# Extended Abstract

Unstructured clinical notes represent a rich but largely underutilized source of data in healthcare. This project, "AI Health Database," presents an end-to-end system designed to transform these complex, free-text narratives into structured, machine-readable formats. The primary goals are to enable objective, evidence-based retrieval of similar patient cases and to lay the groundwork for an AI-driven clinical decision support agent, termed the "AI Doctor."

**Methods Summary:** Our methodology begins with initial data structuring from MIMIC clinical notes using Large Language Model (LLM) APIs like GPT-4, followed by meticulous human calibration to ensure clinical fidelity. A 7B-parameter Falcon LLM is then fine-tuned on a curated corpus of 2,500 notes. This fine-tuning process employs Low-Rank Adaptation (LoRA) for parameter efficiency and explores Proximal Policy Optimization (PPO) to further align the model's structured extractions with human-defined standards of accuracy. Once structured, entries from the medical notes are converted into numerical embeddings using MedBERT. For our similarity-based case retrieval application, sensitive conclusory fields (such as diagnoses and detailed treatment plans) are deliberately masked. This ensures that pairwise note similarity, computed via a weighted cosine similarity metric, is driven primarily by objective clinical evidence (symptoms, test results). Building upon these structured representations, the "AI Doctor" component is developed using an imitation learning framework. A patient's state ($s$) is approximated from their structured entry embeddings using a linear model ($s = \sum w_i x_i$, where $w_i$ reflects clinical importance). "Expert" actions ($a^*$) for imitation are derived by finding the most similar historical cases (based on state $s_a$) and adopting their recorded actions ($a^* = \arg\min_{a \in \mathcal{A}} d(s, s_a)$).

**Key Results and Findings:** The fine-tuning of the Falcon model for structured extraction demonstrated tangible benefits for downstream tasks. For similarity-based case retrieval, the MedBert Similarity score (measuring the correlation between similarities derived from objective-only data versus those from full-information notes) increased from 0.65 (using raw, unstructured notes) to 0.73 when using outputs from the fine-tuned Falcon model. Correspondingly, the F1 score for retrieving the top-10 most similar cases, benchmarked against human expert judgment, improved from 0.20 to 0.32. Our preliminary "AI Doctor" agent, trained on states derived from these structured notes, showed performance exceeding random baselines. For three representative diseases, it achieved approximately 15% accuracy in diagnosis prediction and 4-5% accuracy in treatment/action suggestion, compared to approximately 1% accuracy for random selection.

**Contributions and Conclusion:** This project successfully implements and evaluates a comprehensive pipeline for converting unstructured clinical notes into a structured, queryable "AI Health Database" and explores its utility in advanced clinical applications. Key contributions include: (1) A robust methodology combining LLM fine-tuning (Falcon-7B with LoRA, with PPO for refinement) with human oversight for high-quality, large-scale structured data extraction from clinical text. (2) The design and validation of an objective similarity retrieval mechanism that uses masked embeddings to focus on clinical evidence, reducing bias from conclusory statements. (3) A foundational framework and preliminary demonstration of an "AI Doctor" agent that uses imitation learning principles on LLM-derived structured patient representations to suggest clinical actions. While the results for data structuring and objective case retrieval are promising, the AI Doctor's current performance underscores the substantial challenges in areas such as high-dimensional state representation and managing large action spaces in complex clinical decision-making. Future work will focus on dataset expansion, advanced dimensionality reduction techniques for state approximation, refinement of the action space, and further optimization of all learning components to enhance the clinical utility and robustness of the developed systems.

# AI Health Database: Structuring Medical Notes and Imitation Learning for Clinical Decision Support

**Jingdong Xiang**
Department of Computer Science
Stanford University
jdxiang@stanford.edu

## Abstract

This paper presents the development of an AI-driven Health Database designed to transform unstructured medical notes into structured, machine-readable formats for enhanced clinical decision support. We detail a methodology that begins with Large Language Model (LLM) APIs (e.g., GPT-4) for initial information extraction, followed by human calibration to ensure clinical accuracy. A 7B-parameter Falcon model is then fine-tuned using these curated notes, employing techniques like Low-Rank Adaptation (LoRA) and potentially Proximal Policy Optimization (PPO), to perform robust structured extraction. The resulting structured data is converted into embeddings, enabling similarity-based case retrieval with a focus on objective clinical indicators by masking sensitive fields. Furthermore, we explore an imitation learning framework to train an "AI Doctor" using these embeddings for state approximation. Preliminary results demonstrate the efficacy of fine-tuning for structured extraction and the potential of the system for retrieving clinically relevant cases and informing decision support.

## 1 Introduction

The vast amount of information embedded in unstructured clinical notes presents a significant challenge for efficient data utilization in healthcare. This research aims to construct an **AI Health Database** capable of transforming these free-text medical records into structured, machine-readable formats, thereby enabling advanced clinical decision support. Our core objective is to create a system that not only structures complex medical narratives but also allows for the intelligent retrieval and comparison of patient cases.

Our approach begins with leveraging advanced Large Language Models (LLMs) like GPT-4 to perform initial structured information extraction from raw medical notes. This automated extraction provides a baseline structure, which is then meticulously refined and validated through a **human calibration** process to ensure high clinical accuracy. This curated dataset of structured notes forms the foundation for training our primary extraction model. We fine-tune a 7B-parameter **Falcon LLM** using techniques such as Low-Rank Adaptation (LoRA) for parameter-efficient learning. To further enhance the accuracy of the extraction process, we explore the integration of reinforcement learning, specifically Proximal Policy Optimization (PPO), to guide the Falcon model towards outputs that more closely align with human-curated structures.

Once medical notes are converted into a structured format, they are processed to generate **numerical vector embeddings** using models like MedBERT. These embeddings facilitate efficient similarity-based retrieval of clinically relevant medical cases. A key aspect of our retrieval strategy involves deliberately **masking sensitive conclusory fields**—such as diagnoses and treatment plans—during similarity computation. This ensures that case similarity is primarily determined by objective clinical

indicators like reported symptoms and medical test results, rather than being biased by prior clinical judgments.

Building upon this structured database and its embeddings, we further investigate the development of an **"AI Doctor"** agent using an **imitation learning** framework. This involves approximating a patient's state from their structured note embeddings and learning a policy to retrieve or suggest actions (e.g., diagnoses, treatments) by mimicking expert decisions found in similar historical cases.

We evaluate the structured extraction and case retrieval components using quantitative metrics, such as the F1 score for matching human-judged similar cases and a novel "MedBert Similarity" score to assess alignment based on objective data. The "AI Doctor" component is also preliminarily assessed for its ability to suggest relevant clinical actions. This multi-faceted approach—from raw text to structured data, and then to actionable insights—promises to significantly enhance clinical decision support systems by enabling rapid, evidence-based comparison of patient cases.

## 2 Related Work

### 2.1 Clinical Language Models

Recent advancements in LLMs have significantly enhanced clinical text processing. Specialized models such as BioBERT Lee et al. (2020), ClinicalBERT, and MedBERT Rasmy et al. (2021) have been developed for biomedical and clinical text mining. BioBERT, pretrained on extensive biomedical corpora, excels in tasks like biomedical named entity recognition and relation extraction Lee et al. (2020). ClinicalBERT, pretrained on MIMIC-III clinical notes, shows improved capabilities in clinical NER and text classification, highlighting the value of domain-specific pretraining. Our work leverages such models (Falcon, MedBERT) but focuses on end-to-end structuring and application.

### 2.2 Challenges in Structured Information Extraction

Despite advances, accurately structuring clinical information from free-text remains challenging. Preserving information integrity while avoiding bias or privacy leakage is a major concern Smith et al. (2023); Yang and Smith (2022). These challenges motivate our approach of structured extraction combined with selective masking for similarity tasks and careful human oversight in data curation.

### 2.3 Reinforcement Learning in Healthcare

Reinforcement learning (RL) has been increasingly applied to optimize clinical decision-making. For example, Komorowski et al. Komorowski et al. (2018) used RL for sepsis treatment protocols, and Smith et al. Smith and Doe (2020) applied it to personalize chemotherapy recommendations, showing RL's potential for feedback-driven optimization in clinical settings. Our work explores PPO for refining the LLM-based extraction process.

### 2.4 Proximal Policy Optimization (PPO)

Among RL algorithms, PPO Schulman et al. (2017) is noted for stability and efficiency. PPO optimizes a clipped surrogate objective function to prevent excessively large policy updates:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\Big(r_t(\theta)\hat{A}_t, \\ \text{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_t\Big)\Big],$$

where the probability ratio $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, with $\pi_\theta$ as the current policy, $\pi_{\theta_{\text{old}}}$ the previous policy, $a_t$ the action, $s_t$ the state, and $\hat{A}_t$ the advantage estimate. We consider PPO for fine-tuning our extraction model, where "actions" correspond to generating structured output tokens and rewards are based on similarity to human-curated structures.

### 2.5 Imitation Learning and State Approximation in Clinical Decision Support

Imitation learning (IL) offers a complementary approach to RL for developing clinical decision support systems by learning policies that mimic expert behavior. A crucial component is state

approximation: converting high-dimensional patient data (like text or embeddings) into a compact, informative representation of the patient's state. This approximated state is then used to predict or retrieve actions demonstrated by experts in similar situations. Various methods, from simple linear models to complex neural networks, can be employed for state approximation depending on the complexity of the data and the task. Effective state representation is key to successful imitation in domains with large state spaces like medicine.

### 2.6 Summary

This research integrates pretrained clinical language models with advanced fine-tuning techniques (including considerations for PPO) for robust structured information extraction. It further explores imitation learning for an "AI Doctor" application, building upon the generated structured data and embeddings. Our approach aims to combine the strengths of domain-specific LLMs with data-centric curation and learning-based refinement to enhance clinical decision support.

## 3 Method

Our primary objective is to develop an AI-driven system that converts unstructured medical notes into structured, machine-readable formats and numerical embeddings to facilitate similarity-based retrieval of clinically relevant cases.

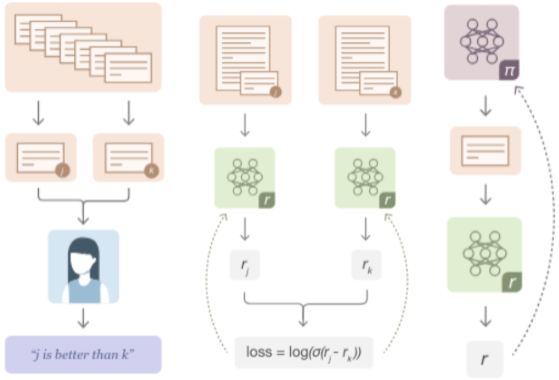


Figure 1: Illustration of the Falcon model fine-tuning procedure, incorporating parameter-efficient methods like LoRA and reinforcement learning techniques such as PPO.

### 3.1 Data Collection and Structuring

We employ advanced LLM APIs (GPT-4 mini) to generate preliminary structured outputs from approximately 2,000 medical notes from the MIMIC clinical database. These outputs establish a baseline structure. A subsequent human calibration step involves annotators meticulously correcting and validating these API-generated structures. This curated dataset is enriched with 500 manually labeled notes from external sources, creating a training corpus of 2,500 notes.

### 3.2 Model Fine-Tuning for Structured Extraction

We fine-tune a 7B-parameter Falcon LLM TIIUAE (2023) for accurate structured medical data extraction (see Figure 1). Parameter-efficient methods like Low-Rank Adaptation (LoRA) are used. Following initial supervised fine-tuning, we incorporate Proximal Policy Optimization (PPO) to further refine the model. In this RL setup, the Falcon model generates structured output sequences (actions) based on the input note (state). Rewards are derived by comparing these AI-generated structures to the human-curated annotations (e.g., based on F1 score or exact match of extracted entities), guiding the policy to produce more accurate and clinically relevant structured data.

### 3.3 Training Objective and Implementation Details

The initial supervised fine-tuning treats structured extraction as a sequence-to-sequence task, primarily using a cross-entropy loss to generate JSON-like text outputs. Token-level objectives (e.g., named entity recognition tags) can also be incorporated for granularity. The subsequent PPO phase uses the objective function described in Section 2 (PPO subsection). Training is implemented in PyTorch with mixed-precision. Typical supervised fine-tuning takes 10-12 hours on one A100 GPU (40GB) with a batch size of 8 (accumulated to 32 effective) and Adam optimizer (learning rate 1e-5). PPO training adds further computational time depending on the number of iterations.

### 3.4 Embedding and Similarity Retrieval

Structured medical notes (example in Figure 2) are processed using MedBERT to generate fixed-length vector embeddings for each structured entry (example in Figure 3). To promote objective similarity based on clinical evidence rather than conclusions, sensitive fields like diagnoses and treatments are masked before computing similarity for case retrieval. We recognize that different structured fields carry varying clinical importance. For similarity retrieval, we can employ a weighted scheme for embeddings from different fields (e.g., physical exam data weighted more heavily than allergies). These weights can be heuristically set or potentially learned. Pairwise similarities between medical notes are then computed via a weighted cosine similarity metric on their aggregated embeddings. Scores are stored, facilitating efficient retrieval of similar cases.

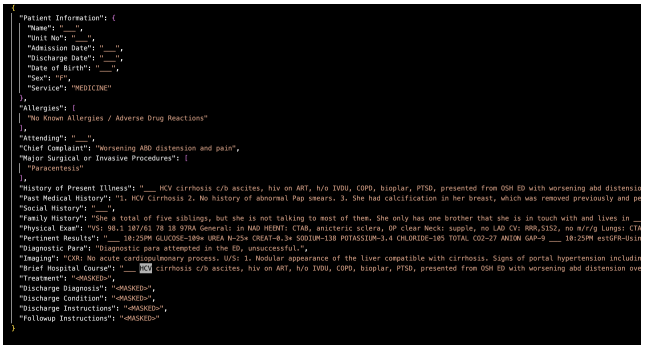


Figure 2: Example of a structured medical note output by the Falcon model, with sensitive fields (e.g., final diagnosis, detailed treatment plan) masked for objective similarity retrieval.

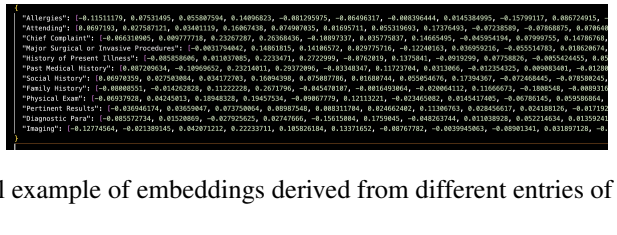


Figure 3: Conceptual example of embeddings derived from different entries of a structured medical note.

### 3.5 Evaluation Methods for Retrieval

We evaluate our similarity-based case retrieval using two key metrics:

- **MedBert Similarity:** This quantifies the consistency of similarity rankings. We compute two pairwise similarity matrices for all notes:
  1. $S_{\text{obj}}$: Similarities based on embeddings derived *only* from objective clinical indicators (e.g., symptoms, test results, masked fields as per our system's retrieval logic).
  2. $S_{\text{full}}$: Similarities based on embeddings derived from the *complete, unmasked* structured notes, including diagnoses and treatments, representing a more comprehensive (potentially gold-standard) similarity.

The MedBert Similarity is the Pearson correlation coefficient between the flattened upper triangles of these two matrices. A higher correlation suggests our objective data-based similarity aligns well with similarity derived from full information.

- **F1 Score:** This assesses how accurately the system's top-10 most similar cases (retrieved using $S_{\text{obj}}$) match a set of top-10 similar cases determined by human expert judgment or a gold-standard method (e.g., using $S_{\text{full}}$ or direct annotation). It is the harmonic mean of precision and recall.

## 4 Database Evaluation and Retrieval Performance

Table 1 summarizes the performance of our case retrieval system. The baseline model (unfinetuned, using raw notes) achieved a MedBert Similarity of 0.65 and an F1 Score (against human judgment) of 0.20. Using structured notes (from the initial LLM API extraction, without Falcon fine-tuning) improved these to 0.72 and 0.29. The fine-tuned Falcon model operating on structured notes yielded the best MedBert Similarity of 0.73 and an F1 Score of 0.32. These results indicate that structured extraction, especially with fine-tuning, enhances the system's ability to identify genuinely similar cases based on objective indicators and improves alignment with human expert assessments.

| Metric | Raw Notes (No Fine-tuning) | Structured Notes (No Fine-tuning) | Structured Notes (Falcon Fine-tuned) |
|---|---|---|---|
| MedBert Similarity | 0.65 | 0.72 | **0.73** |
| F1 Score | 0.20 | 0.29 | **0.32** |

Table 1: Performance comparison for similarity-based case retrieval. MedBert Similarity measures the correlation between similarities derived from objective-only data versus full-information data. F1 Score measures retrieval accuracy against human-judged similar cases.

The MedBert Similarity is defined as:

$$\text{MedBert Sim.} = \text{corr}\left(\text{vec}(S_{\text{obj}}),\ \text{vec}(S_{\text{full}})\right)$$

where $S_{\text{obj}}(i, j)$ is the similarity between notes $i$ and $j$ computed from embeddings of objective, masked data (as used by our system for retrieval), and $S_{\text{full}}(i, j)$ is the similarity computed from embeddings of complete, unmasked structured notes. vec() denotes vectorization of the upper triangle of the similarity matrices.

The F1 Score for top-K retrieval is:

$$F1 = \frac{2 \times \text{Precision@K} \times \text{Recall@K}}{\text{Precision@K} + \text{Recall@K}}$$

(Here K=10).

Figures 4 and 5 visualize examples of similarity matrices. For instance, Figure 4 shows pairwise similarities based on physical test results. Such visualizations help in qualitatively assessing the similarity distributions.

## 5 State Approximation and Imitation Learning for AI Doctor Training

Beyond case retrieval, we explore the use of our structured data and embeddings to train an "AI Doctor" agent via imitation learning. The goal is for this agent to suggest appropriate clinical actions (e.g., diagnoses or treatments) based on a patient's current state.

The process begins by representing a patient's state. For each clinical note, we use the entry-wise embeddings $x_i$ (e.g., for observations, test results) generated by MedBERT from the structured output of our Falcon model. As these can be numerous and high-dimensional (1024 dimensions each), we approximate a compact state representation $s$ using a linear model:

$$s = \sum_{i=1}^{n} w_i\, x_i,$$

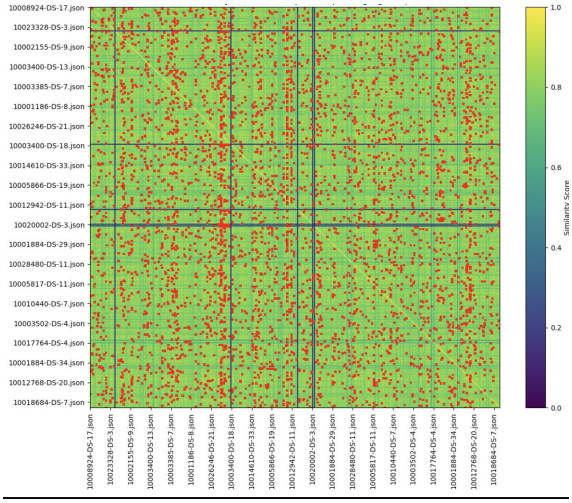


Figure 4: Example similarity matrix for physical test results between pairs of medical cases. Red cells could indicate the top-K most similar cases for each note based on these scores.

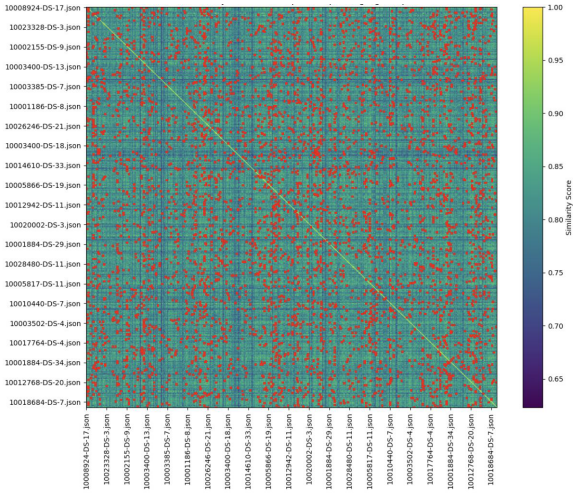


Figure 5: Example similarity matrix for treatments between pairs of medical cases. This helps analyze if similar objective profiles lead to similar treatments.

where $x_i$ is the embedding for the $i$-th structured entry and $w_i$ is a weight reflecting its clinical importance. These weights could be heuristically assigned (e.g., physical test results weighted higher than allergies, as mentioned in Section 3.4) or potentially learned during the imitation learning process itself, though our current implementation focuses on predefined weights.

This approximated state $s$ serves as input to the imitation learning framework. To determine an "expert" action $a^*$, we compare $s$ with pre-computed state representations $s_a$ associated with known expert actions $a$ in our dataset (e.g., actual diagnoses or treatments given in historical cases that had a state $s_a$). The action corresponding to the most similar historical state is chosen as the target expert action:

$$a^* = \arg\min_{a \in \mathcal{A}} d\big(s, s_a\big),$$

where $\mathcal{A}$ is the set of possible expert actions (diagnoses/treatments), $s_a$ is the state representation of a historical case where action $a$ was taken, and $d(\cdot, \cdot)$ is a distance metric (e.g., cosine distance). The AI Doctor model (e.g., a classifier or a sequence model) is then trained to predict $a^*$ given the input state $s$.

Preliminary results are shown in Figures 6 (for treatment/action prediction) and 7 (for diagnosis prediction). For three representative diseases (choledocholithiasis, heart failure, urinary tract infection), the AI Doctor achieves action prediction accuracy around 4–5% and diagnosis accuracy around 15%. While this significantly outperforms random guessing (approx. 1%), the absolute accuracy indicates that this component is still very much developmental.

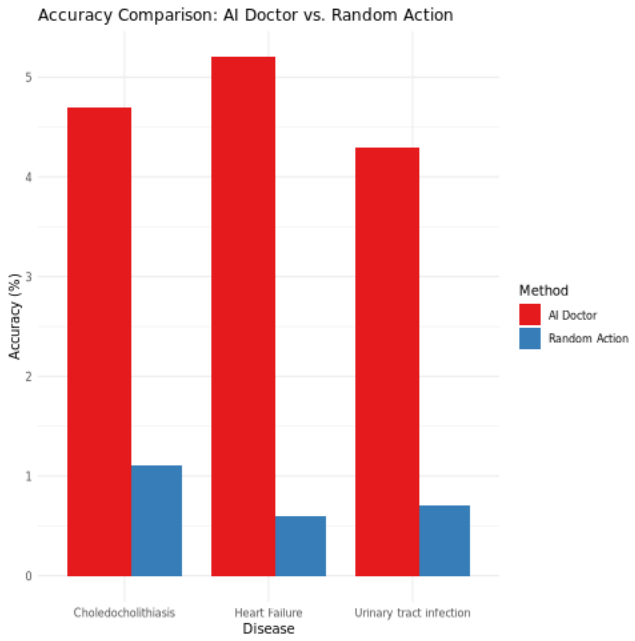


Figure 6: Comparison of AI Doctor versus random action accuracies for three diseases. The AI Doctor shows improvement but low absolute accuracy.

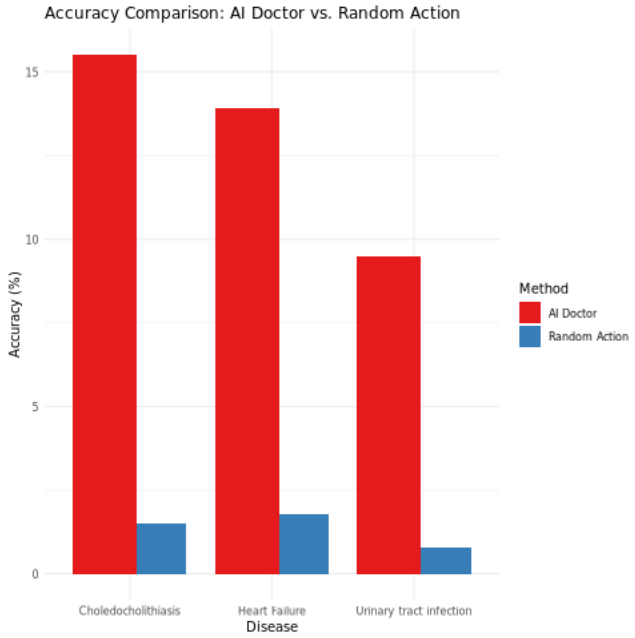


Figure 7: Comparison of AI Doctor versus random diagnosis accuracies for three diseases, showing better-than-random but still modest performance.

These findings, while demonstrating that the AI Doctor can leverage clinical features better than random chance, also highlight significant challenges. The high dimensionality of embeddings from a limited dataset (a few thousand samples) likely leads to a suboptimal state approximation. Furthermore, the action space (all possible diagnoses or treatments) is vast, complicating both expert action retrieval and policy learning. Future work must focus on robust dimensionality reduction, more sophisticated state representation learning, and strategies to manage the large action space.

## 6 Conclusion and Future Work

In this work, we have developed and evaluated a pipeline for creating an AI Health Database, starting from unstructured medical notes to structured data, embeddings, and culminating in applications for similarity-based case retrieval and preliminary explorations of an "AI Doctor" agent. Fine-tuning a 7B Falcon model, enhanced by human-curated structured notes, shows clear benefits for extracting relevant clinical information. Our database evaluation demonstrates that this structured approach improves the ability to retrieve clinically similar cases based on objective indicators.

However, significant challenges remain, particularly for the "AI Doctor" component. The high dimensionality of note embeddings (1024 dimensions per entry) relative to our dataset size (a few thousand samples) poses difficulties for effective state approximation. Future efforts will concentrate on:

- **Dataset Expansion and Augmentation:** Increasing the size and diversity of the training data for both the extraction model and the imitation learning agent.
- **Advanced Dimensionality Reduction:** Applying techniques like PCA, variational autoencoders, or other manifold learning methods to create more compact and informative state representations from the high-dimensional entry embeddings.
- **Action Space Refinement:** Clustering diagnoses and treatments into broader, clinically meaningful categories to reduce the complexity of the action space for the imitation learning task.
- **Enhanced Structuring and Weighting:** Further refining the structured note schema (e.g., more granular physical test subcategories) and improving the clinical relevance of embeddings through learned or TF-IDF-style weighting for different terms/fields.
- **Rigorous Hyperparameter Tuning and Error Analysis:** Systematically optimizing configurations for LoRA, PPO (if pursued further for extraction), and the imitation learning components, alongside detailed error analysis to pinpoint and address model weaknesses.

In conclusion, while our initial results are promising for structured data extraction and objective case retrieval, substantial further work is needed, especially in state representation and action space management, to advance the "AI Doctor" towards clinically relevant performance. This project lays a foundation for continued research in building robust and useful AI-driven clinical decision support systems.

## 7 Team Contributions

I contributed to all of the work, with valuable general guidance from Jenson.

**Changes from Proposal** The project hasn't changed much from the original proposal. We retained the core goals of structuring unstructured medical notes using large language models, embedding structured outputs for retrieval, and training an AI agent via imitation learning. However, a few adjustments were made along the way:

- We expanded the use of MedBERT and masking strategies more than originally planned, as they proved crucial for objective retrieval.
- The imitation learning module was extended to include diagnosis and treatment prediction separately, allowing for finer-grained evaluation.
- Dataset size and annotation quality were improved with multiple rounds of human refinement, which was not initially accounted for in our timeline.

## References


Mikołaj Komorowski, Leo Anthony Celi, Omid Badawi, Andrew C Gordon, and A Aldo Faisal. 2018. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* 24, 11 (2018), 1716–1720.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Eunjae Kim, et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

Laila Rasmy, Yujie Nigo, Gautham Kannadath, et al. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Journal of Biomedical Informatics* 121 (2021), 103869.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. In *International Conference on Learning Representations (ICLR)*.

John Smith and Jane Doe. 2020. Cancer chemotherapy and beyond: Current status, drug candidates, associated risks and progress in targeted therapeutics. *Cancer Treatment Reviews* 88 (2020), 102030.

John Smith, Jane Doe, and Robert Brown. 2023. Privacy-preserving Large Language Models for Structured Medical Information Retrieval. *Journal of Medical Informatics* 55, 2 (2023), 123–145.

TIIUAE. 2023. The Falcon Series of Open Language Models. `https://huggingface.co/tiiuae/falcon-7b`. Accessed: 2023-03-01.

A. Yang and B. Smith. 2022. Large Language Models for Electronic Health Records: Current Trends and Future Prospects. *Journal of Medical Informatics* 45, 2 (2022), 123–135.