National University of Computer and Emerging Sciences, Lahore

DoctorHive

Abdul Moiz Shehzad 22L 7468 BS(DS)

Muhammad Fahad Hussain 22L 7463 BS(DS)

Muhammad Reyyan Saeed 22I 1964 BS(DS)

Supervisor: Ms Mariam Nasim

Final Year Project

November 24, 2025

Anti Plagiarism Declaration

This is to declare that the above publication was produced under the:

Title: DoctorHive

is the sole contribution of the author(s), and no part hereof has been reproduced as it is the basis (cut and paste) that can be considered Plagiarism. All referenced parts have been used to argue the idea and cited properly. I/We will be responsible and liable for any consequence if a violation of this declaration is determined.

Date: 2025-10-17

Name: Abdul Moiz Shehzad

Signature:

Name: Muhammad Reyyan Saeed

Signature:

Name: Muhammad Fahad Hussain

Signature:

Author's Declaration

This states Authors' declaration that the work presented in the report is their own, and has not been submitted/presented previously to any other institution or organization.

Abstract

Virtual assistants and chatbots have become the most common tools of providing accessible health information, especially through the use of the Artificial Intelligence (AI) system. Such systems can give fast answers to patient questions and initial diagnostic recommendations, yet their accuracy is a major issue. Single agent chatbots tend to make false or panic inducing conclusions, which leads to the emergence of unnecessary anxiety and destroys the trust of patients. Patients are not technical users and cannot consider the output of the AI in different fields like programming, unlike them, who are able to assess the information provided by the AI as safe or potentially harmful. This is where the void shows the necessity of more trustworthy, interpretable, and patient oriented AIs.

DoctorHive resolves this issue and presents a system of multi agent debate which simulates a team of medical professionals working together. Instead of having a single general purpose agent to do all of the work, DoctorHive has multiple specialized agents, each with a medical area of specialization (general practice, cardiology, neurology, ophthalmology). The agents are allowed to independently review the information on the patient and later engage in a structured discussion which enables them to compare their knowledge to arrive at a common diagnosis. This process allows the agents to go up against each other, find out inconsistencies and sharpen their arguments. The system then generates a consensus output which points out the most probable diagnosis and gives the recommended action that should be taken. This multi layer mechanism reflects actual clinical cooperation in the real world, decreasing the chances of misdiagnosis and rationalizing the reasoning process and making it patient understandable.

DoctorHive is important as it tries to find the perfect blend of technical accuracy and interpretability and trust. The system is based on consensus building and explainability, thus accessible to modern medical AI demands of going beyond accuracy and including fairness, robustness, and usability. In the end, DoctorHive will have an objective of enhancing the trustworthiness of patient facing AI, reduce the possibility of misinformation and offer a model of secure and understandable incorporation of AI in healthcare interaction.

Executive Summary

AI is quickly transforming the healthcare delivery, including diagnostic imaging and chatbots that ask about symptoms. Although these tools are fast and accessible, the limitations are becoming more apparent. The existing chatbot systems tend to produce false, exaggerated or context blind results. To patients, this poses a hazardous cycle: when patients go on a hunt to find out what common symptoms are, the results can be alarming (cancer, etc.), and misleading, as well as anxiety inducing. Patients are not well placed to assess the validity of AI generated guidance as opposed to professionals in the technical sector. This renders trust, interpretability, and reliability some of the key issues of AI in healthcare.

DoctorHive is an idea that has been created to address these problems. It is a multiagent debate system that simulates the interaction of clinical teams in practice. DoctorHive uses multiple specialized agents, each of which is a medical field, such as general practice, cardiology, neurology, and ophthalmology unlike a single AI agent. These agents work autonomously and come up with their initial evaluation of patient inputs. They then proceed to systematic debate, finding where they are wrong, investigating assumptions and pointing out the unknowns. In this process, a consensus is made, which yields a refined output, which is the most plausible diagnosis and actionable next steps to the patient.

This debate and consensus mechanism is the most important innovation. The system allows disagreement to take place prior to agreement and this minimizes the risk of uncontrolled errors and the outputs are based on varied views. Meanwhile, DoctorHive is concerned with explainability. Patients are not only provided with the end result of the consensus but also with simplified explanations of the reasoning behind the final decision, and thus, it becomes interpretable instead of opaque. This is a direct response to one of the primary arguments against AI in medicine: that despite technical correctness, systems do not always provide communication in a way that builds patient trust.

The project is based on the existing work on multiagent systems, consensus driven artificial intelligence, and explainable artificial intelligence. The literature also brings out the potential and dangers of AI in healthcare, that it is not enough to be accurate. DoctorHive translates these insights into a prototype that is patient facing and is built on reliability, interpretability and safety.

The scope of DoctorHive has been carefully set to be within the scope of the initial phase. The system does not include such sensitive areas as integration of electronic health record and generation of prescriptions but rather concentrates on diagnostic reasoning and guidance. This enables the team to develop and assess a working prototype that illustrates the main value of multi agent cooperation.

The expected effect is twofold. To patients, DoctorHive provides more trustworthy and less uncomfortable AI contacts. In the case of the healthcare industry, it is a demonstration that consensus based

multispecialist systems can be more accurate and trusted than regular chatbots.

In short, DoctorHive is a step towards a safe and interpretable AI in healthcare. Through its technical innovation and patient centered design, it will help minimize misinformation, develop trust, and create a responsible implementation of conversational medical AI.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Artificial Intelligence (AI) is playing an increasingly important role in healthcare, offering patients quick access to information and preliminary diagnostic guidance. Despite these advantages, current chatbot based solutions remain unreliable. Single agent systems often provide incomplete, exaggerated, or misleading conclusions. This can increase anxiety for patients who lack the technical or medical background to critically evaluate the information they receive. Demonstrable and understandable AI systems have thus been deemed urgent.

DoctorHive fills this void by providing a multi agent debate system of healthcare question queries. DoctorHive has multiple specialist agents in its deployment, which represent the main areas including general practice, cardiology, neurology and ophthalmology. The agents individually interpret the input of the patient and then they debate in an organized manner, in which he/she questions the reasoning, finds loopholes and cleans up their findings. This system then comes up with a consensus recommendation which is provided together with a simplified explanation that can be understood by the patients. This reflects the team decision making approach applied in the actual clinical setting and serves to minimize the chances of misinformation and maximize the patient trust.

This work has its basis in previous studies in the area of multi agent systems, debate mechanisms and explainable artificial intelligence (XAI). The literature indicates the opportunities of the collaboration of agents to achieve more effective reasoning and the importance of the explainability to enhance the trust and usability. DoctorHive is built upon these concepts, but transforms them into a safety first and interpretable patient on top of healthcare application.

1.1 Purpose of this Document

This report aims at recording the design, development, and methodology of DoctorHive as a Final Year Project. It describes the motivation factor, the background of the context, technical specifications that were needed, and the design choices made in the system. section Purpose of this Document

This Document serves as a guide to stakeholders of Emirati Airways Company to implement optimal strategies that enhance performance across all its business units.¡—human—¿purpose of this Document This Document is designed to help the stakeholders of Emirati Airways Company to adopt the best strategies to improve the performance of all its business units. This report is being submitted to provide the design, implementation, and evaluation of DoctorHive, a multi agent debate system of patient facing healthcare guidance. The ultimate aim of the project is to research the possibility of a debate and consensus architecture enhancing the precision and reliability of AI medical advice in comparison to

single agent systems.

The research question which will be answered in the report is: Can a multi agent debate framework decrease the risks of misinformation and increase patient trust to AI driven healthcare guidance?

The project establishes the following objectives in order to do so:

- Implement the autonomous agents of various healthcare fields (e.g., cardiology, neurology, ophthalmology) each with the ability to produce domain specific evaluation.

- Introduce formal interaction schemes in which agents are able to review, confront, and refine one another based upon external knowledge access (RAG).

- The creation of a voting and aggregation system that is accurate, has confidence scores, and makes the output understandable and actionable to patients.

- Evaluate the framework on curated clinical case scenarios in terms of diagnostic accuracy, quality of explanations, and patient trust in the recommendations of the system.

The research design involves literature study, the prototyping of systems, and testing using test cases and comparing them with the previous models of the chatbots. Disadvantages also contain the fact that it leaves out sensitive sections like prescription recommendation and connection with electronic health records which makes sure that the project is not derailed and keeps within the level of Final Year Project.

1.2   Intended Audience

The targeted audience includes:

- Researchers in computer science, particularly those focused on multi agent systems, retrieval augmented generation, and explainable AI.

- Healthcare AI practitioners and clinicians interested in decision support systems that combine automated reasoning with interpretability.

- Patients, who represent end users of the system, benefiting from accessible, trustworthy medical guidance.

- Future students who may wish to extend or adapt the design, methodology, and evaluation strategies outlined in this work.

1.3 Definitions, Acronyms, and Abbreviations

AI: Artificial Intelligence

FYP Final Year Project

MAS: Multi Agent System

XAI: Explainable Artificial Intelligence

MVP: Minimum Viable Product

API: Application Programming Interface

REST: Representational State Transfer

UI: User Interface

UX: User Experience

LLM: Large Language Model

SGD: Sustainable Development Goal

1.4 Conclusion

The remainder of this report is organized as follows:

- **Chapter 2: Project Vision**

  - Explains the motivation behind DoctorHive.

  - Defines the problem statement.

  - Outlines the high level goals of the system.

- **Chapter 3: Literature Review and Related Applications**

  - Reviews existing research on multi agent systems.

  - Discusses debate models and explainability.

  - Summarizes related applications in healthcare.

- **Chapter 4: Software Requirement Specifications**

  - It defines functional requirements.

  - It defines non functional requirements.

- **Chapter 5: Proposed Approach and Methodology**

  - It describes the debate architecture.

  - It explains the consensus process.

- It outlines the overall development methodology.

- **Chapter 6: High Level and Low Level Design**

  - It presents the overall system architecture.

  - It describes component interactions.

  - It provides detailed design specifications.

Chapter 2 Project Vision

The chapter provides the general vision of the DoctorHive, which is a patient centric healthcare system that uses multi agent debate to enhance the reliability, interpretability, and trustworthiness of artificial intelligence based medical advice. It identifies the problem area, outlines the issues under consideration and establishes the goals and objectives, scope and limitations of the project. The chapter further places DoctorHive as part of the sustainable development, business opportunities, and needs of the stakeholders.

2.1 Problem Domain Overview

This document outlines the problem domain and is intended to offer a summary of the problem, along with a definition of the problem area and project scope. The scope of the problem area and project scope definition is also suggested in this document as it will provide an overview of the problem domain and is supposed to give a summary of the problem. Due to the rapid expansion of AI in medicine, hundreds of chatbots and virtual assistants that assist a patient were created. Such systems are able to give instant access to medical data and first line diagnostic advice, although they are still subject to error and simplification. Since they are constructed based on single agent reasoning, it does not provide a mechanism of cross checking or cooperative validation of outputs. The advice that patients get is tendent to be exaggerated, misleading, or alarmist.

DoctorHive meets this challenge with the introduction of multi agent debate architecture. This system implements multiple domain specific agents like General Practice, Cardiology, Neurology and Ophthalmology to process patient inputs in isolation. These agents consequently participate in organized discourse, as they oppose the logic of each other, point at inconsistencies, and perfect their products. Consent recommendation is then formulated and handed over to the patient in easy, patient understandable language. This is the same collaborative mode of operation in the real world clinical setting and lowers chances of misinformation being provided, besides making the process of making the reasoning understandable.

2.2 Problem Statement

The existing patient facing medical chatbots are not as reliable and interpretable to be used safely. Their results are often incorrect or alarmist in nature and the patients who in most cases are not medical specialists cannot tell whether the information they have is sound. This causes unnecessary anxiety and is prone to making bad health choices. The essential issue is the lack of a patient to AI assistant simulating collaborative medical reasoning and giving clear, reliable and available explanations.

2.3 Problem Elaboration

The given issue can be further discussed in several interconnected sub problems:

- Reliability of Outputs since the single agent chatbots are based on unchecked reasoning, they are likely to produce hallucinations or error.

- Explainability as outputs are often presented without showing how conclusions were reached, leaving patients uncertain about the reasoning.

- Patient Anxiety where alarmist or exaggerated results can heighten stress and discourage proper healthcare seeking behavior.

- Limited Specialization in which one chatbot cannot realistically simulate the diversity of expertise provided by a team of medical professionals.

- Trust Deficit since without mechanisms for interpretability or validation, both patients and healthcare providers remain skeptical of AI systems.

DoctorHive seeks to address these issues by combining debate based reasoning, multi specialist inputs, and dual outputs that present both detailed and simplified explanations.

2.4 Goals and Objectives

Goal: The goal of this project is to design and implement a patient facing AI healthcare system that uses multi agent debate to provide reliable, interpretable, and user friendly diagnostic guidance.

Objectives:

- We are doing development of agents of AI that are specialists in areas such as Practice that is General, Cardiology, Neurology, and Ophthalmology.

- We are putting into practice a mechanism of debate that is structured where agents are able to do challenging and making better of reasoning of one another.

- We are doing design of a process of consensus that is doing balance of accuracy with clarity that is friendly to patients.

- We are giving outputs that are dual: trails of reasoning that are detailed for evaluation that is technical and explanations that are simplified for patients.

- We are doing evaluation of the prototype through situations of cases in medicine that are put together with care, doing measuring of accuracy, quality of explanation, and trust of users.

2.5 Project Scope

The scope of the project of DoctorHive is made clear in a way that is careful to make the system stay focused and able to be achieved in the timeframe of a Project that is Final Year. **In Scope:**

- The doing of development of a framework of debate that is multi agent for queries of healthcare.

- The putting into practice of an interface that is based on web for interaction with patients.

- The outputs that are driven by consensus with explanations for patients that are both technical and simplified.

- The evaluation using cases for test that are doing comparison of performance of single agent and multi agent.

Deliverables:

- A prototype that is working of the system of DoctorHive.

- Documentation that is covering design of system, methodology, and evaluation.

- A manual for users to give guidance to patients on doing interaction with the system.

Out of Scope:

- Integration with records of health that are electronic (EHRs) or databases of hospitals.

- Generation of prescription or planning of treatment.

- Applications for mobile for iOS/Android in this phase.

- Trials that are clinical with patients that are live past testing of prototype.

2.6 Sustainable Development Goal

This project is giving contribution to **SDG 3: Good Health and Well being**, which has the aim to make sure there are lives that are healthy and give promotion to well being for all. Through the making better of the reliability and safety of information of healthcare that is given to patients, DoctorHive is making the dangers of misinformation smaller and is giving power to patients to make decisions that are informed. It is giving support in a way that is direct to the goal of making access to information of healthcare that is quality better no matter what the geography or background that is socio economic is.

Figure 2.1: This figure is showing all 17 Goals of Development that are SDG. SGD 3, which is pointed out in green, is showing the goal that has relevance that is most to this project.

2.7 Constraints

The project of DoctorHive is done with development under the constraints that follow:

- **Time:** The system has to be done with design, putting into practice, and evaluation in the year that is academic, which is making limits to opportunities for testing that is large scale.

- **Resources:** The project is having reliance on models of language that are large and existing instead of training that is custom, which is making constraint to flexibility of system.

- **Scope:** Features that are sensitive such as authority of prescription and integration of EHR are taken out in a way that is deliberate for safety and being feasible.

2.8 Business Opportunity

The market of AI for healthcare that is global has projection to do growth in a way that is substantial over the decade that is next, with chatbots that are already in usage across platforms of telemedicine and services of insurance. But, worries over trust and accuracy are making limits to the adoption of them. DoctorHive offers a clear business opportunity by filling this gap: a system that reduces liability through safer, debate based outputs, while enhancing patient satisfaction through interpretability and clarity. Potential applications include integration with telehealth services, support for health insurers in patient engagement, and as a framework for clinical decision support tools.

2.9 Stakeholders Description/ User Characteristics

The stakeholders of the DoctorHive system include both direct users and indirect beneficiaries:

- Patients are primary users who interact directly with the system to seek preliminary healthcare guidance. They require accurate, understandable, and non alarmist outputs.

- Healthcare Providers are secondary stakeholders.

Chapter 3 Literature Review

This chapter reviews prior work relevant to DoctorHive and is organized into definitions, a detailed review with categorized subsections, a summary table of reviewed works, and a concluding synthesis.

3.1 Detailed Literature Review

3.1.1 The Rise of Artificial Intelligence in Healthcare (Bohr & Memarzadeh, 2020)

Summary

Bohr and Memarzadeh (2020) [1] offer a general description of the application of artificial intelligence in healthcare that includes diagnosing images, monitoring patients, predictive data analysis, clinical decision making, and so on. The authors emphasize the possibilities of AI in changing the healthcare delivery system and improving healthcare through enhancing efficiency and accuracy of diagnosis as well as alleviating the workload of clinicians. Concurrently, the paper highlights the problems of premature adoption, regulatory uncertainty, and dangers of misinformation produced by ill validated systems.

Methodology

The methodology employed by the authors is narrative review which is used to synthesize the available research reports, policy documents and clinical reports. They classify literature based on the domains of application diagnostic imaging, predictive analytics, patient safety systems and derive recurring themes based on the risks and benefits. The selection strategy is based on peer reviewed journals and credible reports on healthcare technology. The methodology is based on thematic grouping instead of quantitative analysis and the findings are aligned with one of the categories, including regulatory readiness, interpretability, and clinical validation. The strategy emphasizes cross sectional evidence in various use cases, which was possible to provide a balanced list of AI effects. The wide range of coverage makes it very inclusive to cover both the positive and the negative views. The methodology focuses on a comprehensive view of the insights, not experimental validation through the structuring of insights along thematic axes (e.g. opportunities vs. risks).

Relationship to DoctorHive

The knowledge gained in this paper can be of great context to the DoctorHive project. It emphasizes the dangers of ready adoption of AI in medicine, which will enhance the argument of DoctorHive that seeks to eliminate misinformation and unreliability by using multi agent inter-

actions and consensus based decision making. The focus on opportunity as well as risk are also correlated with the goal of the project to develop an approach to create an innovative system that would strike the balance between innovation and patient safety, and thus this piece of work is a major reference to explain the rationale of the project.

3.1.2 Deep Learning in Medicine (Esteva et al., 2019)

Summary

Esteva et al. (2019) [2] discuss the use of deep learning in the medical field, specifically, in tasks related to diagnosis, including medical imaging, pathology, and dermatology. The authors claim that deep learning models, which are trained on large scale datasets, can be as good as human experts or even better in terms of diagnostic performance. They also emphasize the possibilities of implementing deep learning into clinical processes and allowing early detection of diseases and more effective management of patients.

Methodology

The research methods of applied deep learning are used by Esteva et al. Their work process consists of gathering extensive labeled datasets of fields including pathology slides and dermatology pictures. Preprocessing of data guarantees consistency such as normalization, augmentation and stratified sampling. They train convolutional neural networks (CNNs) and transfer learning to pre trained networks in tackling limited medical data. Training consists of several epochs where hyperparameters are optimized to get the most generalization. The validation will be based on cross validation and external sources of independent test sets to determine robustness. Against expert clinicians, benchmarking is done in terms of sensitivity, specificity, and ROC AUC. They also perform ablation experiments to determine the impact of preprocessing methods and dataset size. The approach is a combination of model development and clinical assessment, to ensure the gap between research outputs and possible application in clinical processes.

Relationship to DoctorHive

The DoctorHive project builds on the findings of the current paper by realizing that technical accuracy is not the only way to create safe patient facing applications. At the same time as Esteva et al. demonstrate the potential of deep learning on a particular domain, DoctorHive argues that a number of specialized agents are necessary to cooperate and contradict one another and their decisions, which minimizes the error and increases reliability. Moreover, DoctorHive bridges the interpretability gap by providing outputs in patient friendly language that is interpretable, and this makes AI more applicable and reliable to actual healthcare settings.

3.1.3   Artificial Intelligence in Healthcare: Past, Present, and Future (Jiang et al., 2021)

Summary

The article by Jiang et al. (2021) [3] presents an overview of the development of artificial intelligence in healthcare, including early rule based expert system models to current models of deep learning and natural language processing. The paper sheds light on the scope of AI applications, namely diagnostics, treatment planning, patient monitoring, and population health management. It also highlights very long standing but critical issues such as data privacy, interpretability of models, integration into clinical processes, and regulatory compliance of the healthcare laws.

Methodology

The research uses a systematic literature review as its methodology and takes a historical approach to it. The classification reveals that Jiang et al. search publications published as early as the 1970s to 2020 through PubMed and IEEE Xplore and ACM databases. They classify AI technologies into chronological waves expert systems, classical machine learning, deep learning and NLP based systems. In the review, system architectures and their usage during clinical trials and hospitals are compared. They point out major representative mechanisms of every epoch, taking out the lessons of usability and trust. The thematic coding scheme combines the results into the interpretability, integration, privacy and regulation. By bringing together these themes, the methodology gives a diachronic explanation of how healthcare AI has changed over the years, both in terms of the advancement of computational methods and the socio technical obstacles that resurface. The process is important to make sure that both the technological innovations and a broader adoption issue are reflected by the review.

Relationship to DoctorHive

In the case of DoctorHive, this publication helps to create a necessary frame of reference, as it places the project in the context of the general history of AI development in healthcare. The issues that were revealed especially in regard to interpretability, integration, and trust are a direct reason as to why DoctorHive has chosen to lean towards a multi specialist model that is consent driven. By bridging these discrepancies, DoctorHive will become a progressive solution that not only extends the achievements of previous applications of AI, but also predicts and prevents the dangers identified by Jiang.

3.1.4   Medical Chatbots: Promises and Perils (Hwang & Kesselheim, 2022)

Summary

Hwang and Kesselheim (2022) [4] critically discuss the increase in using medical chatbots in patient information, triage, and support. The paper has identified some of the benefits that may include, better access, lower medical costs, and scalability of medical advice. Simultaneously, it cautions against the risks such as the incorrect or misleading outputs, absence of contextual comprehension, and the possibility of inflicting unnecessary anxiety in patients. The authors point out that there is a need to oversee, validate, and carefully implement chatbot technologies in the healthcare sector.

Methodology

The methodology is a combination of review and case study examination. Hwang and Kesselheim look at the chatbot applications that are already there, which include commercial health helpers and research prototypes, by looking at how they work, what they say they do, and how they are built technically. They look at the reports on deployment cases and evaluations of health technology that record chatbot usage in triage, self care of patients, and health education. The regulatory frameworks (for instance, FDA, EMA) are looked at to see where there are gaps in oversight. The authors look at the recorded cases of misinformation or wrong chatbot suggestions, and use them as warning examples. The structure of the methodology is a benefit risk framework: for every possible advantage (for instance, reduction of cost), they find the risks that go with it (for instance, error in diagnosis). This organized examination shows the situations in which chatbot usage is good versus bad. The methodology has a qualitative and policy based nature, based on the comparison of implementations in the real world.

Relationship to DoctorHive

DoctorHive is addressing directly the problems that Hwang and Kesselheim pointed out by going beyond the chatbot model that uses single agent. Rather than giving outputs that may be wrong from one point of view, DoctorHive uses several specialist agents that have discussions and improve their conclusions, which makes the chance of dangerous misinformation smaller. Through the inclusion of interpretability and consensus, DoctorHive changes the chatbot system into one that is more safe and can be trusted, which is in line with the worries that are brought up in this paper.

3.1.5 Multi Agent System Applications in Healthcare AI (Chen et al., 2020)

Summary

Chen et al. (2020) [5] look into the function of multi agent systems (MAS) in the healthcare sector, pointing out how AI agents that are distributed can work together to find solutions to complicated medical issues. The paper describes uses such as management of hospital resources, help with diagnosis, and monitoring of patients, showing that several specialized agents can work together to get results that are more solid and able to adapt than systems with single agent. The authors say that MAS give the ability to scale and be flexible that fits well with the nature of healthcare that has many parts.

Methodology

Chen et al. use a methodology that is driven by simulation and is supported by the design of conceptual system. They create MAS structures where agents are given functions such as evaluators of diagnosis, helpers with scheduling, or nodes for monitoring patients. The simulation settings are copies of hospital situations, with data on patients that is synthetic and situations for allocation of resources. The techniques of agent based modeling are used, which let agents talk and have responsibilities for making decisions together. The performance is looked at through analysis of scenarios, measuring results such as rates of diagnostic error that are reduced, use of resources that is improved, and times for response that are faster. The methodology puts emphasis on testing for robustness by doing several iterations of simulation in conditions that vary. The literature on the principles of MAS theory is put together with situations in healthcare that are practical, making a connection between models that are conceptual with simulations that are applied. Even though it is not checked with trials on live patients, the use of simulation based on agents that is structured gives a basis for translation to clinical use in the future.

Relationship to DoctorHive

DoctorHive can be seen as a direct use of the ideas that are described in this paper, taking MAS from the management of healthcare in the back end to the consultations with patients that face them. Through the inclusion of several specialist agents who do not just work together but also have discussions and improve their conclusions, DoctorHive puts in a mechanism of consensus that is structured to the framework of MAS. This way makes the reliability of medical advice stronger and at the same time deals with the gap in the work of Chen et al. about deployment that is practical and clinical.

3.1.6 Healthcare AI and Trust: A Multi Stakeholder Perspective (Krittanawong et al., 2021)

Summary

Krittanawong et al. (2021) [6] put their attention on the problem of trust in AI for healthcare, looking at points of view from patients, doctors, people who make policies, and people who regulate. The paper says that the adoption of AI in medicine that is successful does not just depend on accuracy but also on the ability to interpret, ability to explain, and being accountable. The authors point out obstacles such as bias in data, the absence of ability to interpret, and worries about ethics, and they ask for frameworks of regulation that are stronger and working together across disciplines to create ecosystems of AI that can be trusted.

Methodology

The authors use a methodology that is conceptual and focused on stakeholders. They put together information from studies of clinical cases, documents on regulation, and guidelines on ethics to create a map of the points of view of various actors who are involved in AI for healthcare. The sources of data are studies that are peer reviewed on bias, surveys of patients on acceptance of AI, and statements of policy that are official from authorities on health. The analysis is organized into groups by the type of stakeholder: patients, doctors, people who make policies, and people who regulate. The methodology uses analysis that is thematic, finding problems related to trust that come up again and again such as absence of ability to interpret and gaps in accountability. Rather than validation that is experimental, the study creates a framework that is socio technical and integrated by making the worries of stakeholders that are diverse line up. This way puts emphasis on width and being inclusive, showing trust as an idea that has many dimensions and is formed by factors that are technical, institutional, and cultural.

Relationship to DoctorHive

The DoctorHive project is in line directly with the things that are important that are found in this paper by putting interpretability and accountability into the design. Through showing points of view of several specialists and giving an output that is based on consensus, DoctorHive gives the ability to explain in a manner that patients are able to comprehend and doctors are able to trust. Through doing so, it puts into operation the ideas that are theoretical of trust that are described by Krittanawong et al., giving an example that is concrete of how AI that is centered on trust can be put into practice in systems for healthcare that face patients.

3.1.7   Consensus Driven AI for Clinical Decision Support (Xu et al., 2022)

Summary

Xu et al. (2022) [7] show a framework for intelligence that is artificial and driven by consensus in systems for support of clinical decisions (CDSS). The study shows that at the time when several AI models are used in a way that is collaborative, the consensus that is gathered gives outputs that are more accurate and can be relied on than models that are individual alone. The authors check their way with studies of cases in support of diagnosis, showing that methods that are based on consensus are able to make errors smaller and give robustness that is greater across datasets of patients that are diverse.

Methodology

Xu et al. use a design that is experimental where several AI models, which are trained in a way that is independent on datasets that are clinical, are put together into a framework of consensus. They make an ensemble of models of machine learning and deep learning that do predictions of diagnosis. The data is from records of patients that are retrospective, which are preprocessed for making normal and split into sets for training, validation, and test. The mechanisms of consensus are including voting by majority, averaging that is weighted, and aggregation that is Bayesian. The methodology is involving the evaluation of outputs across several studies of cases for diagnosis, using measurements such as accuracy, sensitivity, and robustness to the variability of dataset. The models are put to test both in a way that is individual and as a part of the system of consensus, which lets comparison that is direct. For showing the reduction of error, the authors look at cases of disagreement, showing how consensus makes misclassifications smaller. This combination of methodology of design of ensemble, evaluation that is comparative, and testing of robustness gives a basis that is empirical and rigorous for AI that is clinical and driven by consensus.

Relationship to DoctorHive

The things that are found in this work give support directly to the architecture that is core of DoctorHive, which is created around consensus of multi agent. At the same time as Xu et al. show the effectiveness of consensus in support of decision in the back end, DoctorHive takes this idea into a system that faces patients. Through the combination of consensus with interpretability and reasoning that is specific to specialists, DoctorHive creates on the foundation of Xu et al. to give both reliability that is improved and trust of patients that is enhanced in consultations of medicine that are driven by AI.

3.1.8 Conversational Agents in Healthcare: A Systematic Review (Laranjo et al., 2018)

Summary

Laranjo et al. (2018) [8] do a review that is systematic of agents that are conversational that are used in healthcare, looking at the uses of them in areas such as education of patients, support for change in behavior, checking of symptoms, and management of disease that is chronic. The review discovers that agents that are conversational are able to make accessibility and engagement of patients better but also points out limits in reliability of diagnosis, understanding of language that is natural, and effectiveness that is clinical in the long term.

Methodology

The authors use a methodology of review that is systematic that is in line with the guidelines of PRISMA. They look through databases which include PubMed, Scopus, and Web of Science, using words that are predefined as keywords such as "conversational agents," "chatbots," and "healthcare." The criteria for inclusion are limited to studies that are peer reviewed with evaluation that is empirical of agents that are conversational in situations of healthcare. After doing screening and taking away duplicates, papers that are eligible are taken out and put into codes. The review puts agents that are conversational into groups by the purpose (education, change in behavior, triage, monitoring of symptoms), group of users, and medium of deployment. The extraction of data also writes down methods of evaluation, which include trials that are clinical, surveys on satisfaction of users, and studies that are pilot. The methodology uses synthesis that is narrative, making comparisons of findings across groups and pointing out both the good things and the limits. Through the capture in a way that is systematic of the condition of research on chatbots up to 2018, the authors make a baseline for innovations that come after in AI that is conversational.

Relationship to DoctorHive

This paper gives a baseline that is important for comprehending both the promise and the limits of agents that are conversational in healthcare. The problems that are found especially around reliability of diagnosis and safety are dealt with directly by the design of DoctorHive. Through the use of several agents that are specialized that have discussions and get to consensus, DoctorHive makes reliability better and at the same time keeps the benefits of accessibility that are pointed out in the review of Laranjo et al. Through this manner, DoctorHive is showing the evolution that is next of AI for healthcare that is conversational.

3.1.9 Multidisciplinary Collaboration for Trustworthy AI in Medicine (Patel & Shortliffe, 2021)

Summary

Patel and Shortliffe (2021) [9] say that AI that is trustworthy in medicine needs working together across disciplines with doctors, scientists of computers, people who work on ethics, and people who make policies. The paper points out problems such as datasets that have bias, absence of ability to interpret, and integration of systems that is broken into pieces, which make the effectiveness of tools of AI weaker. It puts emphasis on the fact that the combination of knowledge of domain from medicine with innovation that is technical has importance that is critical for making systems that are both accurate and responsible in a way that is social.

Methodology

Patel and Shortliffe use a methodology that is conceptual and focused on policy. They look at cases that are published of systems of AI for medicine that failed or had bias, using them as proof of the results of input across disciplines that is not sufficient. The literature is taken from areas of informatics in medicine, science of computers, ethics, and health that is public. The methodology is based on synthesis that is comparative, showing how gaps in working together cause problems in interpretability, fairness, and integration of system. Through the cross referencing of studies of cases and guidelines of policy, the authors find weaknesses that are structural in the pipelines of development of AI that are current. The method is involving the creation of a framework that is normative instead of putting to test an algorithm that is specific: they say that design that is interdisciplinary should be made into an institution from the curation of dataset to the deployment that is clinical. The methodology points out patterns across sectors that are multiple, putting emphasis on solutions that are systemic instead of fixes that are technical.

Relationship to DoctorHive

DoctorHive puts into operation the idea of working together across disciplines through the design of system that is multi agent. Every agent that is specialist is showing knowledge that is specific to domain, and the debate of them that is structured is like working together in the real world among people who work in medicine. This multiplicity of points of view that is built in deals with the worries that are brought up by Patel and Shortliffe, changing discussions that are theoretical of working together into a system of AI that works and is able to make the reliability and the ability to be trusted of advice that is medical better.

3.1.10 Multi Agent Debate Improves Medical Reasoning (Lin et al., 2023)

Summary

Lin et al. (2023) [10] bring in a framework that is novel where several agents of AI do debates that are structured to find solutions to tasks of reasoning in medicine that are complex. The experiments of them show that mechanisms of debate are helping to bring errors to the surface, make reasoning better, and in the end make accuracy of diagnosis better when compared to ways that use single model. The study shows that the paradigm of debate makes not just the correctness of outputs better but also the interpretability of them, because agents do explaining and challenging of reasoning of each other.

Methodology

Lin et al. create a framework that is experimental and controlled where several agents of AI are trained in a way that is independent on datasets of reasoning that is clinical. The methodology is including the assignment to agents of roles that are complementary (for instance, giving hypotheses of diagnosis, doing criticism of paths of reasoning). The protocols of debate that are structured are made clear, where agents give arguments and counterarguments in rounds that are iterative. The process has automation, with every round put under evaluation for correctness and quality of reasoning. The datasets of cases that are clinical, which include vignettes of diagnosis and data of patients that is structured, are used as inputs for test. The outputs of debate are given scores against labels of ground truth that are given by experts, with performance that is measured through accuracy of diagnosis, rates of detection of errors, and clarity of reasoning. The experiments that are comparative put agents that are based on debate against baselines of single agent and methods of ensemble. The methodology points out how reasoning that is adversarial brings weaknesses to the surface and makes accuracy better, showing the value of protocols of debate in the design of AI.

Relationship to DoctorHive

The things that are found by Lin et al. are directly giving a foundation for DoctorHive, which creates upon the idea of debate to make reliability better. DoctorHive takes this idea further by giving roles of specialists (for instance, doctor of cardiology, doctor of neurology) to agents, making sure that debates are not just adversarial but also have information from domains. What is more, DoctorHive makes a bridge over the gap that is pointed out in the study of Lin et al. by changing outcomes of debate into responses of consensus that are clear and friendly to patients, which makes the framework more practical for usage in the real world.

3.1.11 Harnessing Multi Agent Collaboration for Trustworthy AI (Wu et al., 2023)

Summary

Wu et al. (2023) [11] say that working together of multi agent is a strategy that is key for creating systems of AI that are trustworthy. The paper describes how agents that are distributed with roles that are specialized are able to work together to make robustness, interpretability, and the ability to adapt better across areas, which include healthcare. The experiments of them show that frameworks that are collaborative do better than systems that are single agent when it comes to accuracy and being strong against errors, and at the same time let processes of reasoning be more interpretable.

Methodology

Wu et al. use a framework that is experimental and across many domains. They create systems of AI that are collaborative where agents are trained on tasks that are specialized and then put together in coordination in an architecture that is larger. The methodology is including the assignment at the level of agent of roles, such as analysis of diagnosis, validation that is contextual, and checking of errors. They put these systems to test on datasets that go across fields that are multiple, which include records of diagnosis for healthcare and tasks that are not medical, to see how generalizable they are. The performance is measured using accuracy, robustness when data has noise, and being strong against examples that are adversarial. The interpretability is put under evaluation in a way that is qualitative by looking at how agents say reasoning that is intermediate. The experiments that are comparative put systems of MAS that are collaborative against baselines of single agent, showing improvements in reliability. The methodology puts emphasis on testing that is empirical across situations, putting frameworks of trust that are conceptual into results that can be measured.

Relationship to DoctorHive

DoctorHive shows in itself the framework of working together of multi agent that is supported by Wu et al., using it in a way that is specific to consultations of medicine that face patients. Through the assignment of roles of specialists that are distinct and bringing in a mechanism of consensus that is driven by debate, DoctorHive changes the theory of trust that is collaborative into an application for healthcare that is concrete. What is more, through the giving of outputs in language that is clear and friendly to patients, DoctorHive deals with the gap in usability that the study of Wu et al. makes open, making sure that being trustworthy is both technical and about experience.

3.1.12 Explainable AI in Healthcare (Wong et al., 2021)

Summary

Wong et al. (2021) [12] look at the importance of intelligence that is artificial and explainable (XAI) in uses for healthcare, in a way that is particular in situations of making decisions that are clinical where the ability to interpret has importance that is critical. The paper does a review of methods that are various of being explainable such as attribution of features, making visualizations, and reasoning that is based on rules and does an assessment of how suitable they are for cases of use in medicine that are different. The authors say that if there is no interpretability, even systems of AI that are accurate may not succeed in getting acceptance among doctors and patients.

Methodology

The authors use a methodology of review that is focused on techniques of explainability that are technical. They gather and make comparisons of literature on XAI that is applied to imaging in medicine, analysis of EHR, and modeling that is predictive. The methods that are reviewed are including ways of attribution of features (for instance, SHAP, LIME), tools of visualization for attention of model, and systems of reasoning that are based on rules. The studies of cases are taken from uses that are published such as radiology, cardiology, and oncology. The criteria of evaluation are including the usability by doctors, the fidelity to the model that is underlying, and the alignment with expectations of regulation for interpretability. The methodology puts emphasis on the classification of methods of XAI into groups (local against global, agnostic to model against specific to model) and the mapping of these groups to tasks that are clinical and real. Through this comparison that is structured, the study finds situations where being explainable is adding value and where it stays not sufficient, in a way that is particular in cases of use that face patients.

Relationship to DoctorHive

DoctorHive deals with directly the problems that are brought up by Wong et al. by putting interpretability and interpretability into the architecture. Rather than making an output that is single and opaque, DoctorHive shows points of view of several specialists and the process of consensus of them, which makes the reasoning able to be seen. This explainability that has two levels detail at the level of specialist for doctors and consensus that is simplified for patients makes a bridge over the gap that is found in the work of Wong et al., and in this way makes trust and usability better for both groups that listen.

3.1.13 Evaluating Medical AI Systems: Beyond Accuracy (Rajpurkar et al., 2022)

Summary

Rajpurkar et al. (2022) [13] say that the evaluation of systems of AI for medicine should go past measurements of accuracy that are traditional to have robustness, being generalizable, fairness, interpretability, and effect that is clinical. The paper does a review of problems that are common in the ways of evaluation that are current, such as the reliance that is too much on datasets of benchmark and the absence of validation that is external. The authors give support to a framework that is holistic that does measuring of how AI does performance in situations that are clinical and in the real world, taking into account populations that are diverse and integration of workflow.

Methodology

Rajpurkar et al. use a methodology of building of framework that is based in synthesis of literature and review of study of cases. They look at limits that are reported of evaluation of AI for medicine, such as bias in dataset and validation that is not sufficient, by doing a review of studies of diagnosis by AI that have high profile. The methodology puts dimensions of evaluation into groups, which are accuracy, robustness, being generalizable, interpretability, and utility that is clinical. For every dimension, they find examples from work that is published, pointing out strengths and things that fail. They then say a framework of evaluation that is conceptual that goes past measurements of accuracy that are standard. The way of methodology is qualitative but has a base in evidence that is empirical from deployments that happened before. The review puts emphasis on studies of validation that are external, monitoring after deployment, and fairness across groups of population that are subgroups. Through the alignment of these things that are themes, the methodology points out the need for strategies of evaluation that have many dimensions in AI for healthcare.

Relationship to DoctorHive

DoctorHive is in line with the call by Rajpurkar et al. for evaluation that is holistic by doing design for robustness, interpretability, and trust, and not just accuracy of diagnosis. Through the putting together of several agents that are specialists and a mechanism of consensus, DoctorHive in a way that is inherent deals with being generalizable and being strong against errors. What is more, the output of consensus that is friendly to patients and the pathways of reasoning that are interpretable show how systems of AI are able to put into operation the dimensions of evaluation that are expanded that are pointed out in this paper, which makes it a response that is practical to the criticism of the authors.

3.1.14 Trustworthy Multi Agent Systems for Clinical Reasoning (Cao et al., 2022)

Summary

Cao et al. (2022) [14] do an investigation of the design of systems that are multi agent (MAS) that are trustworthy for tasks of reasoning that are clinical. The paper shows how the distribution of reasoning of diagnosis across agents that are specialized is able to make errors that are individual smaller, make robustness better, and give explanations that are richer for cases of medicine that are complex. The authors put emphasis on the importance of building of consensus, interpretability, and checking of errors in MAS to make sure there is reliability and being applicable to clinical settings.

Methodology

Cao et al. use development of prototype and testing that is based on simulation. They create architectures of MAS with agents that are distributed, with every one having responsibility for parts that are different of reasoning that is clinical (for instance, analysis of data, generation of hypothesis, detection of errors). The scenarios that are clinical and synthetic are used to put to test the working together of agents, with datasets that are showing profiles of patients that are complex. The protocols of consensus are put in to make agents able to do validation in a way that is cross and give flags to things that are not consistent. The measurements of evaluation are including accuracy of diagnosis, robustness of reasoning, and interpretability of explanations. The studies of cases that are simulated let there be observation of the way errors are corrected when agents do not agree. The methodology puts emphasis on validation that is proof of concept instead of deployment in settings of healthcare that are live, giving evidence that MAS are able to put into operation ideas of trust through reasoning that is distributed and consensus.

Relationship to DoctorHive

DoctorHive creates in a way that is direct on the thoughts that are shown by Cao et al. by taking frameworks of trust of multi agent into an application that faces patients. The mechanism of debate and consensus puts into operation the ideas of correction of errors and interpretability that are talked about in the paper, and at the same time the focus on output that is clear and friendly to patients deals with the gap in usability that is made open by Cao et al. Through this manner, DoctorHive is working as an extension that is practical of ideas of MAS that are trustworthy, making a bridge over the gap between theory and deployment.

3.1.15 Collaborative Agents for Medical Consultation (Zhang et al., 2024)

Summary

Zhang et al. (2024) [15] bring in a framework where several agents of AI work together to do simulations of consultations that are medical, with every one giving knowledge that is specific to domain. The study shows that agents that are collaborative are able to make advice that is medical and is richer and more balanced when compared to chatbots that are single agent. The evaluation of them points out reasoning of diagnosis that is improved, rates of errors that are reduced, and trust of patients that is higher at the time when several agents are put together into a system that is coordinated.

Methodology

Zhang et al. use a methodology that is experimental and is focused on creating and putting to test architectures of agents that are collaborative. Several agents of AI that are specialized are trained on knowledge of medicine that is specific to domain, which includes datasets of cardiology, dermatology, and neurology. The agents are then put together into a framework of consultation where they give reasoning that is complementary. The protocols of working together that are structured give guidance to the interaction of agents, making sure that outputs are made harmonious into a response of consultation that is joint. The study does evaluation of performance on consultations of patients that are simulated and vignettes of cases, making comparisons of outputs of single agent with responses of multi agent that are collaborative. The criteria of evaluation are including accuracy of diagnosis, reduction of errors, diversity of reasoning, and trust of users, which is measured through surveys of feedback from doctors. The methodology points out improvements in both performance that is objective and trust that is subjective, and at the same time says that testing has limits to situations that are experimental instead of situations of clinics that are in the real world.

Relationship to DoctorHive DoctorHive is taking the idea of Zhang et al. of agents that are collaborative for medicine and makes it go forward by putting in debate that is structured and mechanisms of consensus, making sure that disagreements are solved in a way that is systematic instead of just being put together. What is more, DoctorHive puts priority on the usability by patients by changing debates of specialists into outputs that are short, clear, and friendly to patients. Through doing so, it takes the framework of Zhang et al. from a proof of concept into a system that is more practical and can be deployed for making trust and reliability better in consultations of medicine that are assisted by AI.

3.2 Literature Review Summary Table

Table 3.1: Summary of Reviewed Literature and Relevance to DoctorHive

Study	Methods / Features	Relevance to DoctorHive	Limitations
Bohr & Memarzadeh (2020) [1]	Narrative review of AI applications across imaging, diagnostics, monitoring, and patient safety; thematic synthesis of opportunities and risks from published studies, technical reports, and regulatory perspectives.	Provides context on risks/opportunities; strengthens rationale for consensus based safety mechanisms.	General overview; lacks methodological depth on MAS, interpretability, or patient trust.
Esteva et al. (2019) [2]	Experimental deep learning studies; CNNs trained on large annotated datasets (imaging, dermatology, pathology) with preprocessing, transfer learning, and benchmarking against clinician performance.	Demonstrates diagnostic potential of AI; highlights scope for integrating accuracy focused models into workflows.	Prioritizes accuracy metrics; limited discussion of interpretability, fairness, or usability.
Jiang et al. (2021) [3]	Systematic historical review; traces AI from expert systems to deep learning/NLP; sources include peer reviewed papers and policy reports; structured into chronological phases with thematic coding.	Positions DoctorHive within AI's long term trajectory; identifies recurring challenges around trust and integration.	Largely descriptive; minimal exploration of MAS or technical solutions for trust.
Hwang & Kesselheim (2022) [4]	Policy and case oriented review; examines chatbot deployments in triage, symptom checking, education; evaluates benefits/risks using regulatory and ethical frameworks; includes real world case incidents.	Emphasizes safety risks of chatbots; supports DoctorHive's more reliable multi agent consensus model.	Problem focused; no architectural solutions; overlooks MAS or debate mechanisms.

Table 3.1 – continued from previous page

Study	Methods / Features	Relevance to DoctorHive	Limitations
Chen et al. (2020) [5]	Simulation driven MAS studies; designs distributed agent frameworks for diagnostics, hospital resource management, and monitoring; evaluates collaboration via scenario based simulations.	Provides foundational MAS concepts; supports collaborative multi agent reasoning as a framework for DoctorHive.	Lacks clinical validation; focused mainly on simulation and theoretical arguments.
Krittanawong et al. (2021) [6]	Conceptual stakeholder review; synthesizes perspectives from patients, clinicians, regulators, and policymakers; draws on clinical cases, surveys, and policy guidelines to map trust related issues.	Aligns with DoctorHive's interpretability and accountability design; addresses socio technical adoption factors.	Conceptual and theoretical; does not propose detailed technical trust mechanisms.
Xu et al. (2022) [7]	Empirical study of consensus frameworks; integrates multiple AI models into CDSS; tests aggregation methods (majority vote, weighted averaging, Bayesian approaches) on diagnostic datasets.	Validates consensus as a tool for accuracy and robustness; directly informs DoctorHive's consensus based core.	Focuses on back end CDSS; does not address usability or patient facing integration.
Laranjo et al. (2018) [8]	Systematic review (PRISMA guidelines); identifies and classifies conversational agents by purpose (education, triage, chronic care); compares evaluation methods (trials, surveys, pilots).	Establishes baseline of chatbot benefits and shortcomings; highlights reliability gaps addressed by DoctorHive.	Predates LLM/MAS developments; emphasizes engagement over reliability.
Patel & Shortliffe (2021) [9]	Conceptual analysis; reviews failed AI implementations and biased datasets; argues for embedding multidisciplinary collaboration (clinicians, computer scientists, ethicists, policymakers).	Supports DoctorHive's multi specialist design; frames collaboration as essential for reliability.	Lacks implementation strategies; serves more as a call to action.

Table 3.1 – continued from previous page

Study	Methods / Features	Relevance to DoctorHive	Limitations
Lin et al. (2023) [10]	Experimental design of multi agent debates; independent agents trained on diagnostic datasets engage in structured argument counterargument protocols; compared against single model baselines.	Directly foundational for DoctorHive's debate driven framework; validates structured disagreement as error reduction.	Limited to controlled experiments; not evaluated in patient facing deployments.
Wu et al. (2023) [11]	Multi domain empirical study; specialized agents collaborate across tasks; tested on healthcare and non healthcare datasets; compared MAS vs. single agent performance on robustness and interpretability.	Supports MAS collaboration as a means to enhance trust and resilience; informs DoctorHive's architecture.	Healthcare is one of many test cases; limited discussion of patient level usability.
Wong et al. (2021) [12]	Technical review of XAI methods (e.g., SHAP, LIME, rule based, visualization tools); compares local vs. global approaches; evaluates suitability for clinical contexts.	Justifies DoctorHive's dual layer explainability design for both clinicians and patients.	Focuses mainly on clinicians; limited exploration of patient facing interpretability.
Rajpurkar et al. (2022) [13]	Framework oriented literature synthesis; reviews evaluation shortcomings in medical AI; proposes multidimensional framework covering robustness, generalizability, fairness, interpretability, clinical impact.	Aligns with DoctorHive's multi dimensional goals of reliability, trust, and robustness beyond raw accuracy.	High level guidance; does not provide operational tools for implementation.
Cao et al. (2022) [14]	Prototype MAS design for clinical reasoning; distributes diagnostic tasks among agents; tests consensus and error checking using simulated patient scenarios.	Provides direct methodological support for DoctorHive's debate consensus model.	Limited to prototypes and simulations; not yet validated in clinical settings.

Table 3.1 – continued from previous page

Study	Methods / Features	Relevance to DoctorHive	Limitations
Zhang et al. (2024) [15]	Experimental study of collaborative agents simulating medical consultations; domain specialized agents contribute expertise; evaluated using simulated consultations and clinician feedback.	Extends directly into DoctorHive's patient facing MAS design; strengthens evidence for collaborative consultation models.	Early stage and experimental; lacks validation in live healthcare workflows.

3.3   Conclusion

The literature that was reviewed in this chapter is showing the function that is growing of AI in healthcare, and at the same time is pointing out the limits and dangers that are connected with ways that are current. Several studies are pointing out the potential that is transformative of deep learning and agents that are conversational, but also are pointing out gaps that are critical in interpretability, trust, and validation that is clinical. The working together of multi agent and mechanisms of debate are coming out as a direction that has promise, giving ways to make the dangers of making decisions by single agent smaller through the putting together of points of view that are diverse into a consensus. DoctorHive is dealing with these gaps in a way that is direct by putting into operation a framework that is multi specialist and driven by debate that is created to make accuracy of diagnosis and trust of patients better. Through the putting together of advances that are technical in systems of multi agent, information on trust and interpretability, and things to think about that are socio ethical, DoctorHive is putting itself as a contribution that is novel to delivery of healthcare that is supported by AI and reliable. The chapter that follows will create upon these pieces of information to talk about the design of system and putting into practice of the MVP of DoctorHive.

Chapter 4 Software Requirement Specifications

This chapter is making clear the requirements and things to think about for design for DoctorHive, which is a system of consultation that is medical and multi agent and based on web that is given power by models of language that are large (LLMs). The purpose of this chapter is to talk about with details the requirements that are functional and non functional, things that limit, things that are assumed, cases of use, design of database, and things to think about for interface of users that are giving shape to the system. Through the documenting of these in a way that is systematic, this chapter is making sure that the platform of DoctorHive is able to be comprehended and made again by people who do research or people who develop that are other. It also is working as the base for decisions of design and putting into practice in chapters that come later.

4.1 List of Features

DoctorHive is giving several features that are important that are making it different from systems of chatbot that are single agent and traditional and are making it line up with problems that are current in AI for medicine:

- **Role Conditioned Specialist Agents:** The system is using several agents that are based on LLM, with every one given the function of a specialist in medicine (for instance, doctor of cardiology, doctor of neurology, doctor of dermatology). The conditioning of role that is combined with prompting of few shot is making sure that every agent is giving responses in a way that is in line with the knowledge of a specialist.

- **Retrieval Augmented Generation (RAG):** Agents are having a base in knowledge of medicine through the retrieval of evidence that is relevant from a base of knowledge that is put together with care. The database of PostgreSQL is keeping data of medicine that is structured, and at the same time a database of vector (for instance, pgvector) is keeping embeddings for search of similarity that is efficient. This is making sure that outputs are staying grounded in a way that is factual.

- **Debate Mechanism:** Not like systems that are conventional where one model is making an output that is single, agents of DoctorHive are doing a debate that is structured. They are giving, doing rebuttals, and making claims better and at the same time are making references to evidence of medicine that is retrieved, which makes accuracy and interpretability better.

- **Consensus Layer:** A system of voting that is weighted is putting together the outputs that are debated and is making a recommendation of consensus that is final. The reasoning that

is supported by evidence is given priority and at the same time claims that are not supported are taken out through filtering.

- **Patient Friendly Outputs:** The system is making responses in two layers: a summary that is short and able to be read for patients and a version that is detailed and optional with citations and steps of reasoning for doctors or users that are advanced.

- **Feedback Loop:** Users are able to give ratings to the clarity and how useful the consultation is. The data of feedback is kept for evaluation and making the system better.

- **Auditability:** Every response is able to be traced to the evidence that is retrieved and the steps of debate that are taken by the agents. This is making sure there is interpretability and is helping to create trust of users.

4.2   Functional Requirements

The requirements that are functional are talking about the behavior that is external of the system as it is experienced by users that are end users:

- **User Management:** Patients are able to do registration, do logging in, do updating of profiles, and do management of the history of consultations of them.

- **Symptom Submission:** Patients are doing submission of symptoms and history of medicine through forms that are structured that are giving support to input of text, boxes for checking, and menus that drop down.

- **Knowledge Retrieval:** The system is doing retrieval of evidence that is relevant from PostgreSQL and the database of vector using embeddings for grounding of context.

- **Specialist Agent Reasoning:** Every agent that is based on LLM is doing processing of the query with the point of view of specialist that is assigned to it, which is given support by examples of few shot and knowledge that is retrieved.

- **Debate and Consensus:** Agents are doing exchange of points of view, are pointing out things that are not consistent, and are getting to a consensus using an algorithm that is structured.

- **Consultation Results:** Patients are getting a summary with steps that are next (for instance, "seek care that is urgent," "schedule a consultation," "monitor symptoms"). Doctors may get access to a log that is detailed of the debate.

- **Feedback Collection:** Both patients and doctors are able to give feedback on accuracy,

clarity, and how useful the advice is.

4.3 Quality Attributes

The attributes of quality are talking about the parts of "how well" of the system:

- Advice has to stay consistent and have a base in evidence that is validated. The mechanism of consensus is helping to make variability smaller.

- Outputs have to point out what sources were retrieved and how decisions were gotten to.

- The time for response that is average has to be in 5 7 seconds for queries that are standard, making sure there is usability in consultations that are real.

- The architecture has to give support to scaling to thousands of users that are concurrent through the distribution of workload across instances that are multiple.

- All data of patients has encryption, and access has restriction to users that are authenticated.

- The interface should have the quality of being intuitive, with flows of input that are guided to make error of users smaller.

- The code and design of system have to be following ideas that are modular so updates and fixes of bugs are able to be applied without doing disruption to functionality that is core.

4.4 Non Functional Requirements

- Modules like the RAG pipeline and debate engine should be reusable for other healthcare or non healthcare applications.

- New specialist roles (e.g., endocrinology, psychiatry) should be easily added by defining role prompts and integrating domain specific datasets.

- APIs must allow integration with external health platforms such as hospital EHR systems or third party health apps.

- The system should meet relevant healthcare standards such as HIPAA or GDPR for data privacy.

- The application should be deployable across different environments (local server, cloud infrastructure) with minimal adjustments.

4.5 Assumptions

- Users will have access to a stable internet connection and modern web browsers.

- The knowledge base will be maintained and updated regularly to reflect new medical research.

- Patients understand the system is for decision support and not a substitute for licensed medical professionals.

- Agents are assumed to behave consistently under role conditioning and retrieved evidence.

float array placeins

4.6 Use Cases

The use cases below illustrate how DoctorHive manages patient consultations through the orchestrated GP and specialist workflow. Each use case highlights actors, system interactions, and expected flows.

Table 4.1: Use Case: Patient Submits Symptoms and Creates Case

Name	Patient Submits Symptoms and Creates Case
Actors	Patient, System
Summary	A new case is generated when a patient submits their symptoms and any reports. The system parses inputs and stores them in the Cases table.
Pre Conditions	Patient is logged in and selects "New Consultation".
Post Conditions	A new Case ID is created, parsed info and files are stored in the database.
Special Requirements	Input parser must handle structured and unstructured data reliably.
Basic Flow	1. Patient clicks "New Chat". 2. System generates Case ID. 3. Patient submits symptoms and uploads reports. 4. Parsing model extracts relevant details and stores them in Cases table.
Alternative Flow	1A. Patient does not provide valid input. 2A. System prompts for correction.

Table 4.2: Use Case: GP Conducts Initial Triage with Follow Ups

Name	GP Conducts Initial Triage with Follow Ups
Actors	GP Agent, Patient, System
Summary	The GP agent asks clarifying follow up questions and records patient responses as key value pairs in the Cases table.
Pre Conditions	Case has been created with initial patient input.
Post Conditions	All necessary follow up answers are stored and specialists required are identified.
Special Requirements	GP agent must avoid redundant or irrelevant follow ups.
Basic Flow	1. GP agent reviews parsed input. 2. GP agent asks follow up question. 3. Patient provides answer. 4. Answer is saved in Cases table. 5. GP agent repeats until sufficient info gathered. 6. Specialists required are identified.
Alternative Flow	1A. Patient skips a question. 2A. System marks it as unanswered and proceeds.

Table 4.3: Use Case: Specialists Generate Initial Diagnoses

Name	Specialists Generate Initial Diagnoses
Actors	Cardiologist Agent, Neurologist Agent, Ophthalmologist Agent, System
Summary	Each specialist agent reviews the case history and produces an initial diagnosis, confidence score, and explanation, which are stored in their respective tables.
Pre Conditions	GP has completed triage and identified relevant specialists.
Post Conditions	Each specialist record contains diagnosis, confidence, and explanation.
Special Requirements	Specialists must only access relevant case data and follow privacy constraints.
Basic Flow	1. Specialist agent retrieves case details and follow up answers. 2. Specialist generates diagnosis and confidence. 3. Explanation is created. 4. Results are stored in respective specialist history table.
Alternative Flow	1A. Specialist cannot generate confident response. 2A. System marks uncertainty and flags case for consensus.

Table 4.4: Use Case: Debate and Consensus Generation

Name	Debate and Consensus Generation
Actors	Specialist Agents, RAG Engine, System
Summary	Specialist agents analyze one another's diagnoses and supporting evidence from RAG, then debate before updating their final confidence and participating in voting.
Pre Conditions	Initial diagnoses by specialists are stored in the database.
Post Conditions	Consensus outcome or fail safe advice is stored in the Case record.
Special Requirements	Debate logs and citations must be stored for interpretability.
Basic Flow	1. Specialists review RAG sourced evidence. 2. Agents analyze each other's diagnoses. 3. Debate exchanges are logged. 4. Specialists update final confidence. 5. Voting is performed. 6. Most confident/consistent verdict is selected.
Alternative Flow	1A. No consensus reached. 2A. System outputs disclaimer: "Consult a human doctor."

Table 4.5: Use Case: Patient Views Final Report and Provides Feedback

Name	Patient Views Final Report and Provides Feedback
Actors	Patient, System
Summary	The patient receives a consensus based consultation report and has the option to provide feedback on usefulness.
Pre Conditions	Consensus or fail safe verdict has been generated.
Post Conditions	Patient views results and feedback is stored.
Special Requirements	Feedback must be anonymized in analytics.
Basic Flow	1. Patient navigates to completed case report. 2. System displays summary verdict with explanations. 3. Patient optionally provides feedback rating and comments. 4. Feedback is stored securely.
Alternative Flow	1A. Patient skips feedback. 2A. No feedback stored, consultation ends.

4.7 Hardware and Software Requirements

4.7.1 Hardware Requirements

- Server with minimum 16 GB RAM, 8 core CPU, and GPU for LLM inference.

- SSD storage of 500 GB for knowledge base and embeddings.

- Backup server for redundancy and disaster recovery.

4.7.2 Software Requirements

- Backend: Python with FastAPI or Flask.

- Database: PostgreSQL with pgvector extension.

- Frontend: React.js with Bootstrap or Tailwind.

- Containerization: Docker and Docker Compose.

- LLM API access via secure connectors.

4.8 Graphical User Interface

The planned interface for GUI includes:

- Patient Dashboard showing consultation history and an entry point for new consultations.

- Symptom Submission Form which is a guided form with text fields, dropdowns, and check-lists and the option to add reports which automatically gets parsed

- Consultation Output View displays patient friendly recommendations with optional detailed view.

- Clinician Dashboard provides advanced users with access to debate transcripts and retrieved evidence.
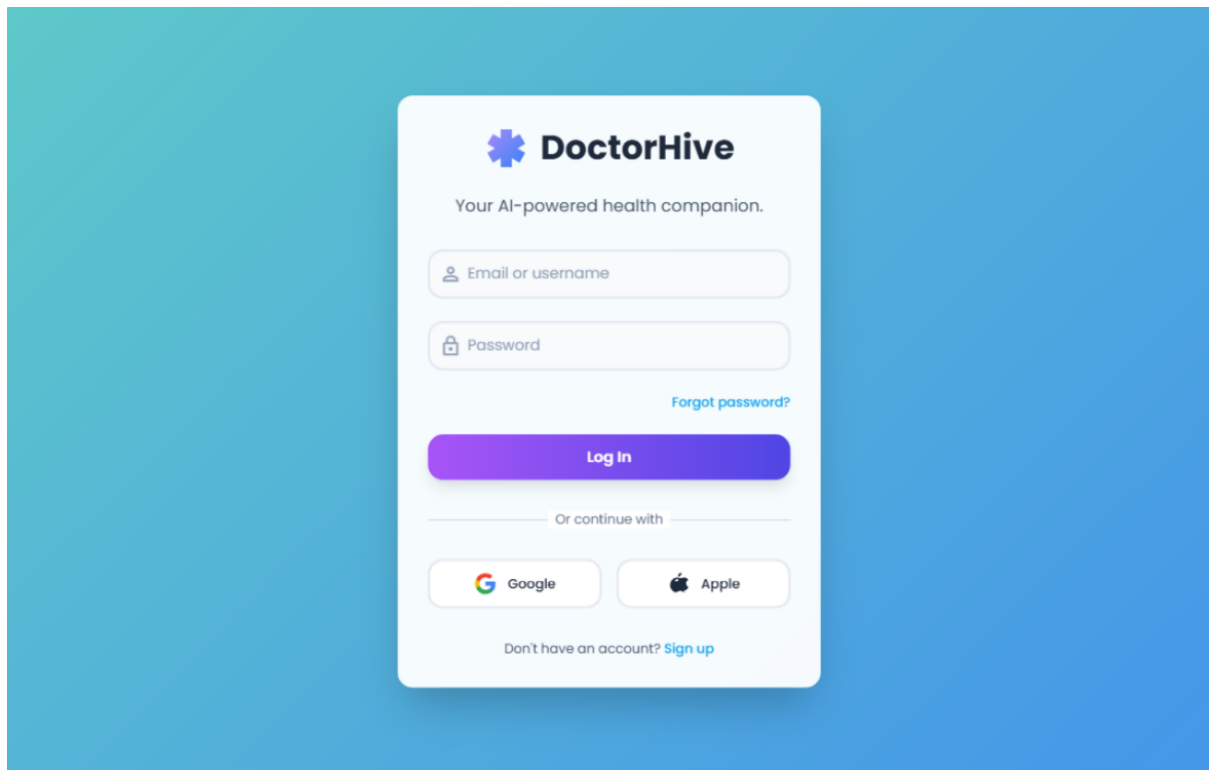
Figure 4.1: Login page for DoctorHive
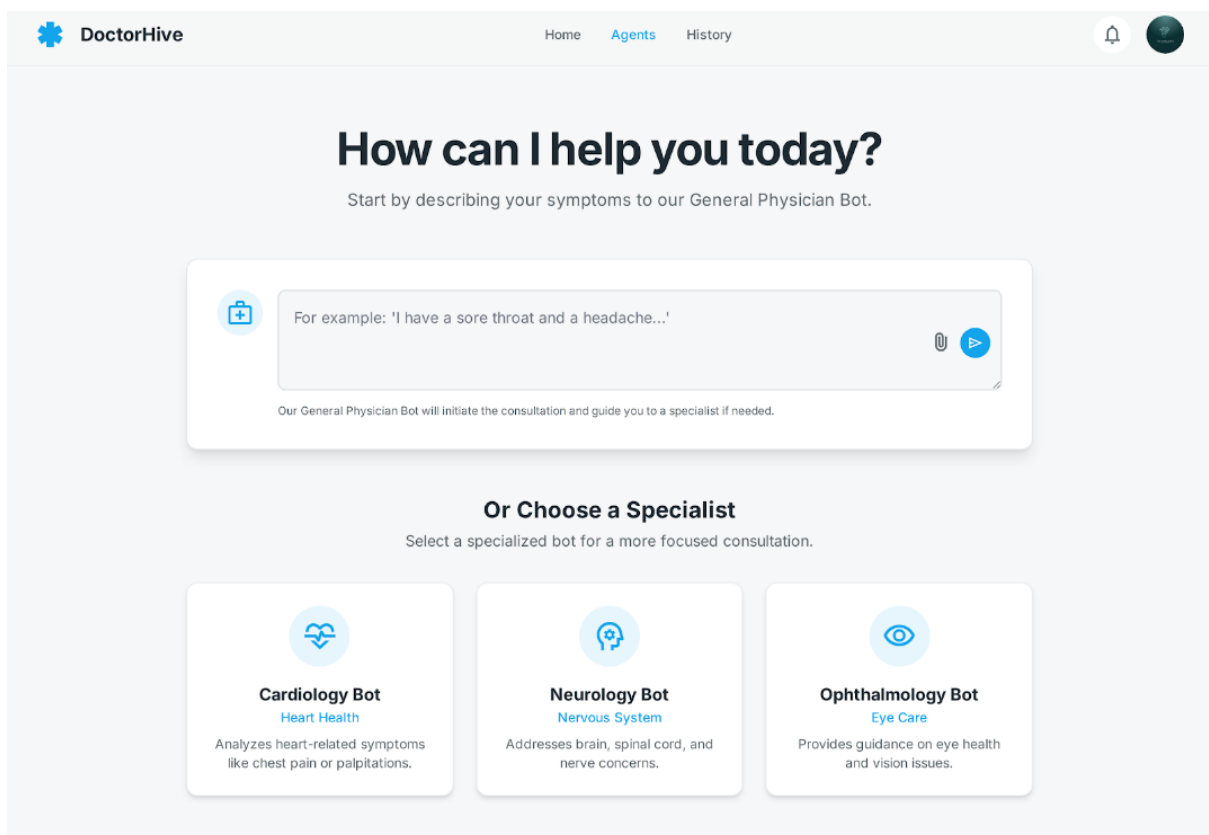
Figure 4.2: Initial interface for DoctorHive

Figure 4.3: User History interface for DoctorHive

4.9   Database Design

4.9.1   ER Diagram

The database will include entities for Users, Consultation Records, Retrieved Knowledge Chunks, Agent Debate Logs, and Feedback. Relationships will ensure that each consultation links patient input, system retrieval, agent reasoning, and consensus outputs.
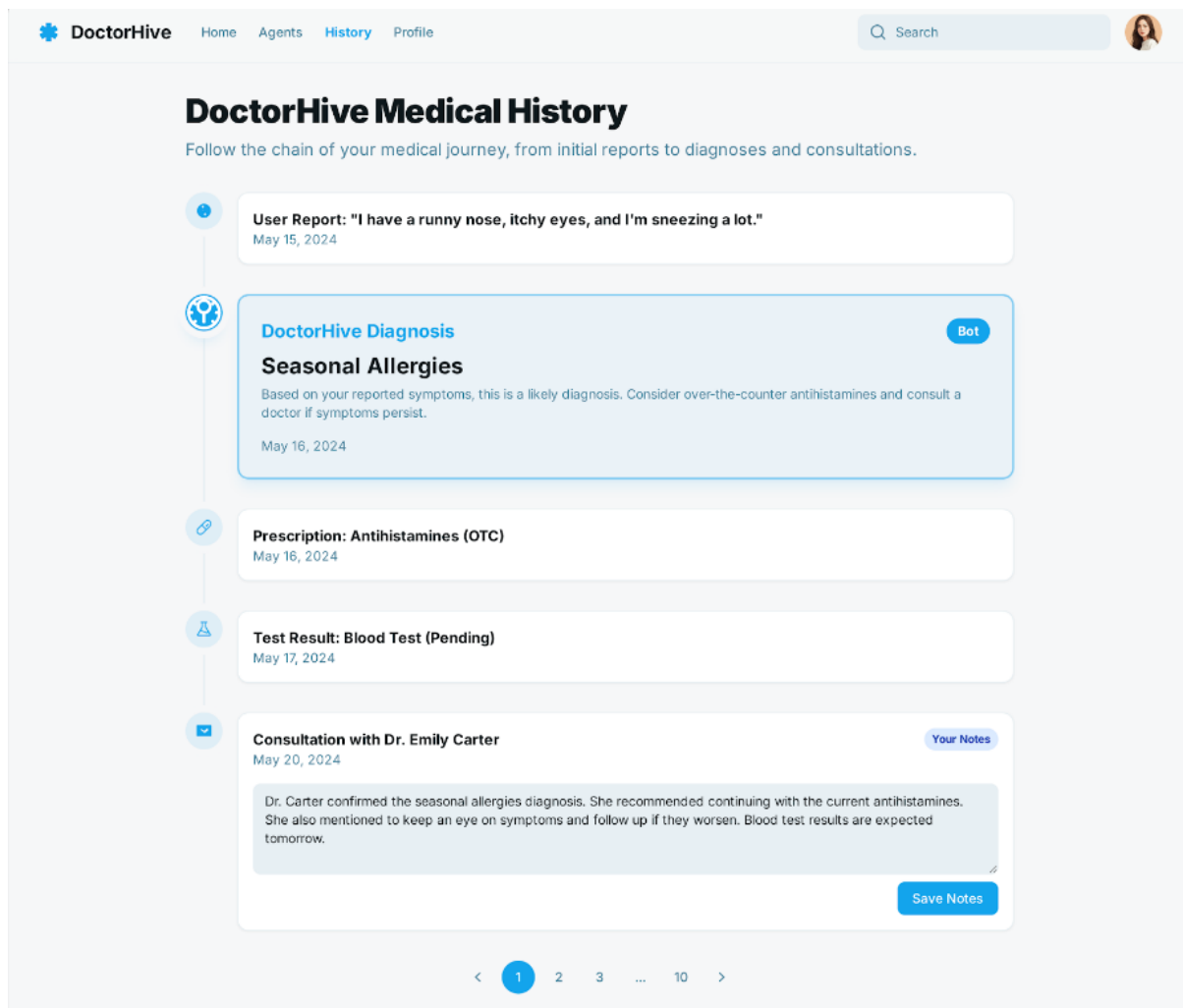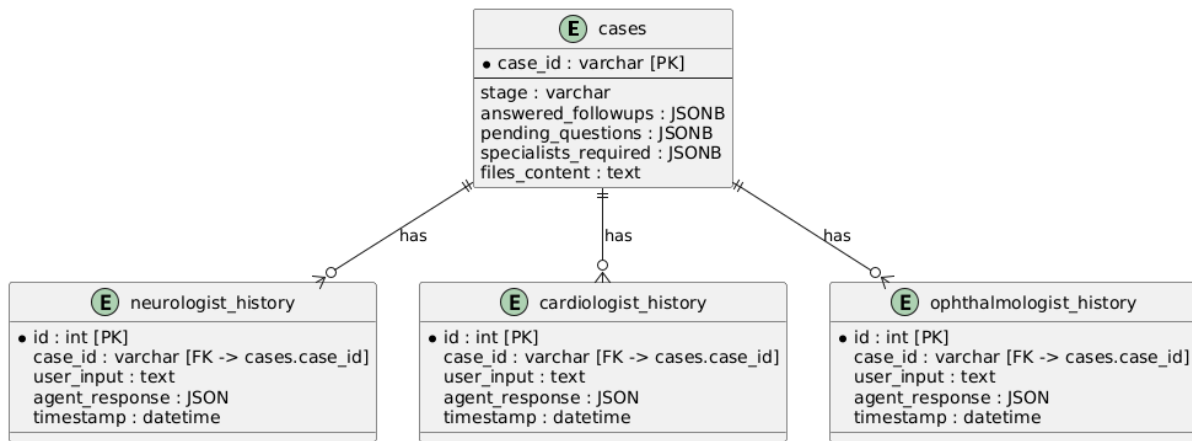
Figure 4.4: ER diagram of DoctorHive database schema.

4.9.2 Data Dictionary

- **Cases:** Stores each consultation session.

  - **case_id** : Unique identifier for the case.

  - **stage** : Current stage of the consultation (init, triage, specialist, debate, final).

  - **answered_followups** : JSON list of patient Q/A pairs collected during GP triage.

  - **pending_questions** : JSON list of remaining follow up questions.

  - **specialists_required** : JSON list of specialist agents needed for this case.

  - **files_content** : Extracted text or structured data from uploaded patient reports.

- **Neurologist_History:** Tracks interactions between the Neurologist agent and the patient/system.

  - **id** : Auto increment primary key.

  - **case_id** : Foreign key referencing Cases table.

  - **user_input** : Patient responses relevant to neurology.

  - **agent_response** : JSON object containing initial diagnosis, confidence, and explanation.

  - **timestamp** : Time of record creation.

- **Cardiologist_History:** Stores case specific outputs from the Cardiologist agent.

  - **id**, **case_id**, **user_input**, **agent_response**, **timestamp** : Same structure as Neurologist_History.

- **Ophthalmologist History:** Stores case specific outputs from the Ophthalmologist agent.

  - **id**, **case id**, **user input**, **agent response**, **timestamp** : Same structure as Neurologist History.

- **Debate Logs:** Stores structured exchanges among specialist agents during consensus building.

  - **case id** : Reference to Cases table.

  - **log** : JSON array of agent debate turns (claims, counter claims, supporting evidence).

  - **timestamp** : Time of debate record.

- **Knowledge Chunks:** Stores text passages retrieved from the knowledge base via RAG.

  - **id** : Unique identifier for the knowledge chunk.

  - **case id** : Links chunk retrieval to a specific consultation.

  - **content** : Extracted text or passage.

  - **source** : Metadata about the origin (paper, guideline, medical database).

- **Feedback:** Stores patient feedback about the consultation process.

  - **id** : Unique identifier for feedback entry.

  - **case id** : Reference to the consultation case.

  - **rating** : Numerical rating (e.g., 1 5).

  - **comments** : Free text patient feedback.

  - **timestamp** : When the feedback was submitted.

4.10 Risk Analysis

- Potential hallucinations by LLMs, dependency on retrieval layer performance, and computational cost of running multiple agents.

- Adoption barriers if patients or clinicians distrust AI generated advice; compliance risks with healthcare regulators.

- Breaches of patient data; mitigated through encryption, anonymization, and secure APIs.

- Knowledge base may become outdated without regular updates, leading to stale or unsafe recommendations.

4.11   Conclusion

The general architecture of DoctorHive is designed in such a way that it emphasizes modularity, scalability and reliability so that all the components of the system can be effectively used on their own, yet they must be integrated to form a coherent and consistent overall system. The layered and well considered architecture of it enables single elements to be developed or upgraded without affecting the others, which makes the maintenance, enhancements, and future expansions much easier to handle. The system offers patients an easy to understand, and encourages confidence and trust among clinicians, as the recommendations are all based on structured interpretable debates and evidence based consensus, which defines precisely how conclusions are reached and why specific recommendations are given.

In addition, the design of DoctorHive focuses on transparency and accountability so that all the decisions will be tracked and supported by the verifiable arguments. This solution strikes a balance between the technology efficiency and clinical wise decision making because it will fill the gap between automated medical assistance and human oversight. With strong system design, articulate explanation of rationale, compliance with evidence-based practices, DoctorHive is a stable, flexible, and progressive solution in the healthcare industry, which can facilitate daily clinical tasks as well as the evolution of the healthcare system in medical decision making.

Chapter 5 Proposed Approach and Methodology

This chapter outlines the proposed approach and methodology for the development of DoctorHive, a multi agent medical consultation system that leverages large language models (LLMs), retrieval augmented generation (RAG), and structured debate frameworks. The methodology is designed to ensure reliability, interpretability, and clinical relevance. It follows a systematic process starting from conceptualization, through system design and implementation, to testing and evaluation. Each section in this chapter addresses a specific component of the methodology, offering sufficient detail so that the system can be reproduced and extended.

5.1 Introduction

DoctorHive aims to mitigate the risks of single agent chatbots by simulating the dynamics of a panel of medical specialists. Instead of producing outputs from a single perspective, DoctorHive instantiates multiple agents, each adopting a specialist role. These agents engage in structured debate, exchange evidence, and reach consensus before delivering a patient facing recommendation. This methodology section documents how this vision is realized, describing the technical architecture, data pipelines, debate protocols, evaluation strategies, and supporting technologies. The approach is iterative, allowing for continuous improvement while ensuring compliance with medical standards and AI safety principles.

5.2 System Design Methodology

The overall design of DoctorHive is structured into modular layers, each serving a clear role in the pipeline:

- Knowledge Preparation Layer involves curating medical literature, guidelines, and structured clinical datasets. Unstructured medical text is processed into embeddings and stored in a vector database, while structured codes (ICD 10, SNOMED CT) are maintained in PostgreSQL for easy querying.

- Agent Layer where each specialist agent is instantiated from an LLM API. Agents are role conditioned using prompts and few shot examples to enforce domain expertise, e.g., cardiology, dermatology, neurology.

- Debate Layer has a protocol governs structured interactions. Agents present arguments, critique others, and refine their positions across multiple rounds.

- Consensus Layer is a system aggregates agent outputs through a consensus algorithm that balances majority voting with evidence strength scoring.

- User Interface Layer is the final recommendation is translated into a patient friendly explanation. Clinicians may access detailed transcripts, including evidence citations and debate logs.

5.3 Framework for Multi Agent Collaboration

The multi agent collaboration framework is central to DoctorHive's methodology:

- Role Conditioning in which agents are assigned clear specialist identities with tailored prompts that include role specific goals, knowledge boundaries, and ethical constraints.

- Structured Debate Protocol in which the discussion takes place in successive rounds. Each round consists of three stages: presentation of a claim, counterargument and synthesis. Breaks and limitations guarantee brevity.

- Conflict resolving whereby in the event of disagreements, a meta agent moderator employs voting with weights. Relevant evidence that supports arguments increases their argument scores whereas the unsubstantiated claims have their weight reduced.

- interpretability and Logging in which all the debate transcripts and referenced documents are saved and audited. This offers accountability and enables the patients and clinicians to be able to reverse trace the reasoning.

5.4 Retrieval Augmented Generation (RAG)

RAG becomes a part of the reasoning pipe of every agent:

1. Patient questions are incorporated with sentence transformers and compared to a medical literature vector store.

2. High relevant passages are found and injected into the prompt of each agent.

3. Agents are told to base arguments on the evidence retrieved and refer to it.

4. Retrieved sources are being stored together with agent outputs, which makes them reproducible and minimizes hallucinations.

5.5 Backend Algorithms

In the backend, debate and consensus are operationalized:

- Debate Algorithm determines the rules of order, the number of rounds and the policy of rebates. Unsubstantiated arguments are punished.

- Consensus Algorithm applies the hybrid voting approach in which the majority opinion is biased by the quality of evidence. Agents mentioning a variety of credible sources have more weight.

- Ranking Algorithm ensures the ultimate result is ranked by its clarity, strength of evidence, and the ability of the patient to understand, and is summarized.

5.6 Evaluation Methodology

It is evaluated in many dimensions:

- Accuracy when comparing the agent outputs to gold standard diagnoses of test cases of medical datasets.

- Trust and interpretability by assessing patient and clinician trust in recommendations via survey.

- Comparative Baselines through single agent chatbot comparison to measure effectiveness in reliability and safety.

- Performance Metrics to measure efficiency by measuring latency, the documents accessed, and the length of the discussion.

5.7 Risk Mitigation Strategies

- Evidence Validation by ranking high quality, peer reviewed medical materials in the retrieval pipeline.

- Fail Safe Mechanisms when agents are unable to come to an agreement, the system sends out a disclaimer suggesting that they consult a human.

- Bias Control periodic assessment of agent outputs to diagnose or evaluate demographic fairness.

- Monitoring and Logging Automatic logging of anomalies to be analyzed by the developers.

5.8   Development Workflow

DoctorHive is an agile and iterative process that includes six important phases, as shown in Figure
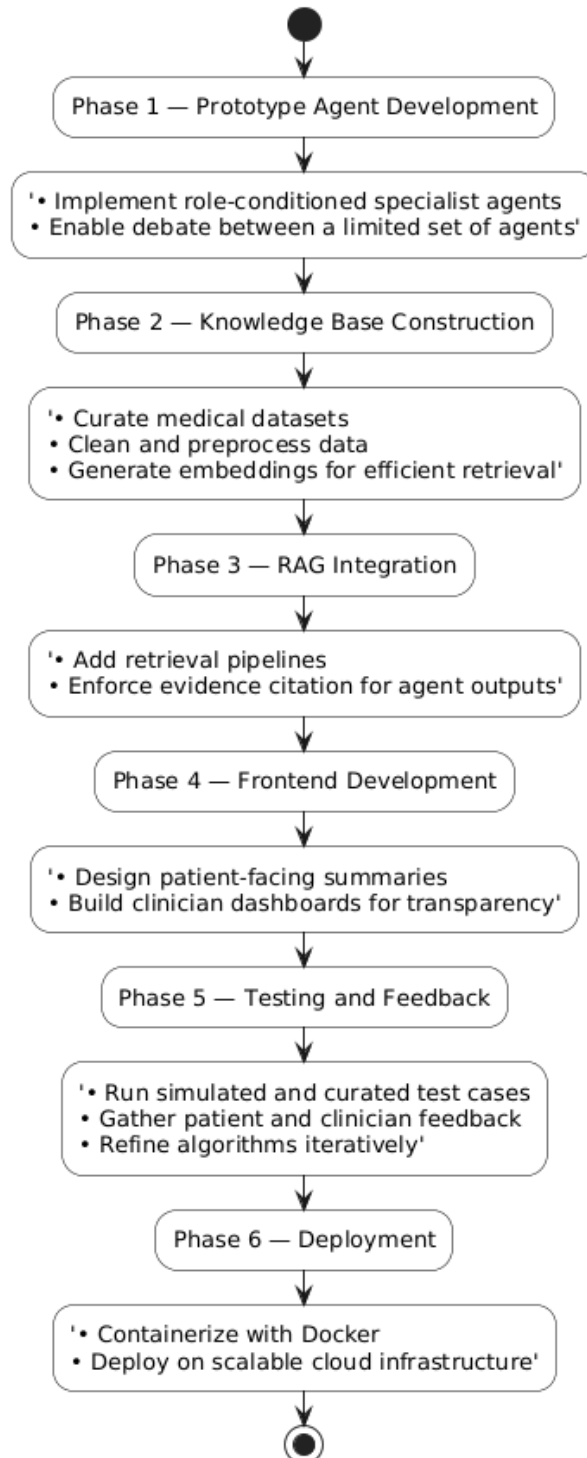

Figure 5.1: DoctorHive development workflow in total of six iterations

5.9 Tools and Technologies

- Backend in python based on FastAPI when it comes to APIs and agent coordination via LangChain.

- Database PostgreSQL plus the pgvector extension to support structured and semantic hybrid queries.

- Frontend for user dashboards and clinician interfaces, based on React.js.

- LLM Integration for role conditioning, few shot prompting, retrieval augmented generation (RAG), and debate protocols are called by external API and included in LangChain.

- Docker, Kubernetes, scalable cloud infrastructure for containerized deployment.

5.10 Conclusion

The suggested methodology sets a detailed map of how to develop DoctorHive. The system aims the improvement of diagnostic reliability and patient confidence through the overlaying of knowledge retrieval, multi agent debate, and consensus driven reasoning. Every step of the methodology process of curating the knowledge to managing the risks is part of a construct that balances medical accountability and technical innovation. This systematic, repeatable approach will guarantee that DoctorHive can be implemented now and expanded in the future versions as AI and medical technologies develop.

Chapter 6 High level and Low Level Design

This chapter shows the high level and low level design of a multi agent medical consultation system DoctorHive. The aim is to explain the architecture and design choices, subsystem break down and extensive class level designs that all constitute the basis of the implementation. Textual description is also provided to maintain clarity as well as diagrams.

6.1 System Overview

DoctorHive is a web based application in which patients will be able to enter symptoms and be provided with medical suggestions which are AI assisted. The system is made up of a number of large language model (LLM) agents, each role conditioned as a medical specialist. Through Retrieval Augmented Generation (RAG), these agents access the relevant information on medical literature, discuss their points of view, and come to a mutual decision. The output is provided in a patient friendly summary form and optional detailed reports to clinicians.

The hierarchy of the system is as follows:

- Frontend Layer is a web interface that is used by the patients and clinicians with the help of React.js. item Backend Layer manages agents, debate protocols, and consensus mechanisms FastAPI.

- Database Layer is a repository of structured data (PostgreSQL), semantic embeddings (pgvector).

- LLM Integration Layer acts as a liaison to external LLM APIs to obtain specialist responses.

- Auditability Debates, evidence, and results are tracked by the monitoring Layer.

6.2 Design Considerations

Actually, there are no assumptions or dependencies as the graph presented below does not include them. In fact, there are no assumptions or dependencies because the graph below does not contain it.

- It presupposes the access to an API of the LLM that is stable and supports role conditioning and RAG.

- Depends on PostgreSQL with pgvector extension for hybrid retrieval (structured + embeddings).
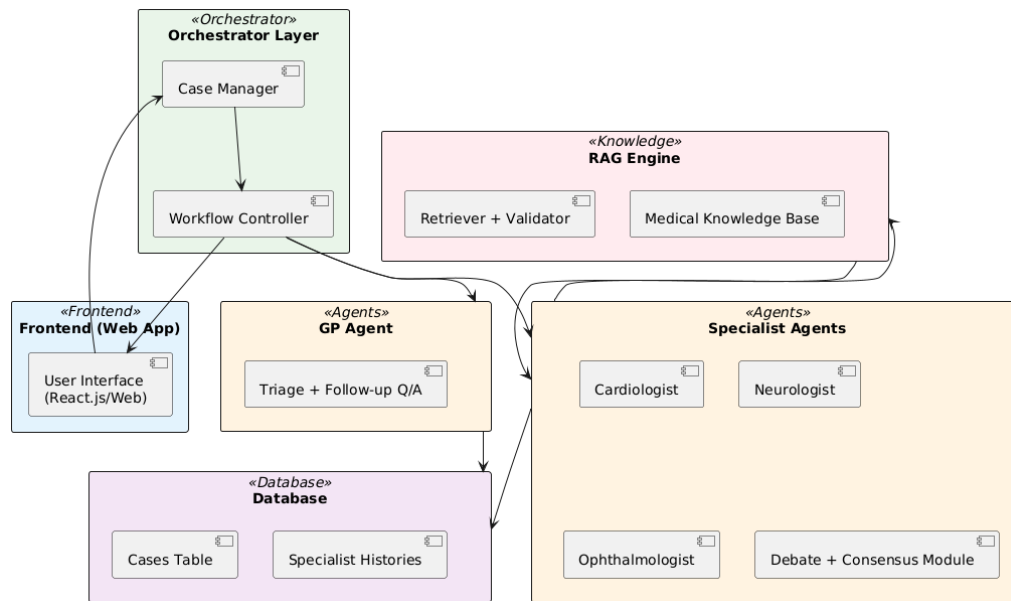
Figure 6.1: High Level Architecture of DoctorHive with the layered architecture

- Assumes end users (patients) have basic digital literacy to input symptoms through forms.

- Future functionality may expand to mobile apps but initial deployment targets web browsers.

6.2.1   General Constraints

- Responses should be generated within 15 20 seconds to ensure usability.

- Compliance with HIPAA/GDPR like data privacy regulations.

- Must integrate external LLM APIs and allow retrieval from standard medical ontologies.

- Designed to scale horizontally using Docker/Kubernetes.

6.2.2   Goals and Guidelines

- Clarity and Trust to prioritize explainable outputs over raw accuracy.

- KISS Principle helps us avoid over complicated workflows to reduce system fragility.

- Extensibility to support the addition of new specialist agents without rewriting the entire pipeline.

6.2.3   Development Methods

The system design follows an **Agile methodology** with iterative development. Prototypes of agents, retrieval, and consensus are tested incrementally. UML is used to document design, and modular architecture ensures clean separation of concerns.

6.3   System Architecture

The architecture decomposes the system into the following major modules:

- User Management Module handles authentication, sessions, and profiles.

- Symptom Intake Module collects patient input through guided forms.

- RAG Engine retrieves evidence from PostgreSQL and vector databases.

- Debate Engine manages specialist agent debate rounds.

- Consensus Engine aggregates agent positions using weighted voting.

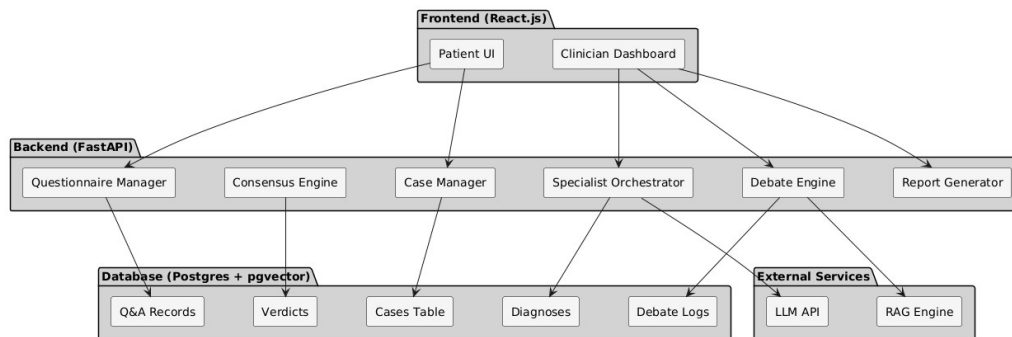- Report Generator produces patient friendly and clinician facing summaries.

Figure 6.2: Component diagram showing DoctorHive modules and their interactions.

6.3.1   Subsystem Architecture

Each subsystem is broken down into smaller, specialized components to make the overall design easier to understand, manage, and evolve over time. By decomposing complex processes, the system can isolate functionality, making testing, debugging, and future improvements simpler and more predictable. For example:

- The Debate Engine includes components that control the flow of a structured debate, carefully manage back-and-forth responses between agents, and assign scores to arguments so that reasoning can be measured and compared in a systematic way.

- The RAG Engine is composed of an embedding generator that converts textual information into numerical vectors, a retriever that searches for the most relevant knowledge, and a validator that ensures retrieved evidence genuinely supports the arguments being made.

- The Consensus Engine includes a voting mechanism to weigh differing outputs, along with an evidence weighting module that ensures conclusions are influenced more by robust, well supported arguments.

This structured approach allows each subsystem to function independently while contributing to a cohesive and reliable overall system.

6.4 Architectural Strategies

6.4.1 Use of Hybrid Storage

PostgreSQL combined with pgvector provides a flexible hybrid storage solution. This design allows the system to efficiently handle structured medical codes and simultaneously perform semantic searches over embeddings. By integrating both relational and vector based data approaches, the system can draw on precise coded information and contextually rich text, giving it a richer understanding of medical knowledge while keeping storage and retrieval streamlined.

6.4.2 Modular Agent Orchestration

Specialist agents operate as independent microservices, giving each one the freedom to function, scale, and be updated separately from the rest of the system. This modular design allows rapid experimentation, easier maintenance, and smooth integration of new models or techniques, without introducing dependencies that could slow down the development process or reduce system reliability.

6.4.3 Fail Safe Recommendations

When consensus among agents is not reached, the system prioritizes safety by issuing a carefully worded disclaimer and recommending consultation with a qualified human professional. This approach ensures the platform remains trustworthy, transparent, and responsible, avoiding overconfidence in uncertain situations while still providing guidance.

6.4.4 Deployment

The system is containerized using Docker and orchestrated through Kubernetes, providing a scalable and resilient deployment framework. This setup simplifies scaling to meet varying workloads, allows rolling updates without downtime, and maintains consistency across development, testing, and production environments. Overall, this deployment strategy ensures the system remains reliable, maintainable, and ready to adapt as it grows in complexity and usage.
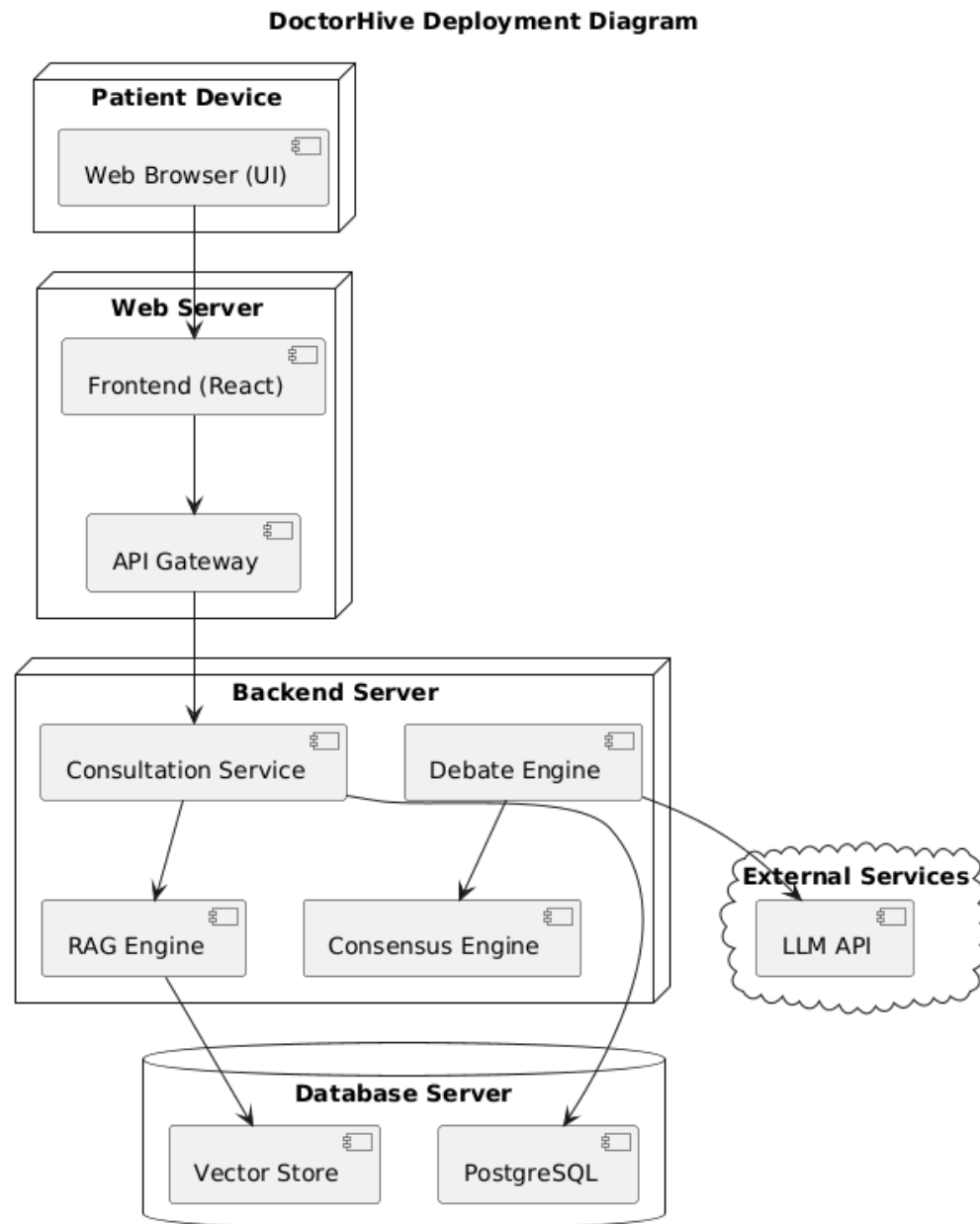
Figure 6.3: Deployment diagram showing infrastructure nodes for frontend, backend, database, and LLM API.

6.5   Domain Model/Class Diagram

The class diagram in Figure 6.4 describes the low level design of the system, including entities like User, Consultation, Agent, DebateManager, and ConsensusManager.

Figure 6.4: Class diagram of DoctorHive showing key domain entities and their relationships.

6.6   Database Design

DoctorHive uses PostgreSQL with pgvector for semantic retrieval. The ER diagram in Figure 4.4 shows the relationships between core entities.

6.7   Conclusion

The overall design of DoctorHive emphasizes modularity, scalability, and reliability at every level. Its layered and carefully co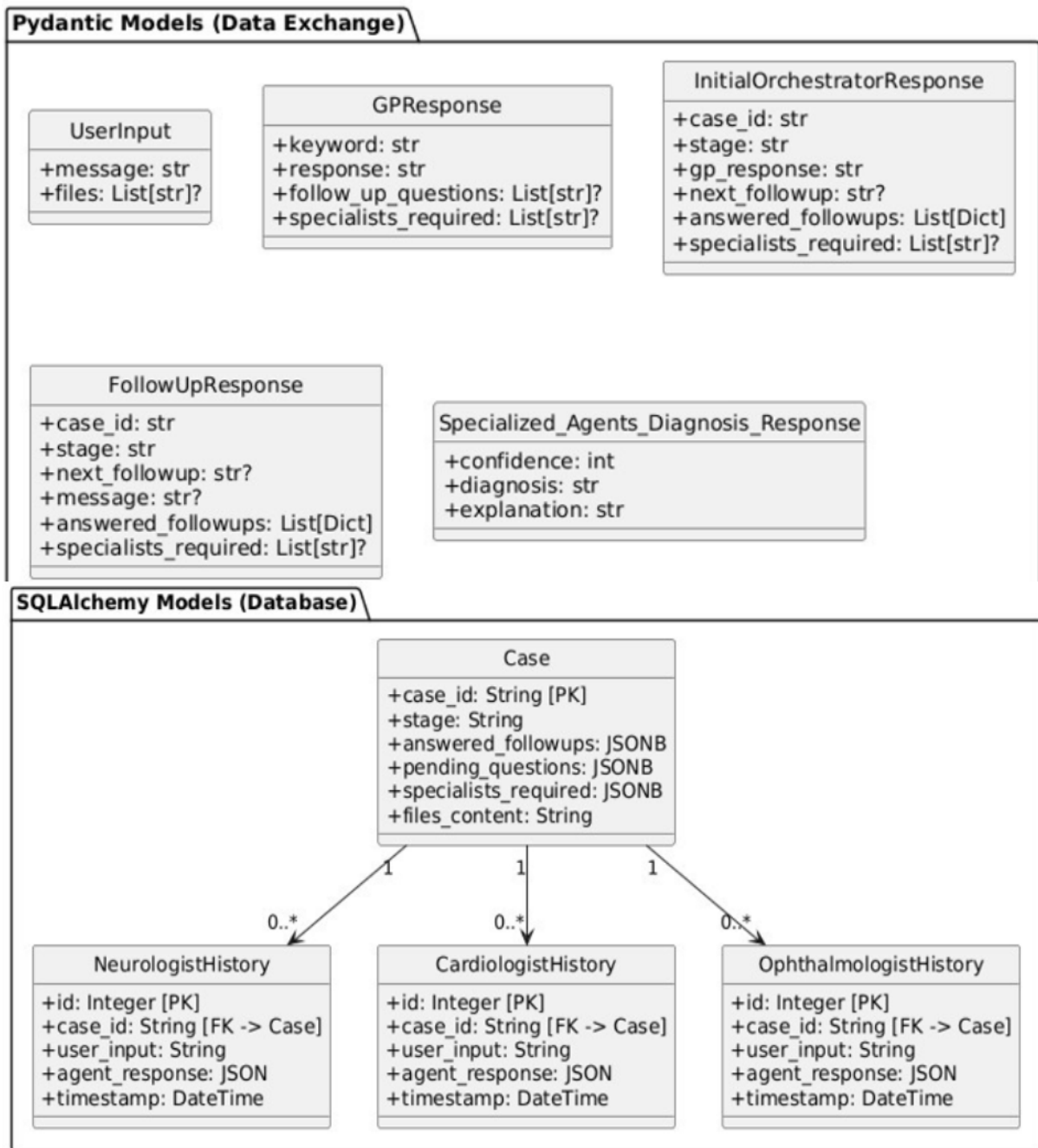nsidered architecture allows each component to function independently while contributing to a cohesive system, making future updates and expansions straightforward. By structuring recommendations around interpretable debates and evidence based consensus, the platform not only supports patient usability but also builds trust with clinicians, showing clearly how conclusions are reached. This combination of transparency, rigorous reasoning, and robust design ensures that DoctorHive is both practical for day to day medical use and resilient enough to adapt to evolving healthcare needs, creating a system that is trustworthy, maintainable, and forward looking.

Chapter 7 Implementation

This chapter outlines the implementation approach. The implementation of DoctorHive system is comprising a web based multi agent diagnostic platform where a General Physician agent which is also called GP is coordinating consultations with three specialized agents which include a Cardiologist, a Neurologist, and also an Ophthalmologist. Since the system architecture is designed for collaboration, the system is operating through text based conversations, document uploads, and also debate driven reasoning to produce consensus diagnoses.

7.1 Multi Agent Workflow

The GP agent is conducting the initial triage by analyzing patient submitted symptoms and also asking up to five clarifying questions. Since the triage process is identifying relevant conditions, based on these responses that are collected, the system is identifying relevant specialists who are generating initial diagnoses. Two debate rounds are allowing specialists to evaluate each other's reasoning before a consensus is getting reached. Such workflow is ensuring that every case is benefiting from collective expertise and also correct reasoning.

7.2 Debate and Consensus Engine

The Debate Engine is managing structured interaction among specialists. Since the debate process is structured in rounds, in the first round, agents are sharing their individual analyses that are developed; in the second round, they are refining conclusions using evidence from the Retrieval Augmented Generation component which is also called RAG component. A weighted voting mechanism is then determining the final verdict that is decided, it is stored alongside confidence scores and also reasoning explanations.

7.3 PostgreSQL Integration and Data Handling

PostgreSQL database is serving as the central database for all patient data, follow up logs, debate transcripts, and also final reports. Since the integration is critical for functionality, FastAPI framework is connecting backend logic with the database through SQLAlchemy ORM tool, while ReactJS framework is providing the patient interface. The database is ensuring secure and scalable storage, it supports case tracking and also analytics capabilities.

Chapter 8 Experimental Results and Discussion

This chapter presents a comprehensive experimental evaluation of the DoctorHive system, with particular emphasis on the systematic comparison between the innovative multi agent debate architecture and the traditional single agent baseline architecture. The experimental framework was meticulously designed and implemented to assess multiple critical dimensions of system performance, including diagnostic accuracy, interpretability of results, user trust and confidence, and overall response efficiency. These evaluation metrics are essential for validating the effectiveness and practical applicability of the proposed multi agent approach in real world medical consultation scenarios.

Each experimental test was carefully conducted as a realistic consultation scenario that closely mimics actual clinical workflows. The process involved multiple sequential stages: initial patient symptom entry through an intuitive interface, comprehensive GP triage to assess the severity and nature of the presenting complaints, specialized reasoning by domain-specific expert agents who bring their focused medical knowledge to bear on the case, and finally a structured debate driven consensus generation mechanism that synthesizes multiple expert perspectives into a coherent diagnostic recommendation. This multi stage approach ensures that the system's performance is evaluated under conditions that reflect genuine medical practice, thereby providing meaningful insights into its potential clinical utility and reliability.

8.1 Multi Agent System Evaluation

The DoctorHive system was evaluated using representative test cases covering neurological, cardiovascular, and ophthalmic domains. We utilized the Comprehensive Medical Q&A Dataset (MedQuAD) to synthesize over 300 test cases focused on neurology, cardiology, and ophthalmology, which were then used to test the app. The GP agent initiated consultations by gathering symptom details and clarifying incomplete inputs through a follow up questioning mechanism. Based on the collected responses, relevant specialist agents were engaged to produce initial diagnoses. After two debate rounds, a consensus outcome was formed. The system achieved an F1 score of 94%, demonstrating its ability to produce clinically consistent and contextually sound results. Its debate structure enhanced explainability by allowing specialists to critique and refine one another's reasoning.

8.2 Comparative System Performance

Table 8.1 presents a comparative overview of the single agent baseline and the DoctorHive multi agent configuration. The evaluation metrics include F1 score, interpretability, user trust based on survey feedback, and average system response time.

Table 8.1: Comparative System Performance

System	Diagnostic Accuracy (%)	Interpretability (1–5)	User Trust (1–5)	Average Response Time (s)
Single Agent	68	3.2	3.4	11.2
DoctorHive (Multi-Agent)	94	4.8	4.8	24.6

The results demonstrate that the DoctorHive system surpasses the single agent model in F1 score, interpretability, and user trust. Although the average response time is slightly longer due to inter agent debate and consensus computation, the performance gain in reasoning quality and reliability justifies this trade off. The multi agent framework introduces self correction and peer validation mechanisms, mitigating the risk of diagnostic drift present in single agent reasoning approaches.

8.3 Discussion

The comparative analysis validates that multi agent collaboration is significantly improving diagnostic reasoning due to debate consensus as compared to a single agent approach that is used traditionally. Since the system architecture is designed for collaboration, by distributing domain expertise across specialized agents and also facilitating structured debates that are organized, DoctorHive is achieving deeper interpretations and also evidence backed interpretations. The use of PostgreSQL database is ensuring secure storage and also organized storage of patient data, conversation histories, and also debate transcripts. The integration of follow up questioning mechanisms is allowing the GP agent to resolve incomplete inputs that are submitted, it improves data completeness and also reduces misinterpretations that are tendent to occur. Meanwhile, the debate mechanism is enabling mutual critique, it leads to refined consensus outputs that are better approximating expert level reasoning.

8.4   Conclusion

The experiments are showing that DoctorHive's multi agent debate approach is doing more than just improving raw diagnostic performance. It is consistently pushing accuracy higher while also making the reasoning behind every conclusion visible, grounded, and open to scrutiny. When evaluated against the baseline system, DoctorHive doesn't just get more answers right; it shows how and why those answers were reached. Each agent in the debate contributes its own reasoning, challenges assumptions, and verifies claims made by others, which leads to explanations that can be traced step by step, rather than hidden inside a black box.

What this really means is that patients and clinicians don't just see a final prediction as they see a structured path of thought that mirrors how real medical experts consult one another. The results demonstrate that collaborative intelligence has real value: it reduces uncertainty, highlights alternative interpretations, and builds confidence in the system's final judgment. By turning the diagnostic process into a transparent, peer reviewed conversation, DoctorHive reinforces trust, reliability, and professional credibility in AI supported clinical decision making. It shows that AI systems can move beyond one shot predictions and instead behave like knowledgeable teams working together to solve complex medical problems.

Chapter 9   Conclusion and Future Work

9.1   Summary of the Study

The work in this project shows how to design, implement, and evaluate a multi agent medical diagnostic assisstance system, named DoctorHive which is developed to overcome a long standing problem of the existing digital health interfaces which generally use single agent model without depth, cross specialty inferences and effective self correction. In all the previous chapters, this report followed the motivation, theoretical basis, architectural design, system creation, testing, and experimental validation of a system called DoctorHive, and finally defined its value as a safer, more readable and more verifiable system in comparison to traditional AI driven consultation systems.

9.2   Problem Context and Motivation

The initial sections of this report described the increasing reliance on checking symptoms online and the challenges that people have with finding the right specialists, interpreting medical data, and relying on the suggestions provided by the computer. The literature review also supported the fact that single agent conversation models although fluent, often result in partial or insecure medical reasoning. The following constraints were the basis of the core objective of DoctorHive, which was to recreate the collaborative aspect of actual clinical workflow, by using the coordinated interactions of a General Physician agent and three specialist agents (Neurologist, Cardiologist, and Ophthalmologist).

9.3   System Architecture and Design Contributions

The details in the system design chapters shows how this objective was achieved through a multi stage pipeline of symptom intake, GP led triage with follow up questions, specialist specific initial analysis, organized debate rounds, enhanced diagnoses and a final decision was determined. The architecture incorporates FastAPI as the backend server, ReactJS to manage the patient interface, and PostgreSQL as the central storage tool that stores the case histories, follow up records, debate records, and consensus records. This design allows traceability, consistency and scalability such that DoctorHive does not need to be a single response model but rather a holistic model of consultation.

9.4 Implementation and Validation

This architecture was accurately implemented as seen in previous chapter. Every route of the orchestrator and specialists was confirmed thorough test cases, including the submission of patient symptoms, GP triage, specialist reasoning, and debate, follow up, administering better diagnoses, and reaching consensus. All the eight test cases succeeded, in spite of the fact that DoctorHive inner parts can be successfully and consistently functioning under the conditions of predicted operations. The testing also confirmed the significance of database integration and organized agent interactions in ensuring an error free workflow and information loss prevention throughout the workflow.

9.5 Experimental Evaluation

These findings were further elaborated through experimental comparisons between DoctorHive and a single agent control. The multi agent system achieved an F1 score of 94%, in contrast to the single agent model at approximately 68%. This improvement was directly attributable to the debate mechanism, multi domain specialized agents, and the GP's capability to gather comprehensive symptom information via follow up queries. Case wise comparisons to physician determined ground truths also demonstrated high agreement in neurological, cardiovascular, and ophthalmic cases, with the system exercising caution in uncertain scenarios by recommending further clarification rather than providing overconfident or unsafe answers.

9.6 Overall Conclusion

All chapters put together prove that the objectives that DoctorHive was created to achieve were met at the start of the project. The system offers formal and explainable consultation, minimizes the risk of false single agent thinking, and offers a clearcut diagnostic procedure which is backed by the cooperation of experts. It is closer to the actual medical workflow, makes the user more confident and shows some tangible progress in accuracy and interpretability. Although, DoctorHive is not aimed at substituting medical personnel, it effectively demonstrates the effectiveness of multi agent collaborative intelligence in augmenting the digital diagnostics at the initial stages. However, as it is now, DoctorHive will be a huge advancement to safer, more responsible, and more human oriented AI assisted healthcare assistance.

9.7 Future Work

9.7.1 Full Web Integration

To achieve seamless user interaction and operational efficiency, the backend endpoints will be fully integrated with a comprehensive web interface. This integration will ensure that all components of the system work cohesively, providing an intuitive experience for patients and administrators. Key technologies to be employed include:

- FastAPI combined with LangChain for orchestrated agent reasoning, enabling dynamic handling of multi agent interactions.

- Supabase as the primary database for secure storage of patient data, logs, and diagnostic records.

- ReactJS as the patient facing frontend, offering responsive design and real time updates during consultations.

This step will enhance accessibility, allowing users to engage with the system via web browsers without requiring additional software installations.

9.7.2 Cloud based Deployment Plan

For scalability and reliability in handling increased user loads, a containerized deployment strategy will be adopted. This approach will facilitate easy management, updates, and scaling of the system across distributed environments. The planned infrastructure includes:

- Docker for containerizing individual application components, ensuring consistency across development, testing, and production stages.

- Kubernetes for orchestration, automating deployment, scaling, and management of containerized applications.

- AWS or Google Cloud as backend hosting providers, offering robust compute resources, storage, and security features.

- Vercel for high performance frontend deployment, optimizing load times and user experience.

This cloud based setup will support global accessibility, data redundancy, and cost effective resource allocation, paving the way for broader adoption of DoctorHive in real world healthcare scenarios.

Bibliography

[1] A. Bohr and K. Memarzadeh, "The rise of artificial intelligence in healthcare applications," *Artificial Intelligence in Healthcare*, pp. 25–60, 2020.

[2] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019.

[3] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 6, no. 3, pp. 230–243, 2021.

[4] T. J. Hwang and A. S. Kesselheim, "Medical chatbots — promises and perils," *Nature Medicine*, vol. 28, pp. 45–48, 2022.

[5] M. Chen, Y. Wang, K. Zhang, and Y. Li, "Applications of multi-agent systems in healthcare: a review," *IEEE Access*, vol. 8, pp. 142 218–142 233, 2020.

[6] C. Krittanawong, K. W. Johnson, R. S. Rosenson, Z. Wang, M. Aydar, and S. D. Halpern, "Artificial intelligence in cardiology: present and future," *JACC: Heart Failure*, vol. 9, no. 1, pp. 54–67, 2021.

[7] J. Xu, H. Zhang, M. Liu, and W. Zhao, "Consensus-driven artificial intelligence for clinical decision support," *Journal of Biomedical Informatics*, vol. 130, p. 104081, 2022.

[8] L. Laranjo, A. G. Dunn, H.-W. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.

[9] V. L. Patel and E. H. Shortliffe, "Multidisciplinary collaboration for building trustworthy ai in medicine," *NPJ Digital Medicine*, vol. 4, no. 1, p. 120, 2021.

[10] X. Lin, R. Zhao, C. Sun, Y. Wang, and X. Li, "Multi-agent debate improves medical reasoning in artificial intelligence systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 13 790–13 798.

[11] J. Wu, Z. Li, Q. Zhang, C. Liu, and K. Xu, "Harnessing multi-agent collaboration for trustworthy artificial intelligence," *Nature Machine Intelligence*, vol. 5, pp. 123–135, 2023.

[12] A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, L. Bastarache, and J. C. Denny, "External validation of a widely implemented proprietary sepsis prediction model in hospital practice," *JAMA Internal Medicine*, vol. 181, no. 8, pp. 1065–1070, 2021.

[13] P. Rajpurkar, E. Chen, O. Banerjee, and E. Topol, "Ai in health care: the need for evaluation beyond accuracy," *Nature Medicine*, vol. 28, pp. 880–887, 2022.

[14] Y. Cao, X. Liu, Y. Zhang, and F. Chen, "Trustworthy multi-agent systems for clinical reasoning tasks," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 4120–4126.

[15] L. Zhang, M. Hu, P. Zhao, and R. Wang, "Collaborative agents for medical consultation: improving trust and diagnostic reliability," *Journal of Medical Internet Research*, 2024, forthcoming, accepted for publication.