CM7070 Final Project, University of London

Preliminary Report

Project Template: Orchestrating AI Models to Achieve a Goal

AI GP Doctor: Multi-Modal Orchestration System for Medical Diagnosis

Table of Contents

Chapter 1: Introduction: The AI GP Doctor

The modern General Practitioner (GP) stands at the confluence of an ever-expanding universe of medical knowledge and a rising tide of patient data. The traditional diagnostic paradigm, reliant on a clinician's personal experience and training, faces unprecedented strain. This project, titled "AI GP Doctor," is born from this challenge. It is a preliminary experiment conceived not to replace the clinician, but to augment their perspective, setting an example for a future where intelligent systems act as capable collaborators in patient care. This report outlines the preliminary draft for this project, following the Orchestrating AI Models project template.

1.1 The Modern Clinician's Dilemma: A Rationale for Augmentation

A GP's diagnostic process is an act of complex, multimodal synthesis. It begins with the patient's narrative, often delivered through spoken language, which must be interpreted for clinical meaning. This is combined with objective data from various sources: structured lab results, unstructured notes from previous encounters, and increasingly, medical images. As the proposal for this project identifies, clinicians must constantly juggle voice, text, and images to form a coherent diagnostic hypothesis. This cognitive load is immense. Doctors' individual experience, while invaluable, is inherently limited.[3] They cannot be expected to recall every rare disease presentation or the latest findings from genomic research for every patient.

The healthcare landscape is simultaneously witnessing a proliferation of powerful but narrow Artificial Intelligence (AI) tools. One AI can expertly detect pneumonia from a chest X-ray, another can transcribe spoken medical notes with high fidelity, and a third can identify cancerous lesions in a dermatological image. However, these tools operate in a standalone system. A GP cannot practically deploy dozens of individual models for a single patient consultation. This fragmentation of AI capability reflects the fragmentation of patient data itself, leaving the ultimate burden of synthesis on the human clinician. The result is a missed opportunity to leverage the full potential of AI to augment human expertise, potentially leading to diagnostic delays or errors. This project is motivated by the conviction that the next frontier in clinical AI lies not in creating more niche models, but in intelligently orchestrating existing ones.

1.2 Project Concept: The AI GP Doctor as an Orchestration System

The AI GP Doctor is envisioned as a multi-modal AI orchestration system designed to simulate the integrated reasoning process of a virtual GP. The core technical concept is to build a prototype system that employs a **Mixture of Experts (MoE)** architecture.[5] In this paradigm, a high-level orchestrator or gating network manages a suite of specialized AI models, the Expert. Each expert is responsible for a specific task, such as speech-to-text conversion, natural language understanding of clinical text, or medical image analysis.

The system's primary function is to receive multimodal inputs representative of a real clinical encounter (e.g., a patient's spoken symptoms, a photograph of a skin condition) and intelligently route this information to the appropriate experts. The outputs from these experts are then synthesized to produce a coherent, context-aware insight or preliminary diagnostic suggestion. This approach directly addresses the limitation of narrowly focused AI systems, creating a single, unified interface that can reason across different data types, much like a human clinician.

1.3 Research Questions and Scope

This project is an exploratory investigation into the feasibility of such a system. The focus is on system integration, not the creation of novel AI models from scratch. We will leverage powerful, pre-trained foundation models where possible. The central research questions guiding this project are:

1. Is it technically feasible to build a stable, modular orchestration system that integrates pre-trained speech, language, and vision models for a simulated clinical use case?
2. Can a Mixture of Experts (MoE) architecture, managed by a central orchestrator, provide a more coherent and contextually rich output than individual models working in isolation?
3. What are the primary challenges and limitations (e.g., latency, inter-model communication, factual accuracy) in developing such a user-facing orchestrated system?

The scope of this preliminary work is to develop a functional software prototype that demonstrates the core workflow, evaluate its performance on a curated dataset of mock clinical cases, and provide a detailed analysis of its strengths and weaknesses. This will lay the groundwork for a more robust system that could, in the future, be developed for real-world clinical evaluation.

Chapter 2: Literature Review

The concept of an AI GP Doctor rests on the intersection of several rapidly advancing fields in artificial intelligence and medical informatics. This review surveys the existing literature on AI orchestration, multimodal data fusion, the application of large-scale AI models in medicine, and the critical role of explainability.

2.1 Orchestration and the Mixture of Experts (MoE) in Medicine

The term AI Orchestration describes the automated coordination and management of multiple AI systems to achieve a high-level goal.[1] While prevalent in enterprise IT for managing complex workflows, its application in medicine is an emerging and critical field. The need arises directly from the proliferation of FDA-approved, single-task AI algorithms. Aidoc, a leader in medical AI, emphasizes that intelligent orchestration is necessary to maximize algorithmic yield by ensuring that every relevant AI is applied to every appropriate patient study, thereby enabling the detection of unexpected or incidental findings.[2, 13] However, a critical review of this commercial work reveals significant limitation**s.** These platforms

are often proprietary, black box systems focused on high-margin, high-volume radiological workflows. Their scalability and cost-effectiveness in the far more diverse and less structured domain of general practice remain unproven. Furthermore, they are typically closed ecosystems, limiting the ability of healthcare systems to integrate third-party or in-house developed models, highlighting the need for more open, flexible architectural frameworks.

The **Mixture of Experts (MoE)** paradigm provides a formal architectural blueprint for such systems. First proposed by Jacobs et al. (1991), MoE models are a form of ensemble learning where multiple expert networks are trained on subsets of a problem space, and a gating network learns to select the best expert(s) for a given input. In recent years, MoE has seen a resurgence with large language models, where it is used to activate only parts of a massive model, improving computational efficiency. In our context, we conceptualize the MoE framework differently: not as a way to partition a single model, but as a way to manage a heterogeneous collection of independent, specialized models. Akira AI, for example, proposes a similar conceptual architecture with a *Master Orchestrator Agent* managing *Domain Specialized Agents*.[12] Our project aligns with this vision, where the orchestrator acts as an intelligent router, a concept critical to making a complex, multi-expert system computationally tractable and logically sound.

2.2 Multimodal Data Fusion in Clinical Diagnostics

The diagnostic process is inherently multimodal. A clinician's final assessment is rarely based on a single piece of information. The literature reflects this reality, with a growing body of research focused on fusing data from different sources to improve diagnostic accuracy. A comprehensive review by Miotto et al. highlights that multimodal AI models can capture complementary information from different data types, leading to better patient stratification and outcome prediction than unimodal approaches.[4] Their work categorizes fusion strategies into early, late, and hybrid methods. Early fusion involves concatenating feature vectors from different modalities before feeding them into a single model, while late fusion combines the outputs of separate models.
The challenge, as identified in our project proposal, is that these fusion techniques are often bespoke and applied to specific disease contexts, such as combining imaging and genomic data in oncology.[12] There is less research on building generalized, flexible orchestration systems that can perform this fusion dynamically for the wide range of conditions seen in a primary care setting. This project aims to address this gap by focusing on the integration layer itself, using an LLM as a sophisticated late-fusion mechanism to synthesize the outputs of different specialist models into a coherent narrative.

2.3 The Role of Large Language and Vision Models in Healthcare

Large Language Models (LLMs) and Vision-Language Models (VLMs) are crucial enablers for the AI GP Doctor concept. Medically-tuned LLMs like Med-PaLM have shown expert-level performance on certain tasks,[7] making them ideal candidates for a central reasoning engine.[6] VLMs provide the essential link between visual data and textual reasoning. However, their application is not without serious risks that the literature is only beginning to grapple with. The problem of *hallucination*[8] is well-documented, but a

deeper challenge is the risk of automation bias, where clinicians may over-trust a plausible-sounding but incorrect AI-generated summary. The process of clinically validating the outputs of these generative models in real-time is a major unsolved problem. A system that cannot make its own uncertainty transparent could be more dangerous than no system at all.

The emergence of large-scale, pre-trained foundation models has been a watershed moment for AI. In medicine, this is most evident with Large Language Models (LLMs) and, more recently, Vision-Language Models (VLMs).

- **Large Language Models (LLMs):** Models like OpenAI's GPT series and Google's PaLM have demonstrated remarkable capabilities. In medicine, specialized variants such as Med-PaLM and GatorTron, which are pre-trained on vast corpora of biomedical literature and clinical notes, have shown expert-level performance on medical licensing exams and in answering clinical questions.[7] Their ability to understand and generate human-like text makes them ideal candidates for tasks like summarizing clinical histories, simplifying complex reports for patients, or, in our case, acting as the central reasoning engine in an orchestration system.[6] However, their use is fraught with risk. The phenomenon of *hallucination* generating factually incorrect information is a critical safety concern that must be addressed.[8]
- **Vision-Language Models (VLMs):** VLMs, such as CLIP and its clinical derivatives, are trained to connect images and text. They can perform tasks like visual question answering ("Is there evidence of a fracture in this X-ray?") or generating a textual report from a medical image. This technology is crucial for our "AI GP Doctor" as it provides the essential link between the visual data a patient might present (e.g., a photo of a rash) and the text-based reasoning of the central LLM.

2.4 The Imperative of Explainability and Trust in Clinical AI

For any AI system to be adopted by clinicians, it must be trustworthy. A "black box" that provides a diagnosis with no justification is unlikely to be used, and could even be dangerous. The field of Explainable AI (XAI) aims to address this by making AI decisions more interpretable. Techniques like Grad-CAM can produce heatmaps that highlight which parts of an image were most important for a model's decision, while LIME (Local Interpretable Model-agnostic Explanations) can explain individual predictions by approximating the complex model with a simpler one locally, but a critique of current XAI methods is that they are often post-hoc and may not fully represent the model's true internal logic. An explanation can itself be misleading. Therefore, In the context of the AI GP Doctor, providing such explanations is not an optional feature; it is a core requirement for building clinician trust and enabling them to validate (or reject) the system's suggestions. The evaluation strategy must therefore include metrics for the quality and utility of these explanations.

Chapter 3: System Design

Based on the project's goals and the findings from the literature review, this chapter details the proposed system architecture for the AI GP Doctor. The design is a direct implementation of the Mixture of Experts (MoE) paradigm [5], prioritizing modularity, intelligent routing, and the advanced synthesis of outputs from specialized AI models. This focus on orchestrating a suite of pre-trained models is a key strategy for building complex, multimodal AI systems in healthcare [1, 2, 13].

3.1 Architectural Overview

The system is designed as a modular, multi-stage pipeline that processes multimodal clinical data to generate a synthesized diagnostic summary. The architecture, as illustrated in the project proposal diagram, comprises five core stages: an input processor, a gating mechanism, a pool of diverse AI healthcare experts, a diagnostic integrator, and an output generator. This structure allows for a sophisticated workflow that mirrors clinical reasoning more closely than a single, monolithic model.

High-Level Data Flow:

1. **Input:** The system accepts multiple data types through a user-facing interface: patient audio recordings, medical images (e.g., chest X-rays), and supplementary text (e.g., clinician's questions, electronic health record snippets).
2. **Preprocessing:** The **Input Processor** immediately transcribes any audio input into clean text.
3. **Routing:** The **Gating Mechanism/Router** analyzes the combination of available inputs (transcribed text, images, specific user queries) and dynamically determines which specialized expert models to engage for analysis.
4. **Expert Processing:** The designated **AI Healthcare Experts**—including models for extractive analysis (ClinicalBERT), generative reasoning (Clinical LLM), and vision-language tasks (LLaVA-Med, BioMedCLIP)—process their assigned data in parallel.
5. **Synthesis:** The **Diagnostic Integrator/Synthesizer** receives the structured outputs from all engaged experts and performs multimodal fusion [4, 11] to construct a single, coherent clinical assessment.
6. **Output:** The **Output Generator** formats this synthesized assessment into a clear and clinically useful report for the end-user.

3.2 Component Deep-Dive

Each component in the architecture is a distinct module with a specialized function.

1. Input Processor (Speech-to-Text Conversion)

- **Model: OpenAI Whisper**.
- **Function:** This initial component's sole responsibility is to perform highly accurate automatic

speech recognition (ASR). It converts the patient's spoken narrative into a text transcript, which serves as a foundational data source for the subsequent text-based analysis by other experts. Its robustness to noise and accents is critical for clinical fidelity.

2. Gating Mechanism/Router (Intelligent Task Allocation)

- **Function:** This component acts as the system's intelligent control unit. It examines the nature of the user's query and the provided data to route tasks to the most appropriate expert(s). For the prototype, this will be a rule-based Python script. For instance:
    - IF query is extractive (e.g., "What symptoms?") AND text_input EXISTS THEN engage(ClinicalBERT).
    - IF query is interpretive (e.g., "Suggest diagnosis") AND text_input EXISTS THEN engage(Clinical LLM).
    - IF image_input EXISTS AND "describe" in query THEN engage(LLaVA-Med).
    - IF image_input EXISTS AND "does this show" in query THEN engage(BioMedCLIP).
    - A more advanced implementation could use a small, trained classifier to learn these routing patterns automatically.

3. AI Healthcare Experts (The Specialists)

This pool contains a diverse set of pre-trained, domain-specific models.

- **ClinicalBERT (The Extractive Expert):**
    - **Function:** This model specializes in understanding clinical text to perform tasks like extractive Question Answering and Named Entity Recognition (NER) [18]. It is ideal for pulling specific, factual data points from the patient transcript, such as symptoms, medications, dosages, and timelines.
    - **Example Output:** A JSON object: {"extracted_entities": {"symptoms": ["dry cough", "fever"], "duration": "3 days"}}.
- **Clinical LLM (The Generative Reasoning Expert):**
  Function: This is a large language model fine-tuned on medical texts and dialogues, such as a derivative of Med-PaLM or GPT-4 trained on medical data [7, 8]. It handles tasks requiring clinical reasoning, such as generating differential diagnoses, summarizing patient cases, or answering complex "what if" questions from the clinician.
    - **Example Output:** A JSON object: {"synthesis": "Based on the stated symptoms of a multi-day fever and dry cough, potential diagnoses include viral bronchitis or community-acquired pneumonia. Further investigation is recommended."}.
- **LLaVA-Med (The Vision-Language Explainer):**
    - **Function:** A Large Language and Vision Assistant for Medicine, this model excels at visual question answering (VQA) and generating rich, descriptive captions for medical

images [6]. It can describe findings in an image in natural language.
    - **Example Output:** A JSON object: {"image_description": "The provided chest X-ray shows evidence of opacity in the left lower lobe, consistent with an inflammatory process such as pneumonia."}.
- **BioMedCLIP (The Vision-Language Classifier):**
    - **Function:** This model is optimized for zero-shot classification by calculating the similarity between an image and a text prompt. It can be used to confirm or deny the presence of specific findings mentioned in the text or query.
    - **Example Output:** A JSON object: {"classification_results": [{"finding": "pneumonia", "confidence": 0.91}, {"finding": "pleural effusion", "confidence": 0.15}]}.

4. Diagnostic Integrator/Synthesizer (Multimodal Fusion)

- **Function:** This is the most critical component for achieving true multimodal intelligence. It receives the structured JSON outputs from all previously engaged experts and performs a sophisticated **late-fusion** step. Its job is to synthesize these disparate pieces of information into a single, cohesive narrative. It will be implemented using a powerful LLM (e.g., GPT-4 or a future equivalent) guided by a meticulously engineered prompt that instructs it to:
    1. Combine findings from text and vision experts.
    2. Highlight areas of corroboration (e.g., cough mentioned in text, opacity seen in image).
    3. Identify and report on any inconsistencies (e.g., text mentions chest pain, but X-ray is clear).
    4. Generate a final summary with an overall confidence assessment. This synthesis is vital for moving beyond a simple collection of findings to a holistic diagnostic suggestion [12, 17].

5. Output Generator

- **Function:** This final module takes the rich, synthesized JSON object from the integrator and formats it into a human-readable report for the user interface. It ensures the final output is well-structured, clear, and clinically actionable.

3.3 Data Flow and Integration Strategy

The modular components will be integrated using a microservices architecture, communicating via a REST API framework. A central orchestration script will manage the API calls. This strategy provides significant advantages:

- **Modularity:** Each expert model can be containerized and managed independently. An update to BioMedCLIP, for example, will not require changes to the Clinical LLM service.
- **Scalability:** If image processing becomes a bottleneck, the vision expert services can be scaled up with more resources without affecting the other components.

- **Flexibility:** This design makes it straightforward to add new experts in the future, such as a model for analyzing structured lab results (e.g., blood tests) or genomic data.

Data will be passed between modules in a standardized JSON format to ensure interoperability. The Diagnostic Integrator is designed specifically to parse a list of these JSON objects from the various experts and use their contents to construct its final, comprehensive summary, which is then passed to the Output Generator.

Chapter 4: Feature Prototype: AI GP Doctor Multimodal Interface

To validate the core technical feasibility of multimodal AI orchestration in clinical settings, a comprehensive functional prototype was developed and rigorously evaluated. This prototype, designated as the "AI GP Doctor," serves as a foundational proof-of-concept for integrating speech recognition, clinical text analysis, and medical image processing within a unified computational framework, thereby addressing the critical gap between theoretical multimodal AI architectures and practical clinical implementation.

4.1 Prototype Conceptual Framework and Theoretical Foundation

The feature prototype represents a methodical implementation of the proposed Mixture of Experts (MoE) architecture, utilizing readily available pre-trained models to demonstrate early-stage feasibility [1, 5]. The fundamental objective centers on simulating authentic primary care consultation workflows through the systematic processing of three distinct input modalities that accurately mirror real-world clinical encounters. This approach is theoretically grounded in the well-established clinical principle that diagnostic accuracy emerges from the synthesis of multiple information streams rather than reliance on isolated data points [4, 11].

The prototype architecture specifically targets three critical input modalities that form the cornerstone of contemporary clinical practice. Audio input processing addresses the fundamental challenge of capturing and interpreting patient voice recordings, which constitute the primary source of subjective clinical information during consultations. The accurate transcription and interpretation of these narrative accounts represents the initial phase of the diagnostic process, requiring sophisticated natural language processing capabilities tailored to medical terminology and clinical context. Text processing functionality builds upon the transcribed audio data, employing advanced clinical question-answering algorithms to systematically interpret and structure patient symptom presentations. This computational approach mirrors the cognitive processes employed by physicians when identifying and prioritizing key clinical information from patient narratives, requiring deep understanding of medical semantics and clinical reasoning patterns.

Medical imaging integration represents the third critical modality, encompassing basic diagnostic analysis

of chest X-rays as a representative example of radiological data integration within the overall clinical assessment framework [6, 10]. This component addresses the essential requirement for visual diagnostic information synthesis, acknowledging that modern clinical practice increasingly relies on the integration of multiple diagnostic modalities to achieve optimal patient outcomes.

The prototype serves as a technical validation platform for the modular orchestration approach detailed comprehensively in Chapter 3, demonstrating how specialized AI models can function cooperatively under centralized coordination systems. This orchestration concept proves particularly critical in healthcare applications, where the complexity of integrating diverse AI solutions requires sophisticated management to ensure seamless interoperability and clinical utility [2, 13]. While necessarily simplified compared to the complete system architecture outlined in the theoretical framework, the prototype successfully demonstrates core workflows involving multimodal data processing and integration, establishing a robust foundation for more comprehensive system development.

4.2 Technical Implementation and System Architecture

The prototype implementation employs Python programming language, leveraging the Gradio framework to provide an accessible web-based interface specifically designed for clinical simulation purposes. The system architecture adheres to modular design principles, integrating three distinct pre-trained models, each optimized for specific modality processing requirements while maintaining seamless interoperability within the unified framework.

The speech-to-text processing module utilizes OpenAI's Whisper-base model, accessed through the Hugging Face Transformers pipeline infrastructure. This component assumes responsibility for the critical initial phase of converting uploaded audio files containing patient symptom descriptions into accurate textual transcripts. The implementation achieves integration through a carefully configured automatic speech recognition pipeline, specifically optimized for clinical audio processing requirements. The Whisper model selection reflects extensive evaluation of available speech recognition technologies, with particular emphasis on accuracy in medical terminology recognition and robustness in various acoustic environments typical of clinical settings.

Clinical question-answering capabilities are implemented through Bio_ClinicalBERT, a specialized language model fine-tuned specifically on clinical text datasets by Emily Alsentzer [16]. This model selection represents a strategic decision emphasizing the critical importance of domain-specific training for achieving optimal accuracy in medical applications [18]. The clinical question-answering module performs sophisticated extractive analysis to interpret and extract specific symptoms and relevant clinical information from transcribed patient narratives. Implementation involves custom pipeline development utilizing AutoModelForQuestionAnswering and AutoTokenizer components from the Hugging Face library ecosystem. The module processes user-provided clinical questions alongside transcribed audio context, generating highly relevant answer spans accompanied by confidence scores that enable clinical

decision-making assessment.

The medical image processing component currently exists as a structured placeholder function returning predefined diagnostic examples, representing a critical area for future development. The planned integration roadmap includes implementation of specialized chest X-ray classification models such as CheXzero or equivalent models from the TorchXRayVision library. These models undergo training on extensive medical image datasets, including the comprehensive MIMIC-CXR dataset [15], enabling identification of diverse pathological conditions. The ultimate functionality of this module encompasses comprehensive medical image processing with diagnostic insight generation, facilitating seamless integration with other data modalities within the unified clinical assessment framework.

4.3 User Interface Design and Interaction Paradigm

The user interface, built with Gradio, prioritizes simplicity while providing three input channels simulating clinical encounters. The interface supports audio uploads (MP3/WAV), accommodating diverse recording equipment typical in clinical environments. Symptom questioning functionality enables clinicians to pose specific inquiries to transcribed narratives, supporting symptom identification, severity assessment, and temporal analysis.

Medical image uploads accommodate standard formats (JPEG/PNG) with built-in validation for appropriate file types and sizes. The system presents processed information through three output channels: transcribed audio results, symptom analysis, and diagnostic interpretations. This parallel presentation enables clinicians to rapidly assess multiple information streams, supporting efficient clinical decision-making.

4.4 Evaluation Methodology and Performance Assessment

The prototype evaluation emphasized technical functionality validation rather than comprehensive clinical accuracy measurement, reflecting the early development stage. Testing incorporated synthetic data and publicly available clinical datasets while maintaining ethical standards for patient data protection.

Speech recognition evaluation demonstrated Whisper-base's effectiveness in transcribing clear audio with high accuracy, showing robustness across varying speaking speeds and tolerance for background noise. The system maintained excellent accuracy for medical terminology, addressing a fundamental clinical requirement. Clinical text processing revealed Bio_ClinicalBERT's effectiveness in extracting symptom information from well-structured questions, validating the strategic use of domain-specific models over general-purpose alternatives [18].

System integration testing confirmed successful simultaneous processing of multiple modalities without computational conflicts, while the modular architecture facilitated independent testing and debugging of individual components. The Gradio interface demonstrated consistent reliability across diverse input

combinations, supporting the platform's suitability for clinical simulation applications.

4.5 Discussion of Findings and Future Development

The prototype evaluation yielded significant insights validating core technical assumptions while identifying critical areas requiring further development. The successful orchestration of pre-trained models for multimodal clinical data processing confirms fundamental technical feasibility of the AI GP Doctor concept [1]. Model selection validation, particularly the superior performance of Bio_ClinicalBERT for medical text analysis, supports the strategic emphasis on domain-specific models for clinical applications.

Current limitations include the placeholder status of image processing functionality, representing the highest priority for future development phases. Limited clinical validation based on synthetic data necessitates comprehensive assessment by medical professionals using authentic clinical scenarios to ensure safety and efficacy [3, 12]. The system's current dependency on well-formulated questions presents potential workflow challenges, as natural clinical inquiry often involves conversational and less structured communication patterns.

Future development priorities encompass functional medical image analysis integration, multimodal fusion capabilities enabling cross-modal insight synthesis, comprehensive clinical evaluation with medical professional engagement, user experience enhancement based on clinical feedback, and advanced language model integration for natural conversational query handling. The AI GP Doctor prototype, despite current limitations, establishes a solid foundation demonstrating core technical feasibility while providing a clear developmental pathway toward a fully functional multimodal clinical AI system with significant potential for enhancing diagnostic accuracy and efficiency in primary care settings.

Chapter 5: Project Workplan and Timeline

Phase 1: Foundation & Setup (Weeks 1-3)

Week 1: Project Setup & Environment

- Set up development environment and version control
- Install required frameworks (Python, Gradio, HuggingFace)
- Create project documentation structure

Week 2: Core Architecture Design

- Implement modular microservices architecture
- Set up REST API framework for inter-component communication
- Create basic logging and error handling systems

Week 3: Input Processing Module

- Integrate OpenAI Whisper for speech-to-text conversion
- Implement audio file upload and preprocessing
- Test speech recognition accuracy with medical terminology
- Create input validation and format standardization

Phase 2: Expert Models Integration (Weeks 4-7)

Week 4: Clinical Text Processing

- Integrate Bio_ClinicalBERT for extractive Q&A
- Implement Named Entity Recognition for symptoms/medications
- Create clinical question-answering pipeline
- Test with synthetic clinical text data

Week 5: Clinical Reasoning Module

- Integrate Clinical LLM (Med-PaLM variant or GPT-4)
- Implement differential diagnosis generation
- Create clinical summarization capabilities
- Design prompt engineering for medical reasoning

Week 6: Medical Imaging Integration

- Integrate LLaVA-Med for image description
- Implement BioMedCLIP for image classification
- Set up chest X-ray processing pipeline
- Create image preprocessing and validation

Week 7: Gating Mechanism Development

- Implement rule-based routing system
- Create intelligent task allocation logic
- Test routing decisions for different input combinations
- Optimize expert selection algorithms

Phase 3: System Integration & Synthesis (Weeks 8-10)

Week 8: Diagnostic Integrator

- Implement multimodal fusion engine
- Create synthesis prompts for combining expert outputs
- Develop confidence scoring system

- Test cross-modal information integration

Week 9: Output Generation & Interface

- Develop Gradio web interface
- Implement structured report generation
- Create user-friendly output formatting
- Add visualization for diagnostic insights

Week 10: End-to-End Integration

- Connect all modules through orchestration pipeline
- Implement complete data flow testing
- Debug inter-component communication
- Optimize system performance and latency

Phase 4: Testing & Evaluation (Weeks 11-13)

Week 11: Technical Evaluation

- Create evaluation dataset with mock clinical cases
- Test individual component performance metrics
- Measure system-level accuracy and consistency
- Conduct hallucination and factual accuracy testing

Week 12: Clinical Validation

- Conduct blinded review of AI vs human reports
- Collect System Usability Scale (SUS) questionnaires

Week 13: Final Analysis & Documentation

- Analyze evaluation results and user feedback
- Document system limitations and future improvements
- Prepare final project report and presentation
- Create deployment documentation and user guides

Chapter 6: Evaluation Strategy

The evaluation of the AI GP Doctor prototype is conceived as a multi-faceted process designed to rigorously assess the performance of its individual components and the clinical utility of the orchestrated system [1, 13]. For any AI system to be adopted in a clinical setting, it must be proven to be not just technically proficient, but also usable, useful, and trustworthy. This chapter outlines a comprehensive evaluation plan that directly corresponds to the prototype's current implementation as detailed in the provided code. The strategy focuses on modular assessment of

functional components, clear identification of placeholder modules, and a user-centric validation of the end-to-end output.

6.1 Technical Performance Evaluation

The initial phase of evaluation establishes a baseline of technical performance through objective, quantitative metrics. This assessment is partitioned into component-level and system-level analyses, reflecting the prototype's architecture.

Component-Level Analysis

This stage will independently benchmark each functional module within the architecture:

- Input Processor: The fidelity of the speech-to-text conversion, which utilizes the *openai/whisper* model, will be measured by its Word Error Rate (WER) against manually created ground-truth transcripts of patient audio files.
- Gating Mechanism/Router: The current prototype employs a rule-based routing mechanism. If text data exists, it activates the text-based experts; if an image is present, it activates the vision experts. Evaluation will therefore consist of a logical validation to confirm that inputs are correctly and consistently routed according to these rules.

AI Healthcare Experts: The performance of each expert will be assessed based on its specific implementation in the prototype:

  - ClinicalBERT: The system uses *emilyalsentzer/Bio_ClinicalBERT* to perform question-answering based on a fixed set of four clinical questions. Its extractive capabilities will be measured using Exact Match (EM) and F1-scores by comparing its answers against a manually annotated dataset, focusing on clinical named entity recognition [18].
  - Clinical LLM: In the prototype, the Clinical LLM's primary analytical function is a basic symptom extraction that searches for a predefined list of keywords such as 'pain', 'fever', and 'cough' within the input text. The evaluation will measure the precision, recall, and F1-score of this keyword-based method against manually identified symptoms.
  - LLaVA-Med and BioMedCLIP: It is critical to note that the *llava_med_analysis* and *biomedclip_classification* functions are currently non-functional placeholders. As such, they cannot be evaluated in the prototype's current state. The evaluation plan will include methodologies (e.g., accuracy, precision, recall, AUC-ROC) to be executed *after* these models, or similar vision-language models, are fully implemented [6, 10].

System-Level Analysis

This analysis assesses the quality of the final, synthesized output from the Diagnostic Integrator, which is a crucial test of the multimodal fusion process [4, 11, 17].

- Diagnostic Integrator: The evaluation will scrutinize the integrator's logic. This includes assessing the accuracy of symptom aggregation from both the Clinical LLM and ClinicalBERT outputs.

Critically, it also requires a clinical review of the rule-based advice generation for home care, medication, and doctor visits, which is generated based on simple keyword checks in the symptoms list.

- Final Synthesized Report: The final patient-friendly reports will be quantitatively compared against human-authored gold-standard reports using ROUGE scores for summarization quality [20]. Furthermore, these reports will be manually scrutinized to calculate a factual consistency score and a hallucination rate, ensuring the system does not generate clinically unfounded information, a known challenge with LLMs in healthcare [7, 19].

6.2 Clinical Utility and User Experience Evaluation

Moving beyond technical metrics, this phase assesses the prototype's real-world value from a clinician's perspective [3, 12]. This will be conducted through a formal user study involving medical professionals, who will be informed of the prototype's current limitations (i.e., the placeholder status of the vision models).

- Blinded Review: For a set of text-based test cases, participating clinicians will be presented with two reports: the final AI-generated summary and a gold-standard human-written summary, without knowledge of the source. Clinicians will rate each report on 5-point Likert scales for criteria including clinical accuracy, completeness, and actionability, providing a direct comparison of the AI's output against the human standard [11].
- Usability and Workflow Integration: After a hands-on session with the Gradio interface, participants will complete the System Usability Scale (SUS) questionnaire to provide a quantitative score of perceived usability [21]. This will be followed by semi-structured interviews to gather qualitative feedback on trust, explainability, and the system's potential fit within clinical workflows [9]. This qualitative data is indispensable for understanding the practical barriers and opportunities for implementing such an orchestration system in a live clinical environment [2, 13].

References:

1. What Is AI Orchestration? | Pure Storage, accessed on June 5, 2025, https://www.purestorage.com/knowledge/what-is-ai-orchestration.html
2. Orchestration in Medical Imaging AI: Maximizing Accuracy, Yield and Unexpected Findings, accessed on June 5, 2025, https://www.aidoc.com/learn/blog/ai-orchestration-in-medical-imaging/
3. AI on AI: Artificial Intelligence in Diagnostic Medicine: Opportunities and Challenges, accessed on June 5, 2025, https://armstronginstitute.blogs.hopkinsmedicine.org/2025/03/02/artificial-intelligence-in-diagnostic-medicine-opportunities-and-challenges/
4. Multimodality Fusion Aspects of Medical Diagnosis: A..., accessed on June 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11672922/
5. Mixture of Experts: Advancing AI Agent Collaboration and Decisions, accessed on June 5, 2025, https://www.akira.ai/blog/mixture-of-experts-for-ai-agents

6. Advancing medical imaging with language models: featuring a ..., accessed on June 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11075180/
7. Large language models in health care: Development, applications..., accessed on June 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11080827/
8. Daily Papers - Hugging Face, accessed on June 5, 2025, https://huggingface.co/papers?q=Med-PaLM
9. Artificial Intelligence in Dermatology: Challenges and Perspectives..., accessed on June 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9674813/
10. AI Imaging & Diagnostics - Google Health, accessed on June 5, 2025, https://health.google/health-research/imaging-and-diagnostics/
11. Development and evaluation of multimodal AI for diagnosis and..., accessed on June 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10748413/
12. The Future of Healthcare: Multimodal AI for Precision Medicine, accessed on June 5, 2025, https://www.akira.ai/blog/multi-modal-in-healthcare
13. 5 Reasons Your Clinical AI Platform Needs Intelligent Orchestration - Healthcare AI - Aidoc, accessed on June 5, 2025, https://www.aidoc.com/learn/blog/5-reasons-your-clinical-ai-platform-needs-intelligent-orchestration/
14. arxiv.org, accessed on June 5, 2025, https://arxiv.org/pdf/2402.06353
15. Machine Learning Datasets - mimic-iv - Papers With Code, accessed on June 5, 2025, https://paperswithcode.com/datasets?q=MIMIC-CXR%2C%20MIMIC-IV
16. arxiv.org, accessed on June 5, 2025, https://arxiv.org/pdf/2002.11379
17. A Comprehensive Review on Synergy of Multi-Modal Data and AI..., accessed on June 5, 2025, https://www.mdpi.com/2306-5354/11/3/219
18. Improving large language models for clinical named entity..., accessed on June 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11339492/
19. arxiv.org, accessed on June 5, 2025, https://arxiv.org/abs/2504.04385
20. Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out*.
21. Brooke, J. (1996). SUS - A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability Evaluation in Industry*. Taylor & Francis.