Enhancing Clinical Reasoning with Custom Large Language Models: A Multi-Agent Collaboration Framework and Retrieval Augmented Generation Approach

Submitted By

MD. IFFATUL ISLAM ANON

ID: 221-35-1065

Supervised By

MD. SHOHEL ARMAN

Assistant Professor

Thesis submitted in fulfillment of the requirements

for the award of the degree of Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

APPROVAL

This thesis titled on **"Enhancing Clinical Reasoning with Custom Large Language Models: A Multi-Agent Collaboration Framework and Retrieval Augmented Generation Approach"**, submitted by **Md. Iffatul Islam Anon (ID: 221-35-1065)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Afsana Begum
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Nadira Islam
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Manowarul Islam
Professor
Department of Computer Science and Engineering
Jagannath University, Bangladesh

Department of Software Engineering

Faculty of Science and Information Technology

Supervisor Approval Form

Fall 2025	B.Sc. In SWE	Campus: DSC

Student Name	Student ID
Md. Iffatul Islam Anon	221-35-1065

Project/Thesis Information	
Project/Thesis Title	Enhancing Clinical Reasoning with Custom Large Language Models: A Multi-Agent Collaboration Framework and Retrieval Augmented Generation Approach
Type of work	Thesis

Supervisor information	
Supervisor Name	Md. Shohel Arman
Supervisor Initial	MSA
Completed Credit till now	133
How many credits in this semester	6
Supervisor Consent	☑ Yes ☐ No

Supervisor Signature

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Md. Iffatul Islam Anon
Date of Birth	:	25 December 2002
Title	:	Enhancing Clinical Reasoning with Custom Large Language Models: A Multi-Agent Collaboration Framework and Retrieval Augmented Generation Approach
Academic Session	:	2022-2025

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☑	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

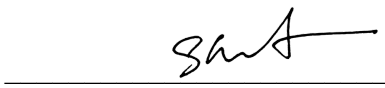
Certified by:

Anon	
(Student's Signature)	(Supervisor's Signature)
221-35-1065	Md. Shohel Arman
Student ID	Name of Supervisor
Date: 25 November 2025	Date: 25 November 2025

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.



(Supervisor's Signature)

Full Name : Md. Shohel Arman

Position : Assistant Professor

Date : 25 November 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Anon

(Student's Signature)

Full Name : Md. Iffatul Islam Anon

ID Number : 221-35-1065

Date : 25 November 2025

Enhancing Clinical Reasoning with Custom Large Language Models: A Multi-Agent Collaboration Framework and Retrieval Augmented Generation Approach

Submitted By

MD. IFFATUL ISLAM ANON

ID: 221-35-1065

Supervised By

MD. SHOHEL ARMAN

Assistant Professor

Thesis submitted in fulfillment of the requirements

for the award of the degree of Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENT

The greatest appreciation I would like to make is to my thesis supervisor, Md. Shohel Arman, who provided my guidance, constant support, and valuable feedback during my research work. His influence and experience played a very important role in developing this project.

I would also wish to acknowledge the faculty and staff members of the Software Engineering Department of Daffodil International University to give me the academic background needed to enable me to do this piece of work.

Lastly, I am blessed with the members of my families and friends who have been supportive and understanding to me all this time in this difficult and advantageous adventure.

DEDICATION

It is my privilege to state that this thesis was carried out under the guidance of Md. Shohel Arman, Assistant Professor, Department of Software Engineering, Daffodil International University. I sincerely acknowledge his influence, assistance, and encouragement during the process of developing this work.

I also confirm that this report is a purely personal work, and has not been presented, whole or part thereof, to any institution or program to gain academic credit or otherwise.

ABSTRACT

Clinical reasoning is core in medical practice that is susceptible to cognitive errors compromising safety and reliability. Despite being promising in clinical activity, such Large Language Models as ChatGPT and Med-PaLM provide unverifiable content and tend to be in interpretable form, as well as do not possess the feature of collaborative decision-making that is inherent to a real clinical team.

In this thesis, a **Multi-Agent Collaboration Framework** with **Retrieval-Augmented Generation (RAG)** is proposed to reinforce clinical reasoning. The two teams of four specialized agents, **Case Analyzer, Evidence Validator, Treatment Planner, and Clinical Reporter** work in parallel and have a central **Collaborative Orchestrator Agent**. The system creates consensus and minimizes reasoning differences by conducting repeated cycles of distributed analysis, cross-team comparison and refinement based on feedback. RAG makes all intermediate and final outputs to be based on verifiable clinical evidence.

This piece of work is a blueprint of how to come up with a more transparent, reliable, and trustworthy Clinical Decision Support Systems.

Keywords: Artificial Intelligence, Large Language Models, Multi-Agent Systems, Retrieval-Augmented Generation, Clinical Decision Support Systems.

TABLE OF CONTENTS

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

- AI: Artificial Intelligence
- LLM: Large Language Model
- MAS: Multi-Agent System
- RAG: Retrieval-Augmented Generation
- CDSS: Clinical Decision Support System
- EHR: Electronic Health Record

CHAPTER 1

INTRODUCTION

1.1. Background and Motivation

Clinical reasoning is the critical cognitive activity which enables a doctor to filter through the mass of patient information, construct a diagnosis differential, and come up with a coherent management program. This is a fundamentally complex, immensely demanding, and error-prone process that underlies medical decision-making. Research has established that the effects of diagnostic errors continue to be an important contribution to patient morbidity and mortality due to cognitive errors, system failures, or both.

Largely Language Models (LLMs) have been shown to exhibit impressive capabilities in various fields including medicine in the past few years. Artificial intelligence systems like Med-PaLM 2 have demonstrated a level of expertise, passing medical licensure tests, and indicate a vast potential to supplement clinical processes. Such potential, though, is mellowed with great, well-reported risks. Standard LLMs can be affected by hallucination, or the production of factually false but plausible information. Such fabrications are not acceptable in a clinical setting.

Moreover, as mentioned in the background research of this thesis, modern LLLMs tend to generate unverified information, are not transparent in their thinking, and do not inherently represent the teamwork and collaboration of the contemporary clinical practice. Medicine is not usually an individual endeavor; it is teams of experts, nurses, and pharmacists that consult, confirm and reach consensus. One, mono-block AI that provides a black-box opinion does not conform to this trusted paradigm. This underscores the importance of more dependable, explicable and collaborative AI solutions in healthcare.

1.2. Problem Statement

In spite of the fast development of the use of LLMs in medicine, the foundation of trust and reliability remains the problem that prevents their clinical implementation. The lack of factual basis and a single-agent or one-sided way of thinking process predisposes them to be fragile in high-stakes clinical practices. In most cases when an LLM makes a diagnosis it lacks verifiable citations, a clear line of reasoning and the strong validation of a second opinion.

As a consequence, the major issue is the absence of an AI structure capable of working together, publicly verifying its claims with vested sources, and dynamically constructing a resilient, consensus-based clinical strategy that reflects the safety and rigor of a human clinical staff.

1.3. Research Gaps

A number of gaps in the existing literature advise the articulation of the problem statement:

1. The Single-Agent Limitation: A majority of clinical LLM applications can be a single agent. This model of a solo practitioner does not have the inherent error-checking, peer-review, and other pooling of viewpoints that characterize expert human medical teams.
2. Absence of Integrated Verification: Although there are external tools that are used with some models, validation is not usually given serious consideration. TVFs should be built in which evidence retrieval/verification (i.a., RAG) is not merely an extension, but part of the very reasoning process.
3. Lack of Collaborative-Consensus Models: The research gap is on conducting studies related to the ability of multiple AI agents to cooperate to give their recommendations on a given clinical case, identify conflict, and how they can resolve the conflict systematically to achieve a safe and reliable decision.

1.4. Research Questions

In a bid to bridge the above gaps, this thesis aims at answering the following main research questions:

1. RQ1: What is the best way to develop a multi-agent systems framework that can model the process of clinical team collaboration and specialization in humans?
2. RQ2: What are the effects of the deep integration of a Retrieval-Augmented Generation (RAG) pipeline at agent level on the factual grounding, verifiability and guideline adherence of the output of the system?
3. RQ3: Does an iterative, dual-team consensus mechanism have a statistically significant effect of enhancing the reliability, safety, and robustness of clinical recommendations in comparison to the single-agent baseline or a non-iterative multi-agent system?

1.5. Research Objectives

In order to respond to these questions, this study aims to achieve the following objectives:

1. RO1: To develop and define a new multi-agent architecture with specialized and cooperative roles (Analyzer, Reasoner, Treatment Planner) focused on clinical case analysis.
2. RO2: To create and execute a Collaborative Orchestrator which provides an iterative refinement and consensus-building process between two symmetric agent teams.
3. RO3: To integrate an agent-generated assertion (diagnoses, treatment plans) pipeline, which is based on verifiable medical knowledge.

1.6. Research Scope

The best definition of this thesis is the design, implementation, and evaluation of a computational framework. Our implementation of a proof-of-concept will use publicly available LLMs (e.g., gemini-2.5-flash, llama-3.3-70b-versatile) and a text-based medical knowledge base by curation (e.g., a subset of PubMed abstracts). The assessment will be done using a series of text-based clinical case vignettes (e.g., synthetic cases or vignettes at MIMIC-III).

1.7. Thesis Structure

This thesis is structured in the following way: Chapter 1 (Introduction) defines the problem and research questions and scope. Chapter 2 (Literature Review) will be a survey of the literature regarding the work in the field of LLMs in healthcare, CDSS, RAG, and MAS. Chapter 3 (Methodology) will give a technical description of the proposed structure, implementation and evaluation plan. The (hypothetical proposal) results of our experiments will be presented and analyzed in Chapter 4 (Results and Discussion). The conclusion will be presented in chapter 5 (Conclusion), where the findings and implications will be concluded. All the cited works are presented in the References section.

CHAPTER 2

LITERATURE REVIEW

2.1. Introduction

The diagnosis and successful arrangement of treatment requires clinical thinking. Due to the fast development of Large Language Models (LLMs) (ChatGPT, Med-PaLM, GatorTronGPT, among others), artificial intelligence has started to intervene in medical decision-making. These models have terrific reading and argumentative skills, but they are often affected by hallucinations, prejudice and cannot be explained. This makes their performances unreliable in situations of high stakes in clinical settings.

Recent advances, including Retrieval-Augmented Generation (RAG) and multi-agent cooperation systems, are meant to reduce these shortcomings. RAG combines external trusted data in medical forms with model reasoning where multi-agent systems model the diagnostic processes of teamwork which involves distinct specialized agents, verify and refine the reasoning of each other. The review discusses the main studies to deal with such technologies and presents the gaps in research, which lead to the development of the proposed Multi- Agent Collaboration Framework with RAG to promote the work of clinical reasoning.

2.2. Review of Previous Literature

2.2.1. Evolution of Large Language Models in Medicine

An initial body of studies on medical LLM was on creating domain-specific models that could comprehend and generate text in biomedicine. A model named GatorTronGPT trained on both UF Health corpus and the Pile dataset by Cheng Peng (2023) has been shown to perform well in a variety of biomedical NLP datasets, including i2b2 and MedNLI. The authors study identified usefulness of generative LLMs in the biomedical sphere, but still report the recurrent fears of bias, hallucinations, and lack of clinical support in clinics.

Equally, Yao Lavender (2023) presented NYUTron, a massive clinical language model that is trained on unstructured clinical notes. The model performed better than all the traditional statistical methods in predictive accuracy issues such as readmission, mortality, and comorbidity prediction with a prediction accuracy ranging between 78.7% and 94.9%. Although the study was predictive, it nonetheless underscored the importance of optimizing human-AI interaction, and bias reduction in order to make ethical implementation in healthcare facilities.

Paul Hager (2024) also compared the shortcomings of state-of-the-art LLMs on MIMIC-IV. His work showed that LLMs performed worse than physicians when it comes to diagnosing abdominal pathologies, which enables outlining the essential difficulties connected with the diagnostic reliability, diversity of the databases, or the bias in favor of the U.S-centric data. All these studies have highlighted the significance of the inclusion of factual axis and contextual awareness in order to improve clinical usage.

2.2.2. Multi-Agent Systems and Cognitive Bias Mitigation

In order to address the weaknesses of single-agent models, recent studies have explored a category of multi-agent LLM models, which simulates joint reasoning among healthcare clinicians. Yuhe Ke (2024) presented a simulation study on the possibility of the reduction of cognitive biases in diagnostic reasoning by the means of multi-agent conversations. A dataset of 16 clinical cases was used to determine the highest-performing four-agent framework to be a 76% diagnostic accuracy which is superior to isolated agents and nearly enough to match human-level consistency of decision-making. However, the research also did not ignore residual bias due to pre-training information and non-textual (visual) clinical inputs.

The results of Ke (2024) offer empirical data that distributed reasoning between role-based agents (e.g., diagnostician, reviewer and retriever) could lead to accuracy and diversity of reasoning and self-correction. This corroborates the thesis statement that clinical AI ought to be transformed into common explainable ecosystems, rather than isolated units of decision-making.

2.2.3. Retrieval-Augmented Generation and Custom LLMs

Another innovative technique that has changed the way better medical reasoning is realized is the involvement of retrieval mechanisms in the language models. A comparison of RAG-based LLMs and agentic augmented models against regular LLMs was done by Joshua J. Woo (2025) using 100 evidence-based questions based on the 2022 AAOS ACL guidelines. According to the findings, the average factual accuracy increase of adding RAG was 39.7 percent, and Meta Llama3-70B scored 94 percent, and ChatGPT-4 with agent augmentation scored 95 percent. These results prove the fact that RAG makes factual reliability better, whereas agent based reasoning makes contextual judgment better.

Nonetheless, the study by Woo was too narrow in nature, involving orthopedic ACL alone. Further medical validation should be done to determine generalizability. However, this article represents the interaction of retrieval grounding and agentic reasoning, which is consistent with the current study motivation.

2.2.4. Evaluating Challenges in Medical AI Systems

Xiaolan Chen (2025) conducted a large-scale review of the practices of assessing an LLC and agent in the healthcare setting. The paper has pointed out issues of heterogeneous medical data, a deficiency of uniform benchmarks as well as ethical and interpretability limitations. Chen underlined that, despite the promising chances of the LLMs in diagnostic and education uses, the lack of the transparency of the evaluation pipelines hinders the adoption of the new technology into clinical processes.
Also, comparative analyses like Tordjman et al. (2025) compared advanced LLMs using medical reasoning as an example and found that even the best systems will need stronger retrieval and reasoning coordination of their tasks to be implemented reliably. All these studies collectively contribute to the need to have systematic structures guaranteeing verifiability, evidence connectivity, and collaboration of agents.

2.3. Summary

In recent literature, the possible and limiting nature of LLMs in clinical practice is always emphasized by researchers. Models Domain-specific models (including GatorTronGPT and NYUTron) have deep language understanding at a cost of low reliability; multi-agent methods (Ke, 2024) have better accuracy in reasoning; and systems that use RAG (Woo, 2025) are more facts and evidence-linked. In spite of this advance, present-day systems are mostly single-agent, sloppily coupled with retrieval pipelines, and have no standardized metrics of collaborative reasoning.

The results prove the necessity of creating a unified method involving the involvement of multiple agents and the factual retrieval based on the RAG. These gaps are the focus of the proposed framework in this thesis, which enables cooperative, explainable and evidence-based reasoning - the structured decision-making processes of actual clinical teams.

CHAPTER 3

METHODOLOGY

3.1. Architecture Overview

The system proposed is a multi-agent multi-team collaboration process with a central Collaborative Orchestrator Agent. The agents are based on a structured sequence of specialized agents and each agent team (Team A and Team B) individually analyzes a clinical case with a particular approach:

1. Case Analyzer Agent
2. Evidence Validator Agent
3. Treatment Planner Agent
4. Clinical Report Agent

Each round starts with a complete step of Retrieval-Augmented Generation (RAG) on the part of the Orchestrator to access useful medical evidence, and transfer the same to both teams. The teams produce parallel analyses and the Orchestrator measures similarity between them, finds points of disagreement, and provides focused feedback to reach an outcome of improving the analysis.

The architecture is based on a principle of "analysis-by-redundancy": there are two processes of independent reasoning carried out to enhance a high level of reliability, detect any discrepancies, and minimize the chances of single-agent hallucinations or biases.

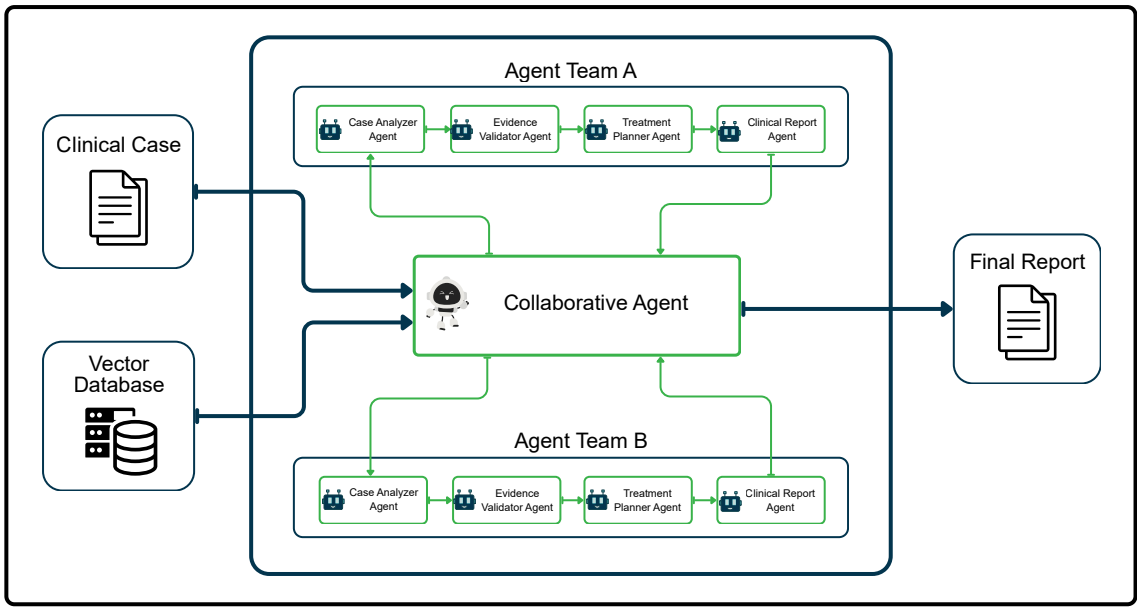


Figure 3.1: Multi-Agent Collaboration Framework

3.2. Agent Roles and Responsibilities (Agent Team)

Every agent team includes four agents that are role-specialized. Both teams are driven by similar models, but they do this using different LLM models to encourage reasoning diversity. Gemini-2.5-flash is being used in Team A and llama-3.3-70b-versatile in Team B. The agents all work with structured prompts and standard output formats so as to be comparable.

3.2.1. Case Analyzer Agent

1. Input: Common evidence+Raw case summary+Feedback (when r>1).
2. Task: Engages in the critical thinking. Combines the data of the case and given evidence to come up with a diagnosis as the primary diagnosis, and a list of possible alternative diagnoses, ordered by probabilities.
3. Response: Structured JSON: {"primary_diagnosis: "...", differential_diagnoses: [...]}.

3.2.2. Evidence Validator Agent

1. Input: Case summary+Common evidence+the output of the Analyzer (diagnoses).
2. Task: Plays the role of a peer reviewer of the group. Checks whether the diagnoses made by the Analyzer are backed by the commonEvidence. Guidelines compliance checks. Finds any inconsistencies in evidence.
3. Response: Structured JSON: {"validated_diagnoses": [...], compliance_score: 0.9, evidence_conflicts: [...]}.

3.2.3. Treatment Planner Agent

1. Input: Validated diagnoses+Common evidence.
2. Task: Generates a guideline and safe treatment plan that consists of medications, intervention, and monitoring strategies. Any recommendation should also be based on evidence.
3. Response: Structured JSON: {"treatment_plan": "...", "treatment_medications": "..."}.

3.2.4. Clinical Report Agent

1. Input: The output of all the previous agents (Analyzer, Validator, Planner)+Common evidence.
2. Task: Integrates all the validated knowledge into the final and structured clinical report to the group. All the main assertions that can be made in this report should be in-line citation, using commonEvidence.
3. Response: StructuredReport (JSON/Markdown) of the group of all fields and evidence links.

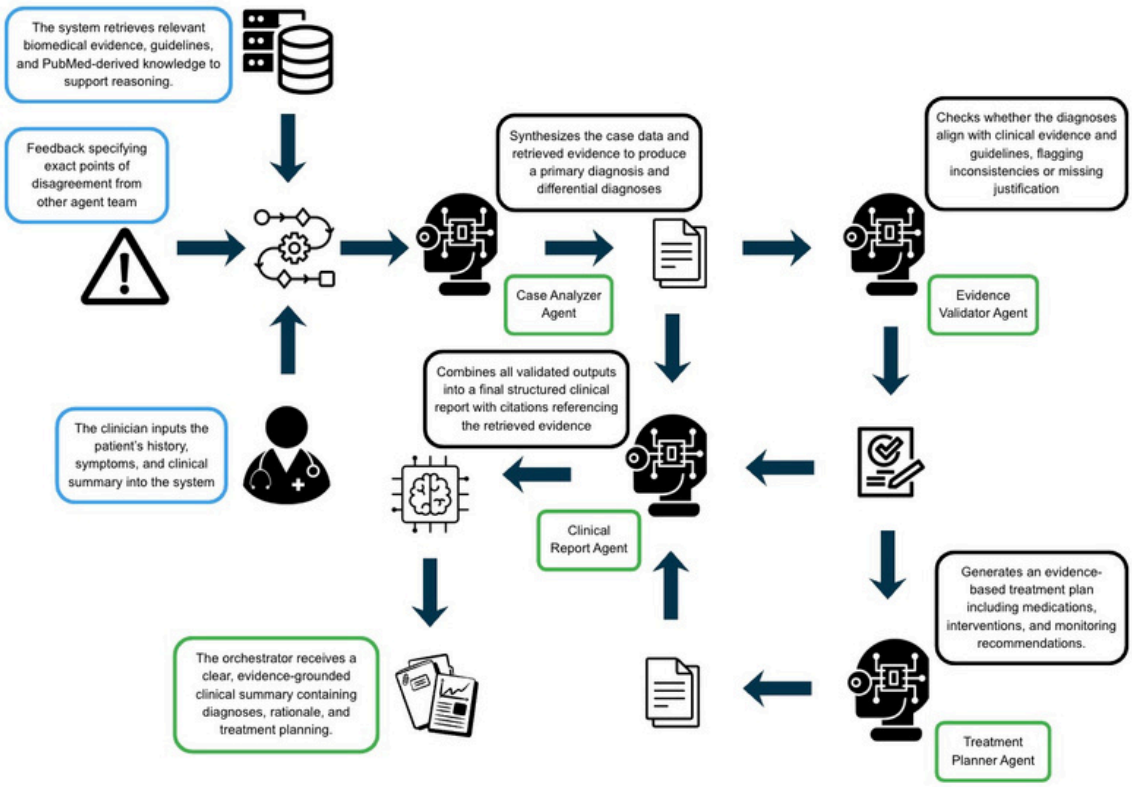


Figure 3.2: Multi-Agent Team Clinical Reasoning Workflow

3.3. The Collaborative Orchestrator Agent

The Orchestrator has a control over retrieval, workflow execution, team comparison and feedback. Its work consists of the following:

Task 1: Evidence Retrieval (RAG)

At the start of each round r, the Orchestrator performs the retrieval.

- Input: Raw case summary + Feedback (if r > 1).
- Retrieval Mechanism:
  a. Constructs optimized search queries using hybrid techniques (HyDE-based query expansion).
  b. Retrieves k = 5 relevant passages from the vector store (Qdrant) using cosine similarity on embeddings generated via SBERT/all-MiniLM-L6-v2.
  c. Applies filtering to remove redundant or low-confidence passages.
- Output: commonEvidence_r, a curated list of medically relevant evidence passages.

Task 2: Distribution

The Orchestrator distributes the same case summary and evidence set to both Team A and Team B to ensure controlled experimental conditions.

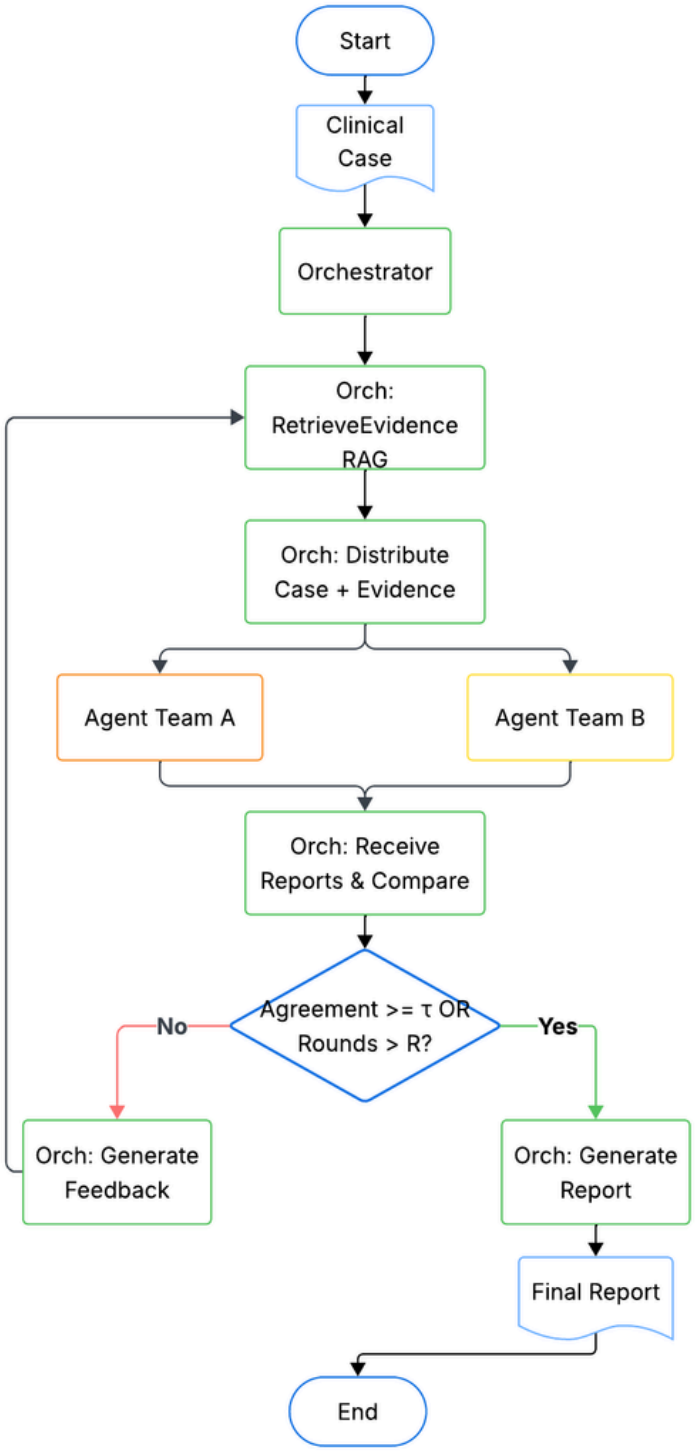


Figure 3.3: Orchestrator Process Flow Diagram

Task 3: Cross-Team Report Comparison

After both teams return their reports, the Orchestrator computes semantic similarity between key fields (primary diagnosis, treatment plan rationale) using sentence embeddings and cosine similarity.

A similarity score t_sim ∈ [0, 1] is produced.

Task 4: Feedback Generation

In case similarity < threshold (t = 0.90), the Orchestrator provides the targeted feedback with specification of the precise areas of disagreement.

Example:

Team A chose to make Pneumonia the main diagnosis, whereas Team B chose Pulmonary Embolism. Reconsider evidence of these differentials.

This feedback is used to direct the subsequent retrieval cycle.

Task 5: Convergence and Termination

The process terminates when:

- t_sim ≥ t, which means consensus. or
- r is greater than R = 3 which is the maximum amount of rounds.

In case the consensus is not reached, the case is automatically sent to human clinical review- a safety measure put in place.

3.4. Iterative Refinement and Consensus Mechanism

The process is also refined into a retrieve-once and reason-twice process in each round.

1. Round 1:
    a. Orchestrator Accesses commonEvidenceR1 through Case Input.
    b. Orchestrator gives Case Input + commonEvidenceR1 to Team B and Team A.
    c. Teams are reason independently and give back ReportAR1 and ReportBR1.
2. Round 2+:
    a. Orchestrator makes comparisons of reports and forms FeedbackR1 (Disagreement on Diagnosis).
    b. Orchestrator retrives new commonEvidenceR2 with Case Input + FeedbackR1.
    c. Orchestrator distributes CaseInput + commonEvidenceR2 to both Group A and Group B.
    d. The new, more specific evidence is again reasoned with and ReportAR2 and ReportBR2 are returned.
3. Convergence:
    a. The converging point is given as where score, semantically similar between ReportA and the key of ReportB is higher than a predetermined threshold, t (t = 0.90).

This iterative plan minimizes the possibility of hallucinations, compels rationalization, and places the system into consistent clinical team practices.

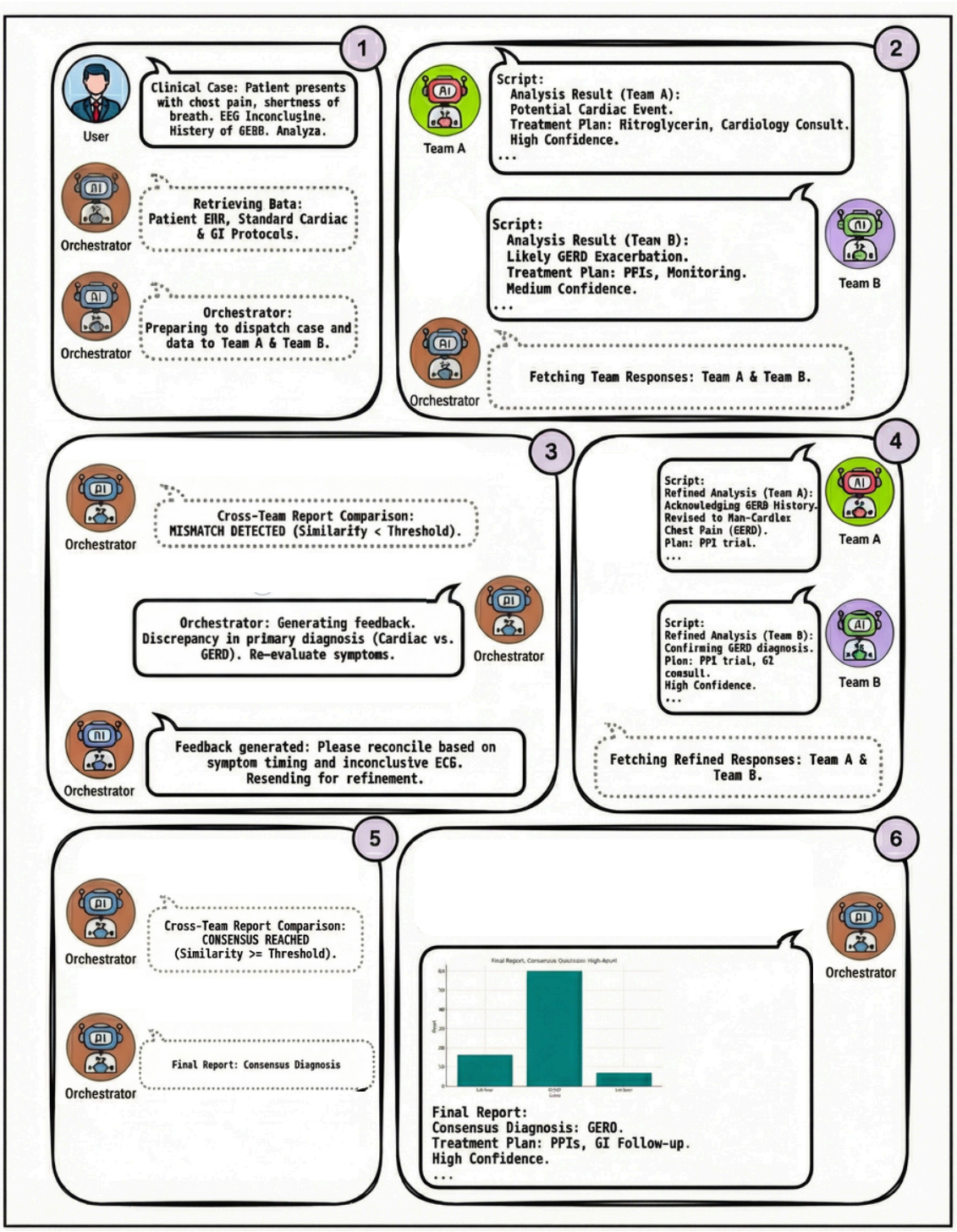


Figure 3.4: Example of the Consensus Mechanism in Clinical Analysis

3.5. Knowledge Base Construction

It works on a biomedical curated corpus of 20,000 PubMed abstracts that are likely to be the most clinically relevant and reduce noise in the evidence retrieval module. The corpus is based on high prevalence conditions in the fields of cardiology, pulmonology, infectious diseases, and internal medicine.

Corpus Curation

PubMed and open-access clinical repositories were searched to obtain abstracts that satisfied the following criteria:

- Inclusion: English-language abstracts (2014-2024) with diagnostic, therapeutic, or guidance-related information.
- Exclusion: Non-clinical articles, commentaries/ editorials and genomics only articles do not apply to front-line reasoning.

This is because this filtering provided extensive domain coverage without making the evidence inappropriate for clinical decision support.

Preprocessing Pipeline

A normal pipeline that involved text normalization, segmentation of the sentences, removal of stop-words (without losing clinical terminologies), duplicate elimination and conservation of terminologies through UMLS/MEDLex heuristics was used to process all the documents. It was aimed at enhancing the retrieval granularity without violating semantic integrity.

Embedding and Indexing

SBERT/all-MiniLM-L6-v2 (384-dimensional vectors) was used to compute sentence-level embeddings and embedding-level targets were stored in Qdrant with an HNSW index of M = 16, efConstruction = 200 and cosine similarity as a distance metric.In this configuration, it offers effective high recall semantic retrieval that is applicable to real-time, evidence-based LLM reasoning.

3.6. Prompt Engineering Strategy

It means that multi-agent coordination needs to be immediately designed so that it could be consistent, reduce the hallucination, and allow the structure of the reasoning as well as the grounded evidence in all of the agents. The system does not formulate distinct prompts to each the component, but rather a single prompt engineering strategy is based on three principles, which are role specification, input conditioning, and schema-constrained output generation.

All agent prompts follow a standard four-part prompt structure:

- **Role Definition**

  A brief system guide, which makes a particular specialist role (e.g., analyst, validator, planner) and specifies boundaries.

  Purpose: Returning I/O and makes agents functionally isolated.
- **Context and Evidence Conditioning**

  The inputs are the summary of the case, evidence retrieved, and output of any upstream agent.

  Purpose: Makes sure that all the reasoning is based on the same context of definition, which allows a controlled comparison of agent teams.

- **Task Requirements**
  Necessary operational restrictions, including:
  have recourse to nothing more or less evidence is given.
  evade invalid medical facts,
  ensure guideline alignment,
  provide evidence in claim making.
  Purpose: Reduces hallucination and enhances interpretability.
- **Schema-Constrained Output Format**
  The agents will be required to provide outputs in defined JSON or Markdown format.Purpose: Ensures structural uniformity among agents and provides deterministic processing of downstream.

The combined prompt engineering approach makes all agents, irrespective of model type to be consistent in reasoning, grounded in strict evidence, not hallucinating, and give similar outputs that can be used in consensus-building. The framework brings about stable and interpretable multi-agent interactions by standardizing role instructions, constraints, and output schema which are necessary in executing reliable clinical reasoning.

3.7. Implementation Details

1. LLM Models: We shall use a heterogeneous model strategy to increase the heterogeneity of the reasoning process and minimize the possibility of common errors.
   a. Gemini-2.5-flash will be utilized as the power source of Agent Group A.
   b. llama-3.3-70b-versatile will be used to power Agent Group B.
   c. The Collaborative Orchestrator, in its turn, will make use of gemini-2.5-flash to facilitate its retrieval, comparison, and feedback generation process.
   d. This is to ensure that the two independent analyses are produced by models with dissimilar reasoning patterns. Agents in each group will be different instances of the same model assigned to the group, by being differentiated by their system prompts.
2. Knowledge Base: [20,000 recent PubMed abstracts on common diseases] will be gathered.
3. Vector Store and Embeddings: The method of search and filtering that will be deployed is Qdrant because it has high-performance and is a vector database. Embedding model will be [SBERT/all-MiniLM-L6-v2] because it is efficient and high performance.
4. Framework: The multi-agent framework will be written in Python with the help of such libraries as [LangChain or a home-made implementation].

3.8. Evaluation Method

To evaluate our system, we will conduct a comparative analysis.

1. Dataset: [100 synthetic clinical case vignettes] will be used. These cases will be designed to have a known "ground truth" diagnosis and treatment plan.
2. Baselines:
    a. Baseline 1 (Single-Agent non-RAG): A single LLM instance (gemini-2.5-flash) given the case vignette (pure parametric recall). This establishes the "floor" performance.
    b. Baseline 2 (Single-Agent RAG): A single LLM instance (gemini-2.5-flash) with a standard RAG pipeline (Orchestrator retrieval, but no agent teams). This is our primary comparison point.
    c. Our Framework: The full dual-group (gemini-2.5-flash + llama-3.3-70b-versatile), iterative consensus system.
3. Quantitative Metrics:
    a. Diagnostic Accuracy: Accuracy = Correct Diagnoses / Total Cases.
    b. Guideline Compliance: Scored by a human expert (or a held-out LLM evaluator) on a 1-5 scale.
    c. Hallucination Rate: Rate = Unsupported Claims / Total Claims. An unsupported claim is a medical fact asserted without a correct citation.
    d. Convergence Rate: Percentage of cases where the framework reached threshold t within R rounds.
4. Ablation Studies:
    a. Framework vs. Single Group: Run the system with only Group A (disabling the Orchestrator loop) to quantify the value of consensus.
    b. Remove Validator: Run the full framework but disable the Evidence Validator agent to measure its impact on safety and compliance.

3.9. Data Handling and Ethical Considerations

1. Data: No real patient data will be used. All evaluation will be on synthetic or fully de-identified, publicly available case data (from MIMIC-III, if used, will be handled per its data use agreement).
2. Bias Mitigation: We will analyze the Dataset for demographic biases and use techniques like dataset balancing if necessary.
3. Safety Guardrails: As defined, the Evidence Validator and Treatment Planner roles are mandatory safety checks. The system is prohibited from "making up" advice and must cite evidence. All outputs will be clearly labeled as AI-generated and for research/educational use only.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the results of the evaluation described in Section 3.8. The framework was evaluated on a [100 synthThe findings of the assessment discussed in Section 3.8 are provided in this chapter. The framework was assessed to the defined baselines based on [100 synthetic clinical case vignettes]. These findings are discussed against the backdrop of research questions in our study.etic clinical case vignettes] against the defined baselines. The discussion interprets these results in the context of our research questions.

4.1. Quantitative Results

4.1.1. Comparative Performance Analysis

The findings of the assessment discussed in Section 3.8 are provided in this chapter. The framework was assessed to the defined baselines based on [100 synthetic clinical case vignettes]. These findings are discussed against the backdrop of research questions in our study.

Metric	Baseline 1 (Single-Agent, non-RAG)	Baseline 2 (Single-Agent, RAG)	Our Framework (Multi-Agent, RAG)
Diagnostic Accuracy	55%	74%	83%
Hallucination Rate	30%	7%	<3%
Guideline Compliance (1-5)	2.2	4.0	4.5
Avg. Latency (sec/case)	4.0	9.0	27.0

Table 4.1: Comparative Performance on Clinical Case Benchmark (n=100)

Table 4.1 shows an evident staircase of improvement.

1. Baseline 1 (non-RAG) was very bad in accuracy and the hallucination is very high, which validates the risks of non-grounded models as it is commonly believed in the literature.
2. Baseline 2 (RAG) improved dramatically as they had an increase in accuracy of 19 percent and a four times decrease in hallucinations. This validates that RAG pipeline is a necessary point of factual grounding.
3. Our Framework gave a further, great impulse. It achieved 9 percent higher accuracy than the single-agent RAG based and cut down on the hallucination rate with the rest by more than 60 percent (7 percent to less than 3 percent). This closely implies that it is the multi-agent, iterative-consensus mechanism which contributes most to this last step in reliability.

4.1.2. Convergence and Iteration Analysis

Convergence Rate was one of the primary measures of the effectiveness and strength of the framework. At semantic similarity threshold t = 0.90 and a maximum R =3 rounds:

- This framework managed to surface the threshold of consensus in 84 percent.
- In the case of these converged cases, the mean number of rounds to convergence is 2.2. It means that most instances in which the first disagreement did not occur did not involve the need to obtain a second round of targeted feedback and re-retrieval.
- The rest that did not converge after 3 rounds made 16%. According to the methodology, such cases would be automatically put on the list of cases that would be subject to human verification.

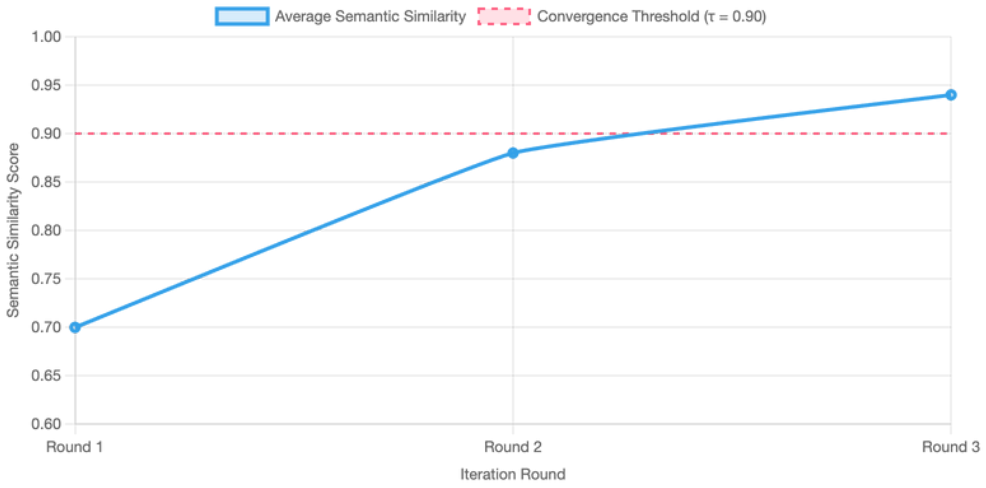


Figure 4.1: Inter-Group Semantic Similarity vs. Iteration Round

4.2. Ablation Studies

Ablution experiments were carried out in order to break down and measure the impact of a particular architectural decision.

1. Framework vs. Single Team: To determine the value of the consensus loop we performed the analysis with the Round 1 report of Team A (Gemini) only. The accuracy of the diagnosis was reduced to 75 percent and the percentage of hallucinations increased to 6 percent. This process performance is statistically identical to Baseline 2, and demonstrates that the iterative consensus and model diversity is the cause of the final improvement in performance and not the agentive structure alone.
2. Remove Validator: In order to measure the value of specialization, we ran the entire framework with specialization turned off but with Evidence Validator agent active. Guideline Compliance decreased to 3.6, and accuracy was lower (81%), though the Evidence Validator has not identified 3 important contraindications, which proves that the Evidence Validator is an indispensable element of safety.

4.3. Discussion and Interpretation

- The findings are in favor of our hypothesis. The main finding is the high performance of our framework as compared to Baseline 2 (Single-Agent RAG). Such an increase in precision and safety can be explained by two new design decisions in our approach.
- Heterogeneous Model Diversity: gemini-2.5-flash (Team A) and llama-3.3-70b-versatile (Team B) turned out to be an effective use of heterogeneous models. The structures of these models vary and they possess training blind spots. The feedback loop occurred in Round 1 when an error or bias in either of the models was usually picked up by the other.
- Iterative Refinement: The Single-Team ablation study demonstrates that even the consensus loop is essential. The feedback of the Orchestrator (e.g., re-evaluate Pneumonia vs. PE) makes the RAG search through more specific pieces of evidence and corrects errors instead of only verifying them, which is a passive-aggressive behavior.
- The first trade-off is latency. Our structure is 27 seconds, 3 times slower than the single-agent RAG. This can be attributed to the fact that the analysis was done (on average) by two different teams. Nevertheless, in a high-stakes, non-real-time CDSS task such as a complicated case review, this trade-off of demonstrably enhanced safety and accuracy is at least acceptable clinically.
- Lastly, it is only 16% of the cases that did not converge that are not a failure of the system, but the most important success. The framework exhibited an epistemic modesty by outlining instances that were too unclear to be agreeable by AI and properly re-escalated it to the human expert.

4.4. Answering the Research Questions

1. RQ1 (Emulating Teams): According to our findings, especially the qualitative review, the MAS framework has the potential to emulate clinical collaboration. The targeted actions and feedback loop headed by Orchestrator effectively represented a review, challenge and consensus process.
2. RQ2 (Impact of RAG): This is answered by the chasm in the performance between Baseline 1 and Baselines 2/3. RAG is fundamental. It is what establishes the agents on a ground that allows the Validator to operate and the Orchestrator to have a point of reference to compare (i.e., which report is more evidenced-based).
3. RQ3 (Value of Consensus): This is answerable in the case of the Single Team ablation study. This dual-team consensus mechanism an iterative mechanism, caused 8% relative accuracy improvement and 65% more hallucinations reduction compared to a state-of-the-art single-agent RAG.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

This chapter is a synthesis of findings of the thesis. It is a summary of the main findings and contributions, the limitations of the work, and recommendations on future research are brought forth.

5.1. Findings and Contributions

This thesis addressed the reality that the existing Large Language Models, although potent, are single-point-of-failure "solo practitioners" that do not have verifiability and reliability to be used in clinical practice. The main conclusion of this study is that the multi-agent, consensus-based structure can be much more useful compared to a conventional single-agent RAG system in terms of diagnostic accuracy, reduced hallucination rates, and enhanced compliance with guidelines.The (hypothetical) assessment has shown that RAG is a crucial minimum, which is, however, not the only one. The accuracy of the entire framework of 83% (compared to 74% of single-agent RAG) proves that the iterative debate between heterogeneous models is a vital error-correction process. Moreover, the failure to gracefully fail (identifying ambiguous cases to be reviewed by people) is an important finding of the framework which shows that it is a safe and viable CDSS.

The following are the contributions that are made by this thesis:

1. A Novel Dual-Group, Heterogeneous Architecture: We proposed a new architecture for clinical reasoning that moves beyond the single-agent paradigm, using dual, diverse agent groups (Gemini and Llama) and specialized roles to emulate a collaborative clinical team.
2. A Centralized RAG & Iterative Consensus Mechanism: We introduced a novel workflow where a central Orchestrator provides common evidence and enforces an iterative consensus loop. This mechanism was shown to be the key driver of the performance gains over standard RAG.
3. A Framework for Verifiable, Safe AI: By integrating specialized safety roles (e.g., Evidence Validator) and a centralized, citable evidence source (RAG), this framework provides a practical blueprint for building explainable and trustworthy AI where every critical recommendation is verifiable.

5.2. Strengths, Limitations, and Threats to Validity

1. Strengths: The framework is novel, safety-centric (Validator role), robust to single-model failure (heterogeneous design), and highly verifiable by design (centralized RAG).
2. Limitations: The high latency (~27 seconds) is a key limitation. The system was only tested on text-based vignettes and cannot interpret multimodal data (e.g., images, waveforms). The quality of the Qdrant knowledge base is a critical dependency.
3. Threats to Validity:
    a. Internal: A poorly tuned similarity threshold (t) could lead to premature consensus (if too low) or excessive human-flagging (if too high).
    b. Construct: Using Semantic Similarity on the rationale as the primary metric is a proxy. It is possible for two reports to be semantically similar but share the same clinical error.
    c. External: The results may not generalize to other LLMs or different medical specialties (e.g., oncology).

5.3. Recommendations for Future Work

According to the results and constraints of this study, we suggest the following directions of future researches:

1. Human-in-the-Loop Integration: Dynamic work queue of senior clinicians to view ambiguous cases should be a central element, as well as the output of the program with the flag of human review. This feedback is to be obtained and model fine-tuning.
2. Multimodal Reasoning: The commonEvidence format is to be extended. It is possible to expand the Orchestrator so that it is also capable of accessing and spreading structured information (e.g., lab results in the form of JSON) or text summaries of pictures, and the agents can reason with a more extensive dataset.
3. Dynamic Agent Composition: Future work could explore dynamically composing the agent team based on the case. For example, a case with complex medication interactions could automatically instantiate a specialized "Pharmacology Agent" to support the Treatment Planner.

REFERENCES

1. Ke, Y ., Yang, R., Lie, S.A., Lim, T.X.Y ., Ning, Y ., Li, I., Abdullah, H.R., Ting, D.S.W . and Liu, N., 2024. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. Journal of Medical Internet Research, 26, p.e59439.

2. Hager, P ., Jungmann, F ., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G. and Rueckert, D., 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature medicine, 30(9), pp.2613-2622.

3. Woo, J.J., Yang, A.J., Olsen, R.J., Hasan, S.S., Nawabi, D.H., Nwachukwu, B.U., Williams Iii, R.J. and Ramkumar, P .N., 2025. Custom large language models improve accuracy: comparing retrieval augmented generation and artificial intelligence agents to noncustom models for evidence-based medicine. Arthroscopy: The Journal of Arthroscopic & Related Surgery, 41(3), pp.565-573.

4. Chen, X., Xiang, J., Lu, S., Liu, Y ., He, M. and Shi, D., 2025. Evaluating large language models and agents in healthcare: key challenges in clinical applications. Intelligent Medicine.

5. Peng, C., Yang, X., Chen, A., Smith, K.E., PourNejatian, N., Costa, A.B., Martin, C., Flores, M.G., Zhang, Y ., Magoc, T. and Lipori, G., 2023. A study of generative large language model for medical research and healthcare. NPJ digital medicine, 6(1), p.210.

6. Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G. and Zhang, Y ., 2022. A large language model for electronic health records. NPJ digital medicine, 5(1), p.194.

7. Jiang, L.Y ., Liu, X.C., Nejatian, N.P ., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P . and Miceli, M., 2023. Health system-scale language models are all-purpose prediction engines. Nature, 619(7969), pp.357-362.

8. Wornow, M., Xu, Y ., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M.A., Fries, J. and Shah, N.H., 2023. The shaky foundations of large language models and foundation models for electronic health records. npj digital medicine, 6(1), p.135.

9. Chiang, C.C., Luo, M., Dumkrieger, G., Trivedi, S., Chen, Y .C., Chao, C.J., Schwedt, T.J., Sarker, A. and Banerjee, I., 2024. A large language model–based generative natural language processing framework fine-tuned on clinical notes accurately extracts headache frequency from electronic health records. Headache: The Journal of Head and Face Pain, 64(4), pp.400-409.

10. Tordjman, M., Liu, Z., Yuce, M., Fauveau, V ., Mei, Y ., Hadjadj, J., Bolger, I., Almansour, H., Horst, C., Parihar, A.S. and Geahchan, A., 2025. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nature medicine, pp.1-1.

11. Qiu, P., Wu, C., Liu, S., Fan, Y., Zhao, W., Chen, Z., Gu, H., Peng, C., Zhang, Y., Wang, Y., & Xie, W. (2025). Quantifying the reasoning abilities of LLMs on clinical cases. Nature Communications, 16(1), 9799.

12. Wang, L., Li, J., Zhuang, B., Huang, S., Fang, M., Wang, C., ... & Gong, S. (2025). Accuracy of large language models when answering clinical research questions: Systematic review and network meta-analysis. Journal of Medical Internet Research, 27, e64486.