HAMAD BIN KHALIFA UNIVERSITY

COLLEGE OF SCIENCE AND ENGINEERING

DEVELOPMENT AND EVALUATION OF AN AGENTIC LLM-BASED RAG-ENABLED FRAMEWORK FOR EVIDENCE-BASED PATIENT EDUCATION

BY

ALHASAN ALSAMMARRAIE

A Thesis Submitted to the Faculty of the

College of Science and Engineering

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

May 2025

ABSTRACT

Patient Education (PE) is crucial for empowering individuals in managing their health, yet significant challenges persist in delivering effective and accessible Patient Education Materials (PEMs), especially for diverse linguistic populations such as Arabic speakers. Large Language Models (LLMs) offer a promising avenue to address these challenges. This thesis details the development and evaluation of the Amal Framework, an LLM-based agentic retrieval-augmented generation (ARAG) system designed to produce evidence-based PEMs, with a specific application for the Arabic language.

The research first establishes a foundational understanding of LLMs, Retrieval Augmented Generation (RAG), prompt engineering, and agentic workflows. It then presents a comprehensive scoping review of the current use of LLMs in patient education, identifying key trends and research gaps, notably the underutilization of RAG and open-source models for non-English PEMs, and the lack of standardized evaluation.

The core of this work lies in the design and implementation of the Amal Framework. This involved creating a robust data pipeline, which included sourcing and processing a substantial corpus of Arabic medical texts from reputable sources. A multi-component generation pipeline was developed, integrating carefully selected open-source LLMs with the ARAG system and a novel Validation Agent (VA) designed to ensure the safety, accuracy, and appropriateness of the generated PEMs.

A rigorous two-part evaluation methodology was devised. The first part assessed the quality of the generated Arabic PEMs across multiple experimental setups, using automated LLM-based scoring and subsequent expert validation on metrics including accuracy, readability, comprehensiveness, appropriateness, and safety. The second part specifically evaluated the performance of the VA in identifying and blocking harmful or unsuitable medical advice using a specialized benchmark dataset.

The findings from these evaluations demonstrate the efficacy of the Amal Framework in generating high-quality, evidence-based Arabic PEMs and the effectiveness of the integrated VA in enhancing safety. This research contributes a novel, systematically evaluated framework for leveraging advanced AI to improve patient education, particularly for underserved linguistic communities, and offers a blueprint for the responsible and effective deployment of LLMs in healthcare. The thesis concludes by discussing the implications of these findings, the limitations of the current work, and directions for future research in this rapidly evolving field.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The completion of this doctoral journey would not have been possible without the steadfast support, guidance, and encouragement of many individuals. I wish to express my deepest gratitude to all those who have played a part, whether big or small, in helping me reach this significant milestone.

First and foremost, I extend my heartfelt thanks to my parents and my entire family. Their unconditional love and unwavering support have been my bedrock throughout this challenging endeavor. They were there for me when I needed them the most, consistently believing in me and offering encouragement during difficult times. They were my number one supporters, and for their constant presence and belief in me, I am eternally grateful.

I am profoundly indebted to my PhD supervisor, Dr. Mowafa Househ. His insightful guidance, continuous encouragement, and unwavering support through all the hardships and complexities of doctoral research were invaluable. It is truly thanks to his mentorship and consistent belief in my work that I was able to navigate this path and ultimately achieve my PhD. His wisdom and patience have been a constant source of inspiration.

I also wish to acknowledge the valuable technical assistance I received from Ali Al-Saifi and Mohamad Bahri. As skilled engineers, their expertise provided limited, albeit highly impactful, technical support at crucial junctures of my research.

DECLARATION

This is to certify that the work described in this thesis is entirely my own, unless otherwise referenced or acknowledged. This work has not previously been submitted for qualifications at any other academic institution. I also declare that in the process of writing this thesis, chatGPT o3 and o3-mini were used as proofreading assistants, for rephrasing statements, checking grammar, and for suggesting improvements to clarity and style. It is important to affirm that these AI tools performed a strictly supplementary role in the development of this thesis. All outputs, suggestions, or text generated by these AI assistants were personally and critically reviewed, edited, and verified by myself. I have ensured that any such AI-generated contributions were used appropriately and that the final content, arguments, analysis, and intellectual contribution of this PhD are entirely my own.

1, June 2025

CHAPTER 1: INTRODUCTION

Patient Education (PE) is a planned educational process that aims to deliver patients or the general public with the necessary facts, advice, or data to improve or maintain their health by prompting positive behavioral changes, treatment adherence, and engaging in self-care Bastable, 2014; Falvo, 2011. PE also encompasses other concepts such as the comprehension of conditions and risks, treatment options and potential outcomes, and matters related to patient-informed consent Coulter and Ellins, 2007; Falvo, 2011.

Patient Education Materials (PEMs) can be delivered through a wide array of formats and media. These range from one-on-one teaching, counseling, and behavioral treatments to printed materials (like brochures), video and audio resources, computer-based programs, smartphone and tablet applications delivering timely information via push notifications, and patient portals integrated into electronic health records Pellisé et al., 2009. While patients may not express strong preferences for specific media, the paramount concern is the quality, reliability, and up-to-date nature of the information content presented in an acceptable and useful form Timmers et al., 2020.

In the past, patient education might have focused primarily on the transfer of biomedical knowledge and advice, essentially instructing the patient on what to do Wittink and Oosterhaven, 2018. However, research indicated this approach was often ineffective and sometimes counterproductive Wittink and Oosterhaven, 2018.

The contemporary patient-centered approach views the patient as an active partner in their healthcare, not merely a passive recipient of diagnostic tests and treatments Pellisé et al., 2009. This paradigm shift necessitates that patient education be tailored to meet the individual's unique needs, preferences, values, and circumstances Wittink and Oosterhaven, 2018. This enables patients to participate meaningfully with their practitioners in deciding on a course of action that suits them best while also being clinically aligned Lataillade and Laurent Chabal, 2020.

Furthermore, effective PE carries significant economic implications. By empowering patients to better manage their conditions and adhere to treatments, education can contribute to reduced healthcare expenses AlSammarraie and Househ, 2025. Evidence suggests it leads to fewer visits to medical facilities and decreased medication use Simonsmeier et al., 2022, potentially lowering the overall cost burden on both the individual and the healthcare system Timmers et al., 2020. Specifically, approaches like PEMs which aim to reduce the long-term costs associated with chronic disease management Lataillade and Laurent Chabal, 2020.

despite its recognized value, faces significant hurdles across multiple levels. Systemically, constraints like insufficient time during clinical encounters Pellisé et al., 2009, limited resources including funding, staffing, and appropriate materials Cutilli, 2020. At the provider level, challenges include gaps in knowledge and skills regarding effective educational techniques Jason, 2021, communication issues such as using medical jargon or failing to verify understanding Askin, 2017. Patient-centric factors also create barriers, most notably low health literacy which limits comprehension and application of information Pellisé et al., 2009, alongside cultural beliefs, language differences, varying learning styles, cognitive or physical limitations, emotional states like stress or anxiety, and socio-economic constraints that impact a patient's ability to engage with and act on educational content Al-Jumaili et al., 2020; Pellisé et al., 2009. These barriers often interact, creating a complex environment that makes delivering consistently effective patient education difficult.

The challenges are further compounded when addressing non-English speaking populations, such as Arabic speakers where medical sources are scarce and are found to be of lower quality Alassaf et al., 2024. Many concerns of misleading information, lack of specificity, and the over simplification of health concepts have historically persisted when evaluating PEMs in Arabic Aldabbagh Dina et al., 2013. Despite that, there has been a growing number of initiatives that target the Arabic speakers to provide higher quality, evidence-based PEMs. Such initiatives are exemplified in the publication of the King Abdullah Bin Abdul-Aziz Arabic Health Encyclopedia (KAAHE) Altuwaijri, 2011, Altibbi's, WebTeb's, and Mayo Clinic's (AR version) web platforms.

Large Language Models (LLMs), such as ChatGPT, Gemini, and Claude, represent a significant technological advancement with the potential to address several challenges in patient education Aydin et al., 2024. These AI systems, trained on vast text datasets, can generate human-like text and perform various tasks relevant to healthcare communication Hamid and Brohi, 2024. Potential applications include generating or drafting new patient education materials Shah et al., 2024 enhancing existing ones by simplifying complex language AlSammarraie and Househ, 2025, translating materials into multiple languages, interpreting medical information for patients, providing lifestyle and medication guidance, and potentially personalizing information based on individual needs. LLMs show particular promise in bridging language barriers and improving health literacy by making information more accessible AlSammarraie and Househ, 2025; Aydin et al., 2024; Shah et al., 2024.

1.1 Thesis Scope and Contributions

This thesis focuses on the development and evaluation of an LLM-based agentic retrieval-augmented generation framework for evidence-based patient educations.

1.1.1 Evidence Based Arabic PEM Data Pipeline

Retrieval Augmented Generation (RAG), an approach that enhances large language model outputs by grounding them with information retrieved from external knowledge sources (details in Chapter 2) Lewis et al., 2020, has its efficacy and trustworthiness heavily reliant on the quality, relevance, and structure of the underlying knowledge base used for retrieval X. Wang et al., 2024a. In the context of generating Patient Education Materials (PEMs) the creation of a sound and verifiable data pipeline is important. This involves not only sourcing appropriate content but also processing it effectively for semantic search and retrieval. Challenges include identifying reliable sources Aldabbagh Dina et al., 2013, selecting embedding models capable of capturing the nuances of of the corpus's terminology, and determining chunking strategies that preserve context while mitigating lost context and minimizing noise S. Zhao et al., 2024.

Existing best practices for RAG often focus on general English language corpora, leaving a gap in established methodologies specifically designed for building and optimizing knowledge bases for Arabic RAG in general and for Arabic medical RAG more specifically. To address this gap and provide a suitable foundation for the Amal framework, this thesis contributed a structured data pipeline designed specifically for creating an evidence based PEM knowledge source prepared for RAG applications. Key contributions within this pipeline include:

1. Expert Guided Data Sourcing and Collection: A methodical process was established to curate a corpus of 52,625 articles from eight reputable Arabic medical websites. These sources, including prominent platforms such as the Arabic version of Mayo Clinic, WebTeb, Altibbi, and the King Abdullah Bin Abdulaziz

Arabic Health Encyclopedia KAAHE among others, were identified through expert consultation. Data was collected using automated tools Firecrawl API and subjected to filtering to remove noise e.g., ads, headers, footers and ensure content quality suitable for PEM generation.

2. Systematic Embedding Model Evaluation and Selection: Given the foundational role of text embeddings in enabling effective RAG, a specific evaluation process was developed to identify a suitable embedding model for the Arabic medical domain. This process began by identifying several candidate models known for multilingual support. To empirically determine the best fit, a testing methodology was employed using sample articles from the project's Arabic medical corpus covering distinct topics and corresponding questions. For each candidate model, embeddings were generated for these texts, and cosine similarities were computed for all article question pairs. The model that demonstrated the most effective differentiation between correctly matched and unrelated pairs, thereby indicating a stronger grasp of semantic relationships in this specific context, was jina-embeddings-v3, which was subsequently selected for the framework.

3. Optimized Chunking Strategy: An informed, Hybrid, sentence based chunking strategy was implemented, moving beyond simple fixed token approaches. Based on statistical analysis of the collected Arabic corpus and RAG best practice recommendations for scientific text, a chunk size of 11 sentences was calculated and applied. This approach was chosen to increase information granularity for retrieval, mitigate bias from over reliance on individual large chunks, and facilitate the synthesis of information from multiple passages, while reducing context fragmentation.

Collectively, these steps constitute a reproducible and evidence driven data pipeline for preparing high quality Arabic medical text for effective use within RAG systems like Amal.

1.1.2 The Generation Pipeline

While a well structured data pipeline provides the necessary foundation, the core generation of PEMs requires a carefully designed process that leverages the capabilities of LLMs while minimizing their inherent limitations. For specialized applications like generating PEMs, merely using a base LLM may not achieve the necessary trustworthiness for patient materials. This is due to the potential for LLMs to produce hallucinations, which are outputs that appear plausible but are factually incorrect, and their documented challenges with source citation, including the generation of fabricated references M. Zhang and Zhao, 2025. Techniques such as prompt engineering, RAG OpenAI, n.d.-c, and the use of AI Agents (Intelligent systems programmed to perform complex tasks such as validation or web-searching) offer pathways to enhance output quality, relevance, and safety Google Cloud, n.d.-b. However, the optimal configuration of these elements, particularly when leveraging open-source LLMs which are more suitable for local deployment in healthcare settings is an area requiring investigation. Compounding this, with the gap in assessment of Arabic LLM-generated PEMs AlSammarraie and Househ, 2025, there was a clear need to develop and evaluate a comprehensive generation and validation pipeline tailored for this specific context.

This thesis contributed the design and systematic evaluation of a multi component generation pipeline specifically for producing Arabic PEMs. This pipeline integrates LLMs, RAG, and an agent based validation step. The key contributions in developing this generation pipeline include:

1. Principled Open Source LLM Selection: A set of criteria was established to identify suitable LLMs, emphasizing open source availability, a parameter count appropriate for potential hospital deployment (less than 32 billion parameters), explicit Arabic language support, and origin from reputable developers. This process led to the selection of twelve distinct LLMs for evaluation, encompassing both general purpose models (like Qwen 2.5, Phi 4, Gemma 2, Mistral Small) and Arabic centric models (like Fanar, Jais, AceGPT).

2. Tailored Agentic RAG (ARAG) System Design: An ARAG system was developed to enhance the factual grounding of the generated PEMs. This system integrates the previously established data pipeline for context retrieval. Upon receiving a user query, the system retrieves the top three most relevant text chunks from the Arabic medical corpus. This retrieved context is then combined with the user query and provided to the candidate LLM to generate an initial response, aiming

to produce more accurate, comprehensive, safe, and contextually relevant PEMs than a base LLM alone.

3. Integration of a Validation Agent (VA) for Safety and Refinement: Recognizing the critical importance of safety in patient facing materials, a VA was incorporated as a distinct subsequent step in the RAG pipeline. This agent, implemented by a second inference call to the same base LLM used for initial generation but guided by a specialized prompt, is tasked with assessing the initial PEM for harmful or unsuitable information. It can perform minor revisions for suitability (e.g., tone adjustment) or, when needed, block responses deemed unsafe, adding a layer of scrutiny before the PEM is finalized.

4. Investigation of Prompt Engineering: The impact of some prompt engineering techniques was also explored within this generation pipeline. Specific prompts were designed for both the initial PEM generation phase and for guiding the VA, aiming to steer the LLMs towards more accurate, appropriately formatted, and safer outputs.

This generation pipeline, through its combination of carefully selected LLMs, context augmentation via RAG, and a dedicated VA, represents a structured approach to producing more reliable Arabic PEMs.

1.1.3 The Evaluation Methodology

Evaluating the output of LLMs, particularly in sensitive applications like PEM generation, presents considerable challenges. For this work, we considered five key evaluation metrics: Accuracy, Readability, Comprehensiveness, Appropriateness, and Safety. The selection of these metrics was supported by expert consultation and findings from a previously conducted scoping review AlSammarraie and Househ, 2025. Furthermore, when employing the VA, the effectiveness of this validation component itself needs rigorous assessment.

To address these needs, this thesis contributed a two part evaluation methodology: one for assessing the generated PEMs and another for evaluating the performance of the VA.

1. **PEM Generation Evaluation Methodology:** A novel two-stage evaluation process was designed to manage the assessment of numerous experimental configurations (48 in total) efficiently while ensuring expert validation.

    - Automated Evaluation: An initial automated assessment was conducted using an LLM (ChatGPT o3-mini) chosen for its Arabic proficiency and ability to provide structured outputs. This stage evaluated all 48 experimental setups, where each of the 12 candidate LLMs was tested under four configurations (base LLM, prompt engineering, ARAG, and ARAG with prompt engineering). Each setup generated responses to 20 expert curated PEM questions (covering general medicine and oncology). These responses were scored on a 1 to 5 Likert-scale across the five key metrics mentioned above. A language validation rule was applied, penalizing non Arabic outputs. This automated stage served to identify the top five performing configurations for further scrutiny.
    - Expert Evaluation: The top five configurations identified in the automated stage were then subjected to manual evaluation by domain experts (an internal medicine specialist and an oncologist). These experts used the identical five metrics, the same Likert scale, and the same ground truth answers as the automated stage to re assess the PEMs. This expert review provided clinical validation of the automated findings and determined the overall best performing LLM setup.

2. **Validation Agent (VA) Evaluation Methodology:** An independent benchmark was developed to specifically assess the capability of the selected LLMs to function as VAs in blocking harmful PEMs.

    - Specialized Test Dataset: A dataset of 50 PEM examples was created, including physician validated instances of harmful medical advice integrated within otherwise safe appearing text. This dataset featured varying severities of harm, from potentially life threatening to unscientific advice, alongside non harmful control PEMs with different tones.
    - Instruction Following Assessment: Each LLM, when acting as a VA, was

instructed via few shot prompting to respond with a specific refusal phrase ("I am sorry I cannot help with this") if it detected harmful input.

- Performance Quantification: Adherence to this instruction was measured by calculating the cosine similarity (using jina v3 embeddings) between the LLM's actual response and the target refusal phrase. A similarity threshold, determined inductively, was used to classify responses as refusals or non refusals. These classifications were then compared against the ground truth to calculate standard performance metrics (True Positives, False Positives, False Negatives, True Negatives, and overall Accuracy) for each LLM as a VA.

This comprehensive evaluation framework allowed for both a broad assessment of different PEM generation strategies and a focused analysis of the critical safety function performed by the VA.

1.1.4 The Amal Framework

The preceding sections have detailed the distinct contributions made in developing a specialized data pipeline for Arabic medical content, a multi component generation pipeline integrating LLMs with RAG and a VA, and a comprehensive two part evaluation methodology. The culminating contribution of this thesis is the Amal Framework, which represents the systematic integration of these individual components into a cohesive and end to end system. Figure 1.1 highlights the architecture of Amal and its three core pillars: the Data Layer, the Generation Layer, and the Evaluation Layer.

The Amal Framework, therefore, is not merely a collection of isolated processes but rather a structured and evaluated approach specifically designed for the purpose of generating trustworthy, evidence-based PEMs. It operationalizes the findings from each development and evaluation stage, from the initial sourcing and preparation of high-quality medical data, through the principled selection and configuration of open-source LLMs within an ARAG architecture incorporating a VA for safety, to the rigorous assessment of both the generated PEMs and the VA's performance.

By bringing these elements together, the Amal Framework offers a novel and validated pathway to address the identified gaps in producing a more reliable and safer PEMs using advanced AI techniques. It provides a blueprint for leveraging LLMs in a responsible and effective manner for this healthcare application.

1.2 Overview

This thesis is structured to systematically present the development and evaluation of an LLM-based agentic retrieval-augmented generation framework for evidence-based patient education. The subsequent chapters will unfold as follows:

Chapter 2: Background provides a comprehensive review of the foundational concepts and technologies underpinning this work. It delves into the architecture and capabilities of Large Language Models, discussing their evolution and the predominant Transformer architecture. Furthermore, this chapter explores various methods employed to enhance LLM performance, with a particular focus on Retrieval Augmented Generation (RAG), including lexical and semantic retrieval techniques, prompt engineering strategies, and the emerging paradigm of agentic workflows.

Chapter 3: Scoping Review on the Use of LLMs for Patient Education presents a systematic scoping review of the current literature. This chapter introduces established readability metrics and PEM evaluation frameworks. It details the methodology employed for the review, including search strategies, inclusion/exclusion criteria, and a novel RAG-inspired data extraction process. The results section presents key findings, including a PRISMA flowchart, terminology standardization, and an analysis of variables related to study demographics, LLM characteristics, prompting techniques, and PEM assessment. The discussion highlights current trends, identifies significant research gaps such as the limited use of RAG, the under-exploration of open-source LLMs for non-English PEMs, and the inconsistency in evaluation methodologies. The chapter concludes by summarizing the review's limitations and suggesting directions for future work in this domain.

Chapter 4: Agentic RAG Design and Evaluation Methodologies details the core contribution of this thesis: the design and implementation of the proposed Agentic Retrieval-Augmented Generation (ARAG) system, specifically tailored for generating Arabic PEMs. This chapter is divided into two main parts. The first part describes the Agentic RAG design, including the data pipeline (data sourcing from reputable Arabic medical websites, filtering, cleaning, embedding model selection, and chunking strategy) and the generation pipeline (selection criteria for candidate LLMs, and the design and role

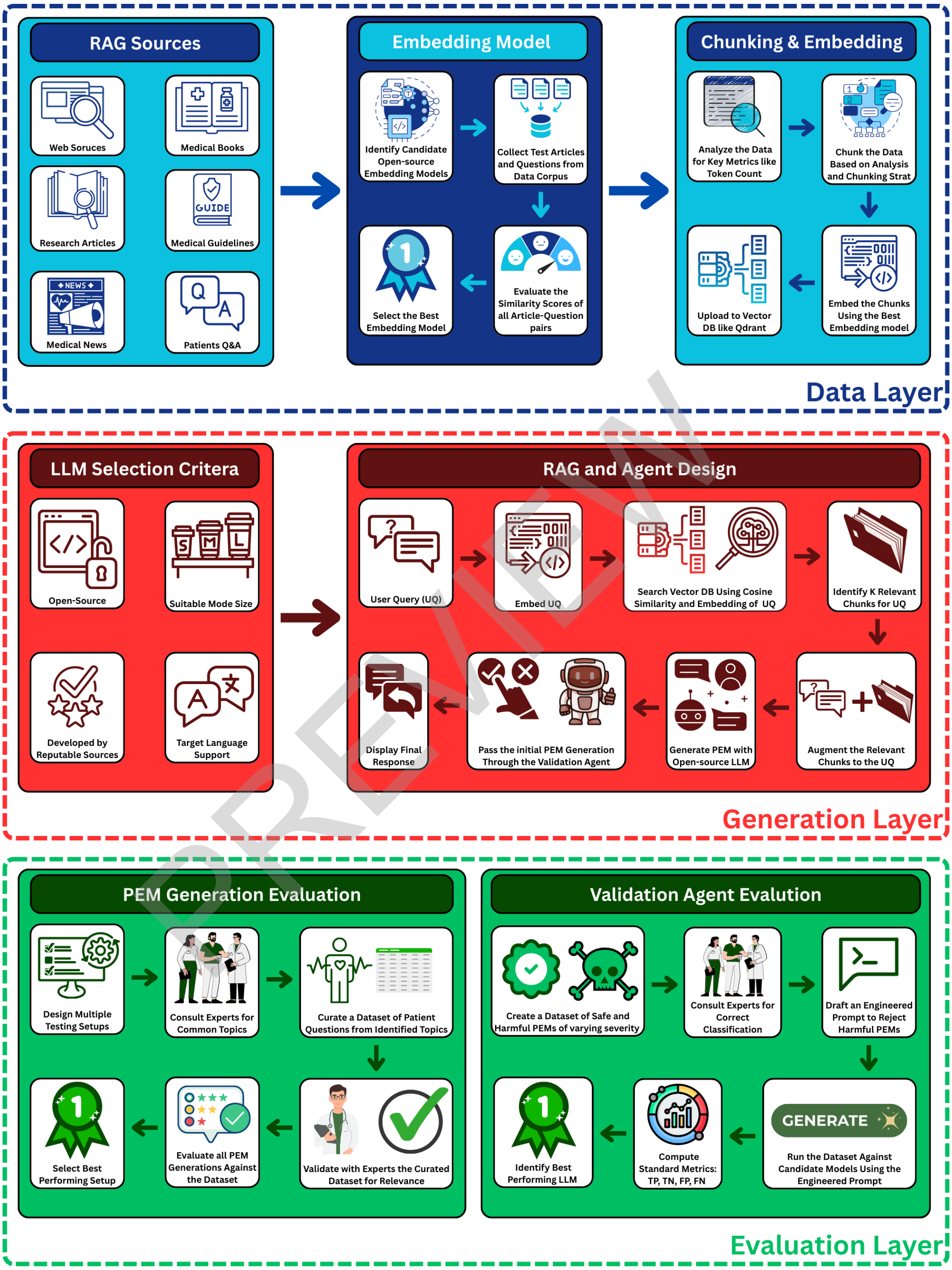
RAG Sources
Web Soruces
Medical Books
Research Articles
Medical Guidelines
Medical News
Patients Q&A
Embedding Model
Identify Candidate Open-source Embedding Models
Collect Test Articles and Questions from Data Corpus
Evaluate the Similarity Scores of all Article-Question pairs
Select the Best Embedding Model
Chunking & Embedding
Analyze the Data for Key Metrics like Token Count
Chunk the Data Based on Analysis and Chunking Strat
Embed the Chunks Using the Best Embedding model
Upload to Vector DB like Qdrant
Data Layer
LLM Selection Critera
Open-Source
Suitable Mode Size
Developed by Reputable Sources
Target Language Support
RAG and Agent Design
User Query (UQ)
Embed UQ
Search Vector DB Using Cosine Similarity and Embedding of UQ
Identify K Relevant Chunks for UQ
Augment the Relevant Chunks to the UQ
Generate PEM with Open-source LLM
Pass the initial PEM Generation Through the Validation Agent
Display Final Response
Generation Layer
PEM Generation Evaluation
Design Multiple Testing Setups
Consult Experts for Common Topics
Curate a Dataset of Patient Questions from Identified Topics
Validate with Experts the Curated Dataset for Relevance
Evaluate all PEM Generations Against the Dataset
Select Best Performing Setup
Validation Agent Evalution
Create a Dataset of Safe and Harmful PEMs of varying severity
Consult Experts for Correct Classification
Draft an Engineered Prompt to Reject Harmful PEMs
GENERATE
Run the Dataset Against Candidate Models Using the Engineered Prompt
Compute Standard Metrics: TP, TN, FP, FN
Identify Best Performing LLM
Evaluation Layer

Figure 1.1: Architecture of the Amal Framework, showcasing its three core pillars: the Data Layer, the Generation Layer, and the Evaluation Layer.

of a Validation Agent for ensuring safety and refinement). The second part outlines the comprehensive evaluation methodologies developed to assess the ARAG system. This includes the PEM generation evaluation (dataset curation, experimental setups, and a two-stage scoring process involving automated LLM assessment followed by expert review across metrics like accuracy, readability, comprehensiveness, appropriateness, and safety) and the Validation Agent evaluation (creation of a specialized dataset with harmful and non-harmful PEMs, and performance scoring based on the agent's ability to identify and block detrimental content).

Chapter 5: Experiment Results will present the empirical findings from the evaluation of the Amal framework and its components. This chapter will detail the performance of various LLMs under different configurations, including baseline performance, the impact of prompt engineering, and the effectiveness of the Agentic RAG system. Results from both the automated and expert evaluations of generated PEMs will be reported, alongside the performance metrics of the Validation Agent in identifying and mitigating harmful content.

Chapter 6: Comprehensive Discussion will provide an in-depth analysis and interpretation of the results presented in Chapter 5. This section will contextualize the findings within the broader landscape of LLMs in healthcare and patient education. It will discuss the implications of the results, the strengths and limitations of the developed Amal framework, and the practical considerations for deploying such systems in real-world healthcare settings, particularly for Arabic-speaking populations.

Chapter 7: Conclusion and Future Work will conclude the thesis by summarizing the key contributions, reiterating the main findings, and discussing their significance. This chapter will also outline potential avenues for future research, including potential enhancements to the Amal framework, exploration of other languages or medical domains, and further investigations into the ethical and practical challenges of using LLMs for patient education.