Project Proposal: LLM-Based Medical Advice Recommender System

Name:Tianyou Xu Zid:z5469582

1. Scope

1.1 Domain & Intended Users

The recommender system will operate in the primary healthcare domain, acting as an intelligent assistant. Its target users are twofold:

1) Patients/General Users: Individuals seeking to understand potential health issues, explore lifestyle modifications, or learn about preventive care options based on their stated health data. The system aims to empower them to have more informed discussions with their healthcare providers.
2) Healthcare Providers: Clinicians, nurses, and medical students can use the system as a decision-support tool to quickly access tailored information, differential diagnoses, or patient education materials based on a patient's profile.

The system's core function is to bridge the gap between complex medical knowledge and individual health needs, providing personalized recommendations for non-emergency situations.

1.2 User Interface, Interaction, and Ethical Considerations

The primary user interface will be a secure, responsive web-based platform, with a mobile-first design philosophy to ensure accessibility for a future mobile app extension.

1) Recommendation Presentation: At the end of each query session, the user will be presented with a curated list of 3 to 5 distinct recommendations. This number is chosen to provide sufficient options without overwhelming the user. Each recommendation will be presented as a "card" containing a clear title, a brief summary, and a "Why this is recommended for you" section that links the advice back to the user's specific inputs.
2) User Interaction Flow:
   - Onboarding & Consent: A first-time user registers and is presented with a clear privacy policy and terms of service. They must provide explicit consent for their anonymized data to be used for model improvement.
   - Profile Input: The user is guided through a structured form to input their health profile, including demographics, current symptoms (with options for severity and duration), known medical history (e.g., chronic conditions, allergies), current medications, and lifestyle factors (diet,

exercise).

- Recommendation Generation: The system processes the input and displays the 3-5 recommendation cards on a results page.
- Feedback Loop: Each card has "Helpful" and "Not Helpful" buttons. Clicking either one can trigger a feedback modal asking for a quick, optional reason (e.g., "Already tried," "Not relevant," "Unclear," "Concerns about side effects"). This qualitative data is invaluable for model tuning.

3) Ethical Safeguards & Disclaimers: A persistent, prominent disclaimer will state: "This tool provides information for educational purposes only and is not a substitute for professional medical advice, diagnosis, or treatment. Always seek the advice of your physician or other qualified health provider with any questions you may have regarding a medical condition." The system will be programmed to detect keywords related to medical emergencies (e.g., "chest pain," "difficulty breathing," "severe bleeding") and immediately display a message urging the user to contact emergency services.

1.3 Mockup Description

1) Homepage: A clean, professional design with a prominent "Login" or "Register" call-to-action. Includes a brief, clear statement of the system's purpose and its limitations.
2) Input Form: A multi-step, user-friendly form. Uses controlled vocabularies (e.g., dropdowns for symptoms) where possible to structure data, but also includes free-text fields for nuanced descriptions. A progress bar guides the user.
3) Results Page: A vertically scrollable list of the 3-5 recommendation cards. Each card is expandable to show more details, including links to evidence-based sources (e.g., established medical journals, reputable health websites).
4) Feedback: A simple pop-up window with a multiple-choice question and an optional comment box to capture user feedback efficiently without causing friction.

1.4 User Feedback & Model Updates

1) Feedback: Both explicit feedback (ratings, survey responses) and implicit signals (e.g., which recommendations are clicked, time spent reading) will be collected. This data is crucial for creating a user-item interaction matrix to power collaborative filtering components.
2) Model Updating: The system will employ a continual learning strategy. User feedback will be used for periodic prompt-tuning of the LLM. The underlying knowledge base of the model will be updated quarterly to incorporate new medical guidelines and research, ensuring continued accuracy.
3) Cold Start Problem:
   - New Users: The cold start problem for new users is mitigated by the

structured input form. The initial set of recommendations will be purely content-based, relying entirely on the user's detailed profile.

- New Items: When new treatments or guidelines emerge, they are initially recommended based on knowledge-based rules and content-based matching before sufficient user interaction data is available.

1.5 Business Model

1) Freemium Subscription Model:
   - Free Tier: Basic recommendation services for individual users. (like GPT or Perplexity)
   - Premium Tier: A monthly/annual subscription fee for advanced features like saving and tracking health history over time, generating detailed summary reports to share with doctors, and integrating with wearable device data (e.g., Fitbit, Apple Watch).
2) B2B (Business-to-Business) Services:
   - Licensing to Telehealth Platforms: Integrating the recommender system as a feature within existing telehealth provider services to assist their clinicians.
   - Partnerships with Clinics: Providing the tool to clinics and hospitals to enhance patient education and engagement, potentially integrating with their Electronic Health Record (EHR) systems for seamless data flow.

2. Datasets

2.1 Candidate Datasets & Statistics

To train and evaluate a robust model, a combination of structured clinical data and conversational medical data is required. All data are collected via hugginface (https://huggingface.co/datasets).

1) MIMIC-IV [1]: This is a Hugging Face version of the large, de-identified MIMIC-IV database from the Beth Israel Deaconess Medical Center. It contains data related to 301,901 patient admissions in intensive care units. This version is structured into multiple tables (e.g., admissions, diagnoses_icd, labevents, prescriptions, chartevents). It provides a deep, structured view of patient journeys in a critical care setting, essential for learning complex relationships between diagnoses, treatments, and patient data.
2) HealthCareMagic-100k-en [2]: This dataset contains 112,165 anonymized patient-doctor conversation pairs in English. Each entry includes a unique ID, the patient's question, and the doctor's answer. The questions cover a wide range of health topics and are often detailed and personal, making this dataset ideal for fine-tuning an LLM to understand user queries and generate advice in an appropriate, conversational tone.

3) HealthSearchQA [3]: This is a curated dataset focused on common medical questions that users search for online. It contains 4,576 question-answer pairs derived from consumer health questions. The dataset is specifically designed to evaluate how well models can find evidence-based answers to these questions. It includes features like question, answer, and evidence, making it valuable for training the system to ground its recommendations in verifiable sources.
4) VL-Health [4]: This is a multimodal dataset containing over 554,116 clinical reports, with each report associated with one or more medical images. This dataset would be reserved for a future version of the project that aims to incorporate visual data (e.g., user-submitted photos of a rash or injury) into the diagnostic and recommendation process.

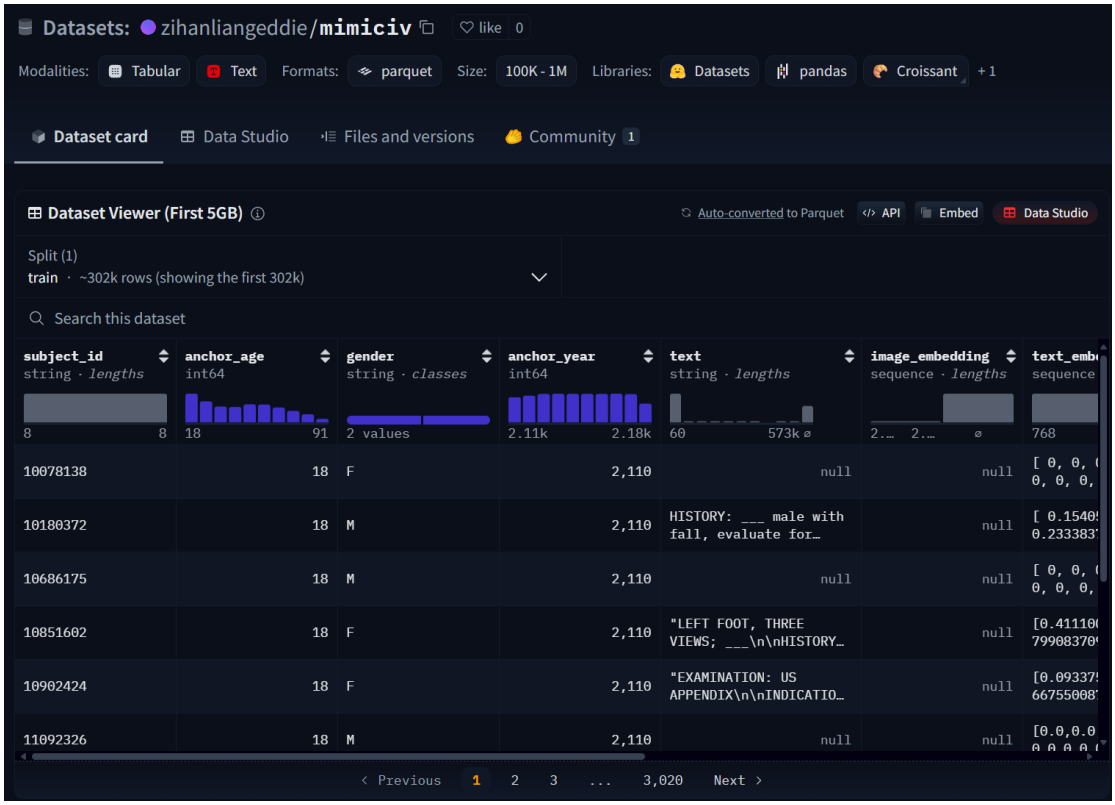

Figure 1 An example of the MIMIC-IV dataset in huggingface

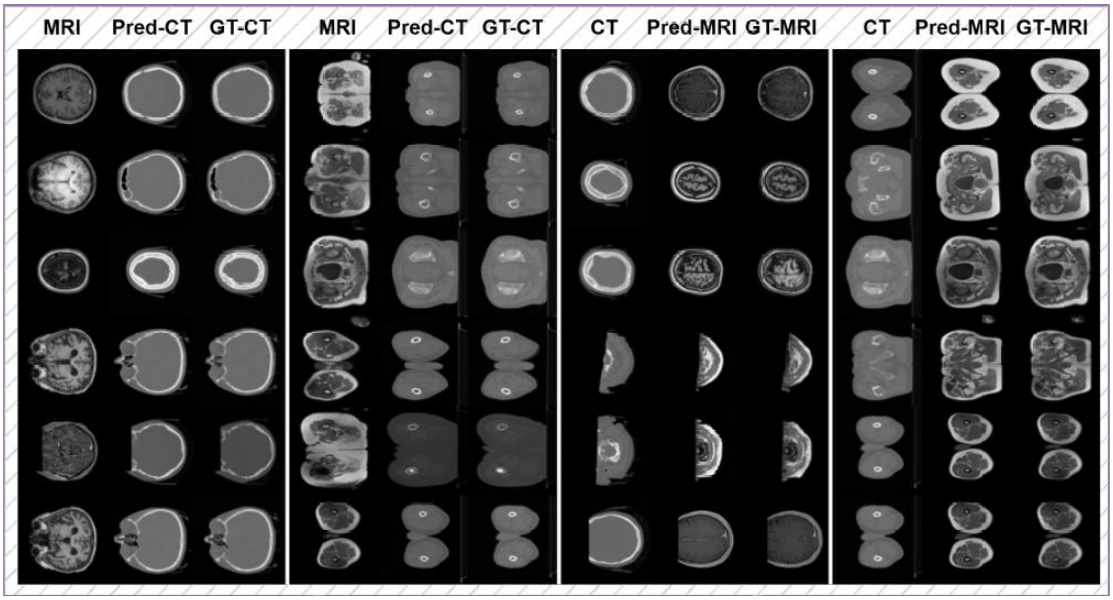

Figure 2 Images of VL-Health dataset

Dataset Name	Number of Samples	Type
MIMIC-IV	301,901	text
HealthCareMagic-100k-en	112,165	text
HealthSearchQA	4,576	text
VL-Health	554,116	text & image

Table 1 Statistics of the collected datasets

2.2 Fields Used

The model will utilize the following data fields from these specific datasets:

1) From MIMIC-IV: subject_id, hadm_id (from admissions), icd_code, icd_version (from diagnoses_icd); drug, dose_val_rx (from prescriptions); and text from clinical notes tables. This structured data will form the basis of the knowledge graph and content-based filtering.
2) From HealthCareMagic & HealthSearchQA: The patient-question or question fields will be used to model user input (symptoms, history). The doctor-answer or answer fields will serve as target outputs for fine-tuning the LLM's advice generation.
3) From User Interaction: Explicit user feedback (ratings), implicit signals (clicks, time-on-page), and the user-provided profile information collected via the input form.

2.3 Limitations of Datasets

While powerful, these datasets have inherent limitations:

1) Limited Scope: MIMIC-IV is from a single hospital system's ICU and may

not represent general primary care scenarios or different demographics. The QA datasets are limited to the questions users chose to ask and the specific doctors who answered, not a complete picture of all possible medical interactions.

2) Sanitized and Unrealistic Nature: The datasets are de-identified and cleaned, which is necessary for privacy but removes some real-world complexity. The data lacks the back-and-forth dialogue often present in a true consultation. Furthermore, HealthCareMagic data, while real, is from an online service and may not have the same diagnostic rigor as an in-person visit.
3) Predefined Task Bias: These datasets, especially when hosted on platforms like Hugging Face, are often pre-processed for specific tasks (like question-answering). This can encourage models that excel at a narrow task but may not generalize well to the holistic, multi-turn, and safety-critical nature of a real recommender system. Overfitting to the specific style of questions and answers in these datasets is a significant risk that must be mitigated through robust evaluation.

3. Methods

3.1 Approach: Hybrid LLM-Based System

The core of the system will be a hybrid model that leverages the strengths of a Large Language Model (LLM) while ensuring safety and personalization through a multi-faceted approach.

1) Content-Based Filtering via LLM Embeddings:
   The user's input (symptoms, history, preferences) and a corpus of medical documents (guidelines, research papers, QA pairs) are encoded into high-dimensional vectors (embeddings) using a pre-trained, medically-tuned LLM. The system then uses cosine similarity to find the documents and existing Q&A pairs most relevant to the user's vector, forming the basis of the recommendation.
   This allows the system to understand the semantic meaning of a user's query beyond simple keywords.
2) Collaborative Filtering with Implicit and Explicit Feedback:
   User feedback is used to build a user-recommendation interaction matrix. We can create embeddings for users based on their profiles and interaction histories. By findng clusters of "similar users" (e.g., users with similar profiles who found the same advice helpful), the system can recommend items that were successful for that peer group. This is a form of item-based or user-based collaborative filtering.
   This adds a personalization layer that goes beyond the user's stated profile, capturing latent preferences.
3) Knowledge-Based Reasoning for Safety and Grounding:
   An external medical knowledge graph (e.g., a subset of SNOMED CT or

UMLS) will be integrated. Before a recommendation generated by the LLM is shown to the user, it is validated against this graph. This "fact-checking" layer can flag potential drug interactions based on the user's medication list, verify that a recommended treatment is appropriate for a given diagnosis, and prevent the LLM from generating factually incorrect or dangerous advice.

This acts as a critical safety rail, ensuring recommendations are grounded in established medical knowledge.

3.2 Variants:

To optimize performance, several variants will be explored:

1) Variant 1: Advanced Prompt Engineering: The structure of the prompt sent to the LLM will be dynamically adjusted based on user type.

   For Patients: "*Based on the following user profile {profile}, explain potential causes and suggest three safe, evidence-based lifestyle changes. Use simple, non-technical language. Frame each suggestion with a clear action and a brief rationale.*"

   For Providers: "*For a patient with this profile {profile}, provide a ranked list of differential diagnoses and suggest corresponding first-line treatment plans, citing relevant clinical guidelines.*"

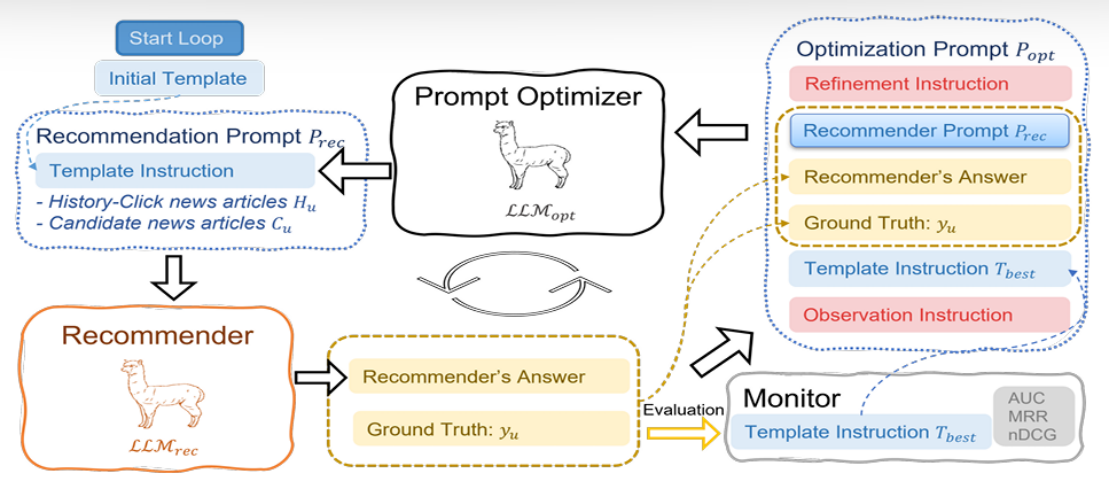

Figure 3 An example of prompt optimization, details can be seen at https://arxiv.org/pdf/2312.10463

2) Variant 2: Fine-Tuning on Medical Dialogues: The base LLM will be fine-tuned on the *HealthCareMagic* and *HealthSearchQA* datasets. This will adapt the model's language and style to be more aligned with medical conversations, improving its ability to generate relevant and empathetic responses.

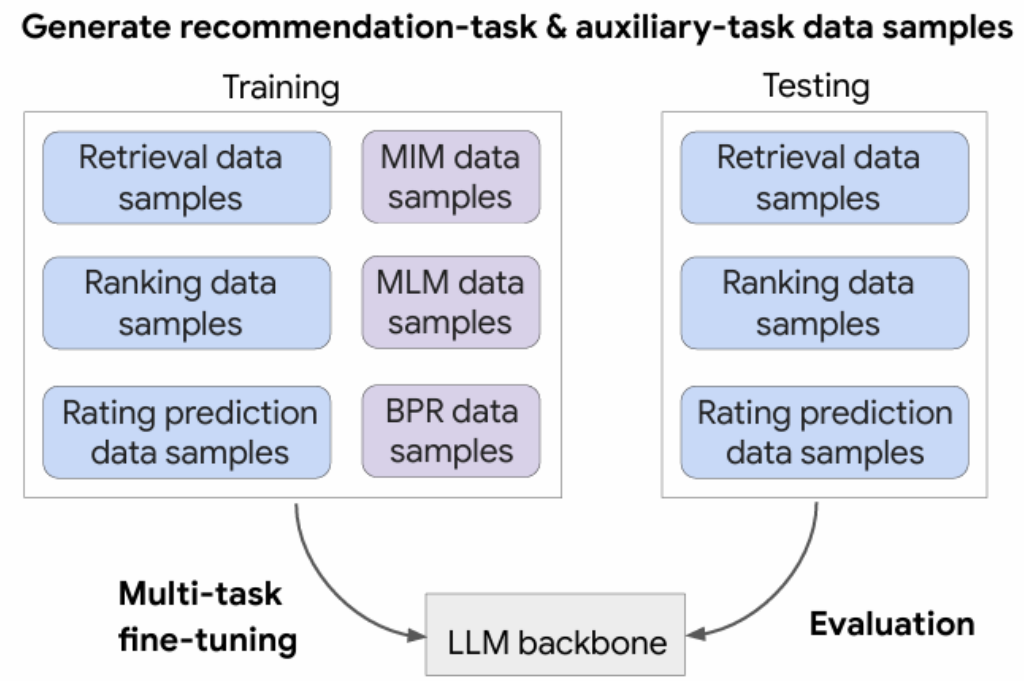

Figure 4 An example of fine-tuning, detailes can be found in https://arxiv.org/pdf/2404.00245

3) Variant 3: Ensemble Hybridization: The final ranking of recommendations will be determined by an ensemble method. This could be a weighted score that combines the relevance score from the content-based component, the personalization score from the collaborative filtering component, and a confidence score from the knowledge-based validation step. This provides a robust final output that balances multiple recommendation signals.

3.3 Justification

LLMs are uniquely suited for this task due to their ability to process and synthesize vast amounts of unstructured text (user input, medical literature) and generate human-readable explanations. This is vital for building user trust. However, LLMs can hallucinate. By hybridizing the LLM with collaborative filtering for personalization and a knowledge graph for safety, we create a system that is not only intelligent and personalized but also trustworthy and evidence-based.

4. Evaluation

4.1 Evaluation Metrics:

A comprehensive evaluation will use a mix of offline model metrics and online user-centric metrics.

Model Metrics (Historical Data):

These metrics will be calculated on a held-out test set from the datasets.

1) Precision@N: Measures the proportion of recommended items in the top-N set that are truly relevant. It answers: "Out of the N items recommended, how many were actually good?"
   Formula:
   $$Precision@N = \frac{Relevant\ items\ \cap\ Recommended\ Items\ in\ Top\ N}{N}$$
2) Recall@N: Measures the proportion of all relevant items that are successfully recommended in the top-N set. It answers: "Out of all the items that were good, how many did we manage to recommend?"
   Formula:
   $$Recall@N = \frac{Relevant\ items\ \cap\ Recommended\ Items\ in\ Top\ N}{Total\ Relevant\ Items}$$
3) Mean Reciprocal Rank (MRR): Measures the quality of the ranking. It is the average of the reciprocal of the rank at which the first relevant item was found for a set of users. It is particularly useful when only one recommendation needs to be "correct".
   Formula:
   $$MRR = \frac{1}{|U|}\sum_{u \in U}\frac{1}{rank_u}$$
   Where $|U|$ is the total number of users and $rank_u$is the rank of the first relevant item for user u.
4) Coverage: The percentage of all possible recommendations in the item catalog that the system is able to recommend. High coverage is important to ensure novelty and avoid only recommending a few popular items.

System/User Metrics (Interaction Data):

These are collected from the user study and live system monitoring.

1) User Satisfaction Score: Measured via a post-session Likert scale survey (1-5) asking users to rate their overall satisfaction with the relevance, clarity, and trustworthiness of the recommendations.
2) Engagement Rate: A measure of user interaction, calculated as the number of users who provide feedback (clicks, ratings, comments) divided by the total number of users who receive a recommendation.
3) Personalization Score: A metric to quantify how tailored the recommendations are. This can be measured by calculating the average dissimilarity (e.g., 1 minus cosine similarity) between recommendation lists for different users. A higher score indicates greater personalization.
4) Tradeoffs: The primary tradeoff is between Precision and Recall. Optimizing for Precision ensures users don't see irrelevant advice, which is key for trust. Optimizing for Recall ensures we don't miss potentially useful advice. For this safety-critical domain, Precision@N will be considered the most important model metric. For system success, the User Satisfaction Score is

paramount.

4.2 Computational Requirements:

1) Latency: Recommendation generation must be near-real time. Since a single LLM inference can be slow, the target is to return results in under 10 seconds. This can be achieved through model optimization (quantization) and efficient hardware (GPUs).
2) Throughput: The system will be deployed on scalable cloud infrastructure (e.g., AWS, GCP) using containerization (Docker) and orchestration (Kubernetes) to handle high concurrency.
3) Model Updates: Full model retraining will be computationally expensive and performed offline quarterly. Fine-tuning and prompt adjustments are less intensive and can be done weekly or bi-weekly during off-peak hours.

4.3 User Study Design:

An informal user study will be conducted to gather qualitative insights.

1) Participants: A small, diverse group of 10-15 participants, including individuals with no medical background and at least two participants with clinical experience (e.g., nursing or medical students) to provide an expert perspective.
2) Methodology:
   - Pre-Study Briefing: Participants are onboarded, the simulated nature of the tool is explained, and consent is obtained.
   - Scenario-Based Tasks: Each participant will be given 3-4 realistic but non-critical patient scenarios (e.g., "A 30-year-old office worker wants to improve their diet and manage stress," "A 55-year-old with a family history of diabetes asks for preventive advice").
   - Interaction: Participants will use the simulated web interface to input the scenario details and receive recommendations. They are encouraged to "think aloud" as they navigate the results.
   - Post-Task Questionnaire & Interview: After completing the tasks, participants will fill out a questionnaire.
3) Feedback Collection: The questionnaire and follow-up semi-structured interview will focus on:
   - Relevance and Accuracy: "How relevant were the recommendations to the scenario?"
   - Trust and Safety: "Did you trust the information? Did you have any safety concerns?"
   - Clarity and Usability: "Was the information easy to understand? Was the interface easy to use?"
   - Actionability: "How likely would you be to follow this advice or discuss it with a doctor?"
   - Open-Ended Comments: General feedback on what they liked, disliked,

and what features they felt were missing. This qualitative data will be crucial for iterating on the system's design and user experience.

5. Limitations

The most significant limitation of this project lies in the profound ethical and safety risks inherent in providing automated medical advice. While the system is intended for educational purposes, there is a substantial danger that an LLM could generate advice that is inaccurate, incomplete, or dangerously incorrect, potentially leading to delayed treatment or adverse health outcomes. This introduces complex questions of liability and accountability, as determining responsibility in the event of harm is not straightforward. Furthermore, there is a considerable risk of user over-reliance, where individuals may treat the system as a substitute for professional medical consultation, misinterpreting its recommendations or ignoring the disclaimers, thereby placing their health at risk. The system's effectiveness is also fundamentally constrained by the nature and scope of its training data. The chosen datasets, while large, are not fully representative of the intended primary care domain. Data from the MIMIC-IV ICU setting reflects acute, critical care scenarios that differ significantly from common health queries, while online question-and-answer platforms lack clinical validation and capture a self-selecting, non-representative sample of the population. This reliance on imperfect data can introduce and amplify inherent demographic, social, and clinical biases, potentially leading the system to provide less effective advice for underrepresented groups and risking the exacerbation of existing health inequities.

6. References:

[1] MIMIC-IV on Hugging Face: zihanliangeddie. "mimiciv." Hugging Face Datasets. Available at: https://huggingface.co/datasets/zihanliangeddie/mimiciv

[2] HealthCareMagic-100k-en on Hugging Face: wangrongsheng. "HealthCareMagic-100k-en." Hugging Face Datasets. Available at: https://huggingface.co/datasets/wangrongsheng/HealthCareMagic-100k-en

[3] HealthSearchQA on Hugging Face: katielink. "healthsearchqa." Hugging Face Datasets. Available at: https://huggingface.co/datasets/katielink/healthsearchqa

[4] VL-Health on Hugging Face: lintw. "VL-Health." Hugging Face Datasets. Available at: https://huggingface.co/datasets/lintw/VL-Health

[5] Liu, Dairui, et al. "RecPrompt: A Self-tuning Prompting Framework for News Recommendation Using Large Language Models." *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2024.