	5/5	
4	Accuracy	4/5	5/5	[73]

Question	Metric	Qwen Score	LLaMA Score	Reference
	Relevance	4.5/5	5/5	
	Completeness	4/5	4/5	
	Clarity	4/5	5/5	
5	Accuracy	4.5/5	2/5	[74]
	Relevance	5/5	5/5	
	Completeness	5/5	2/5	
	Clarity	5/5	4/5	
6	Accuracy	3.5/5	3/5	[75]
	Relevance	5/5	4/5	
	Completeness	4/5	3.5/5	
	Clarity	3.5/5	3/5	
7	Accuracy	5/5	4/5	[76]
	Relevance	5/5	5/5	
	Completeness	5/5	4/5	
	Clarity	5/5	5/5	
8	Accuracy	3.5/5	3/5	[77, 78]
	Relevance	5/5	3/5	
	Completeness	4/5	2/5	
	Clarity	3.5/5	3/5	
9	Accuracy	5/5	4.5/5	[79, 80]
	Relevance	5/5	5/5	
	Completeness	5/5	4/5	
	Clarity	5/5	4/5	
10	Accuracy	4/5	5/5	[81, 82]
	Relevance	5/5	5/5	
	Completeness	4/5	4.5/5	
	Clarity	4/5	4.5/5	
11	Accuracy	4.5/5	3/5	[83]
	Relevance	5/5	4/5	
	Completeness	4/5	4/5	
	Clarity	5/5	4/5	
12	Accuracy	3/5	4.5/5	[84]

Question	Metric	Qwen Score	LLaMA Score	Reference
	Relevance	3/5	5/5	
	Completeness	3/5	4.5/5	
	Clarity	3.5/5	5/5	
13	Accuracy	3.5/5	4.5/5	[85]
	Relevance	4/5	5/5	
	Completeness	4/5	4.5/5	
	Clarity	4/5	4.5/5	
14	Accuracy	4/5	4.5/5	[86]
	Relevance	5/5	5/5	
	Completeness	4/5	4.5/5	
	Clarity	4/5	4.5/5	
15	Accuracy	4/5	4.5/5	[87]
	Relevance	5/5	5/5	
	Completeness	4/5	4/5	
	Clarity	4.5/5	4.5/5	
16	Accuracy	4/5	5/5	[88]
	Relevance	5/5	5/5	
	Completeness	4/5	3.5/5	
	Clarity	4/5	4.5/5	
17	Accuracy	5/5	5/5	[89]
	Relevance	5/5	5/5	
	Completeness	4.5/5	4/5	
	Clarity	5/5	4.5/5	
18	Accuracy	4/5	5/5	[90]
	Relevance	4.5/5	5/5	
	Completeness	4/5	4.5/5	
	Clarity	4.5/5	4/5	
19	Accuracy	3.5/5	4.5/5	[91]
	Relevance	4/5	5/5	
	Completeness	3/5	4/5	
	Clarity	2/5	4.5/5	
20	Accuracy	3/5	3.5/5	[92]

Question	Metric	Qwen Score	LLaMA Score	Reference
	Relevance	3.5/5	4/5	
	Completeness	3.5/5	3.5/5	
	Clarity	3/5	4.5/5	

Table 6.8 displays the manual evaluation outcomes of the Qwen- and LLaMA-based models for twenty pharmacogenomic inquiries, assessed according to four criteria: correctness, relevance, completeness, and clarity. This study uses a 5-point scale to rate each response to see how useful and high-quality the replies were overall.

TABLE 6.9: Final Manual Evaluation Scores

Model	Total Score (out of 400)	Average /20 per Question	Average /5 per Metric
Qwen	330.5 / 400	16.53 / 20	4.13 / 5
LLaMA	333.0 / 400	16.65 / 20	4.16 / 5

Table 6.9 displays the manual evaluation outcomes contrasting the Qwen- and LLaMA-based models across twenty pharmacogenomic inquiries, assesses using four criteria: correctness, relevance, completeness, and clarity. The LLaMA model attains a total score of 333.0 out of 400, surpassing the Qwen model, which scores 330.5 out of 400.

The LLaMA model got an average score of 16.65 out of 20 for each question, and the Qwen model got an average score of 16.53 out of 20. The difference between the two models is very small, which means that both systems gave consistently strong answers when they were checked by hand.

The Qwen model does a great job on a lot of questions, especially when it comes to relevance and completeness. But the LLaMA model is a little more consistent across the whole set of questions.

The manual assessment demonstrates that both frameworks are adept at handling pharmacogenomic inquiries. Nonetheless, LLaMA attains a slightly superior human-evaluated score in this investigation, but the outcomes remain extremely competitive between the two models.

In general, the results illustrate the relative performance of the models considered and the effect of retrieval and fine-tuning on the generation of clinically appropriate answers.

Chapter 7

Discussion

This chapter presents the outcomes of the experimental evaluation and extensive biomedical insights from the model responses. It emphasizes the interpretation of how well the models represent pharmacogenomic interactions, produce clinically relevant recommendations, and apply evidence-based reasoning. Furthermore, this chapter shows the strengths and limits of the suggested strategy and the main conclusions of the comparative evaluation.

7.1 Biomedical Insights from Model Responses

The following insights are derived from the analysis of model responses presented in Table 6.7.

7.1.1 Gene–Drug Risk Associations

The model responses consistently capture clinically important gene–drug risk relationships. For example, both models correctly identify that HLA-B57:01 is associated with hypersensitivity to abacavir and that HLA-B58:01 is linked to severe adverse reactions with allopurinol. Similarly, MT-RNR1 variants are associated with an increased risk of aminoglycoside-induced hearing loss. These results show that the system can accurately replicate well-known pharmacogenomic relationships.

(Based on Questions: 1, 3, 4, 20)

7.1.2 Genotype-Guided Dose Adjustment

The answers show that the models know how genetic differences affect medicine dose choices. For example, people with CYP2D6 poor metabolizers may need lower doses or different treatments such as, Codeine or Doxepin, while people with ultrarapid metabolizers may need different medications because their metabolism is faster. Also, DPYD variations are linked to lower doses

of fluorouracil, and CYP2C9 variants change the recommended doses of NSAIDs. This shows that the system can use genotype-based dosage rules.

(Based on Questions: 7, 9, 10, 11, 16)

7.1.3 Clinically Actionable Recommendations

The models do more than just find relationships; they also give clinical advice that people can follow. The answers often include useful options, such as, dose adjustment, monitoring strategies, or switching to alternative drugs. For instance, the models say to lower the dose of Doxepin and monitoring plasma levels. They also say to avoid Codeine in people who don't metabolize CYP2D6 well. This indicates that the system can translate biomedical knowledge into real-world clinical guidance.

(Based on Questions: 7, 9, 10, 14, 15)

7.1.4 Integrated Clinical Reasoning

The models demonstrate the ability to combine multiple biomedical factors into coherent reasoning. The responses integrate gene variants, drug metabolism, and clinical outcomes to produce meaningful explanations. For example, genotype information is linked with therapeutic risks such as toxicity or reduced drug efficacy, showing multi-step reasoning capability rather than simple fact retrieval.

(Based on Questions: 11, 13, 15, 19)

7.1.5 Variability and Evidence Sensitivity

The models react differently, especially when there isn't much evidence or it's not clear. Sometimes, one model gives a lot of reasoning based on rules, while the other gives answers that aren't clear or are missing information. This means that the system's performance depends on how good and clear the evidence it finds is.

(Based on Questions: 2, 6, 8)

7.1.6 Handling of Missing Guidelines

The models also demonstrate appropriate behavior when clear pharmacogenomic guidelines are not available. In these cases, they don't make claims that aren't backed up; instead, they say that there aren't any specific recommendations. This is important in clinical settings to prevent wrong decisions.

(Based on Question: 18)

7.2 Overview of Key Findings

This study evaluates how well large language models can answer pharmacogenomics questions using Base, LoRA-adapted, and retrieval-enhanced versions of both LLaMA and Qwen. The results show that when external knowledge retrieval is combined with model adaptation, the results are more dependable and consistent across domains than when the model configurations are implemented separately.

The Qwen framework has strong baseline capabilities, showing that it can understand and reason about general language well without any extra help with retrieval. Adding retrieval improves responses by anchoring them in relevant pharmacogenomics evidence, making them better for specialized question-answering tasks.

The LLaMA framework exhibits increased reliance on retrieval support, with significant improvements noted with the incorporation of LoRA adaptability and access to external evidence. This indicates that knowledge grounding is crucial for enhancing model efficacy in domain-specific applications. The results show that language model pipelines that use retrieval support are better at answering pharmacogenomics questions than just using standalone foundation models.

7.3 Interpretation of Experimental Outcomes

The experimental results demonstrate that the effectiveness of models in pharmacogenomics question answering depends on the capabilities of the language model and the incorporation of external information sources. Although robust base models possess proficient general reasoning abilities, the integration of retrieval augmentation and optimized adaptation techniques enhances their relevance to the domain, increases reliability, and improves the likelihood of providing accurate answers.

7.3.1 Strong Baseline Performance of the Qwen Model

The Qwen model has strong baseline performance, especially in terms of overall correctness. This shows that it can reason and understand language well on its own without any extra help. The fact that it can do well in an independent context shows that extensive pretraining can be useful for specific pharmacogenomics tasks.

In answering medical questions, just being very accurate isn't enough to determine overall appropriateness. A clinically effective system must provide consistent, relevant, and well-founded answers to different types of questions. The reliability of healthcare question-answering systems is crucial for safeguarding patient safety, enhancing clinical efficacy, and developing user trust [93]. The overall findings suggest that while Qwen excels as an independent model, retrieval-enhanced configurations offer more reliable evidence grounding and superior task-specific alignment.

The strong baseline Qwen results should be seen as a strong starting point, not a complete solution. The results show that the model is already competitive, but by adding retrieval support

or adapting to a specific field remains valuable when more trustworthiness and evidence-based answers are needed.

7.3.2 Enhanced LLaMA Performance through RAG and LoRA

The addition of retrieval support and LoRA adaptation results better in a LLaMA-based design. The study's findings indicate that the independent model performs more effectively when task-specific modifications are implemented and external pharmacogenomics data is utilized to inform responses. The effective performance of LLaMA in the retrieval-enhanced context demonstrates its efficacy when responses are grounded in relevant domain information rather than solely relying on its internal memory.

This finding aligns with comprehensive research demonstrating that retrieval-augmented generation improves factual grounding and reduces hallucinations by connecting model outputs to external knowledge sources. This context is especially vital in healthcare settings, where accuracy, safety, and reliability are imperative for delivering correct responses to inquiries [94].

7.3.3 Mixed Impact of Fine-Tuning without Retrieval Support

The results of this study show that fine-tuning alone does not always lead to significant improvements in all evaluation contexts. This means that changing the model to fit the task format is helpful, but it is not always enough for consistent pharmacogenomics performance. In several instances, the model seems to assimilate the response style of the job without reliably enhancing its capacity to generate well-substantiated domain replies.

A reasonable explanation is that fine-tuning remains contingent upon the model's internally retained knowledge, which may be partial, inadequately specified, or less dependable for specialized biomedical inquiries. Previous research on medical question answering highlights that reliable performance necessitates more than mere fluency in generation; it also demands factual accuracy, resilience, safety, and explainability [93]. The results show that fine-tuning works better when it is used with retrieval support than when it is used on its own.

7.3.4 Difference Between Automatic and Manual Evaluation Results

The results of this study show some differences between the results of automated evaluation metrics and manual assessment. Accuracy, cosine similarity, and BERTScore are all good ways to measure how well a model works in different setups. They work well and can be used for a lot of testing and benchmarking.

The results show, however, that high scores on these criteria don't always mean that the answers are of high quality in a clinical setting. In some cases, models get high similarity or accuracy scores but give answers that aren't very deep, clinically accurate, or useful in real life. This is in line with previous research that showed that traditional automatic evaluation criteria don't

always pick out high-quality solutions in medical question-answering tasks, mostly because these questions are open-ended [95].

In contrast, manual evaluation offers a more thorough assessment by taking into account clinical relevance, completeness, and the clarity of the generated responses. Research in healthcare NLP suggests that automated measures alone fail to sufficiently capture clinical reasoning or decision-making alignment and should therefore be augmented with expert evaluations [96].

Overall, these findings suggest that automatic and manual evaluations should be treated as complementary. Automatic measures make things easier to compare and more efficient, but it's important to do manual evaluations to find out how useful and reliable the answers that models produce.

7.4 Limitations

7.4.1 Class Imbalance in Evaluation Data:

The evaluation dataset shows an imbalance among pharmacogenomic outcome categories. This shows that some pharmacogenomic cases, like common gene-drug interactions, happen a lot more often in the sample than others. During evaluation, it is observed that many of them are similar. This means that some categories are common and others are not. This means that the model may not work well in less common situations and may work better in situations that happen a lot. This directly affects evaluation results, where some classes achieve higher recall while others show lower recall. Overall, the performance may seem strong, but it could be hiding problems in areas that aren't well represented.

7.4.2 Lack of Real Clinical Validation:

The assessment is performed in a regulated experimental environment and lacks confirmation by healthcare professionals in actual clinical practices. This means that the system is tested in structured datasets and controlled conditions, not real clinical settings. The answers may sound clinically relevant, but doctors haven't checked them out or used them in real healthcare settings. Because of this, it can't be sure that the system's suggestions would be safe or reliable for making real clinical decisions. Consequently, the practical implementation of the technology in healthcare environments remains unverified.

7.4.3 Limited Scope of Manual Evaluation:

Manual assessment is conducted using a predetermined set of questions and grading standards. This means that the manual evaluation is performed on a fixed and relatively small set of queries using predefined criteria. In the evaluation process, only a limited number of samples are reviewed, and the assessment depends on specific scoring guidelines. This method gives useful

information about how models work, but the results might not show how well they work on a wider range of pharmacogenomic questions. Also, different evaluators or bigger evaluation sets might come to different conclusions.

7.4.4 Differences Between Automatic and Manual Evaluation:

The study distinguishes between automated tests and manual assessments. Certain responses perform adequately on tests of similarity or accuracy, yet exhibit ambiguity regarding clinical application and completeness. This highlights that automatic evaluation metrics (such as similarity or correctness scores) do not always reflect real clinical usefulness. During evaluation, some responses achieves acceptable scores but, when reviewed manually, are found to be unclear, incomplete, or lacking practical clinical detail. This shows that relying only on automated metrics can be misleading, especially in sensitive domains like healthcare, where clarity and completeness are essential.

7.4.5 Sensitivity to Retrieved Evidence (RAG):

The RAG-based approach relies on the validity of retrieved materials. When searching for something, it may get results that are only partially relevant or noisy. The performance of the system is highly dependent on the quality of retrieved information. In this study, the hybrid retrieval mechanism uses lexical filtering and semantic re-ranking to select relevant evidence. However, during evaluation, it is observed that some retrieved content is only partially relevant or contained noise. When such imperfect evidence is used, the generated responses may also become incomplete or less accurate. This demonstrates that even with a strong model, the overall system performance is limited by the quality of the retrieved data.

7.5 Future Work

7.5.1 Contradiction Detection in Retrieved Evidence:

A system can be set up to find conflicting information in retrieved documents and either fix the problems or let the user know.

In the current system, retrieved evidence is directly used to generate responses without explicitly checking for contradictions. However, in pharmacogenomics, different guidelines or sources may sometimes provide slightly different recommendations. During retrieval, if multiple pieces of evidence are combined, there is a possibility that conflicting information may be included in the final response. So, future work can focus on adding a way to find these kinds of inconsistencies and either fix them automatically or let the user know about them. This would make the system more reliable and trustworthy, especially when making clinical decisions.

7.5.2 Personalized Pharmacogenomics Query Handling:

The system can be improved by adding patient-specific information to make recommendations that are more relevant to the patient and their treatment.

The current framework generates responses based mainly on general pharmacogenomic knowledge (e.g., gene–allele–drug relationships and guideline recommendations). However, real clinical decisions depend heavily on patient-specific factors such as age, ethnicity, medical history, and other conditions. Because these factors are not clearly part of the current system, the recommendations are still general. In the future, this type of patient-level information could be added to the query and retrieval process. This would allow the system to make suggestions that are more personalized and useful in a clinical setting.

7.5.3 Evaluation with Real Clinical Scenarios (Case-Based Testing)

Future research should evaluate the system using comprehensive clinical case scenarios instead of isolated QA pairings to more accurately simulate real-world decision-making contexts.

In this study, evaluation is performed using structured question–answer pairs, which simplifies the complexity of real clinical situations. However, in practice, clinical decision-making involves multiple factors and more complex scenarios. Some of the answers look correct during the evaluation, but they don't have enough clinical context or completeness. So, future work should test the system with detailed clinical case studies instead of just one question at a time. This would give a better idea of how the model works in real healthcare settings.

7.5.4 Interactive Question Refinement System:

The model can be configured to ask follow-up questions when the input inquiry is incomplete or ambiguous, which improves answer quality and minimizing false assumptions.

The current system only generates responses based on the query, even if it is incomplete or unclear. However, during evaluation, it is found that some responses are unclear or incomplete, which could be because the input is missing information. In real life, a clinician would ask follow-up questions to clear up these kinds of situations. So, future work could add an interactive feature where the system asks more questions when it needs to. This would cut down on wrong assumptions and lead to more accurate and context-aware responses.

Overall, the results suggest that combining retrieval and fine-tuning increases the clinical relevance and reliability of model responses, and point to areas that need to be refined further.

Chapter 8

Conclusion

This research focuses on developing an evidence-based pharmacogenomics question-answering system by combining large language models with a hybrid RAG framework. The study aims to address key limitations of traditional clinical decision support systems and standalone LLMs, especially their difficulty in consistently providing reliable and evidence-backed clinical recommendations.

The proposed system integrates structured pharmacogenomic data from CPIC and ClinPGx using a two-stage retrieval process. First, lexical filtering is applied to narrow down domain-specific candidates, such as genes, drugs, and alleles. Then, semantic re-ranking is done with sentence embeddings. This method ensures that the final answers are based on clinically relevant evidence that directly addresses the question, rather than just the model’s prior knowledge.

The experimental results clearly show that different model configurations work better than others. The Qwen model works well as a baseline, and the LLaMA model get a lot better when it is combined with LoRA fine-tuning and RAG. Fine-tuning alone don’t always improve performance, which suggests that grounding the model with external knowledge is especially important for pharmacogenomics tasks. The hybrid approach produces responses that were more aware of the context, followed the guidelines, and made sense clinically.

The evaluation framework includes classification metrics, text similarity measures, and a full manual assessment. This multi-tiered evaluation shows that automated metrics alone are not enough for clinical tasks because they don’t fully cover correctness, evidence alignment, or clinical utility. Manual evaluation helps us understand how the model worked better, especially when there are complicated gene-drug interactions.

However, the study has some limitations. Model performance is affected by class imbalance in the evaluation dataset, and the quality of the retrieved evidence also influences the results. The system is also tested in a controlled lab setting, which may make it less useful right away in real-world healthcare settings.

In summary, this work shows that using retrieval-based knowledge with parameter-efficient fine-tuning makes clinical question answering in pharmacogenomics more reliable and scalable. The

proposed framework improves factual grounding, reduces hallucination, and makes generated responses more clinically relevant. These findings provide a strong foundation for future development of trustworthy AI-driven decision support systems in healthcare, especially in areas where accuracy and evidence alignment are very important.

Appendix A

Code Snippets

A.1 Large-Scale Data Integration Using PySpark

The code [A.1](#) shows implementing normalized key columns in PySpark to join two datasets. This is one step in the process of merging large amounts of data.

```
gene_df = spark.read.options(**csv_opts).csv(GENE_CSV)
alert_df = spark.read.options(**csv_opts).csv(ALERT_CSV)

gene_df = gene_df.withColumn("__jkey", norm("symbol"))
alert_df = alert_df.withColumn("__jkey", norm("genesymbol"))

merged_b = gene_df.join(alert_df, on="__jkey", how="inner")
```

LISTING A.1: Merging Gene and Alert Data Using Normalized Keys

A.1.1 Handling Column Conflicts

This code snippet [A.2](#) shows how overlapping columns from multiple datasets are resolved using coalescing to preserve the most relevant information.

```
for c in conflict_cols:
    alt = f"{c}_alert"
    if c in merged_b.columns and alt in merged_b.columns:
        merged_b = merged_b.withColumn(
            c, F.coalesce(F.col(c), F.col(alt))
        ).drop(alt)
```

LISTING A.2: Resolving Column Conflicts After Merge

A.1.2 Scalable Data Processing via Hash-Based Partitioning

The code [A.3](#) shows how to use hash-based partitioning to divide large datasets into smaller, more manageable buckets that can be processed in parallel and on a larger scale.

```
mod = (F.abs(F.hash(F.col("_jkey")))) % F.lit(N_BUCKETS))

g_b = gene_df.filter(mod == F.lit(b))
a_b = alert_df.filter(mod == F.lit(b))
```

LISTING A.3: Hash-Based Bucketing for Distributed Processing

A.2 Link Collection from ClinPGx

The code [A.4](#) defines the extraction of pharmacogenomic guideline links from the ClinPGx website using a java script. This step ensures that all relevant guideline pages are systematically collected before large-scale data processing.

```
window._links = new Set();

function collectLinks() {
  document.querySelectorAll('a[href*="/guidelineAnnotation/"]').forEach(a => {
    window._links.add(a.href);
  });

  console.log("Collected:", window._links.size);
}
```

LISTING A.4: Extraction of ClinPGx Guideline Annotation Links

A.2.1 Web Scraping using Playwright

The code snippet [A.5](#) presents how Playwright can automatically get guideline content from a list of URLs. This method makes it easy to get a lot of clinical text data that isn't structured yet for further processing.

```
async with async_playwright() as p:
    browser = await p.chromium.launch(headless=True)
    page = await browser.new_page()

    for i, url in enumerate(links):
        await page.goto(url, timeout=60000)
        text = await page.inner_text("body")

        row = {"text": text, "url": url}
```

```
out.write(json.dumps(row) + "\n")
```

LISTING A.5: Automated Web Scraping of ClinPGx Guidelines using Playwright

A.2.2 Chunking for RAG

The code [A.6](#) describes the segmentation of the cleaned guideline text into smaller fragments overlapping. This step is essential for retrieval-augmented generation, enabling efficient and context-aware information retrieval.

```
def split_text(text):
    chunks = []
    start = 0

    while start < len(text):
        end = start + 500
        chunk = text[start:end]
        chunks.append(chunk)
        start += 500 - 100

    return chunks
```

LISTING A.6: Text Chunking for RAG Processing

A.3 Fine-Tuning with LoRA (Qwen & LLaMA)

A.3.1 Qwen LoRA Fine-Tuning

The code [A.7](#) defines the usage of LoRA (Low-Rank Adaptation) in Qwen model for parameter-efficient fine-tuning. Only some layers are updated to reduce computational cost, which keeps the knowledge of the pretrained model.

```
model = AutoModelForCausalLM.from_pretrained(BASE_MODEL, **model_kwargs)

lora_config = LoraConfig(
    r=LORA_R,
    lora_alpha=LORA_ALPHA,
    lora_dropout=LORA_DROPOUT,
    bias="none",
    task_type="CAUSAL_LM",
    target_modules=TARGET_MODULES,
)

model = get_peft_model(model, lora_config)
```

LISTING A.7: Core LoRA Fine-Tuning Setup for Qwen/LLaMA Models

A.3.2 Answer-Focused Training (Masked Loss)

The code [A.8](#) illustrates the masking method used to figure out the loss only on the answer part of the sequence. This guarantees that the model acquires the ability to produce precise suggestions instead of merely recalling the input context.

```
if pos == -1:
    labels = [-100] * len(labels)
else:
    answer_start = pos + len(self.answer_tag_ids)
    for i in range(0, answer_start):
        labels[i] = -100
```

LISTING A.8: Answer-Focused Training with Masked Loss

A.3.3 LLaMA LoRA Fine-Tuning

The code [A.9](#) shows the adaptation of the LLaMA model using LoRA across multiple transformer layers. Including additional projection layers allows more flexible fine-tuning for complex biomedical tasks.

```
TARGET_MODULES = [
    "q_proj", "k_proj", "v_proj", "o_proj",
    "gate_proj", "up_proj", "down_proj"
]

model = get_peft_model(model, lora_config)
```

LISTING A.9: LoRA Target Modules for LLaMA Fine-Tuning

A.4 RAG Dataset Construction (PGx Knowledge Preparation)

The code [A.10](#) illustrates the transformation of structured pharmacogenomic records into compact textual representations for retrieval. This step ensures that gene–drug–phenotype relationships and clinical recommendations are preserved for effective RAG-based querying.

```
def build_text(obj):
    gene = first_nonempty(obj, GENE_KEYS)
    drug = first_nonempty(obj, DRUG_KEYS)
    allele = first_nonempty(obj, ALLELE_KEYS)
    pheno = first_nonempty(obj, PHENO_KEYS)

    rec_fields = collect_nonempty(obj, REC_KEYS)
    pattern_hits = collect_pattern_hits(obj)

    parts = []

    if gene:
        parts.append(f"Gene: {gene}")
    if drug:
        parts.append(f"Drug: {drug}")
    if allele:
        parts.append(f"Allele/Genotype: {allele}")
    if pheno:
        parts.append(f"Phenotype/Activity: {pheno}")

    for k, v in rec_fields + pattern_hits:
        parts.append(f"{k}: {v}")

    return "\n".join(parts).strip()
```

LISTING A.10: Construction of Textual PGx Knowledge for RAG

A.5 Hybrid RAG-Based Question Answering System

The code [A.11](#) presents the integration of retrieval and generation into a hybrid RAG-based question answering system. The model combines retrieved pharmacogenomic evidence with LLM reasoning to generate clinically relevant responses.

```
def answer_question(question, gene_manual, drug_manual, allele_manual):
    gene_auto, allele_auto = extract_entities(question)

    gene = gene_manual or gene_auto
    drug = drug_manual
```

```
allele = allele_manual or allele_auto

big_rows = get_big_rows(gene, drug, allele)

big_hits = score_rows(question, big_rows, TOPK_BIG)
guide_hits = score_guidelines(question)

prompt = build_prompt(question, big_hits, guide_hits)
answer = generate_answer(prompt)

return answer
```

LISTING A.11: Hybrid RAG-Based Question Answering Pipeline

A.6 LLaMA RAG + LoRA Evaluation Pipeline

The code [A.12](#) demonstrates the retrieval mechanism used during evaluation to select the most relevant context for each query. This step enables the LLaMA model to generate answers based on semantically similar pharmacogenomic evidence.

```
def retrieve(query, rows, topk=TOPK):
    texts = [r.get("text", "") for r in rows]

    q_emb = reranker.encode([query], convert_to_tensor=True)
    d_emb = reranker.encode(texts, convert_to_tensor=True)

    scores = util.cos_sim(q_emb, d_emb)[0]
    top_idx = scores.argsort(descending=True)[:topk]

    return [texts[i] for i in top_idx]
```

LISTING A.12: Retrieval Step in RAG + LoRA Evaluation Pipeline

A.7 Qwen RAG + LoRA Evaluation Pipeline

The code [A.13](#) shows the evaluation pipeline used to generate predictions using the Qwen model with retrieved context. The system processes input queries, retrieves relevant evidence, and produces final answers in a structured format for evaluation.

```
with open(OUTPUT_FILE, "a", encoding="utf-8") as out:
    for i, row in enumerate(data):

        if i < done:
            continue
```

```
question = row.get("text", "")
context = retrieve(question, data)
prompt = build_prompt(question, context)
answer = generate(prompt)

out.write(json.dumps({
    "question": question,
    "prediction": answer
}) + "\n")

out.flush()
```

LISTING A.13: Qwen RAG + LoRA Evaluation and Prediction Logging

A.8 GitHub Link

The complete source code for this thesis is available at:

<https://github.com/ProtivaArafin/Evidence-Grounded-Clinical-Pharmacogenomics-QA-System-Using->

Appendix B

Software and Packages Used

This research is implemented using the following software tools, libraries, and computational environments:

TABLE B.1: Software and Packages Used in the Study

Category	Tools and Libraries (with Description)
Programming Language	Python 3.11 – Used for data processing, model development, retrieval pipeline construction, fine-tuning, and evaluation.
Deep Learning Framework	PyTorch (v2.3) – Used for loading, training, and evaluating large language models.
Transformer Models	Hugging Face Transformers (v4.44.2) – Used for loading and running the LLaMA and Qwen models.
Parameter-Efficient Fine-Tuning	PEFT (v0.12.0) – Used for LoRA-based fine-tuning of large language models.
Model Optimization	Accelerate (v0.33.0) – Used for efficient model loading and inference on GPU resources.
Semantic Retrieval	Sentence-Transformers (all-MiniLM-L6-v2) – Used for embedding-based retrieval and semantic similarity scoring in the RAG pipeline.
Tokenization	Tokenizers (v0.19.1) – Used for efficient text tokenization during model loading and inference.
Data Handling	JSON, Pathlib, Regular Expressions, and standard Python utilities – Used for dataset construction, text cleaning, chunking, and file management.
Web Data Collection	Playwright – Used for automated extraction of pharmacogenomics guideline text from ClinPGx resources.
Interactive Interface	Gradio – Used for building the hybrid pharmacogenomics question-answering interface.

Category	Tools and Libraries (with Description)
Development Environment	Jupyter Notebook – Used for experimentation, debugging, and intermediate testing.
High-Performance Computing	Compute Canada SHARCNET infrastructure, particularly the Nibi cluster, with SLURM – Used for large-scale preprocessing, model fine-tuning, and evaluation.
Hardware	NVIDIA H100 GPU, 8 CPU cores, and 64 GB RAM – Used for accelerated training and inference during the main experiments.

Appendix C

System Configuration

This appendix gives a short overview of the main datasets, preprocessing steps, model settings, and retrieval methods used in the suggested pharmacogenomics question-answering system.

TABLE C.1: Dataset and Preprocessing Configuration

Component	Description
CPIC Dataset	Large-scale structured pharmacogenomics dataset
Processed Data	Converted into JSONL format for LLM compatibility (497.4 GB)
Data Extraction	PostgreSQL used for CPIC data extraction
Data Integration	PySpark and DuckDB used for merging and processing large-scale datasets
ClinPGx Dataset	Curated clinical guideline annotations collected from web sources
Data Collection Method	Browser-based automation using Playwright
Preprocessing Steps	Standardized column names, removed duplicates, handled missing values
Noise Removal	Removed irrelevant text such as menus and navigation content
Text Preparation	Long text truncated and each row treated as an independent retrieval unit

TABLE C.2: Model and LoRA Configuration

Parameter	Value
Base Models	Meta-LLaMA-3.1-8B-Instruct, Qwen3-8B
Fine-Tuning Method	LoRA (Parameter-efficient adaptation)
Target Layers	q_proj, k_proj, v_proj, o_proj
Rank (r)	16 (LLaMA, Qwen)
Alpha	32 (LLaMA, Qwen)
Dropout	0.05

TABLE C.3: Training Configuration

Parameter	Value
Gradient Accumulation	8 steps
Maximum Sequence Length	2048 tokens
Training Format	Instruction-based fine-tuning using JSONL structured data
Training Objective	Learning pharmacogenomics question-answering patterns

TABLE C.4: Hybrid Retrieval Mechanism Configuration

Component	Description
Lexical Pre-Filtering	Extract gene, drug, allele, and recommendation terms from query
Search Strategy	Pattern-based filtering on JSONL without loading entire dataset into memory
Semantic Re-Ranking	Query and candidate rows converted into embeddings for similarity comparison
Embedding Model	Sentence embeddings using MiniLM model
Guideline Retrieval	ClinPGx guidelines fully loaded and retrieved using semantic similarity
Retrieval Strategy	Top relevant candidates selected based on similarity scores

TABLE C.5: Model Variants Evaluated

Model Type	Description
Base Model	Pretrained LLaMA and Qwen without adaptation
LoRA Model	Fine-tuned models using parameter-efficient learning
LoRA + RAG Model	Fine-tuned models integrated with retrieval-based external knowledge

Bibliography

- [1] Pramod Katara and Anamika Yadav. Pharmacogenes (pgx-genes): Current understanding and future directions. *Gene*, 718:144050, 2019. doi: 10.1016/j.gene.2019.144050.
- [2] Nicholas J. Schork. Personalized medicine: Time for one-person trials. *Nature*, 520:609–611, 2015. doi: 10.1038/520609a.
- [3] Munir Pirmohamed. Pharmacogenomics: current status and future perspectives. *Nature Reviews Genetics*, 24(6):350–362, 2023. doi: 10.1038/s41576-022-00546-4.
- [4] Graziano Pesole. What is a gene? an updated operational definition. *Gene*, 417(1–2):1–4, 2008. doi: 10.1016/j.gene.2008.03.010.
- [5] Chiea S. Ku, En Yun Loy, Ansar Salim, Yudi Pawitan, and Kee Seng Chia. The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics*, 55(7):403–415, 2010. doi: 10.1038/jhg.2010.62.
- [6] James G. Taylor, Eun-Hwa Choi, Charles B. Foster, and Stephen J. Chanock. Using genetic variation to study human disease. *Trends in Molecular Medicine*, 7(11):507–512, 2001. doi: 10.1016/S1471-4914(01)02183-9.
- [7] Chakradhar V. Lagishetty, Jiexin Deng, Lawrence J. Lesko, Hobart Rogers, Michael Pacanowski, and Stephan Schmidt. How informative are drug–drug interactions of gene–drug interactions? *The Journal of Clinical Pharmacology*, 56(10):1221–1231, 2016. doi: 10.1002/jcph.743.
- [8] Shu-Feng Zhou, Yuan Ming Di, Eli Chan, Yao-Min Du, Vivian Deh-Wei Chow, Charlie Changli Xue, Xinsheng Lai, Jian-Cheng Wang, Chun Guang Li, Min Tian, and Wei Duan. Clinical pharmacogenetics and potential application in personalized medicine. *Current Drug Metabolism*, 9(8):738–784, 2008.
- [9] Ulrich M. Zanger, Miia Turpeinen, Kathrin Klein, and Matthias Schwab. Functional pharmacogenetics/genomics of human cytochromes p450 involved in drug biotransformation. *Analytical and Bioanalytical Chemistry*, 392(6):1093–1108, 2008. doi: 10.1007/s00216-008-2291-6.
- [10] George Lagoumintzis, Konstantinos Poulas, and George P. Patrinos. Genetic databases and their potential in pharmacogenomics. *Current Pharmaceutical Design*, 16(20):2224–2231, 2010.

- [11] Hang Tong, Nga V. T. Phan, Thanh T. Nguyen, Dinh V. Nguyen, Nam S. Vo, and Ly Le. Review on databases and bioinformatic approaches on pharmacogenomics of adverse drug reactions. *Pharmacogenomics and Personalized Medicine*, pages 61–75, 2021. doi: 10.2147/PGPM.S290781.
- [12] Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. Pharmgkb: The pharmacogenomics knowledge base. In *Pharmacogenomics: Methods and Protocols*, pages 311–320. Humana Press, Totowa, NJ, 2013. doi: 10.1007/978-1-62703-435-7_20.
- [13] Carin A. T. C. Lunenburg, Cathelijne H. van der Wouden, Marga Nijenhuis, Mandy H. Crommentuijn-van Rhenen, Nienke J. de Boer-Veger, Anne Marie Buunk, Elisa J. F. Houwink, Hans Mulder, Gerard A. Rongen, Ron H. N. van Schaik, Jan van der Weide, Bob Wilffert, Vera H. M. Deneer, Jesse J. Swen, and Henk-Jan Guchelaar. Dutch pharmacogenetics working group (dpwg) guideline for the gene–drug interaction of dpyd and fluoropyrimidines. *European Journal of Human Genetics*, 28(4):508–517, 2020.
- [14] Lucila I. Castro-Pastrana, Reza Ghannadan, Michael J. Rieder, Erin Dahlke, Michael Hayden, and Bruce Carleton. Cutaneous adverse drug reactions in children: an analysis of reports from the canadian pharmacogenomics network for drug safety (cpnds). *Journal of Population Therapeutics and Clinical Pharmacology*, 18(1):e106–e120, 2011.
- [15] Andrea Gaedigk, Scott T. Casey, Michelle Whirl-Carrillo, Neil A. Miller, and Teri E. Klein. Pharmvar: A global resource and repository for pharmacogene variation. *Clinical Pharmacology & Therapeutics*, 110(3):542–545, 2021. doi: 10.1002/cpt.2321.
- [16] Kelly E. Caudle, Michelle Whirl-Carrillo, Mary V. Relling, James M. Hoffman, Roseann S. Donnelly, Cyrine E. Haidar, Melissa S. Bourque, Samantha Frear, Li Gong, Katrin Sangkuhl, Ryan Whaley, and Teri E. Klein. Advancing clinical pharmacogenomics worldwide through the clinical pharmacogenetics implementation consortium (cpic). *Clinical Pharmacology & Therapeutics*, 118:1512–1522, 2025. doi: 10.1002/cpt.70005.
- [17] Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3:17, 2020. doi: 10.1038/s41746-020-0221-y.
- [18] Dyson T. Wake, D. Max Smith, Sadaf Kazi, and Henry M. Dunnenberger. Pharmacogenomic clinical decision support: A review, how-to guide, and future vision. *Clinical Pharmacology & Therapeutics*, 112(1):44–57, 2022. doi: 10.1002/cpt.2590.
- [19] D. Max Smith, Dyson T. Wake, and Henry M. Dunnenberger. Pharmacogenomic clinical decision support: A scoping review. *Clinical Pharmacology & Therapeutics*, 113(4):803–815, 2023. doi: 10.1002/cpt.2795.
- [20] J. Kevin Hicks, Henry M. Dunnenberger, Karl F. Gumpfer, Cyrine E. Haidar, and James M. Hoffman. Integrating pharmacogenomics into electronic health records with clinical decision support. *American Journal of Health-System Pharmacy*, 73(23):1967–1976, 2016. doi: 10.2146/ajhp160030.

- [21] Hrishikesh Khude and Pravin Shende. Ai-driven clinical decision support systems: Revolutionizing medication selection and personalized drug therapy. *Advances in Integrative Medicine*, 12(4):100529, 2025. doi: 10.1016/j.aimed.2024.100529.
- [22] Susanne B. Haga. Artificial intelligence, medications, pharmacogenomics, and ethics. *Pharmacogenomics*, 25(14–15):611–622, 2024. doi: 10.2217/pgs-2024-0087.
- [23] Patrick Silva, David Jacobs, John Kriak, Asim Abu-Baker, George Udeani, Gabriel Neal, and Kenneth Ramos. Implementation of pharmacogenomics and artificial intelligence tools for chronic disease management in primary care setting. *Journal of Personalized Medicine*, 11(6):443, 2021. doi: 10.3390/jpm11060443.
- [24] Rahul Mundlamuri, Ganesh Reddy Gunnam, Nikhil Kumar Mysari, and Jayakanth Pujuri. The evolution of ai: From classical machine learning to modern large language models. *IEEE Access*, 2025.
- [25] Raja Vavekanand, Pinja Karttunen, Yue Xu, Stephanie Milani, and Huao Li. Large language models in healthcare decision support: A review. 2024. doi: 10.20944/preprints202407.1842.v1. Preprint.
- [26] Kristian Kersting. Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data*, 1:6, 2018. doi: 10.3389/fdata.2018.00006.
- [27] Cheng He, Ye Tian, and Zhichao Lu. Artificial evolutionary intelligence (aei): evolutionary computation evolves with large language models. *Journal of Membrane Computing*, 7(2): 135–152, 2025. doi: 10.1007/s41965-024-00172-x.
- [28] Chinmay Shripad Kulkarni. The evolution of large language models in natural language understanding. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 1(1):49–53, 2023. doi: 10.51219/JAIMLD/chinmay-shripad-kulkarni/28.
- [29] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024.
- [30] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025. doi: 10.1145/3744746.
- [31] Pinar Ersoy and Mehmet Erşahin. Optimal llm execution strategies for llama 3.1 language models across diverse hardware configurations: A comprehensive guide. *Computational Intelligence and Machine Learning*, 5, 2024.
- [32] Luis Roque. The evolution of llama: From llama 1 to llama 3.1, 2025. URL <https://medium.com/data-science/the-evolution-of-llama-from-llama-1-to-llama-3-1-13c4ebe96258>. Available online.

- [33] Satyadhar Joshi. A comprehensive review of qwen and deepseek llms: Architecture, performance and applications, 2025. Unpublished manuscript.
- [34] Rupesh Phogat, Dheeraj Arora, Pawan Singh Mehra, Jatin Sharma, and Diksha Chawla. A comparative study of large language models: Chatgpt, deepseek, claude and qwen. In *2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*, pages 609–613. IEEE, 2025.
- [35] Liang Zhang, Yang Liu, Yu Luo, Fei Gao, and Jun Gu. Qwen-ig: A qwen-based instruction generation model for llm fine-tuning. In *Proceedings of the 13th International Conference on Computing and Pattern Recognition*, pages 295–302, 2024.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [37] Yunpeng Huang, Jingwei Xu, Junyu Lai, Zixu Jiang, Taolue Chen, Zenan Li, Yuan Yao, Xiaoxing Ma, Lijuan Yang, Hao Chen, Shupeng Li, and Penghao Zhao. Advancing transformer architecture in long-context large language models: A comprehensive survey, 2023.
- [38] Andrea Filippo Ferraris, Davide Audrito, Luigi Di Caro, and Cristina Poncibò. The architecture of language: Understanding the mechanics behind llms. In *Cambridge Forum on AI: Law and Governance*, volume 1, page e11. Cambridge University Press, 2025.
- [39] Haojie Zhang, Ge Li, Jia Li, Zhongjin Zhang, Yuqi Zhu, and Zhi Jin. Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively. In *Advances in Neural Information Processing Systems*, volume 35, pages 21442–21454, 2022.
- [40] Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. *Informatics*, 11(3):57, 2024. doi: 10.3390/informatics11030057.
- [41] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. A systematic review of testing and evaluation of healthcare applications of large language models (llms), 2024.
- [42] Ummara Mumtaz, Awais Ahmed, and Summaya Mumtaz. Llms-healthcare: Current applications and challenges of large language models in various medical specialties, 2023.
- [43] Muhammad Arslan, Hicham Ghanem, Saad Munawar, and Christophe Cruz. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790, 2024. doi: 10.1016/j.procs.2024.10.303.
- [44] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2023.

- [45] Karl Keat, Rasika Venkatesh, Yidi Huang, Rachit Kumar, Sony Tuteja, Katrin Sangkuhl, Binglan Li, Li Gong, Michelle Whirl-Carrillo, Teri E. Klein, Marylyn D. Ritchie, and Dokyoon Kim. Pgxqa: A resource for evaluating llm performance for pharmacogenomic qa tasks. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 229–246, 2024.
- [46] Mullai Murugan, Bo Yuan, Eric Venner, Christie M. Ballantyne, Katherine M. Robinson, James C. Coons, Liwen Wang, Philip E. Empey, and Richard A. Gibbs. Empowering personalized pharmacogenomics with generative ai solutions. *Journal of the American Medical Informatics Association*, 31(6):1356–1366, 2024. doi: 10.1093/jamia/ocae039.
- [47] Mike Zack, Ioan Slobodchikov, Danil Stupichev, Alex Moore, David Sokolov, Igor Trifonov, and Allan Gobbs. Benchmarking large language models for replication of guideline-based pharmacogenomic recommendations. *The Pharmacogenomics Journal*, 25(1):23, 2025.
- [48] Ashley Rector, Keaton Minor, Kamden Minor, Jeff McCormack, Beth Breeden, Ryan Nowers, and Jay Dorris. Validating pharmacogenomics generative artificial intelligence query prompts using retrieval-augmented generation (rag). *arXiv preprint arXiv:2507.21453*, 2025.
- [49] Dhanush Borishetty, Peter Banda, Nikhilesh Andhi, Aakash Desai, Bharath Ram Uppili, Naga Mithil Samudrala, Shravani Shriya Palanki, Dheeraj Reddy Bobbili, and Gayatri Rangarajan Iyer. Pgxrag: A retrieval augmented generation supported pharmacogenomics assistant. *medRxiv*, 2025.
- [50] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.
- [51] Yaejin Jeon, Mi Seon Youn, Sunghoon Kang, Jonghyung Park, Eun Seok Kim, Juyoung Kim, and Ju Han Kim. Hierarchical rag enhances a pharmacogenomic ai assistant in guideline-related queries. *Computers in Biology and Medicine*, 200:111323, 2026. doi: 10.1016/j.compbiomed.2025.111323.
- [52] Stefan Thottunkal, Claire Spahn, Benjamin Wang, Nidhi Rohatgi, Jison Hong, Abha Khandelwal, and Latha Palaniappan. Clinician experiences at the frontier of pharmacogenomics and future directions. *Journal of Personalized Medicine*, 15(7):294, 2025. doi: 10.3390/jpm15070294.
- [53] Clinical Pharmacogenetics Implementation Consortium (CPIC). Cpic: Clinical pharmacogenetics implementation consortium, n.d. URL <https://www.clinpgx.org/cpic>. Accessed: April 13, 2026.
- [54] ClinPGx. The evolution of pharmgkb + cpic, n.d.. URL <https://www.clinpgx.org>. Accessed: April 13, 2026.
- [55] CPICPGx. Home. <https://github.com/cpicpgx/cpic-data/wiki#data-models>, n.d. GitHub repository. Accessed: April 13, 2026.

- [56] Meta AI. Llama-3.1-8b-instruct, 2024. URL <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>. Hugging Face model repository. Accessed: April 14, 2026.
- [57] Menglin Yang, Jialin Chen, Jinkai Tao, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Regina Zhang, Min Zhou, Irwin King, and Rex Ying. Low-rank adaptation for foundation models: A comprehensive review. *arXiv preprint arXiv:2501.00365*, 2024.
- [58] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL <https://arxiv.org/abs/2106.09685>.
- [59] Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- [60] M. Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1–11, 2015. doi: 10.5121/ijdkp.2015.5201.
- [61] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2019.
- [62] Baoli Li and Liping Han. Distance weighted cosine similarity measure for text classification. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 611–618, Berlin, Heidelberg, 2013. Springer.
- [63] Jiacheng Gu, Xiaowei Jiang, Zhen Shi, Hongyi Tan, Xiaoxiao Zhai, Cheng Xu, and Jiafeng Guo. A survey on llm-as-a-judge. *The Innovation*, 2024. doi: 10.1016/j.xinn.2024.100358.
- [64] Confident AI. Llm-as-a-judge metrics, 2024. URL <https://www.confident-ai.com/docs/llm-evaluation/core-concepts/llm-as-a-judge>. Accessed: 2026-04-12.
- [65] Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawajo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–14. ACM, 2024.
- [66] Sentence-Transformers. all-minilm-l6-v2. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>, 2020. Accessed: 2026.
- [67] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harlie, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *npj Digital Medicine*, 5:194, 2022. doi: 10.1038/s41746-022-00742-2.
- [68] Li Fang, Qingyu Chen, Chih-Hsuan Wei, Zhiyong Lu, and Kai Wang. Bioformer: An efficient transformer language model for biomedical text mining. *arXiv preprint arXiv:2302.01588*, 2023. URL <https://arxiv.org/abs/2302.01588>.

- [69] Jean-Baptiste Excoffier, Tom Roehr, Alexei Figueroa, Jens-Michalis Papaioannou, Keno Bressemer, and Matthieu Ortala. Generalist embedding models are better at short-context clinical semantic search than specialized embedding models. *arXiv preprint arXiv:2401.01943*, 2024. URL <https://arxiv.org/abs/2401.01943>.
- [70] ClinPGx. Annotation of epic guideline for abacavir and hla-b. <https://www.clinpgx.org/guidelineAnnotation/PA166104997>, n.d.. Accessed: 16-Apr-2026.
- [71] ClinPGx. Annotation of dpwg guideline for acenocoumarol and vkorc1. <https://www.clinpgx.org/guidelineAnnotation/PA166104938>, n.d.. Accessed: 17-Apr-2026.
- [72] ClinPGx. Annotation of epic guideline for amikacin, dibekacin, gentamicin, kanamycin, neomycin, netilmicin, paromomycin, plazomicin, ribostamycin, streptomycin, tobramycin and mt-rnr1. <https://www.clinpgx.org/guidelineAnnotation/PA166229081>, n.d.. Accessed: 17-Apr-2026.
- [73] ClinPGx. Annotation of epic guideline for allopurinol and hla-b. <https://www.clinpgx.org/guidelineAnnotation/PA166105003>, n.d.. Accessed: 17-Apr-2026.
- [74] ClinPGx. Annotation of dpwg guideline for allopurinol and abcg2. <https://www.clinpgx.org/guidelineAnnotation/PA166264961>, n.d.. Accessed: 17-Apr-2026.
- [75] ClinPGx. Annotation of dpwg guideline for acenocoumarol and cyp2c9. <https://www.clinpgx.org/guidelineAnnotation/PA166104979>, n.d.. Accessed: 17-Apr-2026.
- [76] Gillian C. Bell, Kelly E. Caudle, Michelle Whirl-Carrillo, Ronald J. Gordon, Keiko Hikino, Cynthia A. Prows, Andrea Gaedigk, Jose A. G. Agundez, Senthilkumar Sadhasivam, Teri E. Klein, and Matthias Schwab. Clinical pharmacogenetics implementation consortium (epic) guideline for cyp2d6 genotype and use of ondansetron and tropisetron. *Clinical Pharmacology & Therapeutics*, 102(2):213–218, August 2017. doi: 10.1002/cpt.598.
- [77] Michelle Whirl-Carrillo, Rachel Huddart, Li Gong, Katrin Sangkuhl, Caroline F. Thorn, Ryan Whaley, and Teri E. Klein. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 110: 563–572, 2021. doi: 10.1002/cpt.2350.
- [78] Samantha Breaux. Pharmacogenomics: From experimental design through patient interactions. Master’s thesis, University of British Columbia, 2021. URL <https://open.library.ubc.ca/collections/ubctheses/24/items/1.0398201>.
- [79] J. Kevin Hicks, Katrin Sangkuhl, Jesse J. Swen, Vicki L. Ellingrod, Daniel J. Müller, Kazutaka Shimoda, Jeffrey R. Bishop, Evan D. Kharasch, Todd C. Skaar, Andrea Gaedigk, Henry M. Dunnenberger, Teri E. Klein, Kelly E. Caudle, and Julia C. Stingl. Clinical pharmacogenetics implementation consortium guideline (epic) for cyp2d6 and cyp2c19 genotypes and dosing of tricyclic antidepressants: 2016 update. *Clinical Pharmacology & Therapeutics*, 102(1):37–44, July 2017. doi: 10.1002/cpt.597.

- [80] U.S. Food and Drug Administration. Table of pharmacogenetic associations: Section 1: Pharmacogenetic associations for which the data support therapeutic management recommendations; section 2: Pharmacogenetic associations for which the data indicate a potential impact on safety or response; section 3: Pharmacogenetic associations for which the data demonstrate a potential impact on pharmacokinetic properties only, 2022. URL <https://www.fda.gov/medical-devices/precision-medicine/table-pharmacogenetic-associations>. Accessed 2026-04-17.
- [81] ClinPGx. Annotation of cpic guideline for fluorouracil and dpyd, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166122686>. Accessed 2026-04-17.
- [82] ClinPGx. Annotation of dpwg guideline for fluorouracil and dpyd, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166104939>. Accessed 2026-04-17.
- [83] ClinPGx. Annotation of cpic guideline for celecoxib, flurbiprofen, ibuprofen, lornoxicam and cyp2c9, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166191841>. Accessed 2026-04-17.
- [84] ClinPGx. Annotation of cpic guideline for desflurane, enflurane, halothane, isoflurane, methoxyflurane, sevoflurane, succinylcholine and cacna1s, ryr1, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166303941>. Accessed 2026-04-17.
- [85] ClinPGx. Annotation of rnpqx guideline for hmg coa reductase inhibitors, simvastatin and slco1b1, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166202561>. Accessed 2026-04-17.
- [86] ClinPGx. Annotation of dpwg guideline for siponimod and cyp2c9, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166211021>. Accessed 2026-04-17.
- [87] ClinPGx. Annotation of dpwg guideline for warfarin and vkorc1, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166182841>. Accessed 2026-04-17.
- [88] ClinPGx. Annotation of cpic guideline for vortioxetine and cyp2d6, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166288221>. Accessed 2026-04-17.
- [89] ClinPGx. Annotation of cpic guideline for tropisetron and cyp2d6, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166161955>. Accessed 2026-04-17.
- [90] ClinPGx. Annotation of cpic guideline for chlorpropamide, dabrafenib, gliclazide, glimepiride, glipizide, mesalazine, moxifloxacin, nicorandil, nitrofurazone, probenecid, quinacrine, sodium nitrate, sulfacetamide, tolazamide, trametinib and g6pd, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166279481>. Accessed 2026-04-17.
- [91] ClinPGx. Annotation of cpic guideline for tenoxicam and cyp2c9, n.d.. URL <https://www.clinpgx.org/guidelineAnnotation/PA166192341>. Accessed 2026-04-17.
- [92] ClinPGx. Annotation of cpic guideline for amikacin, dibekacin, gentamicin, kanamycin, neomycin, netilmicin, paromomycin, plazomicin, ribostamycin, streptomycin, tobramycin and mt-rnr1, 2025. URL <https://www.clinpgx.org/guidelineAnnotation/PA166229081>. Accessed: 2026-04-17.

-
- [93] Yinuo Wang, Baiyang Wang, Robert E. Mercer, Frank Rudzicz, Sudipta Singha Roy, Pengjie Ren, Zhumin Chen, and Xindi Wang. Trustworthy medical question answering: An evaluation-centric survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27489–27502. Association for Computational Linguistics, November 2025.
- [94] Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226, 2025. doi: 10.3390/ai6090226.
- [95] Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. Morqa: Benchmarking evaluation metrics for medical open-ended question answering. *arXiv preprint arXiv:2509.12405*, 2025.
- [96] Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Systems*, 2(1):6, 2025. doi: 10.1038/s41746-025-01040-9.