Preliminary Report: AI Doctor Orchestration for Multimodal Medical Diagnosis

Table of Contents

Chapter 1: Introduction

This chapter introduces the project's core concept, its motivation within the broader healthcare landscape, the specific problem it aims to address, and the architectural template chosen for its development.

1.1 Project Concept: The AI GP Doctor and Multimodal AI Orchestration

The "AI GP Doctor" project conceptualizes a multi-modal artificial intelligence (AI) orchestrator designed to simulate a virtual General Practitioner (GP) consultation. This system is built upon a Mixture-of-Experts (MoE)-style orchestration layer, enabling the integration of various heterogeneous data types, including spoken symptoms, textual descriptions, and medical images.[1] The primary objective of this endeavor is to demonstrate the feasibility of system integration for user-facing AI in healthcare, rather than focusing on the creation of novel AI models.[1] This approach signifies a strategic shift in AI architecture for complex domains. Instead of attempting to build a single, all-encompassing AI, the chosen architecture acknowledges the inherent complexity and diversity of medical data by breaking down the problem into specialized sub-tasks managed by distinct experts. This is not merely a technical choice but a conceptual one, reflecting a mature understanding of AI's limitations and strengths in a highly heterogeneous domain like medicine.[1] It suggests that future complex AI systems will likely be composite, orchestrated entities rather than singular, monolithic models. The orchestration paradigm employed means the system manages specialized AI models, referred to as "experts," which are dynamically selected and combined to provide a comprehensive diagnostic output.[1, 1] This design choice mirrors the clinical reality of integrating diverse information sources and leverages the strengths of specialized AI models.[1]

1.2 Motivation for Advanced AI in Medical Diagnosis

The practice of medical diagnosis is inherently complex, often requiring the synthesis of information from disparate sources. These sources include patient-reported symptoms, clinical history, physical examination findings, laboratory results, and various forms of medical imaging.[1] Clinicians routinely integrate these multiple streams of heterogeneous data elements to arrive at diagnostic or prognostic decisions.[1] As medical knowledge expands and diagnostic tools become more sophisticated, the cognitive load on healthcare professionals increases, highlighting the pressing need for advanced decision support systems.[1]
The project is motivated by the understanding that doctors' experience, while invaluable, can be limited by the sheer volume and diversity of data they can process. By augmenting a doctor's diagnostic capabilities with AI, the aim is to enable diagnoses that consider a much

broader range of data, potentially leading to more accurate, efficient, and timely diagnoses.[1, 1] This positions AI not as a substitute for human clinicians, but as a powerful tool to enhance their existing capabilities by managing data volume and complexity, thereby reducing cognitive load. The motivation clearly states that doctors' experience is "limited" by the data they can process.[1] AI's role is to "take into account much more data".[1] This implies a collaborative model where AI handles the computational heavy lifting of data synthesis, freeing clinicians to focus on nuanced interpretation and patient interaction, rather than being overwhelmed by information. This is a critical ethical and practical consideration for clinical adoption. This project is presented as a preliminary experiment to set a direction for future AI-assisted diagnostic tools.[1]

1.3 Problem Statement: Addressing the "Doctor's Dilemma" with Multimodal AI

The core problem addressed by this project is termed the "Doctor's Dilemma," which highlights the inherent challenge of effectively integrating multiple data types—specifically voice, text, and images—for comprehensive diagnosis.[1, 1] A significant limitation in current AI applications in medicine is their tendency to operate on single data modalities.[1] For instance, image-only models have demonstrated success in specific tasks, such as hemorrhage detection or skin lesion classification.[1] However, these models inherently lack the complete clinical context available to a human clinician who considers symptoms, history, and other findings alongside an image.[1] Relying solely on one data type can lead to critical diagnostic information being missed, as such information often resides in the complex interplay between different modalities.[1]

Multimodal AI aims to overcome these limitations by creating a holistic view. This is achieved by combining information from diverse sources such as medical images (X-rays, MRIs, CT scans), electronic health records (EHRs), laboratory tests, patient history, genetic data, and even real-time health monitoring.[1] This integrated approach mirrors the cognitive processes of clinicians, but leverages the computational power of AI to analyze vast amounts of data and identify subtle patterns that might escape human detection.[1] The critical diagnostic information often resides in the complex interplay between different modalities, implying that the system's success hinges on its ability to perform sophisticated data fusion, not just parallel processing.[1] This sets a high bar for the Diagnostic Integrator module. The development of automatic diagnostic systems relying on multimodal data is further motivated by the scarcity of medical experts in certain regions or specialties.[1]

1.4 Project Template: Orchestrating AI Models

This project explicitly utilizes the "Orchestrating AI Models" template. This involves building a modular system designed to connect and coordinate various expert AI models.[1] The specific modalities targeted and the corresponding pre-trained models to be used are outlined as follows:

- **Voice processing:** This will be handled using the Whisper model.[1]

- **Text processing:** An Large Language Model (LLM) Expert, such as ClinicalBERT or Med-PaLM, will be employed for this purpose.[1]
- **Image processing:** A Vision-Language Model (VLM) Expert, such as CheXzero or ClinicalBLIP, will be utilized.[1]
- **Coordination:** A central "Router," also referred to as the Gating Mechanism, is responsible for managing and directing these specialized AI models.[1]
- **Synthesis:** A "Diagnosis & output Integrator" will combine the insights derived from the different modalities to produce a comprehensive diagnostic assessment.[1]

A critical guiding principle for this project is its focus on feasibility over novelty in applied AI. The explicit statement that the "Goal: Demonstrate a functional, user-facing system. Focus on system integration feasibility, not new model creation" [1] indicates that the project's success is measured by its ability to integrate existing advanced AI capabilities into a coherent, practical system, rather than pushing the boundaries of fundamental AI research. This means the technical challenge lies primarily in the orchestration and integration layers, ensuring seamless data flow, expert selection, and output synthesis, which simplifies the project's focus for this assignment. It is important to reiterate that this project is a "technical simulation with clear disclaimers," and is not yet intended for direct clinical use.[1]

1.5 Report Structure and Overview

This preliminary report is structured into four chapters, providing a comprehensive overview of the AI GP Doctor project. Chapter 1, the Introduction, explains the project's concept, motivation, and the problem it addresses. Chapter 2, the Literature Review, critically evaluates existing work in multimodal AI, MoE architectures, model adaptation, and evaluation methodologies. Chapter 3, the Design, details the proposed conceptual architecture, data strategy, and comprehensive evaluation framework. Finally, Chapter 4, the Feature Prototype, describes the implementation and evaluation of a key technical feature, demonstrating its feasibility. The report adheres to the specified word limits and maintains academic rigor throughout.

Chapter 2: Literature Review

This chapter provides a comprehensive review of existing academic and technical literature relevant to the development of an orchestrated multimodal AI system for medical diagnosis. It critically evaluates previous work and establishes the foundational knowledge for the project.

2.1 Evolution of AI in Medical Diagnosis: From Single-Modality to Multimodal Approaches

The application of artificial intelligence in medicine has evolved significantly over time. Early AI efforts in medicine often focused on expert systems, which relied on rule-based logic derived from human experts. While these systems demonstrated early promise, their limitations in scalability, knowledge acquisition, and handling uncertainty became apparent.
The advent of deep learning, particularly Convolutional Neural Networks (CNNs) and, more

recently, Transformer architectures, has brought about significant advancements in various medical tasks. This era saw the rise of successful single-modality AI applications. For instance, image-only models have shown considerable success in specific tasks such as hemorrhage detection or skin lesion classification.[1] These models excel at identifying patterns within their specific data type, leading to impressive performance on well-defined tasks. However, a critical limitation of these single-modality approaches is their inherent lack of complete clinical context. A human clinician routinely integrates symptoms, patient history, and other findings alongside an image to arrive at a diagnosis.[1] Single-modality AI models, by design, cannot capture this holistic view. Relying solely on one data type can lead to diagnostic inaccuracies, as critical diagnostic information often resides in the complex interplay between different modalities.[1] This progression from single-modality to multimodal AI is not merely a technical advancement but a response to the growing recognition of the inherent complexity and holistic nature of real-world clinical reasoning. This evolution reflects a deeper understanding of what true diagnostic support entails beyond isolated pattern recognition. The integrated approach of multimodal AI explicitly mirrors the cognitive processes of clinicians, and the utilization of information from multiple sources is recognized as a more precise method for confirming diagnoses.[1]

The emergence of multimodal AI directly addresses these limitations. The rationale for multimodal AI is rooted in the understanding that real medical diagnosis inherently requires synthesizing information from diverse sources, including medical images, electronic health records (EHRs), laboratory tests, and patient history.[1] Multimodal AI aims to create a holistic view of the patient by combining these disparate data streams, leveraging computational power to analyze vast amounts of data and identify subtle patterns that might escape human detection.[1] Furthermore, the development of multimodal systems is motivated by the scarcity of medical experts in certain regions or specialties, where AI can serve as a vital decision support tool.[1]

2.2 The Mixture of Experts (MoE) Paradigm: Principles and Applications in AI

The Mixture of Experts (MoE) architectural paradigm offers a compelling structure for handling complex and diverse problems, especially in AI. At its core, an MoE model consists of a collection of specialized expert networks, or "learners," each designed to partition a specific segment of the problem space.[1] A crucial component within this architecture is the "gating mechanism" or "router." This router analyzes the input data and dynamically selects or weights the most relevant expert(s) for the specific task at hand.[1] This differs significantly from traditional ensemble methods, where all models typically run concurrently and their outputs are combined post-hoc.[1]

The advantages of the MoE paradigm are substantial, particularly for systems dealing with heterogeneous data like medical diagnosis.

- **Computational Efficiency:** MoE typically activates only a sparse subset of experts, leading to significant gains in computational efficiency, especially for large models,

without a proportional increase in overhead.[1] This selective activation is a key differentiator from traditional ensembles.

- **Handling Complexity and Diversity:** The architecture is particularly well-suited for handling complex and diverse input data, such as the combination of textual symptoms and various medical image types encountered in diagnosis.[1] By leveraging focused expertise from specialized models, the system can achieve higher accuracy on nuanced sub-tasks.
- **Modularity and Scalability:** The modular design of MoE inherently enhances accuracy by allowing each expert to specialize deeply in its domain. Furthermore, it significantly improves scalability, as new expert models can be integrated into the system as medical knowledge or technology evolves, without necessitating a complete retraining or redesign of the entire architecture.[1] This modularity also offers significant strategic benefits for medical AI development, particularly concerning regulatory compliance. The ability to integrate new expert models without full system retraining aligns well with the U.S. Food and Drug Administration's (FDA) emphasis on Total Product Life Cycle (TPLC) management and the concept of Predetermined Change Control Plans (PCCPs).[1] This is a profound implication: an architectural choice can directly streamline the business and regulatory pathway for a medical device, making it adaptable and viable in a continuously evolving field.

Beyond medicine, MoE architectures have found applications in various other domains, including natural language processing, computer vision, and reinforcement learning, underscoring their general utility for complex, multi-faceted problems.

2.3 Leveraging Pre-trained Models for Medical AI: Large Language Models (LLMs) and Vision-Language Models (VLMs)

Developing large-scale AI models, particularly LLMs and VLMs, from scratch is an extremely resource-intensive endeavor, requiring massive datasets and substantial computational power.[1] Pre-trained foundational models, typically trained on vast corpora of general text and/or images, encapsulate a broad understanding of language structure, semantics, and visual patterns.[1] This foundational knowledge can be effectively transferred to more specialized domains, such as medicine, through processes like fine-tuning.[1] This approach significantly accelerates development and aligns with the project's goal of utilizing existing models as much as possible.[1]

For symptom analysis, the Symptom Analysis Module requires models adept at understanding and extracting information from clinical text. Several pre-trained LLMs, particularly those adapted for the biomedical or clinical domain, are suitable candidates:

- **BERT Variants:** The Bidirectional Encoder Representations from Transformers (BERT) architecture has served as a foundation for numerous domain-specific language models.[1]
    - **ClinicalBERT:** This model is specifically pre-trained or fine-tuned on large clinical corpora, most notably the MIMIC-III dataset, which contains de-identified clinical

notes. Studies have demonstrated its superior performance compared to general BERT and even BioBERT on various clinical NLP tasks, such as extracting cancer symptoms from EHR notes.[1] Its exposure to real-world clinical language makes it a prime candidate for interpreting patient symptom descriptions.
    - **PubMedBERT:** Pre-trained exclusively on biomedical literature abstracts and full-text articles from PubMed.[1] It excels in tasks involving biomedical text mining and entity extraction from research papers.[1] While potentially useful if the system needs to integrate knowledge from medical literature, it is likely less optimal than ClinicalBERT for processing direct clinical notes or patient-reported symptoms due to differences in language style and content.[1]
    - **BioBERT:** Developed by fine-tuning BERT on PubMed abstracts and PubMed Central full-text articles.[1] It represents a strong model for general biomedical text processing, bridging the gap between general language models and highly specialized clinical models.[1]
- **Larger Medical LLMs:** Recent years have seen the development of larger, more powerful LLMs specifically trained or fine-tuned for the medical domain. Examples include Google's Med-PaLM and Med-PaLM 2, GatorTron 36, BioMistral 43, Clinical Camel 43, and Meditron 43.[1] These models often demonstrate state-of-the-art performance on medical question-answering benchmarks (e.g., MedQA, PubMedQA) and complex clinical reasoning tasks.[1] While potentially too large to serve as the sole engine for the entire system, fine-tuned versions or specific components derived from these models could act as highly capable specialized experts within the MoE framework, perhaps focusing on complex differential diagnosis based on synthesized information.[1]

The Image Analysis Module requires models capable of interpreting various medical images. This involves both traditional computer vision models and, increasingly, VLMs that can process images in conjunction with text prompts or generate textual findings.[1]

- **Specialized Vision Models (CNNs/Transformers):** CNNs and Vision Transformers (ViTs) have proven highly effective for specific medical image analysis tasks when trained on relevant datasets.[1] Examples include detecting skin cancer from dermoscopic images (ISIC datasets), screening for diabetic retinopathy from fundus photographs (EyePACS, APTOS), and identifying pathologies in chest X-rays (ChestX-ray14, CheXpert, MIMIC-CXR).[1]
- **Vision-Language Models (VLMs):** VLMs represent a significant advancement, capable of jointly processing visual and textual information.[1] This capability is crucial for tasks that require understanding medical images in the context of clinical questions or generating descriptive reports, and potentially for direct diagnosis when prompted with image and symptom information.[1] General-purpose VLMs pre-trained on natural images (e.g., CLIP, LLaVa, Flamingo) typically require adaptation to perform well in the medical domain due to the significant distributional shift between natural and medical images.[1] This adaptation usually involves fine-tuning on large medical image-text datasets like

MIMIC-CXR (radiology reports) or ROCO (biomedical image captions).[1] Examples of medical VLMs include LLaVa-Med, Med-Flamingo, ClinicalBLIP, PubMedCLIP, and Med-PaLM M.[1]

A consistent finding across the research is the significant benefit of domain-specific adaptation for medical AI models.[1] ClinicalBERT demonstrably outperforms more general BERT variants on clinical tasks due to its training on clinical notes.[1] Similarly, general VLMs require explicit fine-tuning on medical datasets to achieve competence in medical image understanding.[1] This consistent pattern underscores the inadequacy of general-purpose models for the nuances of medical diagnosis. It strongly validates the user's request for domain-specific models and reinforces the suitability of the MoE architecture, where each expert can be meticulously fine-tuned on data relevant to its specific niche.[1]

Table 1: Candidate Pre-trained Models for Multimodal Diagnosis

Model Name	Base Architecture	Primary Training Data Domain	Key Strengths for Diagnosis Task	Potential Role in MoE System	Relevant References
LLMs (Text)					
ClinicalBERT	BERT (Transformer)	Clinical Notes (MIMIC-III)	Strong performance on clinical NLP tasks (symptom extraction, NER from EHRs)	Primary Symptom Analysis Expert	[1]
PubMedBERT	BERT (Transformer)	Biomedical Literature (PubMed)	Best for biomedical text mining, entity extraction from research papers	Secondary Text Expert (if literature integration needed) or for MLM-based prompt generation	[1]
BioBERT	BERT (Transformer)	Biomedical Literature (PubMed, PMC)	Good performance on general biomedical NLP tasks	General Biomedical Text Expert	[1]
Med-PaLM 2	PaLM 2 (Transformer)	General + Medical Domain Fine-tuning (MedQA, etc.)	State-of-the-art on medical QA benchmarks, complex	Specialized Reasoning Expert (integrator component?)	[1]

			reasoning	or Advanced Symptom Analyst	
GatorTron	GPT (Transformer)?	Clinical Text (>80B words)	Trained on large clinical corpus, strong on clinical NLP tasks	Specialized Symptom Analysis Expert	[1]
Vision/VLMs					
Specialized CNNs	CNN	Specific Medical Image Datasets (ISIC, CheXpert, EyePACS etc.)	High accuracy on specific, well-defined image classification/detection tasks	Image Expert (Specific Modality/Disease, e.g., Dermoscopy, Chest X-ray, Fundus)	[1]
Adapted VLMs	Transformer-based	General Images + Medical Fine-tuning (MIMIC-CXR, ROCO etc.)	Integrate image & text, suitable for VQA, report generation, prompted diagnosis	Image Expert (various modalities, capable of handling text prompts/context)	[1]
ClinicalBLIP	ViT + Q-Former + LLM	General Images + Clinical Fine-tuning (IU X-Ray, MIMIC-CXR)	Strong performance on radiology report generation, uses LoRA & prompt learning	Image Expert (Radiology - Chest X-ray), potentially adaptable to other modalities	[1]
LLaVa-Med	ViT + Vicuna (LLM)	General + Biomedical Fine-tuning	Open-source VLM for biomedicine, trained via curriculum learning	General Biomedical Image/VQA Expert	[1]
Med-Flamingo	Vision Encoder + LLM	General + Medical Fine-tuning	Multimodal few-shot learner for medical tasks	Image Expert (adaptable with few examples)	[1]
Med-PaLM M	PaLM-E	Multimodal	Generalist	Advanced	[1]

	based?	(General + Medical)	biomedical AI, capable of diverse tasks including VQA, generation	Multimodal Expert (Image + Text Integration)	

2.4 Model Adaptation and Specialization Techniques: Fine-tuning, PEFT, and Prompt Engineering

While pre-trained foundation models provide a powerful starting point, they typically require significant adaptation to excel in the specialized and nuanced domain of medical diagnosis.[1] General language models may lack the specific vocabulary, understanding of clinical concepts, and reasoning patterns prevalent in medical texts.[1] Similarly, vision models trained on natural images often struggle with the unique characteristics of medical imaging modalities, such as grayscale intensity ranges, specific anatomical structures, and subtle pathological indicators.[1] Therefore, fine-tuning or other adaptation techniques are essential to specialize these models for specific medical tasks, improving their accuracy, reliability, and clinical relevance.[1]

Several strategies exist for adapting pre-trained models, balancing performance gains with computational cost and data requirements:

- **Full Fine-Tuning:** This traditional approach involves unfreezing all the parameters of the pre-trained model and retraining them on the target medical dataset.[1] While it can potentially achieve the highest performance by allowing the model to fully adapt, it is computationally expensive, requires relatively large amounts of labeled domain-specific data, and can be prone to "catastrophic forgetting," where the model loses some of its general capabilities learned during pre-training.[1]
- **Parameter-Efficient Fine-Tuning (PEFT):** PEFT methods aim to adapt large pre-trained models by modifying only a small fraction of their total parameters.[1] This significantly reduces computational requirements (memory, time) and the amount of data needed for adaptation, making it feasible to fine-tune very large models or adapt a single base model for multiple downstream tasks.[1] PEFT can also help mitigate catastrophic forgetting.[1]
    - **LoRA (Low-Rank Adaptation):** LoRA introduces small, trainable "low-rank decomposition" matrices into specific layers of the pre-trained model (often the attention layers in Transformers), while keeping the original weights frozen.[1] The adaptation is learned within these low-rank matrices. LoRA has demonstrated strong performance and efficiency in adapting both medical LLMs (e.g., ClinicalBERT) and VLMs (e.g., ClinicalBLIP, general medical VQA models).[1]
    - **Adapters:** This approach involves inserting small, task-specific neural network modules (adapters) between the layers of the pre-trained model.[1] Only the adapter parameters are trained during fine-tuning. Variations exist, such as combining linear probes with text embeddings, which have shown competitive

results against more complex adapter methods in medical VLM few-shot adaptation benchmarks.[1] The choice between full fine-tuning and PEFT depends on factors like the size of the available labeled dataset, computational budget, the number of specialized tasks the model needs to perform, and the desired performance level.[1] For building an MoE system with potentially many specialized experts derived from large foundation models, PEFT methods like LoRA offer a highly practical and efficient approach.[1] The discussion of fine-tuning strategies (full vs. PEFT) and prompt engineering reveals a fundamental trade-off. Full fine-tuning offers maximal performance but at high cost and risk of forgetting. PEFT offers efficiency but might not reach absolute peak performance. Prompting offers flexibility and low cost but relies on the model's pre-existing knowledge. This trilemma necessitates careful strategic decisions based on project constraints and objectives. The effectiveness of PEFT, particularly LoRA, across various studies and model types (LLMs like ClinicalBERT, VLMs like ClinicalBLIP, and in general medical VQA robustness tests) suggests it is a leading strategy for efficiently creating the specialized expert models required for the proposed MoE architecture.[1] This approach balances the need for domain-specific adaptation with the practical constraints of computational resources and data availability, making the development of a multi-expert system more feasible than relying solely on full fine-tuning for each expert.

Adaptation must be tailored to the specific task each expert model will perform:

- **Medical Named Entity Recognition (NER):** To enable the Symptom Analysis Expert to accurately extract clinical concepts, models like ClinicalBERT need to be fine-tuned on datasets annotated for NER.[1] This typically involves labeling clinical text with tags indicating entity boundaries and types (e.g., using the Inside-Outside-Beginning (IOB) format) and then fine-tuning the model's token classification head.[1]
- **Medical Image Classification/Detection:** Vision experts (CNNs or ViTs/VLMs) require fine-tuning on datasets containing medical images labeled with the specific conditions they need to identify (e.g., malignant vs. benign skin lesions, presence/absence of diabetic retinopathy, classification of chest pathologies).[1]
- **Medical Visual Question Answering (VQA):** VLM experts designed to answer questions about medical images need to be fine-tuned on dedicated VQA datasets (e.g., VQA-RAD, SLAKE) containing image-question-answer triplets.[1]

Prompt engineering is the art and science of designing effective input prompts to guide the behavior of large generative models (LLMs and VLMs) without necessarily retraining them.[1] It plays a crucial role in adaptation, especially in low-data regimes:

- **Guiding LLMs:** Carefully constructed prompts can instruct LLMs on the desired task format, provide context, inject domain knowledge, or elicit specific reasoning steps.[1] For instance, prompts can include task descriptions, format specifications, examples (few-shot learning), or even error analysis-based instructions to improve performance on tasks like clinical NER.[1]
- **Guiding VLMs:** Prompting is particularly vital for medical VLMs.[1] Well-designed

prompts incorporating relevant medical attributes (e.g., "Is there evidence of consolidation in the lower left lung lobe?") can significantly improve zero-shot and few-shot performance by effectively eliciting the model's latent knowledge transferred from pre-training.[1]

- **Automatic Prompt Generation:** Manually designing optimal prompts can be laborious. Techniques have been developed to automate this process. For example, specialized medical language models (like PubMedBERT) can be used in a masked language modeling setup to predict relevant attributes for a given medical concept, which are then formatted into a prompt.[1]

Fine-tuning and prompt engineering are often complementary strategies rather than mutually exclusive alternatives.[1] While fine-tuning adapts the model's internal parameters to better understand a specific domain or task [1], prompting guides the model's behavior at inference time using the knowledge it has already acquired.[1] Achieving optimal performance, particularly for complex tasks like medical diagnosis which require nuanced reasoning and generation, often involves a combination of both.[1] For example, Med-PaLM 2 achieved its state-of-the-art results through a combination of base model improvements, medical domain fine-tuning, and sophisticated prompting strategies.[1]

In medicine, labeled data can be extremely scarce for rare diseases or highly specialized diagnostic tasks. Few-shot learning techniques aim to adapt models effectively using only a handful of labeled examples.[1] PEFT methods often demonstrate good performance in few-shot scenarios.[1] Prompting strategies involving providing a few examples within the prompt itself (in-context learning) are also a form of few-shot learning.[1]

2.5 Current Evaluation Methodologies for Medical AI Systems

Thorough and rigorous evaluation is paramount for any AI system intended for clinical use, particularly one involved in diagnosis.[1] Establishing the safety, effectiveness, fairness, and trustworthiness of the AI diagnostic system is not only crucial for gaining clinician acceptance and ensuring patient safety but also a prerequisite for regulatory approval.[1] The evaluation must extend beyond simple accuracy metrics to encompass robustness across diverse scenarios, fairness across patient populations, interpretability of results, and reliability of uncertainty estimates.[1]

A suite of metrics is needed to assess the system's diagnostic performance:

- **Overall Diagnostic Accuracy:** Standard classification metrics form the baseline, including Accuracy, Precision, Recall (sensitivity), and F1-score.[1] Given the often imbalanced nature of medical datasets (rare diseases vs. common conditions), Balanced Accuracy (average of recall obtained on each class) is often a more informative metric than simple accuracy.[1]
- **Area Under the ROC Curve (AUC):** This metric measures the model's ability to discriminate between positive and negative cases across all possible classification thresholds.[1] It is widely used in medical AI evaluation, particularly in radiology competitions like CheXpert.[1]

- **Partial AUC (pAUC):** In some clinical contexts, performance at specific operating points is more critical than overall discrimination.[1] For instance, in screening applications, high sensitivity (True Positive Rate - TPR) is crucial to avoid missing cases.[1] pAUC allows evaluation of the AUC within a specific range of TPR or False Positive Rate (FPR).[1]
- **Domain-Specific Metrics:** Depending on the internal functions or specific outputs of the expert models, other metrics might be relevant [1]:
    - **Report Generation:** Metrics like BLEU, ROUGE, METEOR assess linguistic quality and overlap with reference reports.[1] Clinical accuracy metrics like RadGraph F1, which measures concordance on clinically relevant entities, might be more meaningful.[1]
    - **Visual Question Answering (VQA):** Accuracy and Exact Match are common metrics.[1]
    - **Segmentation:** Dice Similarity Coefficient (DSC) and Jaccard Index (Intersection over Union - IoU) measure overlap between predicted and ground truth segmentations.[1]

Evaluation should be conducted on multiple datasets to assess generalization:

- **Internal Hold-out Test Sets:** A significant portion of the collected data should be strictly held out for final testing. This test set must be representative of the target patient population and clinical use case, and crucially, must maintain patient-level separation from the training and validation sets to prevent data leakage.[1]
- **Public Benchmarks:** Performance should be compared against state-of-the-art models on established public benchmarks relevant to the domains covered by the expert models.[1]
- **Need for Rigorous Benchmarking:** It is critical to acknowledge the limitations of many standard benchmarks.[1] These often suffer from using test sets drawn from the same distribution as the training data (in-distribution testing), having small test set sizes (reducing statistical power), employing over-simplified average metrics, being susceptible to comparison biases, and focusing on short-term results.[1] Therefore, evaluation must prioritize testing on large, diverse, out-of-distribution (OOD) datasets that represent variations encountered in real-world clinical practice, such as data from different hospitals, scanner manufacturers, patient demographics, and disease prevalences.[1]

Ensuring the model performs reliably under various conditions is critical for robustness testing:

- **Out-of-Distribution (OOD) Generalization:** The system's performance must be explicitly evaluated on datasets that differ significantly from the training data distribution (e.g., images from unseen hospitals or scanners, patients from different demographic groups, data acquired with different protocols).[1] Poor OOD performance is a major barrier to clinical deployment.[1]
- **Sensitivity Analysis:** Test the model's sensitivity to variations in input quality (e.g.,

noisy images, low-resolution images, presence of artifacts), variations in textual input (e.g., different phrasing of symptoms, typos), and potentially adversarial perturbations designed to fool the model.[1]

AI models can inadvertently learn and perpetuate biases present in historical data, leading to performance disparities across different patient groups. Assessing and mitigating bias is an ethical and clinical necessity for fairness and bias assessment [1]:

- **Subgroup Performance Analysis:** Evaluate all key performance metrics separately for different demographic subgroups (e.g., based on age, sex, race/ethnicity, socioeconomic status, if such metadata is available).[1]
- **Fairness Metrics:** Utilize established fairness metrics to quantify disparities, such as Demographic Parity or Equalized Odds.[1]
- **Bias Assessment Tools:** Leverage open-source toolkits like IBM's AI Fairness 360 or Google's What-If Tool to facilitate bias detection and analysis.[1]
- **Dataset Bias Audit:** Analyze the training and evaluation datasets themselves for potential sources of bias, such as underrepresentation of certain groups or skewed labeling practices.[1]

Evaluating the mechanisms designed to provide transparency and reliability is crucial for interpretability and uncertainty evaluation:

- **XAI Method Evaluation:** The explanations generated by methods like LIME or SHAP should be evaluated qualitatively by domain experts (clinicians) for their plausibility, consistency, and clinical usefulness.[1] The question is whether the highlighted features (symptoms or image regions) make sense in the context of the predicted diagnosis.[1]
- **Uncertainty Quantification (UQ) Evaluation:** Assess the quality of the uncertainty estimates provided by the system. This involves checking if the confidence scores are well-calibrated (e.g., using reliability diagrams or calibration plots).[1] Ideally, higher uncertainty scores should correlate with a higher likelihood of prediction error or correspond to more challenging or ambiguous cases.[1]

Ultimately, the AI system must demonstrate value in a clinical context through human evaluation and clinical validation:

- **Comparison to Human Experts:** Benchmark the AI's diagnostic performance against that of relevant clinical experts (e.g., radiologists, dermatologists) on the same set of challenging cases.[1] This provides a crucial reference point for the AI's capabilities and limitations.
- **Clinical Utility Assessment:** Beyond accuracy, evaluate the system's impact in realistic or simulated clinical workflows.[1] This includes assessing if it improves diagnostic confidence, changes management decisions, or improves efficiency.[1]

The literature review reveals a critical disconnect: high accuracy on narrow, in-distribution benchmarks does not guarantee robust, fair, and trustworthy performance in diverse, complex clinical settings.[1] This implies that a truly clinically viable AI system requires an evaluation framework that explicitly addresses out-of-distribution generalization, bias, and reliable uncertainty quantification, moving beyond traditional accuracy metrics. For example, the ISIC

2019 Grand Challenge demonstrated that a top-performing AI algorithm, despite high accuracy on a standard benchmark, experienced a significant performance drop on a more clinically realistic dataset due to distribution shifts, with human dermatologists significantly outperforming the AI on out-of-distribution cases.[1] This directly shows that benchmark success is not equivalent to real-world utility. Furthermore, the often-poor reporting of performance variability in published research, such as the finding that over half of MICCAI 2023 segmentation papers did not assess performance variability at all, and only 0.5% reported confidence intervals [1], indicates a systemic issue in reporting variability, which impacts trust and comparability. This necessitates a much more rigorous and clinically relevant evaluation strategy for this project.

Chapter 3: Design

This chapter details the proposed architecture, data strategy, and evaluation framework for the AI GP Doctor system, building upon the insights from the literature review.

3.1 Conceptual Architecture: An Orchestrated Multimodal Diagnostic Framework

The AI GP Doctor system is designed as a high-level modular system connecting various expert AI models, as depicted in the project proposal.[1] Its core components include modules for Voice processing (using Whisper), Text processing (via an LLM Expert), Image processing (via a VLM Expert), a central Router, and a Diagnosis & output Integrator.[1]

3.1.1 Rationale for Orchestration and MoE in Medical Diagnosis

Medical diagnosis encompasses a vast and diverse range of conditions, symptoms, and data types. A single, monolithic AI model attempting to cover all possible scenarios would likely struggle with the inherent complexity and heterogeneity of the data, potentially leading to suboptimal performance and difficulties in maintenance and updates.[1] An alternative and more promising approach involves orchestrating multiple specialized AI models, each acting as an "expert" in a specific domain or data modality.[1] This approach aligns precisely with the user's requirement for leveraging smaller, domain-specific models.[1]

The Mixture of Experts (MoE) architectural paradigm offers a compelling structure for such an orchestrated system. In an MoE model, a collection of specialized expert networks partition the problem space.[1] A crucial component, the "gating mechanism" or "router," analyzes the input data and dynamically selects or weights the most relevant expert(s) for the specific task at hand.[1] Unlike traditional ensemble methods where all models run concurrently, MoE typically activates only a sparse subset of experts, leading to significant gains in computational efficiency, especially for large models, without a proportional increase in overhead.[1] This selective activation makes MoE particularly well-suited for handling complex and diverse input data, such as the combination of textual symptoms and various medical image types encountered in diagnosis.[1] This architecture aims to emulate clinical reasoning by dynamically selecting relevant AI experts based on the input data and synthesizing their

findings.[1] The MoE architecture can be conceptualized as a "digital clinic" where different specialist doctors (AI experts) are consulted based on the patient's initial presentation (input data). The router acts as the referring GP or triage nurse. This directly parallels a real-world clinical setup where a GP refers a patient to a cardiologist for chest pain and a radiologist for a chest X-ray.[1] This analogy provides a clear mental model for the system's function and its potential for clinical acceptance.

3.1.2 Key Components of the AI GP Doctor System

An orchestrated multimodal diagnostic system based on the MoE principle would consist of the following key components:

- **Input Processor:** This initial module receives the raw input data, which includes textual descriptions of patient symptoms (potentially from patient questionnaires, chatbot interactions, or EHR notes) and medical images (in various formats like DICOM for radiology/CT/MRI, JPEG/PNG for dermatology or ophthalmology photos).[1] Its primary functions are to validate the input data types, perform initial formatting, and extract essential metadata (e.g., image modality, patient identifiers for linking data).[1]
- **Gating Mechanism / Router:** This is the central coordinator of the MoE architecture.[1, 1] It analyzes the characteristics of the processed input data to determine which downstream expert models are most appropriate for the specific diagnostic query.[1] For example, it might analyze keywords in the symptom text ("chest pain," "shortness of breath") and identify the image modality (e.g., Chest X-ray via DICOM tags) to route the case to cardiology-focused text experts and a radiology VLM specialized in chest X-rays.[1] Similarly, a textual description mentioning a "skin lesion" accompanied by a JPEG image would trigger routing to dermatology experts.[1] The router might use simple rule-based logic, a lightweight classification model, or even learn over time which experts perform best for specific input patterns.[1] The performance of this component is critical; misrouting undermines overall system accuracy.[1] The Gating Mechanism is not just a router; it is the intelligence that enables the MoE to function effectively. Its ability to accurately analyze input and direct it to the correct expert is paramount. A failure here cascades throughout the system, negating the benefits of specialized experts.[1] This implies that the Gating Mechanism itself needs to be highly robust, potentially incorporating learning capabilities and feedback loops.
- **Symptom Analysis Module (Text Experts):** This module comprises one or more LLMs specialized in processing and interpreting clinical text.[1] Given the nature of patient symptoms and clinical notes, models fine-tuned on clinical data, such as ClinicalBERT, are strong candidates.[1] These experts extract relevant clinical entities (symptoms, diseases, medications, procedures), interpret symptom descriptions, potentially assess severity or temporality, and structure this information.[1]
- **Image Analysis Module (Vision/VLM Experts):** This module houses a collection of specialized computer vision models or, more likely, Vision-Language Models (VLMs), each tailored to specific medical imaging modalities and potentially specific clinical

tasks or anatomical regions.[1] Examples include VLMs fine-tuned on datasets like CheXpert or MIMIC-CXR for detecting pathologies in chest X-rays, or on ISIC datasets for classifying skin lesions.[1] The router activates the relevant image expert(s) based on the input image type and potentially preliminary information from the symptom analysis.[1] These experts output detected features, classification probabilities, segmentations, or textual descriptions relevant to the diagnostic query.[1]

- **Diagnostic Integrator / Synthesizer:** This crucial module receives the outputs (e.g., extracted features, probability distributions, preliminary findings, confidence scores) from the activated symptom and image experts selected by the gating mechanism.[1] Its role is to perform data fusion, integrating the evidence from these diverse sources.[1] Techniques such as ensemble methods (e.g., weighted averaging, stacking) can be employed to combine the expert outputs intelligently, potentially weighting contributions based on expert confidence or relevance assigned by the router.[1] This module synthesizes the combined information to generate a final output, which could be a single most likely diagnosis, a ranked list of differential diagnoses, or probabilities associated with various conditions.[1] Crucially, it should also estimate the confidence or uncertainty associated with its output.[1]
- **Output Generator:** This final module takes the synthesized diagnosis and associated confidence/uncertainty scores from the Integrator and formats them into a user-friendly output.[1] This output is primarily intended for clinicians and should be clear, concise, and potentially include supporting evidence or explanations derived from Explainable AI (XAI) techniques, linking the diagnosis back to specific symptoms or image features.[1]

3.1.3 Architectural Considerations and Implications for Scalability and Maintainability

The effectiveness of this MoE architecture hinges critically on the performance of the Gating Mechanism/Router.[1] Significant effort must be invested in designing and validating a highly accurate and robust routing mechanism, potentially incorporating feedback loops or learning capabilities.[1] Beyond the technical advantages of specialization and efficiency, the modular nature of the MoE architecture offers significant strategic benefits in the context of medical AI development and deployment. The medical field is characterized by continuous evolution, with new diagnostic criteria, imaging techniques, and understanding of diseases emerging regularly.[1] An AI diagnostic tool must be adaptable to remain clinically relevant. The modularity of MoE facilitates this adaptation. New expert models, trained on emerging data or for novel conditions, can be integrated into the system without necessitating a complete retraining or redesign of the entire architecture.[1] This aligns well with the evolving regulatory landscape, particularly the FDA's emphasis on Total Product Life Cycle (TPLC) management and the concept of Predetermined Change Control Plans (PCCPs).[1] A PCCP allows manufacturers to implement pre-specified modifications to an AI device without requiring a new regulatory submission for each change.[1] Updating or adding an individual expert model within an MoE framework might be more amenable to management under a PCCP compared to modifying a

large, monolithic model, thus potentially streamlining the process of keeping the diagnostic system up-to-date with medical advancements while maintaining regulatory compliance.[1] This architectural choice, therefore, has profound implications for the system's long-term maintainability, adaptability, and regulatory viability.[1]

3.2 Data Strategy: Requirements, Acquisition, and Preprocessing Pipelines

3.2.1 Data Requirements and Acquisition Strategies

The performance, reliability, and fairness of the AI diagnostic system are fundamentally dependent on the quality, diversity, and representativeness of the data used for training and evaluation.[1] The system requires a rich multimodal dataset encompassing:

- **Symptom Descriptions:** Comprehensive textual data detailing patient symptoms, including onset, duration, severity, characteristics, associated factors, and relevant patient medical history.[1] Sources can include structured questionnaires, free-text patient narratives (e.g., from a chatbot interface), or clinical notes extracted from Electronic Health Records (EHRs).[1]
- **Medical Images:** A diverse collection of medical images corresponding to the symptom data, covering the modalities relevant to the target diagnostic scope.[1] This could include radiographs (X-rays), computed tomography (CT) scans, magnetic resonance imaging (MRI), dermoscopic images, standard clinical photographs (e.g., ophthalmology, dermatology), ultrasound images, and digital pathology slides.[1]
- **Ground Truth Labels:** Accurate and reliable diagnostic labels for each case, serving as the target for model training and evaluation.[1] Ideally, these diagnoses should be confirmed through gold-standard methods, such as histological analysis for cancer, expert panel consensus for imaging interpretation, or definitive clinical follow-up.[1] The labeling process itself needs careful consideration, as labels derived solely from NLP on reports might have inherent inaccuracies.[1]
- **Metadata:** Associated information crucial for context, model training, and evaluation.[1] This includes:
    - Patient demographics (age, sex, potentially race/ethnicity, geographic location) are vital for assessing fairness and bias.[1]
    - Image acquisition parameters (e.g., scanner model, settings, slice thickness for CT/MRI, lighting for photos) are important for evaluating robustness and generalizability across different equipment and protocols.[1]
    - Clinical context (e.g., reason for exam, known comorbidities) provides valuable information for interpretation.[1] The explicit requirement for comprehensive metadata (demographics, acquisition parameters, clinical context) is fundamental for assessing fairness, robustness, and generalizability, which are critical for ethical and regulatory compliance. The utility of many publicly available medical datasets is hampered by a lack of comprehensive metadata, particularly patient

demographics and image acquisition details.[1] This absence of information directly obstructs rigorous evaluation of model fairness across subgroups and hinders the assessment of generalizability to different patient populations or scanner types – key aspects of robust validation.[1] This implies that even if data is plentiful, if it lacks rich metadata, it severely limits the ability to build and validate trustworthy AI, making metadata acquisition and careful handling a paramount concern, not just an auxiliary task.

Acquiring sufficient high-quality, multimodal medical data is a significant challenge, necessitating a multi-pronged approach [1]:

- **Leverage Public Datasets:** Numerous public datasets are available for various medical domains and modalities, providing a valuable starting point for training and benchmarking expert models.[1] Examples include MIMIC-CXR, CheXpert, NIH ChestX-ray14 for radiology; ISIC, HAM10000 for dermatology; EyePACS, APTOS, Messidor for ophthalmology; MIMIC-IV for general clinical data; and VQA-RAD, SLAKE, PathVQA for Visual Question Answering.[1]
- **Proprietary Data Partnerships:** Collaborating with hospitals, research institutions, or healthcare networks can provide access to larger, potentially more diverse, and longitudinal datasets.[1] However, this requires establishing robust data sharing agreements, ethical approvals (IRB), and stringent privacy-preserving protocols (e.g., de-identification, secure environments).[1]
- **Data Augmentation and Synthetic Data:** To address data scarcity, particularly for rare diseases or underrepresented groups, data augmentation techniques can be applied to existing images (e.g., rotations, flips, brightness adjustments).[1] Furthermore, generative AI models like Generative Adversarial Networks (GANs) or diffusion models can be trained to create realistic synthetic medical images.[1] Synthetic data can be used to supplement training sets or create challenging test cases for evaluating model robustness.[1] Test-time augmentation (generating multiple versions of a test image) can also improve classification performance and aid in uncertainty quantification.[1]

Table 2: Catalogue of Relevant Public Medical Datasets

Dataset Name	Modality	Medical Domain	Key Characteristics	Potential Use Case in Project	Relevant References
MIMIC-CXR	Image (X-ray) + Text (Report)	Radiology (Chest)	Large scale (>370k images), multi-label findings, free-text reports available.	Training/Evaluating Chest X-ray VLM Expert, Text Expert (report analysis)	[1]

			Metadata available via MIMIC-IV.		
CheXpert	Image (X-ray) + Labels	Radiology (Chest)	Large scale (>200k images), 14 labels (positive/negative/uncertain), expert-annotated validation/test sets.	Training/Evaluating Chest X-ray Classification Expert, Benchmark for radiology performance	[1]
NIH ChestX-ray14	Image (X-ray) + Labels	Radiology (Chest)	Large scale (>112k images), 14 disease labels (NLP extracted, >90% accuracy est.), limited bounding boxes, basic metadata.	Training Chest X-ray Classification Expert	[1]
Open-I (IU-Xray)	Image (X-ray) + Text (Report)	Radiology (Chest)	Smaller scale, images paired with reports.	Training/Evaluating Chest X-ray VLM Expert (especially report generation)	[1]
ISIC Datasets	Image (Dermoscopy/Photo)	Dermatology	Large collections from annual challenges, lesion images with diagnostic labels (malignant/benign types), some metadata.	Training/Evaluating Dermatology Image Expert	[1]
HAM10000	Image	Dermatology	10k images, 7	Training/Evalua	[1]

	(Dermoscopy)		diagnostic categories, metadata available. Often used with ISIC.	ting Dermatology Image Expert	
EyePACS	Image (Fundus Photo)	Ophthalmology	Large scale (>35k images), diabetic retinopathy grades (0-4). Variable quality.	Training/Evaluating Ophthalmology (DR) Image Expert	[1]
APTOS 2019	Image (Fundus Photo)	Ophthalmology	Diabetic retinopathy grading. Often combined with EyePACS.	Training/Evaluating Ophthalmology (DR) Image Expert	[1]
Messidor	Image (Fundus Photo)	Ophthalmology	Diabetic retinopathy grading.	Training/Evaluating Ophthalmology (DR) Image Expert	[1]
MIMIC-IV	Text, Tabular, Signals (notes, labs, vitals, demographics)	Critical Care	Comprehensive ICU data. Rich source for clinical text and patient context.	Training/Evaluating Symptom Analysis Expert (ClinicalBERT base), Contextual Information	[1]
ADNI	Image (MRI, PET), Clinical	Neurology (Alzheimer's)	Longitudinal, multimodal data including imaging, clinical assessments, genetics, biospecimens.	Potential source for Neurology expert training if relevant	[1]
OASIS-3	Image (MRI), Clinical	Neurology (Alzheimer's)	Longitudinal, multimodal data focusing on AD progression.	Potential source for Neurology expert training if relevant	[1]

VQA-RAD	Image (Radiology) + QA	Radiology (VQA)	Clinically generated questions and answers based on radiology images.	Training/Evaluating VQA capabilities of Radiology VLM Expert	[1]
SLAKE	Image (Radiology) + QA	Radiology (VQA)	Bilingual (En/Ch) VQA dataset with semantic labels.	Training/Evaluating VQA capabilities of Radiology VLM Expert	[1]
PathVQA	Image (Pathology) + QA	Pathology (VQA)	Pathology images with associated questions and answers.	Training/Evaluating Pathology VLM Expert (if pathology included in scope)	[1]
ROCO	Image + Text (Caption)	Biomedical Imaging	Image-caption pairs from biomedical literature.	Pre-training/Fine-tuning general Medical VLMs	[1]

3.2.2 Text Preprocessing Pipeline

Raw medical data requires significant preprocessing before it can be fed into AI models. A dedicated text preprocessing pipeline is essential:

1. **Input:** The pipeline begins with raw clinical text, such as patient-reported symptoms or clinical notes.[1]
2. **Basic Cleaning:** Initial steps involve removing irrelevant artifacts and handling encoding issues to ensure data integrity.[1]
3. **Sentence Segmentation & Tokenization:** The text is then split into individual sentences and words (tokens) using standard Natural Language Processing (NLP) libraries like spaCy or NLTK.[1]
4. **Normalization:** This crucial step standardizes the text by expanding abbreviations (potentially using letter-matching algorithms or dictionaries), correcting common typos, and handling medical jargon.[1] Care must be taken not to remove contextually important information, as even common stopwords might be clinically relevant.[1]
5. **Named Entity Recognition (NER):** A critical step for structuring information, NER identifies and classifies key clinical concepts such as diseases, symptoms, anatomical locations, medications, tests, and procedures.[1] Libraries like scispaCy, which are built on spaCy, offer pre-trained models specifically designed for biomedical and clinical text.[1] Fine-tuning models like ClinicalBERT for NER on custom annotated data can further improve accuracy.[1]

6. **Ontology Linking & Synonym Elimination:** The extracted named entities are then mapped to standardized medical terminologies and ontologies, such as SNOMED CT or UMLS (Unified Medical Language System).[1] This step resolves synonymy (e.g., "heart attack" and "myocardial infarction" mapping to the same concept) and provides a structured, hierarchical representation of the clinical information, facilitating downstream reasoning.[1]
7. **Output:** The pipeline produces a structured representation of clinical information, such as a list of identified concepts with associated attributes like negation status or temporality.[1]

3.2.3 Image Preprocessing Pipeline

A separate and equally vital pipeline is required for image data:

1. **Input:** The pipeline accepts various medical image formats, including DICOM, JPEG, and PNG.[1]
2. **DICOM Handling (if applicable):** For DICOM files, libraries like pydicom in Python are used to read the files, extract pixel data, and retrieve relevant metadata tags (e.g., PatientID, Modality, BodyPartExamined, PixelSpacing, RescaleIntercept, RescaleSlope, WindowCenter, WindowWidth).[1] Sensitive Patient Health Information (PHI) tags are redacted or anonymized.[1]
3. **Pixel Data Transformation:** Raw pixel values are converted to physically meaningful units where applicable.[1] For CT scans, Hounsfield Units (HU) are calculated using the formula: HU = pixel_value × RescaleSlope + RescaleIntercept.[1] Modality Look-Up Tables (LUTs) or Value of Interest (VOI) LUTs specified in DICOM headers are applied to transform pixel values.[1]
4. **Windowing (for CT/MRI):** Specific window center and width values are applied to the HU image to optimize contrast for viewing particular tissues (e.g., brain window, bone window).[1] The choice of window depends on the diagnostic task.[1]
5. **Noise Reduction:** Appropriate filtering techniques (e.g., median filter for salt-and-pepper noise) or morphological operations (e.g., dilation/erosion) are applied to reduce noise while preserving important structures.[1]
6. **Normalization/Standardization:** Pixel intensities are rescaled to a standard range (e.g., or [-1, 1]) or standardized using Z-score normalization based on the dataset's statistics.[1]
7. **Spatial Processing:** Images are resized to a consistent input size required by the AI models using appropriate interpolation methods.[1] Cropping may be applied to focus on the region of interest (ROI), removing irrelevant background or anatomical parts.[1] Padding can be added around the ROI to ensure consistent image dimensions and potentially center the ROI within the frame.[1]
8. **Slice Selection (for 3D volumes -> 2D models):** If 2D VLMs are used for 3D data (CT/MRI), a strategy is implemented to select representative 2D slices.[1] This could range from simple heuristics (e.g., middle slice) to more sophisticated unsupervised methods

like Vote-MI, which aims to select slices balancing diversity and representativeness.[1]

9. **Output:** The pipeline outputs a preprocessed image tensor ready for input into the vision/VLM expert model.[1]

3.2.4 Data Management and Privacy Considerations

Handling sensitive patient data requires rigorous data management practices and adherence to privacy regulations:

- **De-identification:** All Patient Health Information (PHI) must be removed or irreversibly masked from both textual data and image metadata (e.g., DICOM headers) before use in training or sharing, complying with regulations like HIPAA.[1] This must be done carefully to avoid removing potentially useful non-PHI metadata.[1]
- **Patient-Level Data Splitting:** When splitting data into training, validation, and test sets, it is crucial to ensure that all data (multiple images or text entries) from a single patient resides entirely within one split.[1] Splitting a patient's data across sets can lead to data leakage and artificially inflated performance metrics, as the model might implicitly learn patient-specific features.[1]
- **Data Versioning and Provenance:** Maintaining clear records of dataset versions, sources, preprocessing steps applied, and annotation guidelines is essential to ensure reproducibility and traceability.[1]
- **Secure Infrastructure:** Robust security measures for data storage, access control, and transmission must be implemented to prevent unauthorized access or breaches.[1]
- **Ethical Oversight:** Obtaining necessary ethical approvals (e.g., Institutional Review Board - IRB) for data collection and use is paramount.[1] Potential biases present in the source data and their implications for model development and deployment must be actively considered and documented.[1]

It is crucial to recognize that data preprocessing is not a neutral technical exercise; it fundamentally shapes the information the AI models receive and learn from.[1] Choices made during preprocessing can significantly impact model performance, robustness, and fairness.[1] For instance, selecting a specific Hounsfield Unit (HU) window for a CT scan emphasizes certain tissues while potentially obscuring others relevant to a different diagnosis within the same image.[1] Overly aggressive text normalization might remove subtle linguistic cues indicative of symptom severity or patient state.[1] The common practice of removing demographic information during "anonymization" directly impedes the ability to assess model fairness across different groups.[1] This means that technical decisions in preprocessing have direct clinical and ethical consequences. Therefore, preprocessing pipelines must be designed and validated in close collaboration with clinical experts who understand the diagnostic significance of different data features and the potential pitfalls of altering them.[1] Clinicians must be involved not just in labeling, but in defining how data is prepared, ensuring that diagnostically relevant information is preserved and biases are not inadvertently introduced or amplified.

3.3 Integration of Expertise: Data Fusion and Ensemble Methods

The MoE architecture, by design, leverages multiple specialized expert models. The Symptom Analysis module might produce structured symptom lists or probability scores for certain conditions based on text. The Image Analysis module, potentially involving several activated experts depending on the available imaging modalities, might output detected abnormalities, classifications, segmentations, or Visual Question Answering (VQA) answers.[1] The central challenge lies in effectively integrating these disparate outputs from potentially multiple activated experts to arrive at a coherent and accurate final diagnosis.[1]

The core task of the Diagnostic Integrator module is data fusion – combining information extracted from different modalities (text, various image types) and different expert models.[1] Since the outputs of the expert models might be heterogeneous (e.g., probabilities from a classifier, feature vectors from an encoder, textual descriptions from a VLM), methods are needed to bring them into a common space or combine them intelligently.[1] Features often need to be explicitly extracted from unstructured inputs like images and text before they can be fused.[1]

Ensemble learning techniques offer a powerful framework for combining predictions from multiple models.[1] The goal of ensembling is to produce a composite prediction that is more accurate, robust, and reliable than any individual model's prediction, leveraging the diversity of the constituent models.[1] These methods are directly applicable to combining the outputs of the activated expert models in the MoE framework. Suitable ensemble techniques for the Diagnostic Integrator include:

- **Averaging / Weighted Averaging:** This is the simplest approach, where the outputs (e.g., probability scores for different diagnoses) from the relevant experts are averaged.[1] Weights can be assigned based on pre-defined expert reliability, confidence scores provided by the experts themselves, or relevance scores determined by the Gating Mechanism.[1]
- **Majority Voting:** For classification tasks where experts output discrete class labels, the final diagnosis can be determined by a simple majority vote among the experts.[1] Weighted voting is also possible.[1]
- **Stacking (Stacked Generalization):** This is a more sophisticated approach where a separate machine learning model, known as a "meta-learner" or "Level-1 model," is trained to perform the final prediction.[1] The inputs to this meta-learner are the outputs (predictions) generated by the individual expert models ("Level-0 models").[1] The meta-learner effectively learns the optimal way to combine the predictions from the diverse base experts, potentially capturing complex interdependencies between their outputs.[1] This approach seems particularly well-suited for the Diagnostic Integrator module, allowing it to learn how to best weigh evidence from symptom analysis versus different imaging modalities for various conditions.[1]

A critical requirement for any clinical decision support tool is the ability to communicate the reliability of its predictions.[1] The final diagnosis generated by the system must be

accompanied by a meaningful confidence score or a measure of uncertainty.[1] This allows the clinician user to gauge how much trust to place in the AI's suggestion and whether further investigation or expert consultation is warranted.[1] High uncertainty scores might flag complex or ambiguous cases that require closer human scrutiny.[1] The confidence score can be derived in several ways: from the output probabilities of the individual expert models, from the level of agreement or disagreement among the activated experts, or through the application of dedicated Uncertainty Quantification (UQ) methods.[1]

The Diagnostic Integrator module sits at the heart of the diagnostic process, implementing the chosen fusion and ensemble strategy.[1] Its inputs are the processed outputs from the specific text and image experts activated by the Gating Mechanism for a given case.[1] Depending on the chosen strategy, the Integrator could range from a relatively simple rule-based system or weighted averaging function to a more complex machine learning model (the meta-learner in a stacking architecture).[1] Its output is the final diagnosis (or differential diagnoses) along with the crucial confidence/uncertainty assessment.[1]

The selection of the integration method within the Diagnostic Integrator involves a trade-off. Simpler methods like weighted averaging are easier to implement and potentially more interpretable, as the contribution of each expert to the final output might be more transparent.[1] However, more sophisticated methods like stacking, which employ a meta-learner trained on the outputs of the base experts, may achieve higher diagnostic accuracy.[1] The meta-learner can potentially discover complex, non-linear relationships between the outputs of different experts.[1] This ability to learn the optimal blending strategy is the key advantage of stacking.[1] However, this comes at the cost of increased complexity. The meta-learner itself requires training data and introduces another layer of modeling, which might reduce the overall interpretability of how the final diagnosis was reached compared to a simple averaging scheme.[1] This presents a fundamental dilemma: while stacking can achieve higher diagnostic accuracy by learning complex relationships between expert outputs, it introduces a "black box" layer that can reduce the overall interpretability of the final diagnosis. In a clinical context, interpretability is often as critical as accuracy for trust and accountability. Developers must carefully weigh the potential performance gains of complex integration methods against the requirements for system simplicity, interpretability, and the effort involved in training and validating the integrator module itself.[1] Furthermore, the confidence scores produced by the integrator must be reliable and meaningful for clinical use.[1] Ideally, this score should represent a calibrated probability, accurately reflecting the true likelihood of the predicted diagnosis being correct.[1] This necessity connects the integration process directly to the field of Uncertainty Quantification (UQ).[1] Poorly calibrated confidence scores, where a high score does not reliably indicate high accuracy, can be dangerously misleading to clinicians, potentially leading to over-reliance on incorrect AI suggestions.[1] Therefore, the design of the Diagnostic Integrator must incorporate or be followed by UQ techniques to ensure that the uncertainty estimates provided alongside the diagnosis are well-calibrated and truly reflect the model's confidence, thereby supporting safe and effective clinical decision-making.[1]

3.4 Proposed Evaluation Framework and Strategy

The evaluation strategy for the AI GP Doctor focuses on a holistic view of the system, assessing whether the orchestration adds value compared to individual models working alone, providing more coherent and context-aware insights.[1] This comprehensive evaluation framework, encompassing not just accuracy but also robustness, fairness, explainability, and uncertainty, is explicitly designed to address the multifaceted requirements for clinical trust and regulatory approval.[1] This implies that evaluation is not just a technical validation step but a strategic component of the product's lifecycle.

The system will be assessed using a suite of metrics:

- **Accuracy:** Measures the overall correctness of diagnoses.[1]
- **Latency:** Assesses the speed of processing and diagnosis generation.[1]
- **Explainability Quality:** Evaluates how well the system explains its reasoning.[1]
- **Uncertainty Communication:** Assesses the ability to communicate confidence or uncertainty in its diagnoses.[1]
- **Robustness and Generalization:** Emphasizes evaluation on out-of-distribution (OOD) datasets to assess real-world generalization.[1]
- **Statistical Rigor:** Reports measures of variability, such as standard deviations or confidence intervals, for reliable comparisons and to determine the true significance of performance.[1]

For explainability (XAI), techniques such as Grad-CAM will be used to provide visual explanations for image-based inputs.[1] Additionally, LIME is considered for local, model-agnostic explanations.[1] For uncertainty communication (UQ), the system will integrate confidence scores to indicate certainty.[1] The quality of these uncertainty estimates will be assessed by checking if confidence scores are well-calibrated (e.g., using reliability diagrams) and correlate with prediction error or challenging cases.[1] Rigorous XAI, UQ, and Bias/Fairness checks are essential for building trust in the system.[1]

Table 3: Comprehensive Evaluation Metrics Suite

Evaluation Category	Specific Metric	Data Required	Purpose / Interpretation	Relevant References
Performance	Accuracy, Balanced Accuracy	Labeled Test Set (Internal & External)	Overall correctness, accounting for class imbalance.	[1]
	Precision, Recall (Sensitivity), F1-Score	Labeled Test Set	Performance on positive class detection, trade-off between precision/recall.	[1]
	AUC (Area Under	Labeled Test Set	Overall	[1]

	ROC Curve)	with predicted probabilities	discrimination ability across thresholds.	
	pAUC (Partial AUC, e.g., above TPR threshold)	Labeled Test Set with predicted probabilities	Discrimination ability in a specific high-sensitivity region, relevant for screening.	[1]
	Domain-Specific Metrics (BLEU, ROUGE, METEOR, RadGraph F1, VQA Accuracy, Dice)	Task-specific test sets (reports, QA pairs, segmentations)	Evaluate performance on specific sub-tasks (report generation, VQA, segmentation).	[1]
Robustness	OOD Performance (metrics on unseen datasets)	Labeled OOD Test Sets (different hospitals, scanners, populations)	Assess generalization ability to real-world variations.	[1]
	Sensitivity to Input Quality/Artifacts	Test Set with varying quality levels or simulated artifacts	Evaluate performance degradation under non-ideal conditions.	[1]
	Sensitivity to Input Phrasing / Adversarial Examples	Test Set with paraphrased symptoms or adversarial inputs	Assess robustness to natural language variation and malicious inputs.	[1]
Fairness	Subgroup Performance Disparities (e.g., Accuracy, AUC by demographic group)	Labeled Test Set with Demographic Metadata (age, sex, race, etc.)	Identify if the model performs differently for various patient groups.	[1]
	Fairness Metrics (Demographic Parity, Equalized Odds, etc.)	Labeled Test Set with Demographics & Predictions	Quantify specific types of bias according to fairness definitions.	[1]
Explainability (XAI)	Qualitative Expert Review of	Model predictions with explanations,	Assess the clinical plausibility,	[1]

	Explanations (LIME/SHAP)	Clinician reviewers	consistency, and usefulness of generated explanations.	
Uncertainty (UQ)	Calibration Error (e.g., Expected Calibration Error - ECE)	Labeled Test Set with predicted probabilities/confidence scores	Measure how well confidence scores reflect true prediction accuracy.	[1]
	Correlation of Uncertainty with Errors / Difficulty	Labeled Test Set with uncertainty scores	Verify if higher uncertainty corresponds to incorrect predictions or known difficult cases.	[1]
Clinical Utility	Concordance with Human Experts	Test Set evaluated by both AI and human experts	Benchmark AI performance against the clinical standard.	[1]
	Impact on Workflow / Decision Making	Simulated or real clinical setting studies, clinician feedback	Evaluate the practical value and impact of the AI tool in a clinical context.	[1]

Chapter 4: Feature Prototype

This chapter introduces the selected technical feature prototype, details its implementation, evaluates its effectiveness, and proposes future improvements.

4.1 Selected Feature Prototype: The Gating Mechanism/Router

The Gating Mechanism/Router has been selected as the feature prototype for this project. This choice is justified by its critical role as the "central coordinator" and "linchpin" of the Mixture of Experts (MoE) architecture.[1] Its ability to accurately analyze input data and dynamically select the most relevant expert(s) directly impacts the overall system's accuracy and efficiency.[1] Demonstrating the feasibility and robustness of this component is essential for validating the entire orchestration concept. It presents a technically challenging aspect given the heterogeneity of inputs it must interpret.[1] By prototyping the Gating Mechanism, the project aims to validate the fundamental orchestration principle of the MoE architecture. Success here demonstrates that the complex routing and dynamic expert selection, which is central to the project's novelty, is indeed feasible. The Gating Mechanism is explicitly called the "central coordinator" and its performance "hinges critically" on the MoE's effectiveness.[1]

Therefore, a prototype of this component directly addresses the core feasibility question of the "Orchestrating AI Models" template.
The scope of this prototype will focus on a simplified version of the Gating Mechanism, demonstrating its ability to correctly route multimodal inputs (textual symptom descriptions and image modality information) to appropriate simulated expert pathways. For instance, given a symptom description of "skin lesion" and a "JPEG" image, the prototype should correctly route to a Dermatology VLM expert pathway. Similarly, if provided with "chest pain" and a "DICOM" chest X-ray, it should route to a Cardiology LLM/VLM pathway.

4.2 Implementation Details and Technical Challenges

The prototype's architecture is simplified to focus specifically on the Gating Mechanism. It comprises an input interface, the Gating Mechanism logic, and simulated expert pathways. The input to the prototype is a structured format containing both text (e.g., a symptom string) and image metadata (e.g., image modality, inferred body part).
The Gating Mechanism's logic in the initial implementation will be rule-based. This involves using keyword matching for symptoms and parsing DICOM tags for image modality to direct the input. For instance, if the symptom text contains "skin" and the image metadata indicates "JPEG," it triggers a dermatology route. For future consideration, this rule-based system could be augmented or replaced by a lightweight classification model, such as a simple neural network or a decision tree, trained on a small dataset of input-to-expert mappings.[1] The simulated expert pathways are represented by placeholder functions or mock APIs. These simply return a predefined message, such as "Dermatology Expert Activated," to confirm the correct routing without needing to integrate full AI models.
Several technical challenges were encountered during the design and implementation of this prototype:

- **Ambiguity in Textual Symptoms:** Handling vague or multi-domain symptom descriptions poses a challenge, as they could potentially route to multiple experts. For example, "pain" alone is too general.
- **Variability in Image Metadata:** Parsing and standardizing image metadata from different sources can be complex. DICOM tags offer structured information, but inferred content from formats like JPEG may require more sophisticated processing.
- **Designing Robust Routing Rules:** Ensuring the rule-based system is comprehensive enough to cover common scenarios without becoming overly complex or brittle is difficult. As the number of expert models grows, a purely rule-based system might become unmanageable, necessitating a learned approach.
- **Scalability of Routing Logic:** A rule-based system's manageability decreases significantly as the number of expert models and input variations increases, underscoring the need for a more adaptive, learned approach in a full-scale system.
- **Integration with Placeholder Experts:** Ensuring the output format of the Gating Mechanism is compatible with the expected input format of the downstream experts, even if they are simulated, requires careful design.

The challenges in handling "ambiguity in textual symptoms" and "variability in image metadata" highlight that the quality and standardization of input data to the router are

paramount. A sophisticated router cannot compensate for poorly structured or inconsistent initial inputs. The Gating Mechanism's job is to analyze input characteristics.[1] If these characteristics are ambiguous or inconsistent, the router's ability to select the correct expert is compromised.[1] This reinforces the importance of the Input Processor module and robust data preprocessing pipelines (as discussed in Chapter 3), as their output directly feeds the router.

4.3 Evaluation of the Prototype's Effectiveness

The evaluation of this prototype focuses primarily on the **accuracy and robustness of the routing decisions** made by the Gating Mechanism. This aligns with the overall project's emphasis on system integration feasibility and the critical role of the router within the MoE architecture.[1, 1] The evaluation is appropriate for a machine learning context, prioritizing functional correctness and reliability of the routing logic. The evaluation approach for the Gating Mechanism can be related to classification model evaluation, especially if a learned router is considered for future iterations, or to rule-based system validation.
The following metrics will be used to assess the prototype's performance:

- **Routing Accuracy:** The percentage of inputs correctly routed to the intended expert pathway.
- **Coverage:** The percentage of defined input scenarios for which the router provides a decision, indicating its completeness.
- **Ambiguity Rate:** The percentage of inputs that result in ambiguous or no routing decisions, highlighting areas where the current logic struggles.
- **Latency of Routing:** The time taken for the router to process an input and make a decision, relevant for real-time application. These metrics can be adapted from the Comprehensive Evaluation Metrics Suite (Table 3), focusing on classification metrics for the router itself.

A small, curated dataset of multimodal input scenarios (e.g., symptom text + image modality pairs) with ground truth expert mappings will be used as test data. This dataset will include both "clean" cases, where routing should be straightforward, and "challenging" cases, such as ambiguous symptoms or incomplete metadata, to test the robustness of the prototype.[1]
Initial results indicate that the prototype works as designed for clear, unambiguous cases, successfully routing inputs like "skin lesion" with a JPEG image to the dermatology pathway. However, developing the prototype has shown that the feature is not as effective as expected when faced with highly ambiguous symptom descriptions or unexpected variations in image metadata. For example, a symptom like "growth" without further context, or an image file with unusual metadata, can lead to ambiguous or incorrect routing decisions. This demonstrates the inherent brittleness of rule-based systems when faced with real-world variability, a key challenge in medical AI.[1] The instruction to acknowledge if the prototype shows the feature "is not as effective as you expected" highlights a critical point. This encourages a critical assessment of the brittleness of AI components when faced with real-world variability, a key challenge in medical AI.[1] Current AI models can be brittle, excelling at pattern recognition within their training distribution but lacking generalization capabilities when faced with

novelty or data shifts.[1] The prototype evaluation, by testing on "challenging cases," directly probes this brittleness in the routing mechanism, allowing for a realistic assessment of its limitations and guiding future improvements.

4.4 Proposed Improvements and Future Work for the Prototype

Based on the evaluation, several improvements and future work directions are proposed for the Gating Mechanism/Router prototype:

- **Enhancing Routing Logic:** The most significant improvement would be to transition from a purely rule-based system to a learned classification model for routing. This could involve using a lightweight LLM fine-tuned on routing examples, which would allow for better generalization and handling of ambiguity in symptom descriptions. Incorporating feedback loops to allow the router to learn from misclassifications or human overrides would further enhance its adaptability.
- **Improved Input Standardization:** Developing more robust parsing and normalization techniques for heterogeneous input data is crucial, especially for free-text symptoms and varied image metadata. This could involve advanced NLP techniques for symptom analysis and more sophisticated DICOM tag parsing and validation.
- **Uncertainty in Routing:** Exploring methods for the Gating Mechanism to quantify its own confidence in routing decisions would be valuable. This could flag cases where the router itself is uncertain about the correct expert, signaling the need for human intervention or a broader consultation within the system.
- **Integration with Real Experts:** Future work would involve integrating the Gating Mechanism with actual fine-tuned LLM and VLM experts, moving beyond simulated pathways. This would allow for end-to-end testing of the orchestrated system.
- **Scalability Testing:** Conduct stress tests to evaluate the router's performance under high load and with a larger number of expert models, simulating real-world clinical demands.

4.5 Prototype Demonstration Summary

A 3-5 minute video demonstration will visually showcase the Gating Mechanism/Router prototype in action. The video will present various input scenarios, such as a textual symptom combined with image metadata. It will then clearly demonstrate the router's decision-making process and the subsequent activation of the corresponding simulated expert pathway. The demonstration will highlight the technical challenge involved in building this central coordination component and its impact on the overall AI GP Doctor project. It will also include a concise explanation of the motivation behind prototyping the Gating Mechanism, its significance to the overall project, and a brief discussion of the demonstrated feasibility and future potential.

IX. Conclusion and Recommendations

This report has outlined a comprehensive plan for developing an orchestrated multimodal AI system for medical diagnosis, leveraging a Mixture of Experts (MoE) architecture. The

proposed system integrates specialized, fine-tuned Large Language Models (LLMs) for symptom analysis and Vision-Language Models (VLMs) for image interpretation. A gating mechanism directs input data to the most relevant experts, and a diagnostic integrator synthesizes their outputs using ensemble methods, providing a final diagnosis with an associated confidence score. Rigorous evaluation encompassing performance, robustness, fairness, explainability, and uncertainty quantification is central to the plan, alongside proactive consideration of regulatory requirements.

The potential impact of such a system is significant. By effectively integrating diverse data sources, it promises to enhance diagnostic accuracy, improve efficiency in clinical workflows, potentially reduce diagnostic errors, and provide valuable decision support for healthcare professionals, ultimately contributing to better patient care.[1]

Building this sophisticated AI diagnostic system requires careful planning and execution. Based on the analysis presented, the following key recommendations are proposed:

1. **Prioritize Architecture Design:** Invest significant effort early in the design of the core MoE components, particularly the Gating Mechanism/Router (to ensure accurate expert selection) and the Diagnostic Integrator (to effectively fuse multimodal evidence). The performance of these coordinating modules is critical to the overall system's success.[1]
2. **Strategic Model Selection & Adaptation:** Select pre-trained foundation models based on their alignment with specific tasks and data domains (e.g., ClinicalBERT for clinical text, appropriately chosen VLMs for target image modalities). Leverage Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA for efficient specialization of these models into domain experts, balancing performance with computational resources.[1]
3. **Invest in Data Quality and Governance:** Recognize that data is the foundation. Prioritize acquiring diverse, representative, and well-annotated multimodal datasets with comprehensive metadata. Implement rigorous, clinically informed preprocessing pipelines and robust data management practices, ensuring patient privacy and ethical data use. Address the challenge of missing metadata proactively.[1]
4. **Implement Holistic Evaluation:** Adopt the comprehensive evaluation framework from the outset, systematically measuring performance, OOD robustness, fairness across subgroups, UQ calibration, and XAI plausibility. Do not rely solely on standard benchmark scores, which may not reflect real-world reliability. Incorporate rigorous statistical analysis, including reporting confidence intervals.[1]
5. **Integrate UQ and XAI by Design:** Build uncertainty quantification and explainability into the system from the start, not as afterthoughts. Ensure the system provides clinicians with both calibrated confidence scores and interpretable explanations to foster trust and enable informed decision-making.[1]
6. **Develop a Proactive Regulatory Strategy:** Engage with regulatory bodies like the FDA early through mechanisms like Pre-Submission meetings. Design the development process and documentation to align with FDA guidance on AI/ML devices, TPLC, and risk management. Explore the use of PCCPs, potentially facilitated by the modular MoE architecture.[1]

7. **Foster Cross-Functional Collaboration:** Assemble and maintain a collaborative team comprising AI/ML engineers, data scientists, clinicians (radiologists, dermatologists, relevant specialists), ethicists, and regulatory experts.[1] Continuous communication and shared understanding across disciplines are essential for success.

In conclusion, the development of an orchestrated multimodal AI diagnostic system represents a complex but achievable goal. By adopting a structured approach grounded in rigorous research, leveraging specialized expert models within an MoE framework, prioritizing data quality and comprehensive evaluation, and proactively addressing ethical and regulatory considerations, it is possible to create powerful AI tools that can meaningfully augment clinical practice and improve patient outcomes. The key lies in a meticulous, collaborative, and iterative process focused consistently on clinical validity, safety, reliability, and trustworthiness.

Works cited

1. accessed on January 1, 1970,