Human-centered AI for Mental Health: From Risk Prediction to Physician-aligned Design Guidelines

by

Syed Muhammad Ibne Zulfiker

A thesis submitted in conformity with the requirements
for the degree of Master's in HSR AI

Institute of Health Policy Management and Evaluation
University of Toronto

Human-centered AI for Mental Health: From Risk Prediction to Physician-aligned Design Guidelines

Syed Muhammad Ibne Zulfiker
Master's in HSR AI

Institute of Health Policy Management and Evaluation
University of Toronto
2025

Abstract

This thesis addresses rising clinician burnout and diagnostic challenges in psychiatry by exploring the use of Artificial Intelligence (AI) to support mental healthcare in two key ways: (1) by developing a predictive model that augments traditional features with Large Language Model (LLM) embeddings from free-text clinical notes to identify patients at risk of mental health disorders, and (2) by formulating design guidelines for communicating AI-generated summaries of patient medical histories in ways that align with physicians' decision-making processes. Leveraging structured and unstructured data from a large Canadian Electronic Medical Record (EMR) dataset, the risk assessment model demonstrates improved predictive performance. To inform the design guidelines, we conduct a comprehensive user needs assessment study to gain insights into physicians' cognitive frameworks and diagnostic workflows. Together, these contributions lay the groundwork for scalable clinical decision support in real-world settings.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Nur Camellia Zakaria, whose guidance and support were instrumental throughout the development of this thesis. I am equally thankful to the members of my supervisory committee, including in particular Dr. Zahra Shakeri, for their insightful feedback and encouragement at every stage of this work.

I am deeply appreciative of the colleagues, clinicians, and collaborators who contributed their time and perspectives to the user studies and data collection processes. Their involvement played a central role in shaping the research presented here.

In the process of writing this thesis, I employed ChatGPT, a large language model developed by OpenAI, to assist with the rephrasing of certain sentences for improved clarity and linguistic flow. All intellectual content, data analysis, interpretation, and original contributions remain entirely my own, and the model was used strictly for language refinement within the scope of academic writing. No content generation, idea development, or analytical assistance was derived from the use of ChatGPT.

Finally, I extend my deepest thanks to my family and friends for their patience, support, and encouragement, especially during the most intense phases of this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Healthcare systems worldwide are grappling with physician burnout, driven in part by escalating workloads and systemic inefficiencies. Healthcare professionals are inundated with administrative tasks and documentation demands that compete with direct patient care for their time. In fact, studies indicate that physicians spend roughly two hours on Electronic Medical Record (EMR) documentation for every hour of face-to-face patient care [69]. This imbalance has been linked to rising burnout rates and job dissatisfaction. A recent survey in a Canadian mental health hospital found that 26% of physicians met criteria for burnout, and among those, the majority (61%) identified the EMR as a key contributor [76]. Excessive documentation and poorly designed digital health systems erode clinicians' well-being and compromise care quality. When doctors are pressed for time and cognitively fatigued, important details can be overlooked. Information overload in electronic records is known to cause clinicians to miss critical information, leading to errors or delayed decisions [37]. The flood of data, countless diagnosis notes, lab results and medication lists can overwhelm providers. This situation has fueled calls to reduce documentation

burdens and improve health record usability [76]. It also creates an opportunity for intelligent systems that can synthesize and summarize patient information, allowing clinicians to focus on patient interaction rather than paperwork.

Nowhere is this challenge more pronounced than in fields like psychiatry where understanding the patient's longitudinal history and psychosocial context is essential. Mental health disorders represent a pressing public health challenge worldwide, with close to a billion people living with a mental illness as reported in Nature Medicine in 2019 [26]. The World Health Organization (WHO) has reported depression to be the leading cause of ill health and disability globally [85]. This data underscores the profound personal, social, and economic toll of mental illnesses. The COVID-19 pandemic has only exacerbated this global mental health emergency, driving up demand for services while health systems struggle with limited capacity [26]. In Canada, the situation mirrors these global trends. In any given year, about one in five Canadians experiences a mental illness [10], and recent data show rising prevalence. For example, the 12-month prevalence of major depression increased from 4.7% in 2012 to 7.6% by 2022 [73]. Mental illness and substance use disorders are now the leading causes of disability in Canada [10]. More than one-third of Canadians with a mood or anxiety disorder report that their mental health care needs are only partially met or not met at all [73]. Even in high-income countries, up to 50% of people with depression receive no treatment [85], due to factors ranging from stigma and under-resourcing to shortages of qualified professionals. These trends highlight an urgent need for innovative solutions to improve mental health assessment and care delivery, especially in contexts like Canada's publicly funded health system, where wait times and access barriers persist.

One critical challenge is the timely identification and assessment of mental health conditions. Like many mental health disorders, depression and anxiety often go undetected or untreated until they reach a point of crisis, by which time preventative

interventions are less effective [26]. Early symptoms can be subtle and may be missed in primary care, especially given that generalist clinicians may not have specialized mental health training [51]. Research shows that early screening can significantly improve outcomes. For example, prompt therapeutic support can reduce the severity and incidence of depression by around 20% in at-risk groups [13]. However, recognizing those at risk in time is difficult amidst the high volume and complexity of patient information that clinicians must process. Modern EMRs span years of visits, diagnoses, medication histories, and free-text clinical notes. Manually reviewing these complex patient records to proactively identify mental health risk is often impractical for busy providers [79]. As a result, opportunities for early intervention are frequently missed. There is a pressing need for tools that can sift through the wealth of data in EMRs and flag patients who may have emerging mental health disorders, enabling clinicians to intervene sooner and more effectively without increasing their workload and/or risking burnout.

1.2 Value Proposition of AI

Within this context, Artificial Intelligence (AI) and Machine Learning (ML) are emerging as promising tools for augmenting health care. The rise of EMRs means that unprecedented quantities of patient data, both structured (e.g. demographics, diagnoses, medications) and unstructured (free-text clinical narratives), are now available to inform care. However, unlocking the full value of this data requires advanced analytical methods. Traditional approaches have struggled to make use of unstructured text, often relying on manual review [48, 32] or coded fields that fail to capture the nuance of a patient's story [34, 89]. Recent advances in Natural Language Processing (NLP), particularly transformer-based Large Language Models (LLMs), offer a transformative capability to extract meaningful information from free-text records

[68, 33]. These models can interpret clinical notes with a depth that approaches human understanding, recognizing context, inferring patient symptoms and history, and converting narrative text into useful representations.

1.2.1 Prediction

Researchers have shown that applying modern NLP to full medical records can enable or improve a range of clinical tasks, from diagnosis prediction to treatment recommendation. In one study published in Nature, a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) based model trained on clinical text achieved far higher accuracy in classifying infection indications than earlier rule-based or keyword methods (F1 0.97 vs. 0.71), illustrating the gains possible when free-text data are effectively leveraged. [90]. In the realm of mental health, leveraging textual notes is especially crucial. Clinical notes may describe subtle behavioral signs, patient mood, social circumstances, and clinician impressions that never get encoded in discrete form. Combining structured and unstructured data can thus provide a much richer picture of the patient. A recent review of depression case identification confirms that using EMR data holistically, incorporating narrative notes via ML/NLP, improves the detection and "phenotyping" of depression compared to structured data alone [28].

Building on these advances, large language models offer a unique opportunity. Unlike earlier models such as BERT, which are constrained by relatively limited input token capacities and require note-by-note encoding followed by aggregation across a patient's history, modern LLMs support substantially larger context windows. This enables the direct encoding of entire longitudinal histories, preserving both semantic richness and temporal sequence information within a single embedding. Such capacity allows for attention-based pooling across a patient's full diagnostic narrative, capturing long-range dependencies and subtle clinical patterns that may span multi-

ple encounters. This shift from fragmented representation to holistic encoding marks a fundamental shift in how clinical text can be leveraged for predictive modeling and decision support in mental health care.

While considerable prior work has focused on developing AI-driven prediction algorithms, an important challenge remains: how do we go from prediction to action? Bridging this gap requires a thoughtful consideration of how information is presented to physicians during clinical practice. There is still a pressing need for established design principles that will guide the presentation of patient health information in a way that aligns with physicians' cognitive frameworks and diagnostic workflows.

1.2.2 Design Guidelines for Presenting AI-driven Summaries

Clinicians typically spend a substantial amount of time piecing together fragments of a patient's history from various notes and records, essentially creating mental summaries to inform care [81]. This process is tedious and, even for experts, prone to oversight in the face of intricate or lengthy records [8, 87]. An automated summarization system that consolidates these fragmented patient information while remaining aligned with existing clinical workflows offers a means of alleviating physicians' cognitive burden.

With the advent of powerful generative models, it has become feasible to have an AI system generate a succinct yet comprehensive summary of a patient's EMR, highlighting the most relevant facts, history, and pending issues. Such a tool could be invaluable in mental health care, for example, when a psychiatrist receives a new patient referral accompanied by years of primary care and counseling notes. Rather than reading dozens of documents, the clinician could review an AI-generated summary that distills the core narrative and critical details of the patient's mental health journey. Early research in this area is highly promising. Generative Pre-trained Transformer (GPT) based LLMs like GPT-3.5 and GPT-4 have demonstrated an

ability to produce coherent and accurate summaries of clinical text. In fact, a rigorous evaluation by Van Veen et al. found that, on multiple clinical summarization tasks, summaries generated by LLMs were judged equivalent or superior to human expert summaries in the majority of cases. In a blinded study, physicians rated 81% of AI-generated summaries as at least on par with summaries written by colleagues, with many AI summaries judged more complete, correct, and concise than the expert versions [81]. This evidence suggests that, when properly adapted and guided, LLMs can meet or even exceed clinician expectations for quality while dramatically reducing the time doctors spend on documentation. By alleviating the documentation workload, such technology could allow clinicians to redirect their energy toward patient interaction, potentially mitigating burnout and improving quality of care.

But current automatic summarization efforts tend to only focus on narrow domains (for instance, many early works only target summarizing radiology reports or Intensive Care Unit (ICU) discharge summaries) [5]. There exists a need for a careful, evidence-driven approach to developing AI summarization tools. In the context of mental health, an additional challenge is determining what information an effective summary should contain and how it should be presented. Mental health professionals may prioritize different details (for example, longitudinal trends, medication trials, social support factors) compared to other specialties. Therefore, a human-centered design is crucial. By studying how clinicians read and summarize mental health records, we can discover what they consider most relevant and the ideal structure of a psychiatric case summary. Incorporating clinician expertise via techniques like think-aloud protocols and thematic analysis ensures that the prompts and chain-of-thought reasoning given to an LLM align with expert mental models. This kind of physician-informed prompt design can guide the AI to emulate the decision-making process of an experienced clinician, yielding summaries that are not only accurate but also clinically meaningful and contextually appropriate for mental health care.

1.3 Goals and Contributions

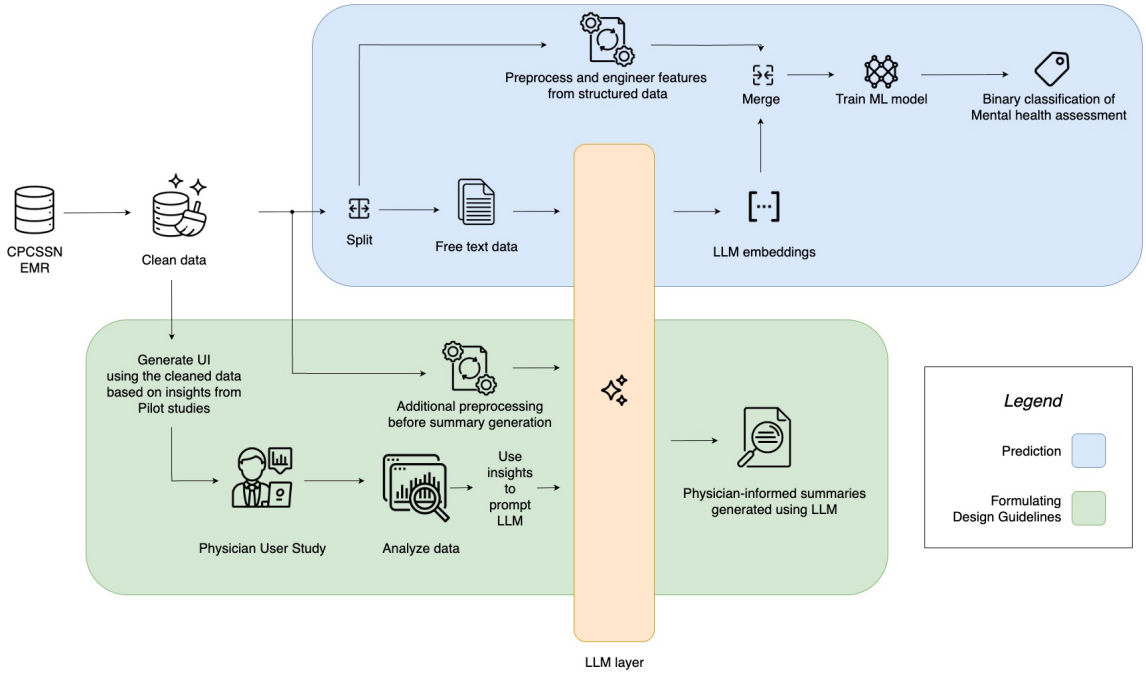


Figure 1.1: High-Level System Architecture

In essence, the convergence of growing mental health needs and increasing clinician burden calls for novel approaches in healthcare AI. There is a clear imperative to develop systems that can harness the data in EMRs, especially free-text clinical notes, to improve mental health outcomes. On one hand, predictive analytics and machine learning can be used to flag patients at risk of mental health disorders earlier than current practices allow. By combining traditional structured health indicators (e.g. demographics, comorbidities, service utilization) with insights gleaned from unstructured text (e.g. notes detailing symptoms, behaviors, and psychosocial concerns), these models can identify subtle patterns indicative of emerging conditions. Automated, continuous surveillance of EMR data has already been shown feasible for problems like incipient mental health crises, and applying this to routine clinical settings could enable timely, preventative care in psychiatry. On the other hand, AI-driven summaries that align with clinicians' cognitive models and

diagnostic reasoning processes can transform how they interact with health information. By providing on-demand, succinct summaries of patients' histories and current status, such technology can reduce information overload and support better clinical decision-making. This research advances both of these avenues in parallel. We focus on the Canadian healthcare context, proposing AI-driven solutions to support mental healthcare on two fronts:

- An early assessment framework for mental health disorders that leverages EMR data, including dimensionality-reduced LLM embeddings from free-text notes, to improve predictive performance over conventional structured-only models.
- A set of design guidelines for automated patient record summarization using large language models, informed by a user study of physicians' needs and reasoning processes.

By addressing the twin challenges of (1) improving early screening performance (by augmenting structured features with LLM-derived embeddings) and (2) establishing design guidelines for automated patient record summarization, our work aims to illustrate how we can adapt modern AI techniques into thoughtful mental healthcare applications that can support clinicians during clinical practice in Canada and beyond. We achieve this through a combination of quantitative evaluation and qualitative needs assessment. A high-level system architecture diagram for the thesis can be found in Figure 1.1.

1.4 Thesis Outline

My thesis statement is as follows:

It is possible to more effectively adapt modern AI techniques for mental health applications that support clinicians by: (1) quantitatively enhancing early assessment performance through the integration of LLM embeddings with structured features, and (2) qualitatively identifying physician' needs and capturing their cognitive and diagnostic processes to ensure that AI-driven summaries are presented in a manner consistent with their practice.

To guide this investigation, my thesis is structured around two central research questions:

- **RQ1:** How does augmenting structured EMR data with LLM embeddings from free-text clinical notes affect the early detection of mental health disorders?
- **RQ2:** How can physicians' needs, along with their cognitive and diagnostic processes, be systematically incorporated into a set of design principles for the automated summarization of patient records?

To address these questions, the remainder of this thesis is structured as follows:

In Chapter 2, *Related Work*, existing literature relevant to machine learning applications on structured EMR data and natural language processing techniques for integrating unstructured clinical notes is reviewed. Particular attention is given to research involving large language models in mental health contexts and their application in clinical summarization tasks. This chapter concludes by identifying key research gaps addressed in this thesis.

Chapter 3, *Study Design and Data Acquisition Methods*, details ethical considerations, introduces our dataset, and explains the data cleaning procedures. It also describes our comprehensive user needs assessment study, including the recruitment of participants, feasibility pilot study, scenario development, interface design, study procedures, and the instruments used for physician surveys.

In Chapter 4, titled *Augmenting Early Mental Health Assessment Using LLM Embeddings*, the methodological approach adopted for preprocessing EMR data, generating LLM embeddings, and training predictive models is presented. The chapter concludes by reporting the evaluation results and offers an analytical discussion highlighting key findings.

Chapter 5, *Formulating Physician-aligned Design Guidelines for AI Summarization*, focuses on the thematic analysis of our physician user needs assessment study. It begins by outlining the analytical approach, followed by a detailed discussion of the themes that emerged and how they inform the formulation of our design guidelines for AI summarization. Subsequently, we discuss the design implications derived from the thematic analysis, including a demonstration of an AI-generated summary produced through prompt engineering that aligns with our proposed design guidelines.

Finally, Chapter 6 concludes the thesis by summarizing major findings, discussing the study's limitations, and proposing avenues for future research to further enhance the applicability of LLMs in clinical practice.

Chapter 2

Related Work

This chapter critically examines relevant literature to contextualize the contributions of this thesis within the broader scope of machine learning and natural language processing applications in healthcare. The objective is twofold: first, to situate this research within established frameworks and methodologies; second, to highlight existing gaps that underscore the need for the proposed system. This thesis specifically focuses on mental health because of the increasing prevalence of mental disorders in recent times and the persistent challenges clinicians face in early detection and management. However, it is important to emphasize that the core methodologies developed herein, leveraging structured and unstructured data via large language models, are agnostic to the specific health domain. Should the clinical focus shift to physical health or other medical specialties, our system would remain equally relevant and valuable. With this broader context established, the subsequent sections of this chapter will explore related work within healthcare with an emphasis on mental health applications using machine learning and natural language processing to effectively illustrate the potential of our proposed systems.

2.1 Machine Learning on Structured EMR Data

Prior work has focused on using structured EMR data such as coded diagnoses, medication records, and encounter histories to predict mental health outcomes. Traditional machine learning models like logistic regression and tree-based ensembles have been popular for these tabular features. In particular, longitudinal Electronic Health Record (EHR) datasets allow models to detect patterns in a patient's history that precede the onset of psychiatric conditions. For example, a recent scoping review (2018–2022) has shown that long-term EHR profiles can predict or detect conditions earlier than formal clinical diagnosis [75]. The authors found that a variety of diseases including mental, behavioral, and neurodevelopmental disorders have been successfully predicted from longitudinal EHRs using machine learning and deep learning models. These models yielded high diagnostic performance and earlier detection compared to usual care, and identified important risk factors from the data. Crucially, using such models for preliminary screening in practice could help personalize prevention and reduce clinician workload.

Raket et al. developed a risk model ("DETECT") using primary and secondary care records of over 145,000 individuals to identify those at risk of a first episode of psychosis [60]. Their work demonstrated that EHR-based risk stratification for serious mental illness is feasible at scale, detecting high-risk individuals before clinicians had documented any psychosis, an example of early identification in practice. Dabek et al. evaluated ML models using longitudinal EMR data to predict the likelihood of developing mental health conditions following the first diagnosis of mild Traumatic Brain Injury (mTBI) [12]. Shao et al. and Ford et al. identified cases of dementia [65, 21]. Fouladvand et al. worked on predicting mild cognitive impairment [22]. These studies illustrate the promise of structured EMR data for mental health prediction, but they also highlight a key limitation: many subtle indicators of mental illness may only appear in clinical narratives rather than in coded fields.

2.2 Integrating Unstructured Clinical Notes using Natural Language Processing

Unstructured free-text notes in the EMR often contain rich context about mood, behavior, and psychosocial factors that structured data may miss. NLP is therefore increasingly used to unlock this information and improve predictive models. Researchers have observed that incorporating textual clinical narratives can significantly enhance the accuracy of diagnostic and prognostic models in mental health. For example, Garriga et al. showed that using clinical notes alongside structured data yields superior performance in predicting mental health crisis relapses [25]. In their large study of 59,750 patients, an ensemble model combining both data modalities outperformed models using only structured EMR fields or only text, underscoring that the two data streams offer complementary signals. Notably, their approach fell back to structured-data-only predictions when notes were absent, but when notes were available (even as little as 10% of weekly notes present), the model's performance improved significantly. This suggests that even sparse narrative information can add "signal" about a patient's mental state that is not captured in coded data.

Advanced NLP methods are employed to convert unstructured text into informative features. Early efforts involved keyword searches or simple bag-of-words representations of clinical text. However, modern approaches leverage pretrained language models (like BERT) to create dense vector embeddings of notes [45]. One practical challenge when using LLM embeddings is their high dimensionality. A single clinical note embedding may be hundreds to thousands of dimensions (for example, 3584 dimensions for Qwen 2.5). Combining multiple notes or multiple dimensions with structured data can lead to extremely large feature spaces, which increases computational complexity and risk of overfitting (the "curse of dimensionality"). Research on multilingual transformers found that up to 90% of embedding dimensions could be

reduced while largely preserving performance on language understanding tasks [31].

2.3 LLM Research in Mental Health

Over the past few years, LLMs have been applied in diverse ways to tackle mental health challenges. Table 2.1 below summarizes representative academic and clinical applications of LLMs in mental health in the past three years and compares them to our study.

Study (Year)	LLM Application	Deployment Type	EMR Integration?	Physician Informed?	Integrates longitudinal patient history as context?	Considers temporal sequence of events?	Diagnosis Specific or General Mental Health?
Liu et al. (2024) [40]	Positive Psychology Chatbot	Proprietary	✗	✗	✗	✗	General
Kim et al. (2024) [38]	MindfulDiary Journaling Assistant	Proprietary	✗	✓	✗	✗	Specific (Major Depressive Disorder)
Sharma et al. (2024) [67]	Self-Guided CBT (Cognitive Restructuring)	Proprietary	✗	✗	✗	✗	General
Mármol-Romero et al. (2024) [47]	Empathic Teen Chatbot	Proprietary	✗	✗	✗	✗	General
Maples et al. (2024) [46]	Emotional Support Chatbot	Proprietary	✗	✗	✗	✗	General
Sharma et al. (2023) [66]	HAILEY Empathy Assistant	Proprietary	✗	✗	✗	✗	General
Perlis et al. (2024) [55]	Clinical Decision Support (Bipolar)	Proprietary	✗	✓	✗	✗	Specific (Bipolar Depression)
Franco D'Souza et al. (2023) [18]	Psychiatric Diagnosis via ChatGPT	Proprietary	✗	✗	✗	✗	General
Cardamone et al. (2025) [9]	Structuring EHR Text	Open Source	✓	✓	✗	✗	General
Llanes-Jurado et al. (2024) [42]	Virtual Human for Therapy/Training	Proprietary	✗	✗	✗	✗	General
Berrezueta-Guzmán et al. (2024) [6]	Robotic ADHD Therapy Assistant	Proprietary	✗	✓	✗	✗	Specific (ADHD)
Lai et al. (2024) [39]	Scalable Mental Health Chatbot	Proprietary	✗	✓	✗	✗	General
Y. Liu et al. (2024) [41]	Treatment Optimization	Proprietary	✗	✓	✗	✗	Specific (Depression)
Our Study	Early assessment and physician-informed design guidelines	Open Source	✓	✓	✓	✓	General

Table 2.1: Comparison of recent LLM-based mental health literature with our study.

As seen in the table, LLM-based applications in mental health can be grouped into a few major areas:

- **Conversational Agents (Chatbots):** Many studies focus on chatbots for therapy, counseling, or support, often aimed at delivering cognitive-behavioral interventions, psychoeducation, or empathetic listening. For example, GPT-based chatbots have been used to conduct positive psychology exercises, help users reframe negative thoughts in a Cognitive Behavioral Therapy (CBT) framework, and engage teens in discussions about depression/anxiety. These systems show promise in improving self-reported well-being and reducing negative effect.

- **Decision Support and Diagnosis:** LLMs have been tested on clinical decision-making tasks in psychiatry. In controlled evaluations, models like GPT-4 have shown an ability to recommend treatments or diagnoses on par with clinicians in case vignettes. This includes selecting medication strategies for bipolar disorder and making psychiatric diagnoses from text descriptions. While promising, these are proof-of-concept studies on simulated cases; real-world deployment would require addressing concerns of accuracy, accountability, and ethical oversight (e.g. avoiding harmful advice).

- **Data Analysis and Prediction:** A growing line of work uses LLMs to analyze unstructured data (clinical notes, social media posts) for mental health insights. Transformer-based models can extract clinically relevant features from text (symptoms, sentiments, etc.) to feed into prediction models. Crucially, combining free-text notes with structured health records has been shown to boost the performance of algorithms predicting outcomes like psychiatric crisis events.

- **Human–AI Collaboration:** Rather than fully autonomous AI, some initia-

tives explore AI assisting humans. The "HAILEY" system, for example, served as a real-time coach to enhance empathy in peer support chats, significantly improving conversation quality when the AI was used. Similarly, LLM-driven virtual patients or role-play agents are being used in training contexts. For example, a GPT-based simulated patient helped medical students practice psychiatric interviews, which students found realistic and useful

2.4 LLMs for Clinical Summarization

Summarizing a longitudinal patient record is a complex task for several reasons. First, patient histories are multi-faceted. They contain diagnoses, medications, lab results, clinical notes, social and family history etc., often spanning many years. Determining what information is "relevant" depends on context and the clinician's judgment. Second, the data is scattered across many EMR sections and notes, so assembling a narrative requires significant chart review. Studies have documented that physicians use specific cognitive strategies to tackle this: for example, they identify key sources of information (like the problem list, recent discharge summaries, or consult notes) and employ temporal reasoning to understand how the patient's condition evolved. In a foundational think-aloud study, Reichert et al. [61] observed 8 clinicians as they summarized patient records and found three primary goals guiding their review: (1) to gather an overview of the active medical problems, (2) to trace the timeline of significant events/interventions, and (3) to synthesize this into an assessment plan for ongoing care. The clinicians prioritized certain data sources (e.g. the most recent notes, problem lists), and they performed cognitive operations like grouping related information and inferring clinical significance. These human insights underscore that summarization is not mere extraction, it involves interpretation and filtering based on clinical relevance. Automated summarization methods historically

struggled with these challenges. Early efforts (2010s and prior) often used rule-based or extractive approaches: for instance, pulling the latest value of key labs, listing active problems, or concatenating the first sentence of recent notes. While helpful, such methods tended to produce disjointed or incomplete summaries and could miss nuances. A review by Pivovarov and Elhadad noted that implementing effective EMR summarizers required surmounting both technical NLP hurdles and aligning with user needs [57].

Over the past five years, the advent of deep learning and large pre-trained language models has radically improved the capabilities for natural language summarization. But a key research question is how to adapt and guide LLMs to generate summaries that meet clinicians' requirements. Recent studies highlight these concerns. For instance, Van Veen et al. evaluated GPT-3.5 and GPT-4 on summarizing clinical texts and conducted a reader study with physicians [80]. They found that when properly adapted to the task, LLM-generated summaries were often preferred over human-written summaries on measures of completeness, factual correctness, and conciseness. The same study noted that achieving this level of performance required careful prompt design and model adaptation. Large language models, with billions of parameters, have rapidly become the state-of-the-art tools for text summarization. Models like GPT-4 and other transformer-based LLMs can capture long-range context and generate human-like prose. In the medical domain, specialized LLMs (e.g. GatorTron [88], BioGPT [44] etc.) have been developed using vast amounts of clinical text, enabling them to understand clinical narratives and instructions. For instance, GatorTron was trained on more than 90 billion words of de-identified notes to create a domain-specific model for clinical NLP tasks. Such models are promising for summarization because they can comprehend the medical context of the input (e.g., recognizing that "HTN" means hypertension or that a list of medications implies certain chronic conditions) and generate summaries in a clinically meaningful way. Recent research in

medical summarization has yielded encouraging results. In one study, five different language models were evaluated on summarizing clinical dialogue transcripts between patients and physicians [23]. A model fine-tuned on clinical dialogue data achieved the highest objective scores (ROUGE metrics for content overlap), while a general model (ChatGPT) scored slightly lower on those but excelled in human-rated qualities like coherence and fluency. This indicates that domain-specific fine-tuning can improve the factual completeness of summaries, whereas general LLMs are very good at producing readable and logically flowing text.

2.5 Research Gap

Previous research clearly demonstrates the promise of using EMR data for early detection of mental health conditions. Structured data provides valuable predictors (utilization rates, comorbidities etc.), and unstructured notes add complementary insights that improve model accuracy [25]. Ensemble approaches have achieved state-of-the-art results in specific prediction tasks like suicide attempts, depression onset, or psychosis risk. However, there remains a gap in the literature that our work addresses: most prior studies target a specific disorder or outcome (e.g. predicting depression vs. no depression), whereas we undertake a more general binary classification of any mental health disorder vs. none. This broad screening approach is important for real-world use, where an automated system could flag patients for any emerging psychiatric condition beyond the specific illness they may be interested in or concerned about during the consultation. Moreover, to our knowledge no published study has yet explored the integration of the latest large language model embeddings (such as Qwen) for EMR-based mental health prediction. Our study is among the first to utilize a cutting-edge open LLM within a hospital data setting for early psychiatric assessment. We further contribute novel analysis on how the dimensionality of LLM

embeddings impacts performance, providing practical insight into model optimization. By combining advanced NLP with traditional ML on routinely collected data, our approach seeks to improve early detection of mental disorders and thus support timelier clinical interventions.

In the realm of clinical text summarization, our work addresses a notable gap: the integration of clinician-centered design, capturing their cognitive and diagnostic processes in practice, with LLM technology. Previous studies have either developed summarization algorithms without substantial input from end-users (risking misalignment with what clinicians actually need), or they have analyzed clinician summarization behavior without leveraging those findings to drive an automated solution. Our approach bridges these two aspects. By conducting a formal user study and deriving thematic insights, we formulate a set of design principles which ensures that the summaries are grounded in real-world physician use cases. By then incorporating these design principles through advanced prompt engineering and chain-of-thought facilitation in a generative model, we demonstrate how user insights can directly inform AI behavior. It is important to clarify that the scope of this thesis was limited to demonstrating the applicability of the proposed design principles through LLM prompt engineering. Evaluating the quality of the responses generated by the language models was explicitly beyond the scope of this work and is reserved for future research.

Our approach is grounded in prior Human–Computer Interaction (HCI) and cognitive research. For example, the study by Reichert et al. directly influenced later designs of EMR summarization interfaces by elucidating how clinicians group and prioritize data [61]. To our knowledge, this is one of the first attempts to employ a think-aloud protocol to guide prompt design for a clinical LLM application. In addition to this, current LLM-based summarizers focus narrowly on individual pieces of the record. Most work targets one note type at a time, for example, summarizing a

single encounter's notes or an individual discharge summary [64]. This is fundamentally different from summarizing an entire longitudinal record. Prior research has not fully addressed how to combine multiple data modalities (free text notes, structured lab test records, medications etc.) from across a patient's timeline into one summary. In practice, clinicians must manually synthesize these data, reading separate lab reports, scrolling through medication lists, and recalling past diagnoses to contextualize the current issues. Our work differs by tackling this integration challenge. Rather than summarizing just the last discharge note or just the problem list, we generate summaries that bridge across data types and time. By doing so, we extend beyond the scope of prior text-only summarization efforts and aim to support use cases like a psychiatrist getting a cohesive patient history at a glance (including medical comorbidities), or a primary care doctor quickly reviewing a new patient's full psychiatric and medical background.

Chapter 3

Study Design and Data Acquisition Methods

For our thesis, we draw upon two principal sources of data: (1) a pan-Canadian EMR repository known as the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) and (2) our user needs assessment data that informs physicians' cognitive models and diagnostic processes in practice. Both of these datasets received University of Toronto Research Ethics Board (REB) approval, and all participants involved provided informed consent for the use of their data. The remainder of this chapter provides a detailed exposition of these two datasets, including methods of acquisition.

3.1 Ethical Considerations

Our research involved working with sensitive clinical data derived from EMRs, necessitating strict adherence to ethical guidelines and institutional oversight. To ensure compliance with national standards for the ethical conduct of research involving human participants, I successfully completed the Course on Research Ethics (CORE 2022), based on the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2). In addition to the TCPS 2 certification, I completed

onboarding training for CPCSSN, which governs access to the dataset. This training focused on protocols for secure data handling and the ethical responsibilities of researchers working with population-level health data. I also undertook the Responsible Conduct of Research (RCR) – Life Science course offered by the Collaborative Institutional Training Initiative (CITI) Program before submitting a detailed ethics application to the University of Toronto Research Ethics Board (REB), outlining the study's methodology, data sources, privacy safeguards, and risk mitigation strategies. This application was reviewed and approved by the REB, confirming that the study meets institutional and regulatory standards for ethical research involving human data. The approved ethics protocol, including reference number and documentation, is included in Appendix B of this thesis.

3.2 CPCSSN Dataset

The primary dataset we utilize in our research is sourced from CPCSSN, a comprehensive, pan-Canadian database comprising primary care EMRs. Established nationwide beginning in 2007, CPCSSN represents Canada's first and largest collaborative EMR database dedicated to primary care, integrating historical records that extend from as early as 1998 through 2015 aggregated from primary care providers operating across eight Canadian provinces and one territory [15, 14]. Participating patients were informed regarding the utilization of their anonymized data, with provisions allowing for exclusion upon request [7]. The dataset comprises longitudinal de-identified records extracted from patient visits, typically updated at quarterly intervals, incorporating a wide array of routine primary care information, including:

- **Demographic Information:** Includes anonymized patient identifiers, age, and sex, capturing individuals predominantly aged between 18 and 90 years.
- **Clinical Observations and Examination Results:** Data points such as

body mass index (BMI) and systolic blood pressure (sBP).

- **Laboratory Test Results (Biomarkers):** Frequently documented biomarkers include fasting blood sugar (FBS), hemoglobin A1c (HbA1c), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (TG), and total cholesterol (TC).
- **Medication Information:** Comprehensive records detailing medication prescriptions, including dosage, frequency, duration, and strength.
- **Encounter Information:** Documentation of encounter dates, types, and associated clinical diagnosis notes.
- **Allergies and Risk Factors:** Description of clinical risk factors and allergies arising from different sources.
- **Medical Procedure Information:** Details on medical procedures including date of surgery.

3.2.1 Data Cleaning

The data cleaning process was conducted in a structured and incremental manner. Key preprocessing steps are summarized below:

- **Outlier Removal:** Implausible values such as negative ages and clinically inconsistent records were identified and excluded.
- **Removing Duplicates:** Duplicate patient entries were eliminated by retaining only the earliest entry per unique PatientID. Similarly, duplicate records in other tables, such as risk factors and medical procedures, were removed.
- **Generating Calculated Features:** Derived variables such as medication duration were computed using available timestamps.

- **Handling Missing Values:** Rows with incomplete key features (e.g., risk factors or medication durations) were discarded.
- **Validating Data Across Tables:** Inconsistencies were corrected, such as resolving conflicts between patient death year and active status. Age calculations were adjusted accordingly for deceased patients.
- **Merging All Tables into a Single Master Dataframe:** Cleaned and preprocessed tables were sequentially joined on PatientID (or EncounterID where applicable) to create a unified dataset representing individual patient records.
- **Standardizing Fields:** Datetime fields were simplified by removing time components, and relevant columns were typecast to standardized string formats for consistency.

Upon completion of all preprocessing steps, the final cleaned dataset comprised 352,161 unique patient records.

3.2.2 Dataset Description

Our analytic cohort comprises 352,161 unique primary-care patients. Women constitute the majority of the population (approx. 59.2%) whereas men account for approx. 40.8%. Records with indeterminate sex are rare (0.01%). The age profile is broad but skewed toward mid and later life. The median age is 56 years (IQR 38–72, mean 55.7 $\pm$ 22.7). Consequently, the dataset captures both working-age adults and a substantial proportion of older adults. Healthcare utilization metrics inform that patients experienced a median of 21 primary-care encounters over the available observation window (IQR 10–42, mean 33.1), reflecting regular contact with the healthcare system. The cohort shows a median of 2 recorded diagnoses overall (IQR 0–5, mean 4.16), with mental-health–specific diagnoses at a median of 0 (IQR 0–2, mean 2.21),

indicating a long-tailed distribution in which a sizeable minority carry multiple psychiatric codes. Domains such as medications (median 8 prescriptions, IQR 2–23) and laboratory tests (median 4, IQR 0–16) further illustrate the heterogeneity of care. In aggregate, these characteristics depict a large Canadian primary-care cohort that spans the adult life course and exhibits substantial variation in service use and clinical burden. Detailed descriptive statistics of the dataset can be observed in Table 3.1.

Domain	**Variable (unit)**	**Median**	**Mean ± SD**	**IQR / n (%)**
Demographics	Male sex (n, %)	—	—	143,617 (40.8%)
	Female sex (n, %)	—	—	208,501 (59.2%)
	NA sex (n, %)	—	—	43 (0.01%)
	Age (years)	56	55.7 ± 22.6	38–72
Healthcare utilization	Primary-care encounters (count)	21	33.1 ± 40.5	10–42
	Encounter frequency/month (count)	0.37	0.63 ± 1.78	0.21–0.65
Clinical complexity	Overall diagnoses (count)	2	4.16 ± 6.67	0–5
	Mental-health diagnoses (count)	0	2.21 ± 5.79	0–2
	Chronic diseases (count)	0	0.53 ± 0.76	0–1
Tests & Interventions	Medications (count)	8	20.99 ± 37.95	2–23
	Laboratory tests (count)	4	12.78 ± 24.04	0–16
	Medical procedures (count)	0	3.32 ± 10.46	0–2

Table 3.1: Core descriptive statistics of the study population (N = 352,161)

3.2.3 Applications of CPCSSN Data in this Thesis

The CPCSSN dataset is used both to train our models for early mental health assessment in Chapter 4 and as the source EMR to be synthesized by physicians for generating summaries in Chapter 5.

3.3 User Needs Assessment Study

3.3.1 Objective

The primary objective of this study is to thoroughly understand physician requirements, expectations, and underlying cognitive processes involved in the summarization of electronic medical records and to evaluate the clinical utility of large language models in automating the synthesis of said summaries. Specifically, pertaining to the scope of this thesis, we aim to gather detailed insights on the components, structures, and content of an ideal EMR summary as identified by practicing physicians. By identifying the key focus areas physicians prioritize when reviewing EMRs, we want to establish foundational requirements to guide the behaviour of LLM-based summarization tools.

In particular, we want to shed light on the following 3 topics:

- **Data Priority:** Using quantitative metrics like number of interactions along with qualitative analysis of think-aloud transcripts, we want to gain insights on the priority of information physicians look out for while synthesizing EMR data. This will help inform us what data to extract and highlight in our LLM summary.

- **Summary Structure:** The repetition of similar structures across the different physician-generated summaries, along with analysis of the post-summary interview will give us a concrete reference on what the ideal LLM summary should

look like.

- **Value Proposition of an AI Summary:** Using metrics like a Single Ease Question (SEQ) and further thematic analysis of post-interview transcripts, we want to highlight which parts of the EMR synthesis and summarization task the physicians find the most mentally taxing and how automating them using LLM summaries might reduce their cognitive burden and generate value.

3.3.2 Recruitment and Participant Summary

We recruited physicians for our user studies through established connections facilitated by a physician faculty member within our department. Leveraging his professional network, he reached out to several colleagues, providing an initial point of contact between prospective participants and our research team. Interested physicians received our detailed study protocol accompanied by documentation of our REB approval via email, explicitly outlining the nature and scope of the research. All participants provided explicit informed consent regarding data collection procedures and the intended use of the collected data, adhering strictly to ethical guidelines.

Our participant cohort exclusively consisted of practicing family physicians. Demographic details of each participating physician, including their specialty, age, years of clinical experience, and languages used in clinical practice, are comprehensively summarized in Table 3.2. In addition, the table also includes a unique identifier code assigned to each physician, which will be used consistently throughout the remainder of this thesis to reference the corresponding participant.

3.3.3 Feasibility Pilot Study

Before initiating our user needs assessment study, we conducted a feasibility pilot study involving a licensed physician with 30 years of clinical experience. The primary

Study	Code	Specialty	Age	Clinical Experience (years)	Languages Used in Clinical Communication
Feasibility Pilot	F1	Family Medicine	58	30	English
User Needs Assessment	P1	Family Medicine	72	44	English
	P2	Family Medicine	60	35	English
	P3	Family Medicine	43	15	English, French, Spanish

Table 3.2: Demographic and Professional Background of Physician Participants.

goal of this pilot was to identify specific physician requirements to optimize our study procedures, with particular emphasis on informing the design of an intuitive interface for presenting the electronic medical records. Additionally, the study aimed to provide insights into the clinical reasoning processes physicians employ when reviewing EMRs, as well as to establish a foundational understanding of the structure, content, and level of detail typically expected in clinically useful patient summaries. The findings derived from this feasibility study informed the subsequent refinement of the study design protocol.

Process

At this stage, we provided the participant with scrollable PDF versions of various tables from the EMRs, as we had not yet developed an interactive user interface for EMR presentation. The physician was then instructed to review the patient records contained within the provided PDF documents and to synthesize the information into clinically meaningful summaries. During this process, the physician was asked to engage in a think-aloud protocol, verbalizing their evaluative judgments, cognitive strategies, and experiential reactions, including aspects they found favorable or unfavorable, their approach to information extraction, elements they perceived as

frustrating, and suggestions for improvement.

Design Revisions

The study yielded key insights into clinical reasoning and the specific needs of physicians when interacting with EMRs. To begin with, the participating physician emphasized the central role of presenting complaints in synthesizing patient medical histories. According to their insights, presenting complaints guide the prioritization and relevance of subsequent EMR information. For instance, they indicated that lab results might hold lower priority for a patient presenting primarily with mental health issues.

The participant highlighted that the static presentation of information posed a significant barrier to identifying relevant information. A crucial component of the EMR was demographic details, which our physician participant stressed should be prominently placed at the top of the user interface. This placement aligns with the natural workflow observed during patient consultations, as physicians routinely assess basic demographic information prior to delving deeper into clinical details. The participant also underscored the necessity for seamless accessibility of critical health information categories. Specifically, details including health conditions, diagnostic notes, medications, examinations, and laboratory results should be made readily available and intuitively navigable. An efficient interface design, according to the physician's experience, necessitates organizing this information chronologically, with the most recent entries displayed first. This helps identify current patient issues and facilitates decision-making. Critically, the participating physician provided us an exemplar of an existing EMR interface used in their clinical practice. This practical reference delineated precise locations and preferred layouts for key information elements within a production EMR system. Leveraging this real-world example served as the primary reference for developing our EMR user interface for our final user

study. This approach ensured that our design aligned closely with established clinical workflows, optimizing the authenticity and practical utility of our user study.

Furthermore, the physician elaborated on the structural elements that constitute an ideal patient summary, providing clarity on their expected sequence and content. According to them, an optimal summary begins with a comprehensive past medical history, encompassing current and past medical conditions, previous surgical procedures, and ongoing medications. This is then supplemented by lifestyle information, which offers additional insights into factors influencing the patient's health status. Subsequently, documentation of patient allergies followed by a clear record of vaccination status completes the holistic view of the patient. And finally, the participant reported challenges associated with information fatigue, emphasizing the risk of overlooking critical diagnoses when tasked with synthesizing multiple patient records concurrently.

3.3.4 Study Design

Our study design comprises two phases. Phase 1 will constitute the main area of focus in this literature and define the scope of this thesis. It will involve extensive interaction with practicing physicians through semi-structured interviews and surveys. These interactions will explore the physicians' cognitive processes when reviewing patient histories within EMR interfaces in order to guide the behaviour of an LLM-based summarization tool. Upon successful completion of Phase 1, future research will include a rigorous evaluation of the LLM-generated summaries. This subsequent phase, namely Phase 2, will leverage blinded randomized evaluations, employing validated assessment tools and questionnaires, complemented by in-depth qualitative analysis. The primary objective at this stage will be to evaluate the LLM-generated summaries against physician-produced standards, ultimately ensuring the practical efficacy of AI-generated EMR summarization in clinical workflows.

To reiterate, the inclusion of Phase 2 in this section is solely intended to present a complete picture of the overarching study timeline as outlined in the REB proposal. The scope of this thesis will focus exclusively on Phase 1, which encompasses the foundational qualitative and quantitative groundwork. Phase 2 is discussed in the study design only to provide contextual framing and to inform the trajectory of subsequent research efforts.

3.3.5 Scenario

Prior to interacting with the EMR interface, the physicians will be presented with a scenario that they've just received a new patient on their list presenting with anxiety issues, and they're preparing for an initial consultation and have a few minutes before they see them. They are being presented with the EMR interface to build an understanding of the patient's medical history. This scenario is influenced by our pilot findings, in which the physician emphasized the critical importance of the presenting complaint as a central guiding factor for synthesizing patient histories, determining relevant diagnostic priorities, and directing clinical decision-making. We also limit each physician to synthesizing a single patient record to mitigate cognitive overload, referring to the pilot findings in which the participant experienced information fatigue when reviewing multiple patient records concurrently.

We selected a patient from our CPCSSN dataset who developed an anxiety disorder at a later stage in their medical history timeline. Notably, this patient was flagged to be at risk of mental health disorders by our early screening model (described in Chapter 4), as part of predictions generated on previously unseen test data. To construct an authentic clinical scenario, we incorporated all available medical history data up to, but explicitly excluding, the first recorded mental health diagnosis. This approach was deliberately chosen to simulate a scenario where the patient presented to the physician with mental health concerns for the first time. It is important to

emphasize that the CPCSSN dataset has been rigorously anonymized, with personally identifiable information, such as patient names, removed prior to extraction. All names and clinical identifiers used in the creation of the EMR interface are fictional and do not correspond to real individuals or clinics. They are used just to mirror our reference production EMR to realistically mimic the actual clinical experience of physicians.

3.3.6 Physician Interface Design

Drawing directly from the insights of our pilot study (Section 3.3.3), all design decisions and user interaction elements were closely modeled on a reference production EMR interface provided by our participating physician. This reference served as the benchmark for ensuring our prototype realistically mirrored the actual clinical experience of physicians. Some screen captures from the finalized user interface are presented in Figure 3.1 for reference.

Prototyping and Implementation using Figma

We utilized Figma, a collaborative design and prototyping tool, to build our physician EMR user interface. Figma's integrated prototyping capabilities allowed for detailed simulation of navigation flows and interactive components, enabling us to effectively mimic user interactions without the need for complex software development. This facilitated rapid iteration of designs informed by the feedback gathered from the pilot.

Structure and Organization of the EMR Interface

In line with physician feedback from our pilot, patient demographic details were prominently displayed in a dedicated standalone section at the very top of our interface, ensuring immediate accessibility. The interface was structured into three

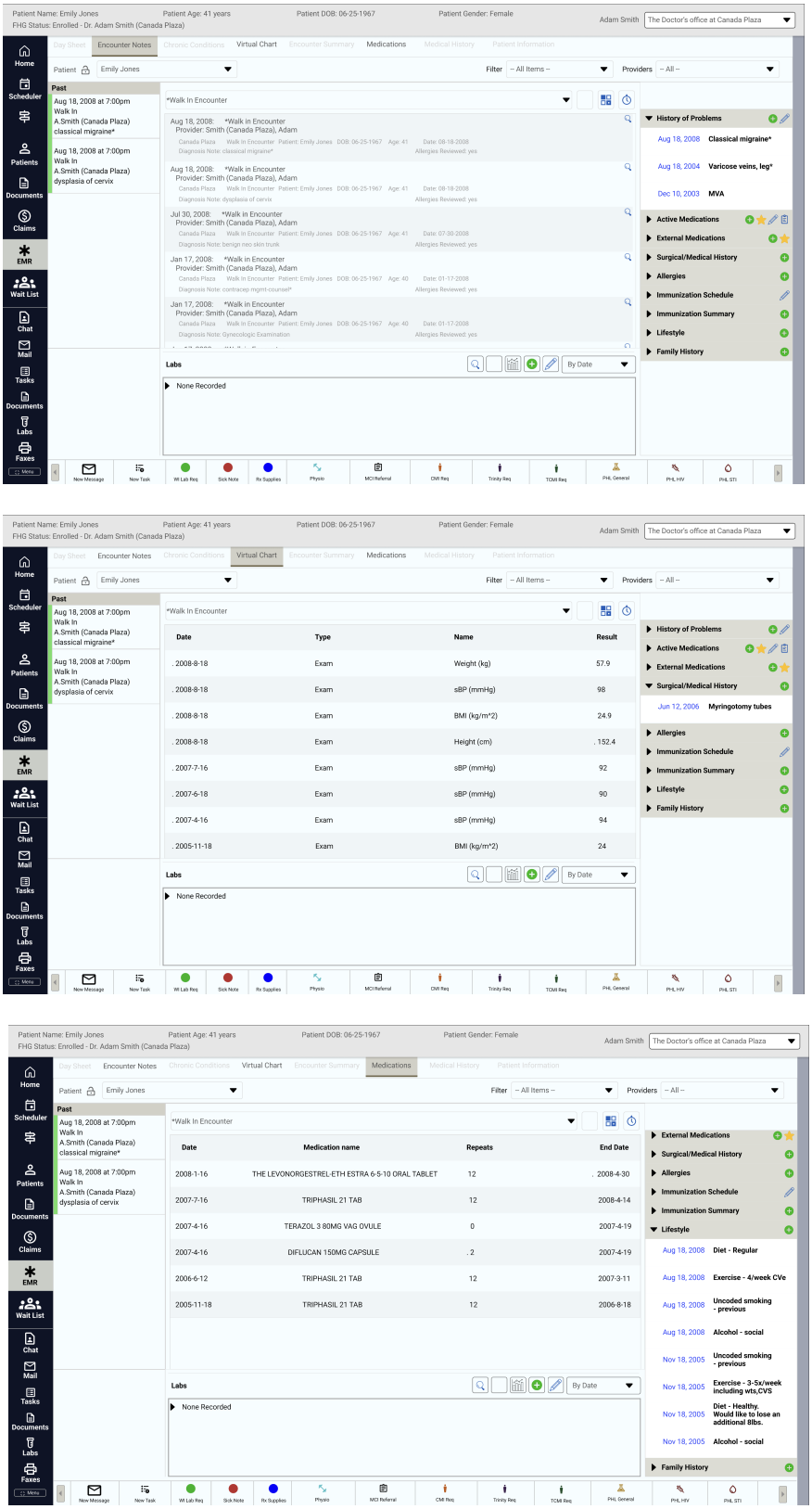


Figure 3.1: Screen Captures from our Physician EMR UI illustrating the different interactive components. The three images demonstrate the three interactive tabs on the center panel along with some of the expandable subsections on the right panel.

main content tabs: Encounter Notes, Virtual Chart, and Medications, each carefully designed to facilitate efficient and targeted information retrieval by physicians.

Each tab consisted of a three-panel layout: a left panel, a central panel, and a right panel. The left and right panels were intentionally kept consistent across all tabs to provide continuous access to the most essential patient information. The left panel consistently displayed the patient's most recent encounter notes, thus allowing constant access to the patient's latest clinical interactions. Conversely, the right panel provided quick reference to the patient profile, including allergies, lifestyle history, and an organized record of past medical problems. To further optimize usability, each of these items in the right panel was presented as collapsible and expandable sections, enabling physicians to manage information density according to their preferences and needs.

The central panel within each tab displayed detailed patient-specific records relevant to that tab's purpose. Specifically, the Encounter Notes tab showcased historical documentation of past medical encounters, the Virtual Chart tab presented past examinations and laboratory test results, and the Medications tab featured a comprehensive list of medications previously prescribed to the patient. All information was presented in the order of latest first.

Supplementary Visual Components

For visual consistency with our reference EMR, we included additional design elements such as a left-side menu bar and a bottom navigation bar. Although these components did not contribute directly to the user tasks assessed during our study, their inclusion was considered necessary to replicate the visual familiarity and realism of actual EMR systems currently in clinical use.

Populating the EMR Interface with Patient Data

The interface was populated with clinical data from the CSCSSN dataset corresponding to the patient described in section 3.3.5.

3.3.7 Study Procedure and Survey Instruments

This section outlines the study protocol and offers a detailed account of all validated survey instruments used in the design of our study. The complete instrument for both phases is provided in Appendix C.

Phase 1

We will initiate the study with a series of individual, semi-structured sessions involving a cohort of practicing physicians (N = 3). The sessions will be conducted online and scheduled to last approximately 30 minutes. Each session will comprise three sequential components.

- **Demographic & Background Survey:** Participants will complete a brief questionnaire that will capture their age, sex, years of clinical experience, medical specialty and other pertinent background information using a survey form.
- **Task-Based EMR Review with Think-Aloud:** For this task, physicians will review a de-identified EMR of a single patient that will include medication lists, encounter diagnosis notes, laboratory results, procedural histories etc. within an interface as described in Section 3.3.6. Before being presented with the interface, they will be given a scenario as detailed in Section 3.3.5. Participants will then be instructed to verbalize their thought processes continuously ("think-aloud") while synthesizing the record into a concise clinical summary, mimicking customary practice during patient encounters. Throughout the task, structured probes (e.g., "What are you thinking right now?", "What are you

hoping to find here?", "Can you take me through your thought process right now?") will be utilized to elicit additional contextual information regarding the participant's cognitive process. Screen activity, along with participants' video feeds, will be recorded to enable the analysis of facial expressions during task engagement; verbalizations will be later transcribed for qualitative analysis. The participants will record their generated summaries in a text field within the survey form.

- **Post-Task Questionnaire:** Immediately after the summarization task, participants will respond to a Likert-style "Single Ease Question" followed by an open-ended interview covering topics such as ideal summary structure, criteria for information sufficiency, mental models for EMR navigation, and data prioritization strategies. Participants will also be asked additional ease-of-use questions aimed at assessing perceived task difficulty and potential areas where an AI-generated summary might provide practical support.

Phase 2

For Phase 2, a second cohort of licensed physicians will appraise the quality and usefulness of the summaries. Before beginning the evaluation, each participant will complete a brief demographic survey identical to that of Phase 1. For each session, the physician will receive the reference longitudinal EMR records and a pair of corresponding summaries, one produced by a clinician in Phase 1 and one generated by our LLM pipeline, presented in randomized order and without provenance labels.

The physicians will then be presented with the following 2 sets of questionnaires, accompanied by contextually relevant open-ended questions designed to elicit deeper qualitative insights.

- **Blind Quality Assessment:** Participants will review the two summaries

alongside the reference EMR. Using a ten-item Physician Documentation Quality Instrument (PDQI) adapted for outputs generated by large language models [77] as a rubric, they will rate accuracy, thoroughness, comprehension, etc. on five-point Likert scales. A forced-choice preference question ("Summary A, Summary B, or both equally") will capture overall favorability.

- **Usefulness & Acceptance Survey:** Physicians will fill out the Unified Theory of Acceptance and Use of Technology (UTAUT) [82] questionnaire (using a seven-point Likert scale) to gauge performance expectancy, effort expectancy, social influence, and facilitating conditions with respect to integrating the AI summaries into their workflows.

3.3.8 Applications of The User Study Data in this Thesis

The data from our user study helped us establish the foundational requirements to generate the design guidelines for prompting an LLM-based summary generation pipeline in Chapter 5.

Chapter 4

Augmenting Early Mental Health Assessment Using LLM Embeddings

This chapter investigates the use of LLM embeddings to enhance the early assessment of mental health conditions from electronic medical records. Our methodology is designed to address three central research questions that guided the design and evaluation of our experiments.

- **RQ1 - Modality Contribution:** What is the incremental benefit of augmenting a structured-data-only classifier with LLM-derived text embeddings for predicting the risk of a mental health diagnosis for a patient?
- **RQ2 - Embedding Dimensionality:** How does classification performance change as we vary the size of the text embeddings (e.g. 128, 256, 512 dimensions), and what is the trade-off between model performance and computational efficiency?
- **RQ3 - Prediction Lead Times:** How does model accuracy evolve when

enforcing different lead times (e.g. predicting a mental health diagnosis 3, 6, or 12 months in advance)?

Figure 4.1 provides a high-level system overview of the early mental health assessment pipeline. The subsequent sections provide a detailed description of each individual module, outlining how they collectively contribute to addressing the central research questions.

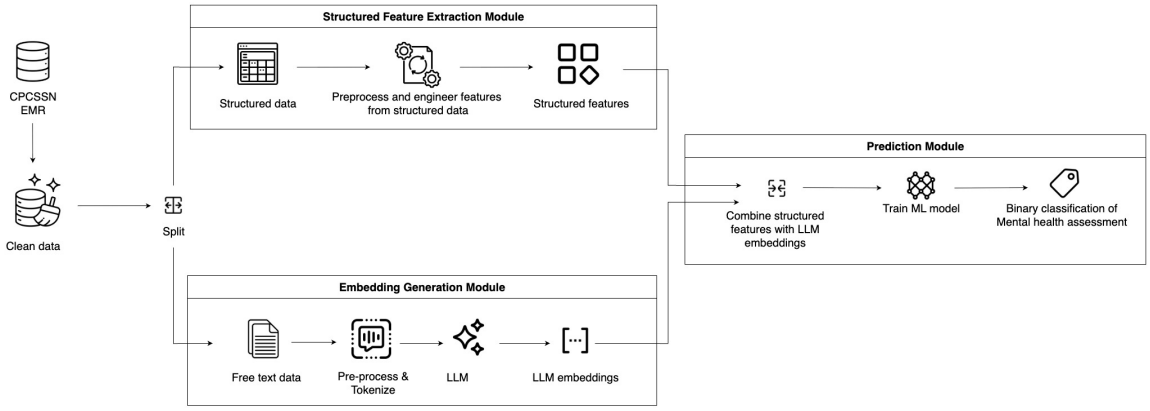


Figure 4.1: System Overview of Early Assessment Pipeline

4.1 Feature Extraction Module

4.1.1 Parameter Settings

In developing our experimental setup, careful selection of key parameters was essential to meaningfully address our research questions. Specifically, we focused on two critical parameters: embedding dimensionality and prediction lead times.

We chose embedding sizes of 128, 256, and 512 dimensions to systematically investigate the balance between model performance and computational efficiency. These dimensions are well-established in existing literature [83, 70], as they have demonstrated strong classification performance while maintaining representational capacity with manageable resource demands. Additionally, to align our predictive model eval-

uations with realistic clinical scenarios, we selected lead times of 3, 6, and 12 months. Patients often visit healthcare providers infrequently or irregularly [1], sometimes with intervals extending up to a year or more. Therefore, testing these lead times allowed us to assess whether accurate early detection of mental health disorders could be achieved despite significant gaps between patient encounters.

4.1.2 Structured Features

The initial cleaned dataset comprised electronic medical records of 352,161 unique patients from the CPCSSN dataset. To establish binary classification labels indicating the presence or absence of mental health disorders, International Classification of Diseases, Ninth Revision (ICD-9) [52] diagnosis codes were systematically evaluated and mapped to Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) categories [29]. The mappings are detailed in Table A.1. Codes corresponding to DSM-5-defined mental health conditions were assigned a positive label (1), thereby indicating the presence of a mental health diagnosis. Patients lacking ICD-9 codes (n = 89,401) were excluded from the analysis, leaving an analytical cohort of 262,760 individuals.

The selection of features was informed by established risk factors for Mental Health Disorders (MHD) as documented in prior epidemiology and psychiatric research [62, 49, 72]. Core demographic variables such as age, sex, and dates of birth and death were included. Additionally, known contributors to MHD, including obesity and tobacco use, were incorporated.

To comprehensively capture patient profiles, derived variables encompassed indicators of health service engagement, such as the frequency of medical encounters and both the count and duration of prescribed medications, as well as abnormal clinical laboratory results. The presence of chronic medical conditions [53] and comorbidities related to both physical and mental health [86, 74, 71, 59, 56, 36, 19, 3, 17]

were encoded. Diagnostic information was obtained through ICD-9 codes linked to each patient record. Chronic diseases were defined in accordance with the Canadian Chronic Disease Surveillance System [58], while physical comorbidities were identified through a comprehensive review of existing literature pertaining to each DSM-5 diagnostic category. For a full list of reviewed sources and the detailed ICD-9 codes employed in the identification process, refer to Appendix A Chronic Diseases and Table A.2.

The datasets underwent further preprocessing. Feature-wise missingness was assessed, and given that $< 1\%$ of observations contained missing values, Multiple Imputation by Chained Equations (MICE) [4] was applied. The imputation model was fitted exclusively on the training set first and then subsequently applied to the testing (holdout validation) set to mitigate data leakage. Multicollinearity was managed by eliminating features exhibiting pairwise Pearson correlations greater than 0.70. Ultimately, each patient's record was characterized by a set of 28 features as detailed in Table A.8.

4.1.3 Text Embeddings

We selected the Qwen 2.5 7B Instruct model for generating our language representations. This open-source model offers a compelling combination of scalability and contextual reasoning capabilities. With a 128K token context window, the model is well-suited to handle the extensive length of EMR notes, many of which exceed the typical context limits of other widely used LLMs. Moreover, Qwen 2.5 is optimized for instruction-following tasks and demonstrates strong performance on structured inputs, which are common in clinical documentation.

For each patient, all relevant diagnosis-related clinical notes were first chronologically sorted from earliest to latest and then concatenated into a single longitudinal text sequence, providing the model with a comprehensive view of the patient's di-

agnostic history. To isolate the predictive signal relevant to future mental health diagnoses, we truncated the notes at specific time points relative to the first recorded mental health diagnosis. For the zero lead time condition, we included all notes up to but not including the first mental health diagnosis. To examine the impact of lead time, we similarly truncated the notes 3, 6, and 12 months in advance of the diagnosis date, thereby simulating progressively earlier windows of prediction.

The resulting text sequences were tokenized and passed through the Qwen 2.5 7B Instruct model to generate fixed-length embeddings. This model produced 3584-dimensional vectors by extracting the final hidden state from the LLM, encoding semantic and contextual features across the full clinical timeline. While this high-dimensional representation captured nuanced textual information, it introduced computational inefficiencies and increased the risk of overfitting in downstream models. To mitigate these issues and systematically explore our embedding dimensionality question, we applied Principal Component Analysis (PCA) [35] to reduce the vectors to 128, 256, and 512 dimensions. PCA aligns well with the demands of our use case. With high-dimensional clinical text embeddings that often exhibit correlated structure, PCA produces orthogonal components that concentrate variance and reduce multicollinearity, thereby improving generalization in downstream classifiers. Moreover, PCA is computationally efficient in practice, with scalable singular value decomposition implementations that make it relatively resource-inexpensive to fit and apply at scale. PCA was initially fitted on the training set, and the resulting transformation was subsequently applied to the test set to ensure consistency and prevent data leakage. The reduced embeddings were then normalized to the [0, 1] range to standardize feature scales across modalities, facilitate convergence during model training, and prevent any single feature from disproportionately influencing the learning process. These reduced embeddings were then concatenated with structured EMR features, forming a unified multimodal input matrix for classification.

This integrated representation allowed us to directly assess the incremental predictive value of LLM-derived embeddings while maintaining scalability and tractability across large clinical corpora.

4.2 Prediction Module

4.2.1 Model Development

To ensure comprehensive benchmarking, we selected six well-established supervised learning algorithms based on their demonstrated efficacy in prior research [2, 20, 24, 27, 50]: `K Nearest Neighbors (KNN)`, `Logistic Regression (LR)`, `ADABoost (ADA)`, `Random Forest (RF)`, `XGBoost (XGB)`, and `LightGBM (LGBM)`. Using multiple model architectures enabled us to compare simple, interpretable approaches (e.g., logistic regression, kNN) against powerful ensemble methods (e.g., random forest, boosted trees) and to assess the trade-offs between them.

To optimize model performance, hyperparameter tuning was performed using a grid search algorithm. For each model, all combinations of hyperparameters were evaluated via 5-fold cross-validation. Model performance during tuning was optimized on ROC-AUC. The search was performed separately for each model using predefined grids tailored to their respective hyperparameters [Appendix A.9]. The best hyperparameter configuration (as determined by the cross-validated ROC-AUC) [Appendix A.10] was selected for final evaluation. A fixed random seed was used throughout to ensure reproducibility.

Given significant class imbalance (205,502 patients without mental health diagnoses versus 48,248 with diagnoses in the training set), random undersampling was employed during training to balance classes. Given the high dimensionality and semantic complexity of the LLM-derived embeddings, we elected not to apply oversampling techniques such as Synthetic Minority Oversampling Technique (SMOTE),

which rely on linear interpolation and may compromise the integrity of the embedding space. Due to the availability of a sufficiently large sample size, we opted for random undersampling to achieve class balance while preserving the fidelity of the learned representations.

We implemented 10-fold stratified cross-validation to compute all performance metrics for comparative analysis. Stratification allowed us to maintain proportional representation of classes across sets. For each fold, the dataset was initially split into training and evaluation subsets. Subsequently, imputation and random undersampling were applied exclusively to the training data. Imputation was performed separately on the evaluation set using parameters derived exclusively from the training data, thereby preserving the integrity of the validation process and again preventing data leakage.

4.2.2 Final Model

Our best performing model was LightGBM with 128-dimensional embeddings at no lead time. The methodology and analysis underlying the selection of this model are elaborated upon in detail in Section 4.4. This configuration was subsequently adopted as the final model for our early assessment pipeline. Finally, we ran this model on the held-out test set to evaluate the generalizability of our solution over previously unseen data.

4.3 Technical Setup

To generate text embeddings from the free-text clinical notes within our EMR dataset, we leveraged high-performance cloud infrastructure via the Lambda Cloud platform. To ensure compliance with data privacy and ethical standards, we needed to deploy our own large language model instances on secure cloud infrastructure rather than

rely on third-party APIs such as those offered by OpenAI or Anthropic. The EMR data used in our study contains sensitive clinical information, and transmitting this data to external vendors, even via encrypted channels, posed unacceptable risks in terms of patient confidentiality, data sovereignty, and institutional data governance policies. By running inference locally on a dedicated cloud instance, we maintained full control over the data pipeline and eliminated exposure to external systems, aligning our methodology with best practices in healthcare data security and responsible AI deployment. Specifically, we deployed a GPU instance equipped with one NVIDIA A100 (40 GB SXM4) accelerator, 30 virtual CPUs, 200 GiB of RAM, and a 0.5 TiB solid-state drive. This setup offered the optimal balance between computational capacity and cost-efficiency.

4.4 Evaluation

4.4.1 Data Split

Our analytical cohort comprised patient records collected from 12 distinct clinical networks, each containing multiple individual sites or clinics. To rigorously evaluate model generalizability and performance on previously unseen data, records from one entire network, encompassing 9,010 patient records, were isolated as a holdout validation set. The remaining 253,750 patient records were allocated for model training and comparison. Importantly, both the training and holdout datasets exhibited comparable demographic distributions and characteristics, as presented in Table 4.1.

4.4.2 Performance metrics

For our analysis, we prioritized ROC-AUC, PR-AUC, and recall metrics due to their distinct relevance in the context of this study. We selected ROC-AUC and PR-AUC for their advantage as threshold-independent measures of classification performance,

Dataset	Variable (unit)	Median	Mean ± SD	n (%) / IQR
Train set	Total population (n, %)			253,750 (96.57%)
	Male sex (n, %)			100,149 (39.47%)
	Female sex (n, %)			153,565 (60.53%)
	NA sex (n)			3
	Age (years)	61	59.35 ± 21.6	43–75
	Has MHD (n, %)			50,355 (19.16%)
	Does not have MHD (n, %)			212,405 (80.84%)
Holdout Test set	Total population (n, %)			9,010 (3.43%)
	Male sex (n, %)			3,302 (36.65%)
	Female sex (n, %)			5,708 (63.35%)
	NA sex (n)			–
	Age (years)	62	61.88 ± 21.15	48–76
	Has MHD (n, %)			2,107 (23.4%)
	Does not have MHD (n, %)			6,903 (76.61%)

Table 4.1: Summary statistics for the train and holdout test sets.

particularly suited to scenarios with class imbalance, such as predicting the onset of mental health disorders. While ROC-AUC provides a global assessment of a model's ability to discriminate between classes across all possible thresholds, it can be overly optimistic in imbalanced datasets due to its reliance on true negative rates. In contrast, PR-AUC offers a more nuanced evaluation by focusing exclusively on the positive class, capturing the trade-off between precision and recall without being diluted by the abundance of true negatives. Recall complements them due to its clinical significance, where identifying true positives (individuals genuinely at risk of developing a mental health disorder) is critically more impactful than minimizing false positives. This clinical prioritization addresses the substantial risks associated with missed or delayed screening of psychiatric conditions. In addition to these metrics, we also report accuracy, precision, and F1 score.

4.4.3 Statistical Test

All statistical comparisons for the model evaluation metrics were conducted using paired t-tests, with pairing performed at the fold level across the 10-fold stratified cross-validation test results. Since all models were trained on identical folds, this methodological approach effectively preserved dependencies between model scores. To mitigate the risk of inflated Type I errors due to multiple pairwise comparisons within each analysis group, the Holm-Bonferroni correction was applied, effectively controlling the family-wise error rate.

4.4.4 Results

RQ1: Modality Contribution

Performance comparisons were conducted between the structured-only (referred to from now on as 'noembed') and embedding-augmented (128-dimensional embeddings,

no lead time) datasets. With the exception of KNN, the noembed dataset consistently performed significantly worse across all evaluation metrics. Within-model pairwise comparisons indicated statistically significant improvements ($p < .001$) favoring the dataset augmented with embeddings for all model architectures. Specifically, ROC-AUC values improved by approximately 5%, PR-AUC values by 3-4% and recall increased by about 2–4%. These enhancements directly address our first research question, providing clear empirical evidence that embeddings generated from free-text notes carry clinically meaningful signals that significantly enhance predictive performance beyond what structured data alone can offer.

It is important to note that, with the exception of KNN and LR, all evaluated model architectures demonstrated consistent high performance in terms of ROC-AUC and recall. However, precision scores remained comparatively lower across models. We considered this trade-off acceptable within the context of our clinical scenario. We prioritized true positive rates over minimizing false positives, given that the consequences of failing to detect individuals in need of mental health assessment are clinically more significant than the implications of incorrectly flagging low-risk individuals. For this reason, we did not tweak the decision thresholds post-training, as doing so would have likely reduced recall in favor of precision.

Across all evaluated metrics, the relative performance of the model architectures can be ordered as follows: KNN < LR < ADABoost < Random Forest < XGBoost < LightGBM, with LightGBM exhibiting the most favorable overall performance. It is evident that tree-based classifiers consistently outperformed traditional linear models such as logistic regression (LR) and non-parametric approaches like KNN. KNN exhibited consistent outlier performance across all configurations. This can be likely attributed to it falling victim to the "Curse of Dimensionality," where, in high-dimensional spaces, the distance metric used by KNN becomes less informative, as data points tend to become equidistant. LR handles dimensionality better than KNN,

but still assumes linear separability, which is rarely true in complex EMR datasets with multimodal inputs. On the contrary, tree-based ensemble methods demonstrate superior capacity to capture nonlinear interactions and handle heterogeneous clinical features. Within the tree-based classifiers, ADABoost likely falls short because of its shallower decision stumps. Overall, LightGBM performed exceptionally well with our high-dimensional PCA-reduced vectors, which corroborates results reported in existing literature [11]. The average aggregated evaluation metrics over 10-fold stratified cross-validation results across all model architectures and configurations are presented in Tables A.11 and A.12 in the Appendix.

Table 4.2 presents a comparative performance analysis of our best-performing model, LightGBM (LGBM), between the noembed and 128-dimensional embedding datasets with no lead times. Statistical significance levels are reported with the noembed configuration serving as the baseline. Following this, in Table 4.3, we provide a performance comparison across the different model architectures trained on the 128-dimensional embedding dataset (0 lead time), with Logistic Regression (LR) designated as the baseline for significance testing.

Model	**Accuracy (%)**	**ROC-AUC (%)**	**PR-AUC (%)**	**Recall (%)**	**Precision (%)**	**F1 (%)**
LGBM noembed [baseline]	68.22	77.95	54.33	73.17	34.49	46.88
LGBM 128-dimensional embedding	**73.47*****	**83.37*****	**57.73*****	**77.59*****	**36.84*****	**49.95*****

Table 4.2: Performance comparison of LGBM with and without embeddings (0 lead time) [noembed serving as baseline].

Significance levels: * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Model (128-dimensional embedding)	**Accuracy (%)**	**ROC-AUC (%)**	**PR-AUC (%)**	**Recall (%)**	**Precision (%)**	**F1 (%)**
LR [baseline]	68.76	70.25	34.21	58.58	29.26	39.03
KNN	**82.17*****	68.37***	31.97***	17.53***	**44.46*****	25.14***
ADA	69.81***	78.13***	45.75***	72.09***	32.61***	44.91***
RF	71.42***	80.73***	51.73***	75.45***	34.55***	47.40***
XGB	73.16***	82.99***	55.78***	77.16***	36.46***	49.52***
LGBM	73.47***	**83.37*****	**57.73*****	**77.59*****	36.84***	**49.95*****

Table 4.3: Performance comparison between different model architectures with 128-dimensional embeddings (0 lead time) [taking Logistic Regression as baseline].

Significance levels: * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

RQ2: Embedding Dimensionality

Across our experiments, we find marginal differences in performance between low- and high-dimensional LLM embeddings. In fact, lower-dimensional embeddings slightly outperform their higher-dimensional counterparts. This suggests that even compact embeddings retain sufficient expressive power to encode relevant semantic and clinical information from textual corpora. But on the other hand, this observation may also highlight limitations inherent to the PCA-based dimensionality reduction process. Specifically, as the number of retained components increases, they tend to encode progressively lower-variance signals, potentially amplifying noise and irrelevant correlations. This phenomenon can lead to overfitting, wherein models become sensitive to spurious patterns rather than clinically meaningful indicators. Further research needs to be conducted comparing PCA with other dimension reduction techniques to come to a concrete conclusion.

Table 4.4 presents a comparative performance analysis of LightGBM across dif-

ferent embedding dimensions with no lead times. Statistical significance levels are reported with the 128-dimensional embedding serving as the baseline.

Model (dimensions)	**Accuracy (%)**	**ROC-AUC (%)**	**PR-AUC (%)**	**Recall (%)**	**Precision (%)**	**F1 (%)**
LGBM (128) [baseline]	**73.47**	**83.37**	**57.73**	77.59	**36.84**	**49.95**
LGBM (256)	73.31**	83.24**	57.66*	**77.82**	36.71*	49.88
LGBM (512)	73.11***	83.04***	57.21**	77.41	36.45***	49.56***

Table 4.4: Performance comparison of LGBM with different embedding dimensions (0 lead time) [taking 128-dimensional embeddings as baseline].

Significance levels: * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

RQ3: Prediction Lead Times

Investigating the impact of different lead times (zero, three, six, and twelve months) revealed no clear advantage for any lead time regarding the primary evaluation metrics (ROC-AUC, PR-AUC and recall). Differences among these lead times were predominantly minor and largely statistically non-significant. However, precision exhibited a significant decreasing trend with each incremental increase in lead time, from zero to three months, three to six months, and six to twelve months, respectively. This decline in precision was consistent across all model architectures and statistically significant ($p < .001$), which subsequently affected the F1 score negatively in a similarly significant manner.

This degradation in precision, and the corresponding drop in F1 score, is expected. As models attempt to make predictions further in advance of a potential mental health diagnosis, the amount of accessible clinical information becomes pro-

gressively limited. Consequently, the model is more prone to issuing false positives, which lowers precision. This finding aligns with clinical expectations. Early prediction inevitably involves a trade-off between sensitivity and specificity due to reduced observable symptomatology or contextual cues.

Table 4.5 presents a comparative performance analysis of LightGBM (with 128-dimensional embeddings) across different lead times. Statistical significance levels are reported with 0 lead time serving as the baseline.

Model (lead time in months)	**Accuracy (%)**	**ROC-AUC (%)**	**PR-AUC (%)**	**Recall (%)**	**Precision (%)**	**F1 (%)**
LGBM (0) [baseline]	**73.47**	**83.37**	**57.73**	77.59	**36.84**	**49.95**
LGBM (3)	72.40***	82.71**	57.44**	77.69	31.58***	44.90***
LGBM (6)	72.39***	82.79*	57.08**	**77.87**	30.48***	43.81***
LGBM (12)	72.55***	82.96**	57.30**	77.69	28.89***	42.12***

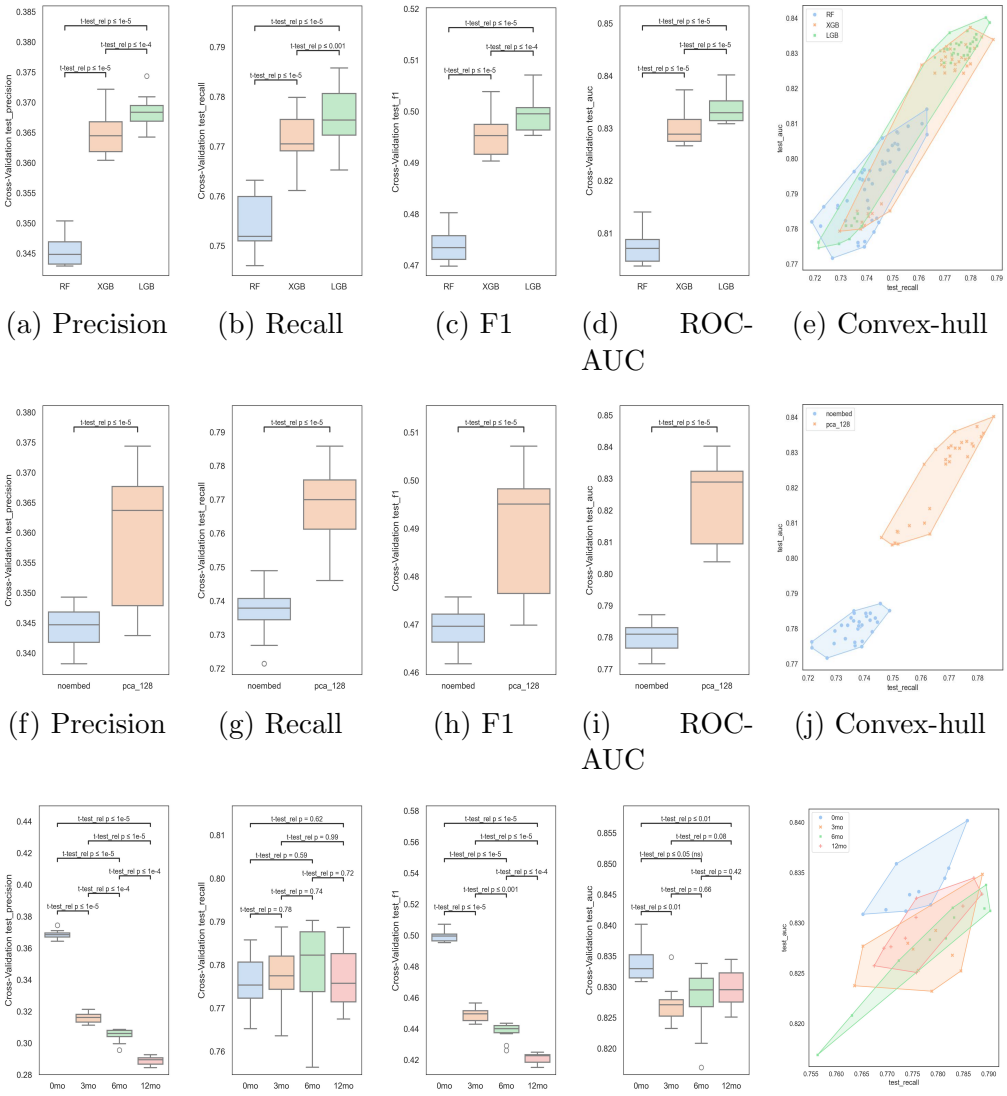
Table 4.5: Performance of LGBM with 128-dimensional embeddings for different prediction lead times [taking no lead time as baseline].

Significance levels: * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Additionally, to facilitate a clearer understanding and visual comparison of model performance across key metrics, we generated distribution plots alongside a two-dimensional convex hull representation, as illustrated in Figure 4.2.

Evaluation on the Held-out Test Set

Our best performing model, LightGBM with 128-dimensional embeddings at zero lead time, was evaluated on the held-out test set, yielding performance metrics comparable to those observed during model development, with an overall ROC-AUC of 83.23%, PR-AUC of 57.37% and a recall of 77.65%. These results substantiate the model's



(a) Precision (b) Recall (c) F1 (d) ROC-AUC (e) Convex-hull

(f) Precision (g) Recall (h) F1 (i) ROC-AUC (j) Convex-hull

(k) Precision (l) Recall (m) F1 (n) ROC-AUC (o) Convex-hull

Figure 4.2: Model Performance Comparisons. Panels (a–d) compare our three top performing models (Random Forest (RF), XGBoost (XGB), and LightGBM (LGBM)) on the 128-dimensional dataset with 0 lead time. Panels (f–i) present the performance uplift from no embeddings to incorporating 128-dimensional embeddings to our feature set for LightGBM on 0 lead time. Panels (k–n) illustrate the performance of our best-performing model (LGBM) across 0-, 3-, 6-, and 12-month lead times for 128-dimensional embeddings. Panels (e), (j), and (o) provide convex hull summaries for each respective row.

prediction capacity for early mental health assessment and confirm its generalizability to previously unseen data. A comprehensive summary of all evaluation metrics comparing the 10-fold cross-validation training results against the performance on the held-out test set is provided in Table 4.6.

Dataset	**Accuracy (%)**	**ROC-AUC (%)**	**PR-AUC (%)**	**Recall (%)**	**Precision (%)**	**F1 (%)**
Train set 10-fold CV	73.47	83.37	57.73	77.59	36.84	49.95
Holdout test set	73.51	83.23	57.37	77.65	36.88	50.01

Table 4.6: Performance comparison between the training (cross-validation) and hold-out test sets on LightGBM (128-dimensional embeddings, 0 lead time).

Chapter 5

Formulating Physician-aligned Design Guidelines for AI Summarization

This chapter presents our approach to understanding physician needs to better align LLM responses for EMR summarization. Our overarching goal is to establish physician-informed design guidelines for prompts that not only generate clinically useful summaries but also reflect the information-seeking behaviors and cognitive workflows of physicians. To realize this goal, we conducted a detailed user needs assessment study employing primarily qualitative thematic analysis, supplemented by quantitative analyses focusing on interaction patterns. To test the findings from our analysis in practice, we demonstrated an informed prompt engineering process, where we tailored a summarization prompt to align with physician feedback. We also described the additional preprocessing steps undertaken to prepare the EMR data for summarization along with the technical setup employed to deploy a locally hosted LLM for the experiment.

As previously outlined in Section 3.3, this study aims to elicit physician perspec-

tives across three key domains:

1. The prioritization of clinical data within EMRs to understand patient profiles during consultation
2. The preferred sequence and format for AI-generated patient summaries
3. The perceived value and utility of AI-generated summaries in improving clinical workflows

5.1 Methodology

5.1.1 Analysis Approach

Our thematic analysis involved a rigorous, multi-step approach to ensure methodological transparency and analytical rigor. First, we transcribed the recorded interviews conducted with participating physicians verbatim. Following transcription, we developed a codebook containing codes aligned explicitly with our research objectives. Each code was accompanied by a clear, concise definition to facilitate reliable application and consistency among coders.

Following this, two independent coders applied the developed codebook to the transcripts. To assess intercoder reliability, Cohen's Kappa statistic [84] was computed both overall, across all coded units, and individually for each predefined code. The overall Cohen's Kappa calculated for the two independent coders was 0.90, indicating a very high level of intercoder agreement. Furthermore, to address discrepancies for codes where intercoder agreement fell below the established reliability threshold ($\kappa < 0.8$), a third independent rater conducted consensus coding. This step involved carefully reviewing contested segments and determining final code assignments through discussion and re-evaluation.

Finally, themes were developed from the refined set of codes using a semi-deductive analytical process. This process was guided by our three primary research objectives, facilitating the systematic extraction of meaningful insights directly relevant to our study aims. A detailed account of the developed codebook and the corresponding thematic mappings is provided in the subsequent section.

5.1.2 Codebook and Thematic Mapping

To address our three central research questions, we developed a comprehensive codebook that systematically categorized physician responses from the think-aloud and post-task interview. Each code was designed to capture specific patterns in how physicians seek, prioritize, and synthesize patient information within an EMR, as well as their preferences for summary structure and their perceptions of / need for AI-generated summaries. Codes were then grouped under broader thematic categories aligned with our study's analytical framework: Information Seeking Behavior, Summary Structure, and Value Proposition of AI.

The frequency of each code's occurrence was recorded to highlight the relative salience of specific behaviors and viewpoints across participants. Table 5.1 presents the detailed list of codes, their corresponding themes, and the number of repetitions observed during analysis. Building on this codebook and thematic mapping, the subsequent Results section delves into each of the three overarching themes in detail.

Theme	**Code**	**#of repetitions**	**Intercoder agreement** (κ)
Information Seeking Behavior	seeks demographics	3	1
	seeks patient summary	5	1
	seeks past medical history	3	1
	seeks medications	7	1
	seeks allergies (context dependent)	1	1
	seeks hospitalizations	2	1
	seeks labs and exams (context dependent)	7	0.917
	seeks encounter notes	8	0.927
	seeks lifestyle factors	1	1
	seeks chronic illnesses	1	1
	filters for presenting complaint	10	0.804
	prefers realtime review	4	1
Summary Structure	summary needs demographics	3	1
	summary needs presenting complaint	6	0.781
	summary needs past medical history	6	0.641
	summary needs chronic illnesses	3	1
	summary needs medications	4	1
	summary needs medical procedures	3	1
	summary does not need labs	1	1
	summary does not need lifestyle information	1	1
	summary does not need allergies	3	1
	summary does not need medical procedures	1	1
	summary needs allergies (context dependent)	1	1
	summary needs encounter notes	3	0.654
	summary needs labs and exams (context dependent)	1	0.661
	summary needs hospitalizations	1	1
	prioritizes recent info	10	1
	prefer structured summary format	3	1
Value Proposition of AI	confused by data	3	1
	unfamiliarity with the interface	4	1
	finds value in an AI summary	4	0.883
	existing EMR critique	8	0.927

Table 5.1: Themes, codes, corresponding no. of repetitions and intercoder agreement (Cohen's Kappa κ) for all codes from the qualitative analysis.

5.2 Results

5.2.1 Theme One: Information Seeking Behavior

Prioritization of Demographics and Past Medical History

A prominent and recurrent pattern across physician interviews was the instinctive prioritization of patient demographics and past medical history during the initial stages of EMR review. Physicians consistently anchored their clinical orientation by first establishing a basic understanding of the patient's identity by looking through the demographic information section at the top of our EMR interface (detailed in Section 3.3.6), before delving into current complaints or the patient's medical case history. This is also supported by the high intercoder agreement ($\kappa = 1$) for the codes "seeks demographics" and "seeks past medical history" in Table 5.1. This initial framing appeared to serve as a mental scaffold, allowing them to contextualize subsequent information and guide their decision-making throughout the encounter. One physician exemplified this behavior by stating:

"Initially I'm looking at the top. I see her name. She's 41. She's female." [P1]

This demonstrated how demographic identifiers were the first touchpoints of engagement with the EMR. Following demographic orientation, physicians rapidly shifted their attention towards understanding the patient's historical medical context, emphasizing the need to quickly ascertain clinically significant past events. This is seen in the follow-up remark by the same physician:

"I want to know about her key [past] history ... chronic illnesses, major diagnoses, surgical and medical hospitalizations." [P1]

Another physician echoed this sentiment, immediately focusing on the patient's historical health data and demographic details, starting their EMR review saying:

"She takes some medications. She was seen for migraines. She was born in 1967 ..."

[P3]

Collectively, these examples reveal a shared information seeking strategy among physicians in which demographic and key historical data namely chronic illnesses, major diagnoses and, surgical and medical hospitalizations are central to the cognitive workflow of EMR navigation. The act of front-loading this information reflects an effort to orient oneself to facilitate an informed interpretation of the ongoing or current clinical issues.

Filtering By Presenting Complaint

Following the initial review of demographic details and past medical history, physicians systematically shifted their attention toward seeking information that directly corroborates the patient's presenting complaint. As detailed previously in Section 3.3.5, the scenario presented to our participants involved a patient whose presenting complaint was anxiety. This sub-theme captures the physicians' tendency to filter large volumes of clinical data through the lens of the current presenting issue, thereby narrowing their information-seeking behaviour to what was immediately relevant for the encounter at hand. This is evidenced by the high frequency with which the code "filters for presenting complaint" appeared during our physician interviews, as shown in Table 5.1.

Interaction patterns within the EMR interface revealed a clear prioritization of two key data sources for this purpose: past encounter notes and current or historical medications. Figure 5.1 illustrates the relative frequency of interactions (clicks/scrolls) across the various interactive components of our EMR UI as introduced in Section 3.3.6. As can be seen, more than 50% of the participants' interactions during the study were concentrated within the encounter notes and medications tabs. Among these two, the encounter notes tab exhibited the highest engagement, with physicians

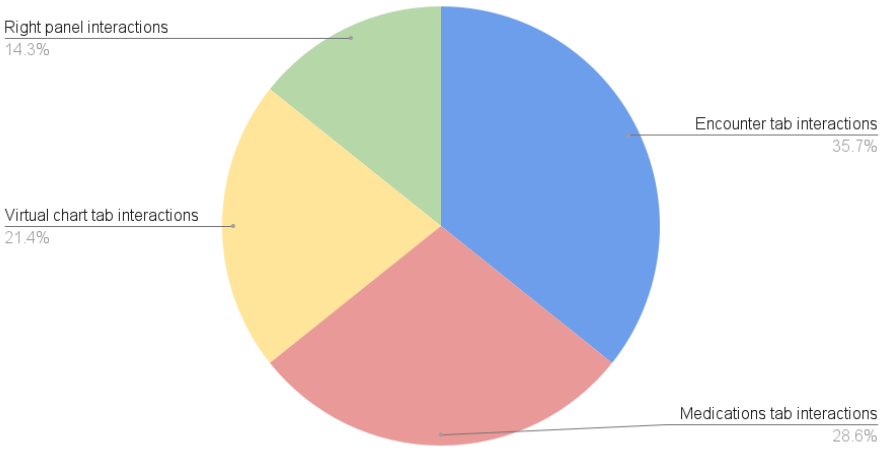


Figure 5.1: Relative frequency of interactions across interactive components of the EMR UI.

dedicating at least one-third of their total EMR review time on this tab alone. However, these tabs were not reviewed exhaustively but were instead selectively scanned for signals pertaining to the presenting issue. Physicians employed a goal-directed strategy, evaluating whether these data sources contained content relevant to the presenting complaint of anxiety and disregarding extraneous details. As one physician remarked:

"Quick scan of the (encounter) chart for other visits for anxiety ... shows nothing related." [P2]

The clinical reasoning underlying this behaviour was explicitly stated by another participant, who emphasized the importance of the presenting complaint in guiding their information seeking process:

"Well ... my main focus with the patient is the reason why they came in that day, right?" [P3]

These quotes illustrate a deliberate filtering process, whereby information irrelevant to the current concern is skipped to reduce cognitive load and maintain task focus.

However, this filtering may inadvertently lead physicians to overlook cues that are clinically significant but not explicitly linked to the presenting complaint of anxiety. For instance, details that do not mention anxiety directly, such as sleep disturbances, substance use, or psychosocial stressors may be bypassed, despite their potential relevance to diagnostic reasoning or treatment planning.

Lab Results and Allergies are Context Dependent

In contrast to the consistently prioritized review of demographics, past medical history, and presenting complaint, the examination of lab results and allergies was markedly more variable and contingent on clinical context. Physicians did not report routinely reviewing these data categories during every encounter. Instead, they engaged with them selectively, depending on the nature of the visit or the specific clinical decision being made. These reports were further substantiated by the physicians' interaction and engagement patterns observed during the study. Notably, only one of the three participating physicians interacted with the allergy section at any point during their EMR review. The Virtual Chart tab containing lab results received minimal attention across participants as well, with average engagement durations falling below seven seconds.

Some physicians described a preference for deferring lab review during the patient interaction itself. One physician noted:

> *"I get most of my benefit doing that (reviewing lab results) with the patient while I'm in the room with them and able to share my screen..." [P2]*

This indicates that lab data is often more actionable and relevant when discussed in real-time with patients, rather than pre-loading it. Others described a more task-oriented approach, choosing to review lab results only when prompted by a clinical concern. For example, one participant explained:

"I don't need to see the list presented to me of most recent labs and things because a lot of the time I don't need that. That was a one and done kind of thing. I would go into that later or try to find it if I needed it if required so." [P2]

This illustrates a just-in-time strategy, where lab data is accessed on demand, such as when an abnormal value (e.g., elevated HbA1c) might substantiate a suspicion of an undiagnosed condition like diabetes, but is otherwise not actively sought out during routine review.

Similarly, allergy information was not reviewed as a matter of course but was instead consulted in a highly specific clinical context: during the act of prescribing. As one physician succinctly stated:

"Allergies. I usually don't look at allergies until I'm about to prescribe a medication. Then I will look at allergies to make sure that I'm not going to prescribe something they're allergic to." [P1]

These accounts underscore that lab results and allergy information are not uniformly prioritized during EMR review. Rather, their relevance is evaluated dynamically, shaped by the clinical scenario and the physician's intent. Such context-dependent information seeking reflects a nuanced understanding of when data is likely to impact clinical reasoning, and when it can be safely deferred or omitted without compromising care.

5.2.2 Theme Two: Summary Structure

Recency

A key element physicians emphasized in their expectations for an ideal AI-generated summary was the prioritization of recent information. Across interviews, there was a clear consensus that clinical relevance is tightly linked to recency, as corroborated

by the high frequency of the code "prioritizes recent info" demonstrated in Table 5.1. What is happening now, or what has happened recently, is often more valuable than more remote historical data when preparing for or engaging in a patient encounter. As such, physicians expressed a strong preference that summaries foreground the latest clinical developments, including recent visits, current medications, and any ongoing issues that may be actively managed. Some physician quotes that encapsulated this sentiment are:

"I'm most interested in what is most current." [P1]

"I'm going to be looking back one or two visits....The last year is the most important..." [P2]

The last comment reveals not only a preference for recency but also a mental heuristic that physicians often use to bound their review of the record to the past one or two encounters or the preceding year.

Temporal filtering is a central organizing principle in physicians' mental models of clinical relevance. For LLM–generated summaries to be clinically useful, they must mirror this prioritization by elevating the most recent and ongoing elements of a patient's health record.

Preference of Structure Over Paragraphs

In envisioning the format of an ideal AI-generated summary, physicians consistently expressed a strong preference for structured outputs rather than free-form narrative paragraphs. The code "prefer structured summary format" appeared consistently during physician interviews and exhibited high intercoder agreement ($\kappa = 1$) (Table 5.1). The need for clarity, efficiency, and rapid information retrieval in clinical environments shaped their expectations for summaries that are modular, easily scannable,

and organized into distinct sections or bullet points. One physician articulated this preference directly, stating:

"Yeah, I would like it (the summary) separated into sections." [P1]

While another physician emphasized a bulleted format:

"Like bullet point. So not paragraphs, but just like ... can you summarize this patient in eight lines like 8 short statements?" [P2]

This comment captures both the preferred format and the desired level of conciseness. Rather than requiring physicians to extract key details from dense narrative text, a succinct, bulleted summary delivers core insights in a readily digestible form. Together, these perspectives reveal a clear expectation for AI-generated summaries to reflect a clinician-oriented structure.

Information Priority

Beyond the content and format of an ideal AI-generated summary, physicians were particular as to the sequence in which information should be presented. This subtheme captures the importance of aligning the structure of the summary not only with what information is included, but also with the order in which it appears.

Across interviews, physicians commonly articulated a preferred order that began with demographic information, followed by past medical history, recent clinical encounters, and current medications. Lab results and allergy information were consistently mentioned as being less universally relevant, and instead seen as context dependent (useful only in specific scenarios such as managing chronic conditions or prescribing medications). Table 5.2 provides a detailed breakdown of the information sequencing preferences expressed by each participant.

The table illustrates the individualized yet largely convergent ordering preferences across participants. Despite minor variations, the overall trend reflects a shared

Participant ID	Summary Information Priority				
	1	**2**	**3**	**4**	**5**
F1	Demographics	Past medical history	Encounters	Medications	Lab results (context dependent)
P1	Demographics	Past medical history (chronic illnesses)	Medical Procedures	Medications (active medications)	Lab results (abnormal)
P2	Past medical history (past diagnosis)	Encounters	Medications	Allergies (when prescribing)	Lab results (context dependent, specifically looks for trends)
P3	Demographics	Encounters	Medications	Lab results	Allergies (when prescribing)

Table 5.2: Sequence of information elements expected in a good summary, as reported by each participant (priority decreases left-to-right).

cognitive model among physicians. This insight provides actionable guidance for structuring AI-generated summaries in a way that mirrors and supports physicians' real world information seeking behaviour.

5.2.3 Theme Three: Value Proposition of AI

Interface Complexity

As part of our post-summary survey, participants completed a Single Ease Question (SEQ) rating, which is a 7-point scale where higher scores indicate greater ease of task completion (see Section 3.3.7 for details). The average score was 5.67, suggesting that the task of reviewing and summarizing the medical records was relatively manageable. Despite this, physicians expressed significant frustration with the design and usability of current EMR systems, highlighting how these shortcomings directly impair clinical efficiency and decision-making.

Several physicians described their current EMR interfaces as unintuitive and unnecessarily cumbersome, resulting in avoidable friction during clinical encounters. Codes like "confused by data" and "unfamiliarity with the interface" came up often

and with high interrater agreement ($\kappa = 1$) (Table 5.1). One physician noted:

"I find it (the EMR they routinely use) super hard and the filtering options are also complex and not easy to do ... I find the workflows (for information retrieval) in it really clunky and difficult ... even the encounter notes are a mess..." [P2]

This sentiment illustrates how even routine tasks such as locating a prior encounter note or information retrieval in general become time-consuming due to poor interface design and fragmented navigation. Another physician explained how these inefficiencies influence their behaviour, often discouraging more thorough data review:

"If there was more information, if I had to scroll through multiple windows to see everything, click into notes—that really becomes a lot more time consuming, right? ... I will avoid doing that if I don't need to." [P3]

This remark highlights a concerning dynamic: the structure of current EMRs not only delays information access but can actively deter physicians from exploring potentially relevant data due to information overload, especially under time constraints. As a result, clinical decisions may be made with incomplete context, not due to lack of data, but due to the barriers involved in accessing it. These critiques underscore a key motivation for integrating AI into clinical workflows: the potential to alleviate information overload and streamline the retrieval of relevant patient data.

Moreover, in the initial survey questionnaire, participants reported spending, on average, just under six minutes reviewing EMRs during typical clinical encounters. This self-reported estimate was subsequently corroborated during the summary generation task, wherein participants spent slightly more than six minutes on average to review the EMR and produce a corresponding summary. Notably, the integration of AI-generated summaries offers a transformative shift in this workflow: by automatically synthesizing relevant patient information, such tools can substantially reduce the time required for EMR review thus improving efficiency.

Need for AI Summaries

This sub-theme captures the ways in which participants envisioned the integration of AI, particularly LLM–based summarization tools as a means of enhancing their workflow, reducing cognitive burden, and supporting more efficient clinical decision-making. The potential of AI to minimize friction in navigating the EMR was emphasized repeatedly as referenced by the code "finds value in an AI summary" in Table 5.1. One physician expressed this clearly:

> *"OK, it would be great if I didn't have to look through all this stuff and if when I open the chart, I immediately saw the summary that we just described..." [P1]*

Others highlighted the efficiency gains such a tool could offer, while also acknowledging the prerequisite of accuracy and trust. As one physician noted:

> *"The AI summary piece would probably be helpful in speeding that (the EMR review) process up as long as you could trust it to be accurate" [P2]*

Our third physician also reflected on the cognitive and practical benefits of AI summarization, noting:

> *"I think with [an AI] summary tool, focusing on that approach that I kind of outlined and generating that kind of summary, I personally would find that to be helpful ... I mean, it might just save me a few clicks in terms of having to go through and collect all that information myself and then sort of storing it in my memory while I talk to the patient, right? ... That's the value proposition." [P3]*

Together, these perspectives converge on a shared vision: AI-generated summaries can enhance both clinical efficiency and patient care by reducing the time, effort, and mental overhead of physicians.

5.2.4 Study Implications

The findings from our thematic analysis revealed critical insights into how physicians seek, interpret, and structure clinical information during patient encounters. These insights informed the development of physician-aligned design guidelines aimed at enhancing the usefulness and clinical relevance of AI-generated summaries. To illustrate how these design guidelines translate in practice, we present a demonstration using a set of two summaries generated by a locally hosted LLM, showcasing outputs produced with and without the integration of our empirically derived guidelines.

Corroborating Presenting Complaint Before Seeking Additional Contextual Cues in EMR

Theme one highlighted that physicians consistently began by orienting themselves with demographic information. It also emphasized the salience of the presenting complaint, which helped physicians scope and filter relevant information (notably recent encounters and ongoing medications) from the EMR. However this introduced a notable limitation: filtering for the presenting complaint may inadvertently lead physicians to overlook cues that are clinically significant but not explicitly linked to the chief complaint. This insight revealed a potential opportunity for AI-generated summaries to surface relevant, yet non-obvious, contextual information that clinicians might otherwise miss. Furthermore, physicians emphasized the need to ground their understanding in the broader medical history of the patient, which includes chronic illnesses, major diagnoses, and histories of surgical or medical hospitalizations. In contrast, lab results and allergy information were seen as contextual, primarily referenced when prescribing medications or corroborating a differential diagnosis. These data elements were not regarded as universally essential for a generalized summary, as physicians noted that such information is often more appropriately reviewed during real-time decision-making in the consultation itself.

Structure and Recency as Key Priorities

Theme Two informed us that physicians preferred structured summaries over free-form paragraphs. Participants expressed that information should be organized into clearly delineated sections, possibly using bullet points. This theme also reinforced the importance of recency in information prioritization, especially recent encounters and medications. Our analysis also provided a clear sequence of information priorities: after demographic information, the presenting complaint, and past medical history, physicians consistently identified encounter notes and medications as the most critical sources of clinical insight. This prioritization was substantiated by interaction data, as detailed in Section 5.2.1, over 50% of all user interactions with the EMR interface occurred within the encounter notes and medications tabs. These insights provided a direct blueprint for the sequencing of sections in our prompt design.

Overcoming Navigational Barriers with Streamlined Retrieval

Theme three revealed how interface complexity affected physicians' information retrieval process and exacerbated the problem of information overload, key areas that AI integration may hope to solve. Physicians viewed AI integration as a potential remedy for these challenges, expressing optimism that AI-generated summaries could streamline their review process and improve efficiency in accessing clinically salient information. This optimism was substantiated by our empirical observations as well. On average, physicians required approximately six minutes to manually review a complete EMR, a task that an AI-generated summary can accomplish in mere seconds.

Furthermore, beyond improved summarization, AI also offers an opportunity to reimagine the EMR experience through intelligent augmentation, personalization, and automation. A key limitation of existing EMRs, is the lack of adaptability to varying clinical needs. Interfaces are often "one-size-fits-all," offering limited support for tailoring information displayed based on task context, user role, or patient complexity.

AI-driven interfaces, in contrast, hold the potential to dynamically adapt to physician intent. For instance, predictive modeling could anticipate what types of data a physician is likely to need based on the presenting complaint, past interactions, or even current workflow stage, and foreground that information accordingly.

Additionally, participants described the act of locating specific information as unnecessarily tedious, with poor organization and non-intuitive search functionality contributing to information avoidance. NLP tools, particularly those leveraging LLMs, can be integrated to support semantically aware search across the EMR. Rather than requiring exact keyword matches or manual navigation through rigid templates, physicians could query the system conversationally (e.g., "Show me this patient's recent cardiac-related events") and receive contextually relevant results, thereby reducing the friction associated with information retrieval.

Taken together, these opportunities point to a broader vision of AI integration in EMRs that goes beyond static summaries to encompass adaptive, anticipatory, and context-sensitive support. Physicians in our study did not just express a desire for automation, rather, they outlined a need for systems that better align with their cognitive workflows, that reduce unnecessary friction, and that ultimately allow them to focus more on patient care.

Shaping AI Prompts Using Design Guidelines

To operationalize these insights, our final prompt uses a chain-of-thought approach to guide the LLM through the summary generation process step-by-step. The prompt proceeds as follows:

```
You are an experienced physician. Given the longitudinal EMR information presented
to you, generate a structured summary of the patient's records (ideally using bullet
points). Include the sections outlined below.
Organize the summary using the following headings, in order:
1. Demographic Information (include age, sex, and the presenting complaint)
2. Past Medical History (include chronic illnesses, major diagnoses, and surgical
and medical hospitalizations)
3. Most Recent Encounters and Medications (emphasize content relevant to the
presenting complaint)
```

The initial segment of the prompt is informed by established best practices and prior literature [54, 63, 81] including the assignment of a role to the LLM (in our scenario, an experienced physician). This is followed by constraining the summary to a physician-informed predefined structure, enforced through explicitly defining the headings in the prompt. The sequence and content of these headings are derived from the insights generated through our thematic analysis. Key considerations include the explicit inclusion of the presenting complaint, clear delineation of relevant elements within the past medical history, and an emphasis on recency and relevance to the presenting complaint in the selection of prior encounters and medications.

The subsequent two subsections outline (i) the additional preprocessing steps necessary to transform the EMR data into a format suitable for input to an LLM, and (ii) the technical infrastructure employed to test our prompt in order to demonstrate its effect on an LLM's generated output before and after physician-informed prompt refinement.

Additional Preprocessing: To prepare the patient data for summarization, the final step in preprocessing involved transforming our cleaned, integrated EMR dataset into a temporally ordered narrative suitable for input to a language model. Each row in the dataset was converted to a short natural language sentence using domain-specific templates. These sentences were concatenated chronologically to produce a comprehensive narrative timeline for each patient. This narrative format preserved

both the sequence and content of events, enabling large language models to reason over the full patient history. Figure 5.2 provides a comprehensive representation of the overall preprocessing pipeline for summarization input.

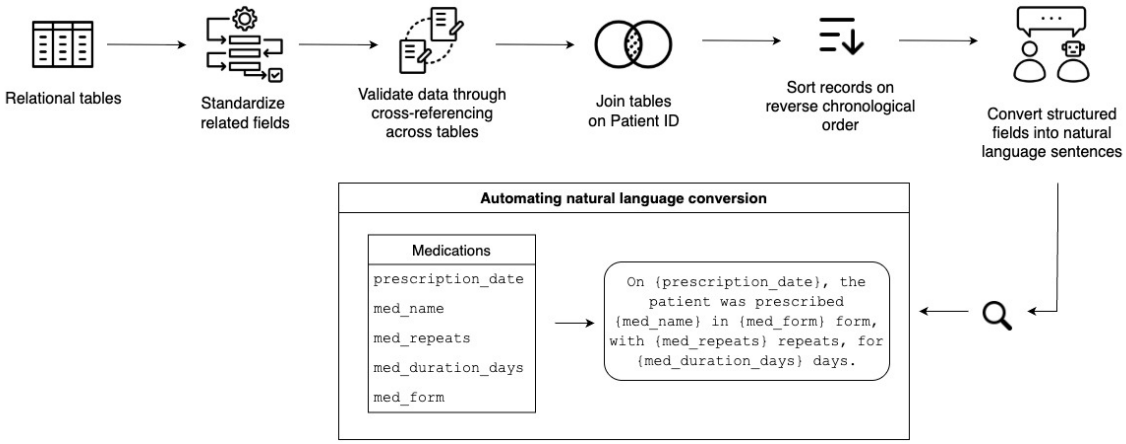


Figure 5.2: Overall Summary Preprocessing Pipeline

Technical Setup: To test our patient-informed prompt design, we locally deployed an LLM operating with an extended context window. Specifically, we utilized the Qwen 2.5 7B Instruct model, the same model used in Chapter 4 to generate our LLM embeddings, compiled for inference via llama.cpp, a lightweight C++ backend for efficient LLM execution. This configuration was hosted on secure, local infrastructure and exposed through a private RESTful endpoint. All inference requests comprising the full narrative history and a guiding prompt were routed internally, ensuring that no patient data ever left the local environment. This design explicitly addresses privacy and data sovereignty concerns that are often cited as barriers to the adoption of generative AI in clinical settings. All summaries were generated using a temperature setting of 0.1, consistent with best practices established in prior literature [81]. A low temperature value constrains randomness in the model's output, resulting in more consistent and predictable summaries across similar inputs. Minimizing variation helps ensure that the generated content remains focused, reproducible, and clinically reliable, an important consideration for physicians who need to trust the system's

reasoning and avoid unexpected deviations in patient summaries.

Summarization Outputs Pre- and Post- Physician-Informed Prompt Engineering

Below lists three summaries. The first written manually by a participating physician during our user study (referred to as the "S1: Manual Physician-written Summary"), the second using a prompt adapted from Van Veen et al. [81], which has been widely cited in literature for its effectiveness in producing coherent clinical summaries (referred to as the "S2: Conventional AI-generated Summary") and the third using the custom prompt informed by the design guidelines derived from our thematic analysis, as detailed in Section 5.2.4 (referred to as the "S3: Physician-informed AI-generated Summary").

S1: Manual Physician-written Summary

48 year old Female. Patient is presenting for an anxiety issue.

Quick scan of chart for other visits for anxiety in the last three to four visits shows nothing related.

Visits have been for migraine, cervical dysplasia and skin lesions.

A quick assessment of medications shows no psychoctive drugs, and patient summary is not elucidating any red flags for anxiety.

S2: Conventional AI-generated Summary

Prompt: You are am expert medical professional. Summarize the medical records into a concise clinical note.

Response: The medical records document various health conditions including classical migraine and dysplasia of the cervix, along with risk factors such as previous smoking and social alcohol use. The patient has been prescribed contraceptive medication multiple times for extended periods and has undergone gynecological examinations and procedures like myringotomy tubes.

S3: Physician-informed AI-generated Summary

Prompt: Included in Section 5.2.4 Shaping AI Prompts Using Design Guidelines

Response:

Demographic Information

Age: 48 years old

Sex: Female

Presenting Complaint: Anxiety

Past Medical History

Chronic Illnesses:

- Classical migraine (diagnosed on 2008-08-18)
- Dysplasia of cervix (diagnosed in multiple instances)

Major Diagnoses / Surgical and Medical hospitalizations:

- Varicose veins, leg (onset in 2004)
- Motor vehicle accident (MVA) (in 2003)
- History of high-risk sexual behavior and HPV infection

Most Recent Encounters and Medications

Encounters:

- On 2008-08-18, the patient was diagnosed with classical migraine*.

Medications:

- Oral contraceptive pills like TRIPHASIL 21 TAB were prescribed on multiple occasions from 2005 to 2007.

Additional Notes

The patient has a detailed health record with multiple gynecological issues and past medical history. The recent encounter on 2008-08-18 highlighted the presence of classical migraine. Given her history, it is important to consider this diagnosis in relation to her current presentation of anxiety, as both conditions can have overlapping symptoms or be influenced by similar factors such as stress and hormonal changes.

As we can observe from the three summaries above, both the physician-written summary (S1) and the physician-informed summary (S3) start with the demographic information and the presenting complaint, whereas the conventional summary (S2) omits this critical contextual information. Furthermore, S1 and S3 are organized into clearly delineated sections, but S2, by comparison, presents information as a single, monolithic paragraph. S1 and S3 also adhere to the information priorities expressed

by physicians during our study and include past medical history, information regarding recent encounters, and medications filtered for the presenting complaint. S2, although it includes some of this information (albeit without any structure), also includes irrelevant details such as risk factors. S2 also does not tailor its response to the presenting complaint. Additionally, S3 adds a separate section at the end, going over some of the relevant medical history of the patient that might be overlooked by the physician because it is not directly related to the presenting complaint of anxiety but may still be worth considering given overlapping symptoms or shared contributing factors.

Chapter 6

Conclusion, Limitations and Future Work

6.1 Conclusion

This thesis presents two distinct yet complementary contributions aimed at enhancing clinical workflows in primary care settings, with a particular emphasis on mental health.

First, we demonstrated the value of augmenting structured EMR features with embeddings derived from LLMs applied to unstructured clinical notes. Our findings indicate that incorporating these embeddings significantly improves the early detection performance of mental health disorders. Specifically, we observed an approximate 5% increase in ROC-AUC, a 3-4% increase in PR-AUC and a 2–3% improvement in recall when embeddings were included, highlighting their utility in capturing nuanced patient information not easily represented in structured data alone. Additionally, we found that lower-dimensional PCA embeddings performed as well as, or in some cases better than, higher-dimensional embeddings, suggesting that compact representations may be more effective and computationally efficient. We also uncovered a trade-off

between prediction lead time and performance, where increased lead times led to reduced precision.

Second, we conducted a user needs assessment study to gain insights on physician requirements, expectations, and underlying cognitive models involved in the review and synthesis of EMRs. By integrating physician needs into the formation of chain-of-thought prompts, we laid down a set of design principles that prioritize information salience, clinical reasoning, and usability. Beyond proposing prompt structures, our work establishes a methodological foundation for gathering and synthesizing physician feedback, a process that can be extended to other clinical specialties and electronic medical record interfaces. This contribution is particularly critical in ensuring that AI-generated outputs align with real-world physician workflows and cognitive models.

Taken together, these contributions support our broader vision for the future of AI-augmented healthcare. One in which AI models not only detect risk earlier and more accurately but also present information in ways that align with the needs and workflows of clinicians. Our early assessment framework enables the identification of patients at risk for mental health disorders, even in cases where such disorders are not the primary presenting complaint. Our summarization pipeline can leverage those same risk signals to surface relevant clinical information using a physician-informed structure, ensuring that physicians are presented with summaries that are contextually meaningful.

By reducing information overload and streamlining the clinical decision-making process, our contributions have the potential to alleviate clinician burnout, improve the efficiency of patient consultations, and most importantly, enhance the quality of care delivered to patients.

6.2 Implications of Findings

Our findings carry a number of implications for both methodological approaches in predictive modeling and the broader design of clinical decision-support systems.

First, it is important to note that the scope of our early assessment study did not include an investigation of individual feature importance. This decision was deliberate, as our primary objective was to examine whether augmenting structured EMR features with LLM-derived embeddings could improve predictive performance. By ensuring that both the baseline (no-embedding) model and the embedding-augmented model utilized the exact same set of structured features, we were able to establish a controlled setup to isolate the incremental benefits of introducing LLM embeddings. Moreover, existing studies in comparable clinical contexts have already conducted detailed analyses of structured feature importance, and these findings informed our own choice of features. Nevertheless, our second study on physician-informed design guidelines provides valuable insights into which types of information physicians consider most salient when reviewing EMRs. These physician preferences can serve as critical data points in refining future iterations of predictive modeling, particularly in guiding the prioritization of structured features to more closely align with clinical reasoning and decision-making practices.

Second, our early assessment study did not attempt to interpret what information is being encoded within the LLM embeddings themselves. While embeddings demonstrably improved predictive performance, the underlying representational mechanisms remain opaque. However, the complementary findings from our design guidelines study offer a way forward. Specifically, the physician-informed principles on what constitutes relevant and clinically useful information can be leveraged to design more targeted embedding generation pipelines. Instead of indiscriminately encoding all available free-text notes, future approaches could explicitly prioritize the extraction of clinically salient sections, thereby increasing both the interpretability and utility

of embeddings in medical prediction tasks. Such alignment between embedding design and physician information needs represents a promising avenue for bridging the gap between performance-driven improvements and clinically grounded model interpretability.

6.3 Limitations

While the findings of this thesis contribute meaningfully to the growing body of work on using structured and unstructured EMRs for both prediction and summarization in mental healthcare, several limitations must be acknowledged.

Augmenting Early Mental Health Assessment Using LLM Embeddings

In our first study, we utilized Principal Component Analysis (PCA) as the dimensionality reduction technique primarily due to its computational efficiency and effectiveness in preserving global variance within the dataset. However, PCA is inherently a linear dimensionality reduction method. Consequently, our analysis does not explore or compare PCA against more sophisticated non-linear techniques such as Uniform Manifold Approximation and Projection (UMAP), or neural network-based methods like autoencoders, which, although computationally demanding, may potentially capture more complex, non-linear relationships in the data.

Additionally, our investigation focused exclusively on embeddings derived from the Qwen 2.5 architecture, primarily due to its large context window, strong instruction-following capabilities, and relatively low resource requirements. But with the rapid advancement of LLMs, more specialized models trained exclusively on medical data are emerging, displaying comparable or superior characteristics. Thus, the comparative performance of embeddings generated from other LLM architectures remains unexplored. Including additional models could provide valuable insights into the

generalizability and robustness of our results.

It is also important to note that our study design was retrospective, inherently limiting our predictive models to findings based on associations rather than causal inferences. Nevertheless, due to the large sample size utilized in this study, the results maintain a high likelihood of robustness and reliability.

Physician-Informed Summarization of Patient Records Using LLMs

The second study also presents certain constraints. Specifically, the feedback extracted from our user study, which informed the prompt design for LLM-based summary generation, was exclusively gathered from family physicians. This choice potentially limits the generalizability of our findings, as medical professionals from other specialties may have distinct preferences or requirements regarding summary structure and content. Furthermore, our user studies utilized a single EMR interface for data presentation. As EMR systems can vary significantly in layout, functionality, and information hierarchy, this design choice may also influence physician interactions and expectations. Consequently, extending the user study to incorporate a broader range of medical specialties and diverse EMR platforms could significantly enhance the applicability and comprehensiveness of our results. Another limitation relates to the EMR interface we developed for the study. While modeled on a production EMR, several components of our demo interface were static rather than interactive. Although most of these design simplifications did not affect the study objectives, one missing feature, namely, the ability to click on an encounter to expand it for additional details, may have constrained physicians' ability to gather richer insights from the records. Future studies should account for such functionality to more closely approximate real-world EMR use.

6.4 Future Work

Building upon the findings and methodological foundation laid in this thesis, several directions emerge for future exploration.

Exploring Alternative Dimensionality Reduction Techniques

Firstly, future studies should investigate alternative dimensionality reduction methods beyond PCA, including UMAP, t-SNE, or neural autoencoders. Furthermore, an important baseline that remains to be established involves training predictive models using the full-dimensional embeddings directly, without applying any dimensionality reduction. Doing so will offer valuable insights into the potential information loss introduced by dimensionality reduction and will aid in better evaluating trade-offs between computational efficiency and model accuracy.

Exploring Additional LLM Architectures

Additionally, while this thesis utilized embeddings derived specifically from the Qwen 2.5 7B Instruct model, future research should assess a broader range of LLM architectures, including specialized models trained explicitly on medical data. Comparative studies involving such models could yield insights into the domain specificity required for effective clinical representation learning. Moreover, as these models mature and demonstrate instruction-following capabilities akin to general-purpose LLMs, they may offer improved alignment with clinical language and context.

Expanding User Studies Across Specialties and EMR Platforms

This thesis has laid the foundation for conducting structured user studies with physician participants, demonstrating effective ways of capturing and integrating physician feedback into the behavior of LLMs through carefully engineered prompts. Expanding

this approach, future studies should extend the participant pool beyond family physicians to include specialists across various medical fields. Such research could yield more nuanced insights, thereby enabling tailored prompt designs and summary generation pipelines that align closely with the unique informational needs and clinical workflows of different specialties.

Additionally, future investigations should incorporate a wider array of EMR interfaces, as the design and layout of these systems may significantly influence physician interaction patterns, information prioritization, and summary expectations. Exploring varied EMR platforms will help ensure that prompt engineering and summary generation strategies remain robust and adaptable across diverse clinical settings.

Evaluating the Clinical Utility and Quality of Generated Summaries

While we established comprehensive design guidelines for developing summarization prompts informed by physician needs, the summaries produced using these guidelines have yet to undergo rigorous evaluation. Future research should specifically investigate the clinical utility, quality, accuracy, and reliability of these physician-informed summaries. Systematic comparative evaluations against existing summarization methods or manually crafted physician summaries would offer deeper insights into their relative efficacy and potential impact on clinical workflow efficiency and patient care outcomes. Such empirical validations are essential to ensure that the developed summarization prompts meaningfully contribute to enhanced clinical decision-making and are effectively integrated into healthcare practice.

Fine-Tuning LLMs with Physician Expertise

Moreover, future work could advance beyond prompt engineering to include direct fine-tuning of LLMs with physician-derived annotations and feedback. This fine-tuning approach has the potential to better capture subtle physician preferences and

cognitive processes, potentially leading to enhanced summarization accuracy, clinical relevance, and user satisfaction.

Appendix A

Tables

Table A.1: ICD-9 codes to DSM-5 category mapping

DSM-5 category	**Associated ICD-9 codes**
Neurodevelopmental Disorders	'299', '315', '317', '318', '319', '314', '307'
Schizophrenia Spectrum and Other Psychotic Disorders	'295', '297', '298'
Bipolar and Related Disorders	'296'
Depressive Disorders	'311'
Anxiety Disorders	'300', '313'
Trauma- and Stressor-Related Disorders	'308', '309'
Sleep-Wake Disorders	'327', '347'
Sexual Dysfunctions	'302'
Substance-Related and Addictive Disorders	'305', '303', '304', '291', '292'
Personality Disorders	'301'
Neurocognitive Disorders	'294', '333', '331'
Feeding and Eating Disorders	'307'

Chronic Diseases

Tables A.3, A.4, A.5, A.6 and A.7 include the ICD-9 codes of all chronic diseases as defined by the Canadian Chronic Disease Surveillance System [58] as used in our study.

Table A.2: Physical comorbidities associated with DSM-5 categories

DSM-5 Category	Physical Comorbidities	Reference Publication
Anxiety Disorders	Allergies, Asthma, Back pain, Bowel disease, Bronchitis, Cataracts, Emphysema, GI ulcers, Heart disease, Migraine, Rheumatoid arthritis	El-gabalawy et al. 2011 [19]
Anxiety Disorders	Angina pectoris, Colitis, Constipation, Essential (primary) hypertension, GI Ulcer, Gallbladder problems, Hepatitis, Liver disease, Mitral valve prolapse, Myocardial infarction	Hä et al. 2003 [30]
Bipolar and Related Disorders	Essential (primary) hypertension, Thyroid disorders	Lu et al. 2022 [43]
Depressive Disorders	Acute coronary syndrome, Cancer, Diabetes, Essential (primary) hypertension, Stroke	Kang et al. 2015 [36]
Personality Disorders	Angina pectoris, Arteriosclerosis, Chest pain, Diabetes, Essential (primary) hypertension, GI Ulcer, Gastritis, Myocardial infarction, Rheumatoid arthritis, Stroke, Tachycardia	Quirk et al. 2015 [59]
Shizophrenia Spectrum and Other Psychotic Disorders	Blindness, Constipation, Diabetes, Dyspepsia, Epilepsy, Irritable bowel syndrome (IBS), Liver disease, Parkinson's disease, Viral hepatitis	Smith et al. 2013 [71]
Sleep-Wake Disorders	COPD, Chronic pain, Congestive heart failure, Diabetes, End-stage renal disease, Fibromyalgia, Gastroesophageal reflux, Myocardial infarction, Parkinson's disease, Rash and other nonspecific skin eruption, Rheumatoid arthritis, Spinal pain	Dikeos et al. 2011 [16]
Substance-Related and Addictive Disorders	Asthma, COPD, Chronic pain, HIV/AIDS, Traumatic brain injury, Tuberculosis, Viral hepatitis	UNODC 2022 [78]
Substance-Related and Addictive Disorders	Asthma, COPD, Cancer, Chronic kidney disease, Diabetes, Essential (primary) hypertension	Wu et al. 2018 [86]

Table A.3: ICD-9 codes for Musculoskeletal diseases

Disease	**ICD-9 Code**
GOUT	274
GOUTY NEPHROPATHY	274.1
GOUT WITH OTHER MANIFESTATIONS	274.8
UNSPECIFIED GOUT	274.9
DIFFUSE DISEASES OF CONNECTIVE TISSUE	710
ARTHROPATHY ASSOCIATED WITH INFECTIONS	711
CRYSTAL ARTHROPATHIES	712
ARTHROPATHY ASSOCIATED WITH OTHER DISORDERS CLASSIFIED ELSEWHERE	713
RHEUMATOID ARTHRITIS AND OTHER INFLAMMATORY POLYARTHROPATHIES	714
OSTEOARTHROSIS AND ALLIED DISORDERS	715
OTHER AND UNSPECIFIED ARTHROPATHIES	716
OSTEOPOROSIS	733

Table A.4: ICD-9 codes for Neurological diseases

Disease	**ICD-9 Code**
PRESENILE DEMENTIA	290.1
SENILE DEMENTIA, DEPRESSED OR PARANOID TYPE	290.2
SENILE DEMENTIA WITH ACUTE CONFUSIONAL STATE	290.3
ARTERIOSCLEROTIC DEMENTIA	290.4
OTHER SENILE AND PRESENILE ORGANIC PSYCHOTIC CONDITIONS	290.8
UNSPECIFIED SENILE AND PRESENILE ORGANIC PSYCHOTIC CONDITIONS	290.9
EPILEPSY	345
GENERALIZED CONVULSIVE EPILEPSY	345.1
PETIT MAL STATUS	345.2
GRAND MAL STATUS	345.3
PARTIAL EPILEPSY, WITH IMPAIRMENT OF CONSCIOUSNESS	345.4
PARTIAL EPILEPSY, WITHOUT MENTION OF IMPAIRMENT OF CONSCIOUSNESS	345.5
INFANTILE SPASMS	345.6
EPILEPSIA PARTIALIS CONTINUA	345.7
OTHER EPILEPSY	345.8
UNSPECIFIED EPILEPSY	345.9
PARKINSON'S DISEASE	332

Table A.5: ICD-9 codes for Cardiovascular diseases - Part 1

Disease	**ICD-9 Code**
HEART FAILURE	428
LEFT HEART FAILURE	428.1
ESSENTIAL HYPERTENSION	401
SPECIFIED AS BENIGN	401.1
ESSENTIAL HYPERTENSION NOT SPECIFIED AS MALIGNANT OR BENIGN	401.9
SECONDARY HYPERTENSION	405
SPECIFIED AS BENIGN	405.1
SECONDARY HYPERTENSION NOT SPECIFIED AS MALIGNANT OR BENIGN	405.9
ACUTE MYOCARDIAL INFARCTION	410
OTHER ACUTE AND SUBACUTE FORMS OF ISCHAEMIC HEART DISEASE	411
OLD MYOCARDIAL INFARCTION	412
ANGINA PECTORIS	413
OTHER FORMS OF CHRONIC ISCHAEMIC HEART DISEASE	414
ANEURYSM OF HEART	414.1
ISCHAEMIC HEART DISEASE OTHER	414.8
ISCHAEMIC HEART DISEASE UNSPECIFIED	414.9
SUBARACHNOID HAEMORRHAGE	430
INTRACEREBRAL HAEMORRHAGE	431
OTHER AND SPECIFIED INTRACRANIAL HAEMORRHAGE	432
SUBDURAL HAEMORRHAGE	432.1
UNSPECIFIED INTRACRANIAL HAEMORRHAGE	432.9

Table A.6: ICD-9 codes for Cardiovascular diseases - Part 2

Disease	**ICD-9 Code**
OCCLUSION AND STENOSIS OF PRECEREBRAL ARTERIES	433
CAROTID ARTERY	433.1
VERTEBRAL ARTERY	433.2
MULTIPLE AND BILATERAL	433.3
OTHER OCCLUSION AND STENOSIS OF CEREBRAL ARTERIES	433.8
UNSPECIFIED OCCLUSION AND STENOSIS OF CEREBRAL ARTERIES	433.9
TRANSIENT CEREBRAL ISCHAEMIA	435
ACUTE BUT ILL-DEFINED CEREBROVASCULAR DISEASE	436
OTHER AND ILL-DEFINED CEREBROVASCULAR DISEASE	437
OTHER GENERALIZED ISCHAEMIC CEREBROVASCULAR DISEASE	437.1
HYPERTENSIVE ENCEPHALOPATHY	437.2
CEREBRAL ANEURYSM, NONRUPTURED	437.3
CEREBRAL ARTERITIS	437.4
MOYAMOYA DISEASE	437.5
NONPYOGENIC THROMBOSIS OF INTRACRANIAL VENOUS SINUS	437.6
OTHER CEREBROVASCULAR DISEASE	437.8
UNSPECIFIED CEREBROVASCULAR DISEASE	437.9
LATE EFFECTS OF CEREBROVASCULAR DISEASE	438

Table A.7: ICD-9 codes for Chronic Respiratory diseases

Disease	**ICD-9 Code**
ASTHMA	493
INTRINSIC ASTHMA	493.1
UNSPECIFIED ASTHMA	493.9
BRONCHITIS, NOT SPECIFIED AS ACUTE OR CHRONIC	490
CHRONIC BRONCHITIS	491
MUCOPURULENT CHRONIC BRONCHITIS	491.1
OBSTRUCTIVE CHRONIC BRONCHITIS	491.2
OTHER CHRONIC BRONCHITIS	491.8
UNSPECIFIED CHRONIC BRONCHITIS	491.9
EMPHYSEMA	492

Table A.8: Feature descriptions used in the early assessment framework

Feature Name	**Description**
Sex_Bin	Binary representation of patient sex (Male = 0).
Age	Continuous variable representing patient age.
Num_Risk_Factors	Number of identified health risk factors.
Total Overall_Diagnoses	Total count of recorded diagnoses.
Total_Encounters	Count of patient healthcare interactions.
Lab_Risk_Proportion	Proportion of abnormal lab test results based on clinical thresholds.
Num_Chronic_Diseases	Count of chronic diseases identified.
Total_Med_Procs	Total number of medical procedures undergone by the patient.
Status_Active_Bin	Binary indicator if patient status is active.
Status_Deceased_Bin	Binary indicator if patient is deceased.
Status_Inactive_Bin	Binary indicator if patient status is inactive.
Status_Unknown_Bin	Binary indicator for unknown patient status.
Diabetes_Bin	Binary indicator for presence of diabetes.
Essential primary hypertension_Bin	Binary indicator for essential hypertension.
COPD asthma emphysema bronchitis_Bin	Binary indicator for respiratory conditions including COPD, asthma, emphysema, or bronchitis.
Viral hepatitis_Bin	Binary indicator for presence of viral hepatitis.
Chronic pain_Bin	Binary indicator for presence of chronic pain.
Rheumatoid arthritis_Bin	Binary indicator for rheumatoid arthritis.
Myocardial infarction_Bin	Binary indicator for myocardial infarction history.
Parkinson_s disease_Bin	Binary indicator for Parkinson's disease.
Stroke_Bin	Binary indicator for stroke history.
Angina pectoris_Bin	Binary indicator for angina pectoris.
Liver disease_Bin	Binary indicator for liver disease.
Migraine_Bin	Binary indicator for migraines.
Constipation_Bin	Binary indicator for chronic constipation.
Num_PhysComorb	Count of physical comorbidities calculated from ICD-9 codes.
LongTermMeds_Num	Number of medications taken for longer than 30 days.
ShortTermMeds_Num	Number of medications taken for less than 7 days.

Table A.9: Hyperparameter Grid for Each Model

Model	**Hyperparameter**	**Values**
kNN	n_neighbors	[5, 10, 20, 50, 100, 200]
	weights	['uniform', 'distance']
	metric	['euclidean', 'manhattan', 'minkowski']
	algorithm	['auto', 'ball_tree', 'kd_tree', 'brute']
Logistic Regression	penalty	['l1', 'l2', 'elasticnet', None]
	C	[0.01, 0.1, 1, 10, 100]
	solver	['liblinear', 'lbfgs', 'saga']
	max_iter	[100, 200, 500]
Random Forest	n_estimators	[100, 300, 500, 1000]
	max_depth	[None, 10, 20, 50]
	bootstrap	[True, False]
XGBoost	n_estimators	[100, 500, 1000]
	learning_rate	[0.01, 0.1, 0.3]
	max_depth	[3, 5, 10]
ADABoost	n_estimators	[50, 100, 200, 500, 1000]
	learning_rate	[0.01, 0.1, 1.0]
LightGBM	num_leaves	[30, 50, 100, 500]
	learning_rate	[0.01, 0.1, 0.3]
	n_estimators	[100, 500, 1000]
	max_depth	[-1, 10, 20]
	min_child_samples	[10, 20, 50]
	subsample	[0.5, 0.7, 1.0]

Table A.10: Best Hyperparameters Selected by GridSearchCV

Model	**Hyperparameter**	**With Embeddings**	**Without Embeddings**
kNN	n_neighbors	100	100
	weights	'distance'	'distance'
	metric	'manhattan'	'manhattan'
	algorithm	'auto'	'ball_tree'
Logistic Regression	C	1	0.01
	max_iter	500	500
	penalty	'l1'	None
	solver	'saga'	'saga'
Random Forest	n_estimators	1000	1000
	max_depth	20	20
	bootstrap	False	True
XGBoost	n_estimators	1000	500
	learning_rate	0.01	0.1
	max_depth	10	5
ADABoost	n_estimators	1000	1000
	learning_rate	1.0	1.0
LightGBM	num_leaves	500	100
	learning_rate	0.01	0.01
	n_estimators	1000	1000
	max_depth	20	-1
	min_child_samples	50	20
	subsample	0.5	0.5

Table A.11: Average aggregated evaluation metrics over 10-fold stratified cross-validation results across all model architectures by Embedding Dimensionality

Model	Dataset	Accuracy	ROC-AUC	PR-AUC	Recall	Precision	F1
KNN	no embeddings	66.18%	72.91%	34.32%	67.15%	31.86%	43.22%
	128 dims	82.17%	68.37%	31.97%	17.53%	44.46%	25.14%
	256 dims	81.86%	64.15%	28.76%	13.82%	40.99%	20.62%
	512 dims	79.40%	56.07%	29.21%	16.34%	31.33%	21.19%
LR	no embeddings	67.51%	65.53%	31.55%	49.83%	29.46%	37.03%
	128 dims	68.76%	70.25%	34.21%	58.58%	29.26%	39.03%
	256 dims	68.73%	70.23%	34.33%	58.69%	29.25%	39.05%
	512 dims	68.52%	70.01%	34.07%	58.70%	29.08%	38.89%
ADA	no embeddings	62.15%	72.13%	41.52%	70.57%	29.58%	41.68%
	128 dims	69.81%	78.13%	45.75%	72.09%	32.61%	44.91%
	256 dims	69.88%	78.14%	45.98%	71.97%	32.65%	44.92%
	512 dims	69.84%	78.12%	44.87%	71.78%	32.59%	44.83%
RF	no embeddings	67.69%	77.80%	48.24%	73.90%	34.16%	46.72%
	128 dims	71.42%	80.73%	51.73%	75.45%	34.55%	47.40%
	256 dims	70.68%	79.68%	51.47%	74.39%	33.72%	46.41%
	512 dims	69.84%	78.57%	50.29%	73.06%	32.79%	45.26%
XGB	no embeddings	68.18%	78.31%	51.37%	73.99%	34.57%	47.12%
	128 dims	73.16%	82.99%	55.78%	77.16%	36.46%	49.52%
	256 dims	73.07%	82.87%	55.37%	77.54%	36.43%	49.57%
	512 dims	72.69%	82.71%	55.01%	77.54%	36.05%	49.22%
LGBM	no embeddings	68.22%	77.95%	54.33%	73.17%	34.49%	46.88%
	128 dims	73.47%	83.37%	57.73%	77.59%	36.84%	49.95%
	256 dims	73.31%	83.24%	57.66%	77.82%	36.71%	49.88%
	512 dims	73.11%	83.04%	57.21%	77.41%	36.45%	49.56%

Table A.12: Average aggregated evaluation metrics over 10-fold stratified cross-validation results across all model architectures by Lead Times

Model	Lead Time (mo.)	Accuracy	ROC-AUC	PR-AUC	Recall	Precision	F1
KNN	0	82.17%	68.37%	31.97%	17.53%	44.46%	25.14%
	3	82.30%	66.45%	33.46%	22.98%	33.68%	27.28%
	6	81.68%	66.08%	33.42%	25.33%	30.52%	27.67%
	12	80.76%	65.97%	32.53%	29.35%	27.14%	28.18%
LR	0	68.76%	70.25%	34.21%	58.58%	29.26%	39.03%
	3	69.03%	70.16%	33.82%	58.27%	25.28%	35.26%
	6	68.92%	69.94%	33.68%	57.50%	23.97%	33.83%
	12	69.10%	69.96%	33.61%	58.12%	22.65%	32.59%
ADA	0	69.81%	78.13%	45.75%	72.09%	32.61%	44.91%
	3	69.31%	77.73%	44.48%	72.01%	28.12%	40.44%
	6	69.54%	77.87%	43.88%	72.14%	27.25%	39.56%
	12	69.76%	78.13%	43.80%	72.37%	25.85%	38.09%
RF	0	71.42%	80.73%	51.73%	75.45%	34.55%	47.40%
	3	70.60%	80.21%	51.01%	75.59%	29.65%	42.53%
	6	70.50%	80.26%	50.39%	75.54%	28.57%	41.47%
	12	70.88%	80.62%	50.46%	75.77%	27.25%	40.08%
XGB	0	73.16%	82.99%	55.78%	77.16%	36.46%	49.52%
	3	72.09%	82.41%	55.27%	77.59%	31.29%	44.59%
	6	71.97%	82.51%	54.88%	77.42%	30.05%	43.29%
	12	72.19%	82.64%	54.92%	77.46%	28.56%	41.73%
LGBM	0	73.47%	83.37%	57.73%	77.59%	36.84%	49.95%
	3	72.40%	82.71%	57.44%	77.69%	31.58%	44.90%
	6	72.39%	82.77%	57.08%	77.87%	30.48%	43.81%
	12	72.55%	82.96%	57.30%	77.69%	28.89%	42.12%

Appendix B

REB approval

Human Participant Ethics Protocol Submission

CONFIDENTIAL

0 - Identification

RIS Human Protocol Number
48505

Protocol Title
Understanding Physician EHR Workflow Needs and Validating AI-Driven Summarization Tool

Protocol Type
Investigator Submission

Applicant Information

Applicant Name
Nur Camellia Zakaria

Rank / Position
Asst Professor

Department / Faculty
Dalla Lana School of Public Health - Dalla Lana Sc

Business Telephone
416-978-1242

Extension

Email Address
CAMELLIA.ZAKARIA@UTORONTO.CA

Collaborators/Co-Investigators

Name	Department	Email	Phone	Designation	Alt Contact
Karim Keshavjee	Inst of Health Policy$ Mgmt & Evaluation	karim.keshavjee@utoronto.ca	647-405-9756	Co-Investigator	
Syed Muhammad Ibne Zulfiker	IHPME	syedmuhammad.ibnezulfiker@mail.utoronto.ca	6476564329	Co-Investigator & Alt	X

Projected Project Dates

Estimated Start Date
4-May-25

Estimated End Date
3-May-26

2 - Location

Location of the Research: ☒ University of Toronto ☐ Other Locations

Administrative Approval/Consent

Administrative Approval/Consent Needed: ○ Yes ◉ No

Community Based Particatory Research Project? ○ Yes ◉ No

Other Ethic Boards Approval(s)

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Another Institution or Site involved? ○ Yes ◉ No

3 - Agreements and Reviews

Funding

Project Funded? ○ Yes ◉ No

Explain why no funding is required

a preliminary work that feeds into a larger grant application

Agreements

Funding/non-funding Agreement in Place? ○ Yes ◉ No

Any Team Member Declared Conflict of Interest? ○ Yes ◉ No

Reviews

☐ This research has gone under scholarly review by thesis committee, departmental review committee, peer review committee, or some other equivalent

☐ This research will go under scholarly review prior to funding

☒ This review will not go under a scholarly review

4 - Potential Conflicts

Conflict of Interest

Will researchers, research team members, or immediate family members receive any personal benefit? ○ Yes ◉ No

Restrictions on Information

Are there any restrictions regarding access to, or disclosure of information (during or after closure)? ○ Yes ◉ No

Researcher Relationships

Are there any pre-existing relationships between the researchers and the researched? ○ Yes ◉ No

Collaborative Decision Making

Is this a community based project - i.e.: a collaboration between the university and a community group? ○ Yes ◉ No

5 - Project Details

Summary

Rationale

Describe the purpose and scholarly rationale for the project

Studies show that doctors spend more time reviewing EHRs than with patients. This impacts physician-patient interactions and has a negative impact on the quality of care. With the recent advancements in large language models, it is now possible to condense key information from an EHR record into an easily digestible summary.
Our proposed solution will summarize and consolidate relevant patient information using a large language model. Physicians can use this summary to quickly get a grasp of a patient's medical history before an encounter. This will save physicians time and allow for improvement in patient-physician interactions.
We propose a novel, explainable and privacy-preserving generative AI approach to creating summaries from patients' electronic health records. Our reasoning model will give us insights on an otherwise black box approach to how an LLM formulates its response. The large language model will be deployed on-premises to address any concerns associated with calling external third-party APIs on sensitive health information. Our proposed investigational approach assesses the quality, efficacy and usefulness of our solution.
Our primary objective is to study the quality, efficacy and usefulness of our proposed generative AI solution to create summaries of patient medical histories.
The secondary objectives include:

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

1. Establishing a physician-generated baseline summary of patient EHR records.
2. Identifying the most critical pieces of information pertaining to medical record summaries.
3. Measuring the value proposition/time saved through automating this process using AI.

Methods

Describe formal/informal procedures to be used

The primary goal of this study is to measure the quality, efficacy and usefulness of our proposed generative AI solution to create summaries of patient medical histories. To achieve this goal, we will conduct a 2-phase user study, first to establish a baseline for how EHR summaries should be and what the most relevant information are to physicians before a patient encounter. And secondly, to compare how close an LLM-generated summary comes to the established baseline.
Phase 1:
The first phase of the study will involve an interview with the first cohort of physicians going through a list of demographic and general experience questions. This will be followed by multiple identical tasks. Each task involves the physician going through an EHR record of a patient consisting of 20 to at most 100 entries. The physician will then be asked to summarize the information in the EHR records, as usually done during a patient encounter. The physicians will be given 15 minutes to complete the task. This will be followed by a 10-minute post-summary generation questionnaire.
The expected duration of the first phase is 30 minutes, including 5 minutes for the introductory demographic and general experience questionnaire, followed by 15 minutes to complete the tasks (as many summaries as can be generated within the allocated time) and another 10 minutes for the follow-up questionnaire.
Phase 2:
The second phase of the study will also involve an interview and multiple tasks for the second cohort of physicians. Each task will see the participants go through a pair of summaries, one of which is LLM-generated and the other physician-generated [from phase 1]. The participants will then be asked to rate the LLM-generated summary based on a standardized questionnaire.
For the second phase, the expected duration is 35 mins, with 5 mins for the pre and post interview questionnaire (altogether) and 30 minutes to complete the tasks (as many summaries as can be evaluated within the allocated time).
Copies of questionnaires, interview guided, and/or other instruments used
Refer to Appendix A [Questionnaire]

Copies of questionnaires, interview guided and/or other instruments used

Document Title	Document Date
Appendix A Questionnaire	2025-04-15

Clinical Trials

Is this a clinical trial? ○ Yes ◉ No

6 - Participants and Data

Participants and/or Data

What is the anticipated sample size of number of participants in the study? 30

Describe the participants to be recruited, or the individuals about whom personally identifiable information will be collected. List the inclusion and exclusion criteria. Where the research involves extraction or collection personally identifiable information, please describe where the information will be obtained, what it will include, and how permission to access said information is being sought.

The population to be studied is a group of physicians who deal with EHR records on a regular basis who provided informed consent.
Inclusion criteria:
- Physicians with experience using EHR systems
- Physicians willing to participate
- Physicians with a minimum of 1 year of practice in the field
- Physicians who are board-certified and licensed

Exclusion criteria:
- Physicians with little to no experience with EHRs
- Physicians who did not provide consent

Beyond the consent form, it is important to highlight that no personally identifiable information will be collected as part of the data collection process other than age and medical specialty of the participant. We will not be recording their names, email addresses, and staff identification. The only data collected about the participants will be their age, medical specialty, years of clinical experience, the languages they speak and their typical patient pool. The survey materials that we collect will be assigned with a random identifier. We encourage participants to remember / hold this information in a safe and convenient place should they need to contact us for future concerns.
In more detail, participants' confidentiality will be maintained during data analysis and publication/presentation of results using the following means:
1. Each participant will be assigned a random participation ID (e.g., a unique six-digit number) at the beginning of the study.
2. The researchers will save the data files by the participation ID, not by the participant's name or email address.
3. Only members of the research group will view the research data, while data with identifiable information will only be accessible by the principal investigator.
4. All data files will be stored in a secured data server accessed only by authorized researchers involved in this research project.
5. Results of data analysis in publications or presentations will be done in an anonymized and aggregated manner.

In the event a participant requests removal from the study, they will need to provide the research team with their participation ID. Accordingly, participants will

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

need to state if they want all data collected to be deleted and/or excluded from our analysis and reporting.
At this preliminary stage, we target a sample size of 30. With 10 participants for phase 1, and 20 for phase 2.

Is there any group or individual-level vulnerability related to the research that needs to be mitigated (for example, difficulty understanding consent, history of exploitation by researchers, or power differential between the researcher and the potential participant)? ○ Yes ◉ No

Recruitment

Is there recruitment of participant? ◉ Yes ○ No

Recruitment details including how, from where, and by whom

The recruitment will primarily involve direct engagement through word of mouth. The PI, project team members, and potentially other engaged colleagues may encourage participation informally by discussing the study with peers and colleagues.
The RA will spearhead the recruitment efforts. The PI will answer any initial questions potential participants may have about the study. The research assistant or other project team members may also help disseminate information through word of mouth. Their role will be to spread awareness and encourage participation.
Other academic or administrative staff may assist informally, especially if they have access to specific groups of participants who are relevant to the study (e.g., through research groups or administrative circles).

Is participant observation used? ◉ Yes ○ No

Participant Observation Details

In this study, participant observation will be employed to record and analyze the time it takes for participants to complete specific tasks under natural working conditions. During each observation session, researchers will discreetly monitor participants as they engage in these tasks, noting the precise start and end times using digital timing tools to ensure accuracy. This structured observational approach will provide qualitative insights into the context and behaviors that might affect performance. All data will be collected in accordance with ethical standards, safeguarding participant confidentiality and ensuring that informed consent is obtained prior to observation. This method allows for a comprehensive understanding of performance efficiency and offers valuable feedback on workflow processes in real-world scenarios.
Refer to Appendix B [Recruitment letter]

Will translation materials be used/required? ○ Yes ◉ No

Attach copies of all recruitment posters, flyers, letters, email text, or telephone scripts

Document Title	Document Date
updated recruitment letter to clarify study duration length at different phases, option to remove data at any time after study.	2025-06-09

Compensation

Will the participants receive compensation? ○ Yes ◉ No

Non Compensation Description

No direct compensation will be provided to the participants. Study outcomes may provide new knowledge on an optimal generative AI model to summarize patien

Is there a withdrawal clause in the research procedure? ◉ Yes ○ No

Is compensation affected when a participant withdraws?

In this study, no compensation is provided to participants; therefore, the issue of compensation being affected by withdrawal is not applicable.

7 - Investigator Experience

Investigator Experience with this type of research

Please provide a brief description of the previous experience for this type of research by the applicant, the research team, and any persons who will have direct contact with the applicants. If there is no previous experience, how will the applicant and research team be prepared?

Our research team is composed of eligible and experienced professionals who have a strong background in conducting user studies. Each member has successfully completed TCPS2 and/or CITI (Collaborative Institutional Training Initiative) training, ensuring that we adhere to the highest ethical standards in human subject research. This training equips us with the necessary knowledge about informed consent, data privacy, and participant rights, enabling us to conduct our studies responsibly and effectively.

Are community members collecting and/or analyzing data? ○ Yes ◉ No

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

8 - Possible Risks and Benefits

Possible Risks

Potential Risk Details:

Physical Risks	○ Yes	◉ No
Psychological/emotional Risks	○ Yes	◉ No
Social Risk	◉ Yes	○ No
Legal Risk	◉ Yes	○ No

Risk Description

The collected data from participants carry a potential privacy risk of unauthorized access or data breaches.
To mitigate privacy and security risks, each patient's data will be anonymized, ensuring no personally identifiable information is included in the data set. In more detail, participants' confidentiality will be maintained during data analysis and publication/presentation of results using the following means:
1. Each participant will be assigned a random participation ID (e.g., a unique six-digit number) during the briefing session.
2. The researchers will save the data files by the participation ID, not by the participant's name or email address.
3. Only members of the research group will view the research data, while data with identifiable information will only be accessible by the primary investigator.
4. All data files will be stored in a secured data server accessed only by authorized researchers involved in this research project.
5. Results of data analysis in publications or presentations will be done in an anonymized and aggregated manner.
6. In the event a participant requests removal from the study, they will need to provide the research team with their participation ID. Participant data will be removed upon withdrawal.

Potential Benefits

Benefit Description

There is no direct benefit to the participants. Study outcomes may provide new knowledge on an optimal generative AI model to summarize patient medical history by synthesizing their electronic health records.

9 - Consent

Consent Process Details

The participants will be given time to carefully read the consent form (refer to Appendix C [Consent Form]), after which they will provide written consent. The study will only proceed once consent has been obtained.

Uploaded letter/consent form(s)

Document Title	Document Date
updated study title, clarifed study duration length at different phases, and clarified option to remove data at any time after study.	2025-06-09

Is there additional documentation regarding consent such as screening materials, introductory letters etc.: ○ Yes ◉ No

Uploaded letter/consent form(s)

Will any information collected in the screening process - prior to full informed consent to participate in the study - be retained for those who are later excluded or refuse to participate in the study? ○ Yes ◉ No

Is the research taking place within a community or organization which requires formal consent be sought prior to the involvement of the individual participants ○ Yes ◉ No

Are any participants not capable (e.g.: children) of giving competent consent? ○ Yes ◉ No

10 - Debriefing and Dissemination

DeBrief

Will deception or intentional non disclosure be used? ○ Yes ◉ No

Will a written debrief be used? ○ Yes ◉ No

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Do participants/communities have the right to withdraw their data following the debrief? (•) Yes ( ) No

Withdrawal Process Details

Participants will be informed by the research investigator that they may, at any stage of the study, choose to skip providing specific data, discontinue participation in the research procedures, or withdraw from the study entirely. Participant data will be removed upon withdrawal. Should you decide to withdraw from the study after we have collected your data at any time, you have the option of removing your data. However, you will have to inform our research team with your random identifier assigned.

Information Feed Back Details following completion of a participants participation in the project

1. Participants have the option to request their search data only by providing us with their random identifier.
2. We will be using the information we gathered, particularly the anonymized research data in aggregated forms as part of our conference report, which will be accessible upon successful publication.

Procedural details which allow participants to withdraw from the project

Discontinuation criteria will include participants' requests to withdraw from the study at any time in order to attend to immediate patient needs or other clinical obligations. In such cases, participation will cease immediately to prioritize patient well-being.

[ ] Not Applicable

What happens to a participants data and any known consequences related to the removal of said participant

During study: Participants can terminate at any time and data collected in real-time will be deleted immediately.
Post-study: The participant will need to provide the random identifier we have issued to them. Note that without the random identifier, we will not be able to match the research data for deletion. Hence, they will be advised to retain this information even after the study in case they change their mind on providing the data to us.

[ ] Not Applicable

List reasons why a participant can not withdraw from the project (either at all or after a certain period of time)

Not applicable

[ ] Not Applicable

11 - Confidentiality and Privacy

Confidentiality

Is the data confidential? (•) Yes ( ) No

Will the confidentiality of the participants and/or informants be protected? (•) Yes ( ) No

List confidentiality protection procedures

A consent form reflecting the participant's information will be locked in a secured cabinet for 10 years and will not reflect the random identifier we will issue them with. The principal investigator will be the only one managing this record.

Are there any limitations on the protection of participant confidentiality? ( ) Yes (•) No

Is participant anonymity/confidentiality not applicable to this research project? ( ) Yes (•) No

Data Protection

Describe how the data (including written records, video/audio recordings, artifacts and questionnaires) will be protected during the conduct of the research and subsequent dissemination of results

We will not be collecting personally identifiable information, particularly audio and video recordings at any stage.
Text-based personally identifiable information will be contained to only the consent form and demographic information like age and medical specialty. The consent form will not link the participant's issued random identifier. Only aggregated and anonymized results will be used for dissemination.

Explain for how long, where and what format (identifiable, de-identified) data will be retained. Provide details of their destruction and/or continued storage. Provide a justification if you intend to store identifiable data for an indefinite length of time. If regulatory requirements for data retention exists, please explain.

All records will be maintained by the PI. To enable evaluations and /or audits from the REB, the PI agrees to keep study related records, all regulatory documents, and copies of all case report forms. To comply with Health Canada Regulations, the PI will retain the records for 10 years.

Will the data be shared with other researchers or users? ( ) Yes (•) No

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

12 - Level of Risk and Research Ethics Board

Level of Risk for the Project

Group Vulnerability: Low

Research Risk: Low

Risk Level: 1

Explanation/Justification

Explanation/Justification detail for the group vulnerabilty and research risk listed above

The proposed user study, while carrying a minor potential privacy risk of unauthorized access or data breaches, adheres to all necessary precautions to minimize this risk. The study is designed to focus objectively on assessing the quality, efficacy and usefulness of the proposed generative AI solution to create summaries of patient medical histories and does not involve any sensitive or invasive questions. Given these factors, the level of research risk is classified as low, and there is no indication of increased vulnerability among the participant group.

Research Ethics Board

REB Associated with this project: Health Sciences

13 - Application Documents Summary

Uploaded Documents

Document Title	Document Date
Appendix A Questionnaire	2025-04-15
updated recruitment letter to clarify study duration length at different phases, option to remove data at any time after study.	2025-06-09
updated study title, clarifed study duration length at different phases, and clarified option to remove data at any time after study.	2025-06-09

14 - Applicant Undertaking

I confirm that I am aware of, understand, and will comply with all relevant laws governing the collection and use of personal identifiable information is research. I understand that for research involving extraction or collection of personally identifiable information, provincial, federal, and/or international laws may apply and that any apparent mishandling of said personally identifiable information, must be reported to the office of research ethics.

As the Principal Investigator of the project, I confirm that I will ensure that all procedures performed in accordance with all relevant university, provincial, national, and/or international policies and regulations that govern research with human participants. I understand that if there is any significant deviation in the project as originally approved, I must submit an amendment to the Research Ethics Board for approval prior to implementing any change.

☒ I have read and agree to the above conditions

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

RIS Protocol Number: 48505

Approval Date: 22-Jun-25

PI Name: Nur Camellia Zakaria

Division Name:

Dear Nur Camellia Zakaria:

Re: Your research protocol application entitled, "Understanding Physician EHR Workflow Needs and Validating AI-Driven Summarization Tool"

The Health Sciences REB has conducted a Delegated review of your application and has granted approval to the attached protocol for the period 2025-06-22 to 2026-06-21.

This approval covers the ethical acceptability of the human research activity; please ensure that all other approvals required to conduct your research are obtained prior to commencing the activity.

Please be reminded of the following points:

- An **Amendment** must be submitted to the REB for any proposed changes to the approved protocol. The amended protocol must be reviewed and approved by the REB prior to implementation of the changes.
- An annual **Renewal** must be submitted for ongoing research. Renewals should be submitted between 15 and 30 days prior to the current expiry date.
- A **Protocol Deviation Report** (PDR) should be submitted when there is any departure from the REB-approved ethics review application form that has occurred without prior approval from the REB (e.g., changes to the study procedures, consent process, data protection measures). The submission of this form does not necessarily indicate wrong-doing; however follow-up procedures may be required.
- An **Adverse Events Report (AER)** must be submitted when adverse or unanticipated events occur to participants in the course of the research process.
- A **Protocol Completion Report** (PCR) is required when research using the protocol has been completed.
- If your research is funded by a third party, please contact the assigned Research Funding Officer in Research Services to ensure that your funds are released.

Best wishes for the successful completion of your research.

OFFICE OF RESEARCH ETHICS
McMurrich Building, 12 Queen's Park Crescent West, 2nd Floor, Toronto, ON M5S 1S8 Canada
Tel: +1 416 946-3273 ● Fax: +1 416 946-5763 ● ethics.review@utoronto.ca ● http://www.research.utoronto.ca/for-researchers-administrators/ethics

Appendix C

Full survey questionnaire

C.0.1 Demographic and general experience questionnaire

- Refer to the full questionnaire in Table C.1.

Table C.1: Demographic and General Experience Questionnaire.

Question	**Response (free text)**
1. What is your *medical specialty*?	
2. Where are you currently practicing? If not, where did you use to practice?	
3. *Age* (years)	
4. *Gender*	
5. How many *years of clinical experience* do you have?	
6. What languages do you use to communicate with your patients during your professional practice?	
7. How much time do you typically spend reviewing patient records before a consultation?	

C.0.2 Phase 1: Post-summary generation questionnaire

- Refer to the full questionnaire in Table C.2.

Table C.2: Post-Summary Generation Questionnaire.

Question
1. **Single Ease Question (SEQ):** "Overall, how difficult or easy was it to perform this task?"
1 **2** **3** **4** **5** **6** **7** Very Difficult Neutral Very Easy

C.0.3 Phase 2: Quality evaluation questionnaire

- Which of the two summaries do you prefer? [Summary A, Summary B, or both equally]
- The adapted PDQI-9 questionnaire. Refer to the full questionnaire in Table C.3.

C.0.4 Phase 2: Usefulness questionnaire

- We adopted UTAUT validated scale for our usefulness analysis. The questionnaire uses a 1-7 Likert scale, where 1 is "strongly disagree" and 7 is "strongly agree". Refer to the full questionnaire in Table C.4.

Table C.3: Adapted Physician Documentation Quality Instrument (PDQI).

Attribute	**Score**					**Description of Ideal Note**
1. Accurate	*Not at all*				*Extremely*	The note is true. It is free of incorrect information.
	1	2	3	4	5	
2. Thorough	*Not at all*				*Extremely*	The note is complete and free from omission and documents all of the issues of importance to the patient.
	1	2	3	4	5	
3. Useful	*Not at all*				*Extremely*	The note is extremely relevant, providing valuable information and/or analysis.
	1	2	3	4	5	
4. Organized	*Not at all*				*Extremely*	The note is well-formed and structured in a way that helps the reader understand the patient's clinical course.
	1	2	3	4	5	
5. Comprehensible	*Not at all*				*Extremely*	The note is clear, without ambiguity or sections that are difficult to understand.
	1	2	3	4	5	
6. Succinct	*Not at all*				*Extremely*	The note is brief, to the point, and without redundancy.
	1	2	3	4	5	
7. Synthesized	*Not at all*				*Extremely*	The note reflects the author's understanding of the patient's status and plan of care.
	1	2	3	4	5	
8. Internally consistent	*Not at all*				*Extremely*	No part of the note ignores or contradicts any other part.
	1	2	3	4	5	
9. Free from Hallucination	*Not at all*				*Extremely*	The note is free of hallucination and only contains information verifiable by the EMR.
	1	2	3	4	5	
10. Free from Bias	*Not at all*				*Extremely*	The note is free of bias and contains only information verifiable by the transcript and not derived from characteristics of the patient or visit.
	1	2	3	4	5	
Total Score						

Table C.4: UTAUT Questionnaire: Constructs and Associated Items.

Construct	**Variable / Construct**	**Question**
Performance Expectancy	PE 1: Perceived usefulness	Using the system in my job would enable me to complete tasks more quickly.
	PE 2: Job fit	Use of the system can increase the quantity of output for the same amount of effort.
	PE 3: Relative advantage	Using the system enhances my effectiveness on the job.
Effort Expectancy	EE 1: Perceived ease of use	My interaction with the system would be clear and understandable.
	EE 2: Ease of use	Overall, I believe the system is easy to use.
	EE 3: Complexity	Working with the system is so complicated, it is difficult to understand what is going on.
Social Influence	SI 1: Subjective Norm	People who influence my behavior would think that I should use the system.
	SI 2: Subjective Norm	People who are important to me would think that I should use the system.
	SI 3: Social factors	In general, the organization would support the use of this system.
Facilitating Conditions	FC 1: Perceived behavioral control	I have the resources necessary to use the system.
	FC 2: Perceived behavioral control	Given the resources, opportunities and knowledge it takes to use the system, it would be easy for me to use.
	FC 3: Compatibility	I think that using the system fits well with the way I like to work.

Bibliography

[1] Asrar Aldadi, Kathryn A Robb, and Andrea Williamson. Factors influencing multiple non-utilised healthcare appointments from patients' and healthcare providers' perspectives: a qualitative systematic review of the global literature. *BJGP open*, 8(4), 2024.

[2] Pedro Alves, Carl D Marci, Chandra J Cohen-Stavi, Katelynn Murray Whelan, and Costas Boussios. A machine learning model using clinical notes to estimate phq-9 symptom severity scores in depressed patients. *Journal of Affective Disorders*, 2025.

[3] PV AshaRani, Mohamed Zakir Karuvetil, Tan Yeow Wee Brian, Pratika Satghare, Kumarasan Roystonn, Wang Peizhi, Laxman Cetty, Noor Azizah Zainuldin, and Mythily Subramaniam. Prevalence and correlates of physical comorbidities in alcohol use disorder (aud): a pilot study in treatment-seeking population. *International Journal of Mental Health and Addiction*, 21(4):2508–2525, 2023.

[4] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[5] Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ette, Jamil Zaghir, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogr-

lic, and Christian Lovis. Scientific evidence for clinical text summarization using large language models: Scoping review. *Journal of Medical Internet Research*, 27:e68998, 2025.

[6] Santiago Berrezueta-Guzman, Mohanad Kandil, María-Luisa Martín-Ruiz, Iván Pau de la Cruz, and Stephan Krusche. Exploring the efficacy of robotic assistants with chatgpt and claude in enhancing adhd therapy: Innovating treatment paradigms. In *2024 International Conference on Intelligent Environments (IE)*, pages 25–32. IEEE, 2024.

[7] Richard Birtwhistle, Karim Keshavjee, Anita Lambert-Lanning, Marshall Godwin, Michelle Greiver, Donna Manca, and Claudia Lagacé. Building a pan-canadian primary care sentinel surveillance network: initial development and moving forward. *The Journal of the American Board of Family Medicine*, 22(4):412–422, 2009.

[8] Sue Bowman. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in health information management*, 10(Fall):1c, 2013.

[9] Nicholas C Cardamone, Mark Olfson, Timothy Schmutte, Lyle Ungar, Tony Liu, Sara W Cullen, Nathaniel J Williams, and Steven C Marcus. Classifying unstructured text in electronic health records for mental health prediction models: Large language model evaluation study. *JMIR Medical Informatics*, 13(1):e65454, 2025.

[10] Centre for Addiction and Mental Health. Mental Illness and Addiction: Facts and Statistics, 2025. Accessed on 13 June 2025.

[11] Aditya Choudhary, Sarthak Pawar, and Yashodhara Haribhakta. Efficient malware detection with optimized learning on high-dimensional features. *arXiv preprint arXiv:2506.17309*, 2025.

[12] Filip Dabek, Peter Hoover, Kendra Jorgensen-Wagers, Tim Wu, and Jesus J Caban. Evaluation of machine learning techniques to predict the likelihood of mental health conditions following a first mtbi. *Frontiers in neurology*, 12:769819, 2022.

[13] Christopher G Davey and Patrick D McGorry. Early intervention for depression in young people: a blind spot in mental health care. *The Lancet Psychiatry*, 6(3):267–272, 2019.

[14] Azam Dekamin, MIM Wahab, Aziz Guergachi, and Karim Keshavjee. Fius: Fixed partitioning undersampling method. *Clinica Chimica Acta*, 522:174–183, 2021.

[15] Azam Dekamin, MIM Wahab, Karim Keshavjee, and Aziz Guergachi. High cardiovascular disease risk-associated with the incidence of type 2 diabetes among prediabetics. *European Journal of Internal Medicine*, 106:56–62, 2022.

[16] Dimitris Dikeos and Georgios Georgantopoulos. Medical comorbidity of sleep disorders. *Current Opinion in Psychiatry*, 24(4):346–354, 2011.

[17] Mehmet E Dokucu and C Robert Cloninger. Personality disorders and physical comorbidities: a complex relationship. *Current opinion in psychiatry*, 32(5):435–441, 2019.

[18] Russell Franco D'Souza, Shabbir Amanullah, Mary Mathew, and Krishna Mohan Surapaneni. Appraising the performance of chatgpt in psychiatry using 100 clinical case vignettes. *Asian Journal of Psychiatry*, 89:103770, 2023.

[19] Renée El-Gabalawy, Corey S Mackenzie, Shahin Shooshtari, and Jitender Sareen. Comorbid physical health conditions and anxiety disorders: a population-based exploration of prevalence and health outcomes among older adults. *General hospital psychiatry*, 33(6):556–564, 2011.

[20] Wei Feng, Honghan Wu, Hui Ma, Yuechuchu Yin, Zhenhuan Tao, Shan Lu, Xin Zhang, Yun Yu, Cheng Wan, and Yun Liu. Deep learning based prediction of depression and anxiety in patients with type 2 diabetes mellitus using regional electronic health records. *International Journal of Medical Informatics*, page 105801, 2025.

[21] Elizabeth Ford, Philip Rooney, Seb Oliver, Richard Hoile, Peter Hurley, Sube Banerjee, Harm van Marwijk, and Jackie Cassell. Identifying undetected dementia in uk primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. *BMC medical informatics and decision making*, 19:1–9, 2019.

[22] Sajjad Fouladvand, Michelle M Mielke, Maria Vassilaki, Jennifer St Sauver, Ronald C Petersen, and Sunghwan Sohn. Deep learning prediction of mild cognitive impairment using electronic health records. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 799–806. IEEE, 2019.

[23] David Fraile Navarro, Enrico Coiera, Thomas W Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. Expert evaluation of large language models for clinical dialogue summarization. *Scientific Reports*, 15(1):1195, 2025.

[24] Xiaoyi Raymond Gao, Marion Chiariglione, Ke Qin, Karen Nuytemans, Douglas W Scharre, Yi-Ju Li, and Eden R Martin. Explainable machine learning aggregates polygenic risk scores and electronic health records for alzheimer's disease prediction. *Scientific reports*, 13(1):450, 2023.

[25] Roger Garriga, Teodora Sandra Buda, João Guerreiro, Jesús Omaña Iglesias, Iñaki Estella Aguerri, and Aleksandar Matić. Combining clinical notes with

structured electronic health records enhances the prediction of mental health crises. *Cell Reports Medicine*, 4(11), 2023.

[26] Roger Garriga, Javier Mas, Semhar Abraha, Jon Nolan, Oliver Harrison, George Tadros, and Aleksandar Matic. Machine learning model to predict mental health crises from electronic health records. *Nature medicine*, 28(6):1240–1248, 2022.

[27] Stephania Ruth Basilio Silva Gomes, Malcolm von Schantz, and Mario Leocadio-Miguel. Predicting depressive symptoms in middle-aged and elderly adults using sleep data and clinical health markers: A machine learning approach. *Sleep Medicine*, 102:123–131, 2023.

[28] Allison Grothman, William J Ma, Kendra G Tickner, Elliot A Martin, Danielle A Southern, and Hude Quan. Case identification of depression in inpatient electronic medical records: Scoping review. *JMIR Medical Informatics*, 12:e49781, 2024.

[29] Martin Guha. Diagnostic and statistical manual of mental disorders: Dsm-5. *Reference Reviews*, 28(3):36–37, 2014.

[30] Martin C Härter, Kevin P Conway, and Kathleen R Merikangas. Associations between anxiety disorders and physical illness. *European archives of psychiatry and clinical neuroscience*, 253(6):313–320, 2003.

[31] Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Camacho. Exploring dimensionality reduction techniques in multilingual transformers. *Cognitive Computation*, 15(2):590–612, 2023.

[32] Waleed Javaid, Shaelyn Cavanaugh, Christina Lupone, Telisa Stewart, Tasaduq Fazili, and Benjamin White. Fixed vs. free-text documentation of indication for antibiotic orders. In *Open Forum Infectious Diseases*, volume 4, page S325, 2017.

[33] Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362, 2023.

[34] Sameer S Kadri, Yi Ling Lai, Sarah Warner, Jeffrey R Strich, Ahmed Babiker, Emily E Ricotta, Cumhur Y Demirkale, John P Dekker, Tara N Palmore, Chanu Rhee, et al. Inappropriate empirical antibiotic therapy for bloodstream infections based on discordant in-vitro susceptibilities: a retrospective cohort analysis of prevalence, predictors, and mortality risk in us hospitals. *The Lancet Infectious Diseases*, 21(2):241–251, 2021.

[35] Nandakishore Kambhatla and Todd K Leen. Dimension reduction by local principal component analysis. *Neural computation*, 9(7):1493–1516, 1997.

[36] Hee-Ju Kang, Seon-Young Kim, Kyung-Yeol Bae, Sung-Wan Kim, Il-Seon Shin, Jin-Sang Yoon, and Jae-Min Kim. Comorbidity of depression with physical disorders: research and clinical implications. *Chonnam medical journal*, 51(1):8–18, 2015.

[37] Saif Khairat, Jennifer Morelli, Marcella H Boynton, Thomas Bice, Jeffrey A Gold, Shannon S Carson, et al. Investigation of information overload in electronic health records: Protocol for usability study. *JMIR Research Protocols*, 14(1):e66127, 2025.

[38] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. Mindfuldiary: Harnessing large language model to support psychiatric patients' journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2024.

[39] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. Supporting the demand on mental health services with ai-based conversational large language models (llms). *BioMedInformatics*, 4(1):8–33, 2023.

[40] Ivan Liu, Fangyuan Liu, Yuting Xiao, Yajia Huang, Shuming Wu, and Shiguang Ni. Investigating the key success factors of chatbot-based positive psychology intervention with retrieval-and generative pre-trained transformer (gpt)-based chatbots. *International Journal of Human–Computer Interaction*, 41(1):341–352, 2025.

[41] Yang Liu, Xingchen Ding, Shun Peng, and Chengzhi Zhang. Leveraging chatgpt to optimize depression intervention through explainable deep learning. *Frontiers in psychiatry*, 15:1383648, 2024.

[42] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. Developing conversational virtual humans for social emotion elicitation based on large language models. *Expert Systems with Applications*, 246:123261, 2024.

[43] Chenyue Lu, Di Jin, Nathan Palmer, Kathe Fox, Isaac S Kohane, Jordan W Smoller, and Kun-Hsing Yu. Large-scale real-world data analysis identifies co-morbidity patterns in schizophrenia. *Translational psychiatry*, 12(1):154, 2022.

[44] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[45] Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2022, page 719, 2023.

[46] Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. Loneliness and suicide mitigation for students using gpt3-enabled chatbots. *npj mental health research*, 3(1):4, 2024.

[47] Alba María Mármol-Romero, Manuel García-Vega, Miguel Ángel García-Cumbreras, and Arturo Montejo-Ráez. An empathic gpt-based chatbot to talk about mental disorders with spanish teenagers. *International Journal of Human–Computer Interaction*, 41(7):3957–3973, 2025.

[48] Dorothy McCoy, Rani Sebti, and Arpi G Kuyumjian. An evaluation of selected indications and appropriateness of ampicillin/sulbactam, an unrestricted antimicrobial, at a single center. *Pharmacy and Therapeutics*, 42(3):189, 2017.

[49] Briana Mezuk, Vicki Johnson-Lawrence, Hedwig Lee, Jane A Rafferty, Cleopatra M Abdou, Ekeoma E Uzogara, and James S Jackson. Is ignorance bliss? depression, antidepressants, and the diagnosis of prediabetes and type 2 diabetes. *Health Psychology*, 32(3):254, 2013.

[50] Fateme Nateghi Haredasht, Sajjad Fouladvand, Steven Tate, Min Min Chan, Joannas Jie Lin Yeow, Kira Griffiths, Ivan Lopez, Jeremiah W Bertz, Adam S Miner, Tina Hernandez-Boussard, et al. Predictability of buprenorphine-naloxone treatment retention: A multi-site analysis combining electronic health records and machine learning. *Addiction*, 119(10):1792–1802, 2024.

[51] David Nickson, Caroline Meyer, Lukasz Walasek, and Carla Toro. Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. *BMC medical informatics and decision making*, 23(1):271, 2023.

[52] Government of British Columbia. Diagnostic code descriptions (icd-9), 2017. Accessed: 2025-06-27.

[53] Government of Canada. Canadian chronic disease surveillance system-an overview, 2021. Accessed: 2025-06-27.

[54] OpenAI. Best practices for prompt engineering with the openai api, 2023. Accessed: 2025-08-12.

[55] Roy H Perlis, Joseph F Goldberg, Michael J Ostacher, and Christopher D Schneck. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology*, 49(9):1412–1416, 2024.

[56] Robert H Pietrzak, Risë B Goldstein, Steven M Southwick, and Bridget F Grant. Physical health conditions associated with posttraumatic stress disorder in us older adults: results from wave 2 of the national epidemiologic survey on alcohol and related conditions. *Journal of the American Geriatrics Society*, 60(2):296–303, 2012.

[57] Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

[58] Public Health Agency of Canada. The canadian chronic disease surveillance system – an overview, May 2021. Accessed July 14, 2025.

[59] Shae E Quirk, Renée El-Gabalawy, Sharon L Brennan, James M Bolton, Jitender Sareen, Michael Berk, Andrew M Chanen, Julie A Pasco, and Lana J Williams. Personality disorders and physical comorbidities in adults from the united states: data from the national epidemiologic survey on alcohol and related conditions. *Social psychiatry and psychiatric epidemiology*, 50:807–820, 2015.

[60] Lars Lau Raket, Jörn Jaskolowski, Bruce J Kinon, Jens Christian Brasen, Linus Jönsson, Allan Wehnert, and Paolo Fusar-Poli. Dynamic electronic health record

detection (detect) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *The Lancet Digital Health*, 2(5):e229–e239, 2020.

[61] Daniel Reichert, David Kaufman, Benjamin Bloxham, Herbert Chase, and Noémie Elhadad. Cognitive analysis of the summarization of longitudinal patient records. In *AMIA Annual Symposium Proceedings*, volume 2010, page 667, 2010.

[62] Tessa Roberts, Georgina Miguel Esponda, Dzmitry Krupchanka, Rahul Shidhaye, Vikram Patel, and Sujit Rathod. Factors associated with health service utilisation for common mental disorders: a systematic review. *BMC psychiatry*, 18:1–19, 2018.

[63] Elvis Saravia. Prompt engineering guide. https://github.com/dair-ai/Prompt-Engineering-Guide, 2022. Accessed: 2025-08-12.

[64] Arne Schwieger, Katrin Angst, Mateo De Bardeci, Achim Burrer, Flurin Cathomas, Stefano Ferrea, Franziska Grätz, Marius Knorr, Golo Kronenberg, Tobias Spiller, et al. Large language models can support generation of standardized discharge summaries–a retrospective study utilizing chatgpt-4 and electronic health records. *International Journal of Medical Informatics*, 192:105654, 2024.

[65] Yijun Shao, Qing T Zeng, Kathryn K Chen, Andrew Shutes-David, Stephen M Thielke, and Debby W Tsuang. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC medical informatics and decision making*, 19:1–11, 2019.

[66] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, 2023.

[67] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2024.

[68] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[69] Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760, 2016.

[70] Varun Sivamani. Matryoshka embeddings: Compact dimensions, maximum impact. https://medium.com/@varunsivamani/matryoshka-embeddings-a2ce94b3d56a, June 2025. Medium article.

[71] Daniel J Smith, Julie Langan, Gary McLean, Bruce Guthrie, and Stewart W Mercer. Schizophrenia is associated with excess multiple physical-health comorbidities but low levels of recorded cardiovascular disease in primary care: cross-sectional study. *BMJ open*, 3(4):e002808, 2013.

[72] Xiuli Song, Qiang Zheng, Rui Zhang, Miye Wang, Wei Deng, Qiang Wang, Wanjun Guo, Tao Li, and Xiaohong Ma. Potential biomarkers for predicting depression in diabetes mellitus. *Frontiers in Psychiatry*, 12:731220, 2021.

[73] Statistics Canada. Mental disorders in Canada, 2022, 2023. Accessed on 13 June 2025.

[74] Brendon Stubbs, Ai Koyanagi, Nicola Veronese, Davy Vancampfort, Marco Solmi, Fiona Gaughran, André F Carvalho, John Lally, Alex J Mitchell, James Mugisha, et al. Physical multimorbidity and psychosis: comprehensive cross sectional analysis including 242,952 people across 48 low-and middle-income countries. *BMC medicine*, 14:1–12, 2016.

[75] Laura Swinckels, Frank C Bennis, Kirsten A Ziesemer, Janneke FM Scheerman, Harmen Bijwaard, Ander de Keijzer, and Josef Jan Bruers. The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: scoping review. *Journal of Medical Internet Research*, 26:e48320, 2024.

[76] Tania Tajirian, Brian Lo, Gillian Strudwick, Adam Tasca, Emily Kendell, Brittany Poynter, Sanjeev Kumar, Po-Yen Chang, Candice Kung, Debbie Schachter, et al. Assessing the impact on electronic health record burden after five years of physician engagement in a canadian mental health organization: Mixed-methods study. *JMIR Human Factors*, 12:e65656, 2025.

[77] AA Tierney et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. nejm catalyst innovations in care delivery 5 (3). https. *doi. org/10.1056/CAT*, 23, 2024.

[78] United Nations Office on Drugs and Crime. Comorbidities in drug use disorders: No wrong door, March 2022. Accessed July 14, 2025.

[79] Duy Van Le, James Montgomery, Kenneth C Kirkby, and Joel Scanlan. Risk prediction using natural language processing of electronic mental health records

in an inpatient forensic psychiatry setting. *Journal of biomedical informatics*, 86:49–58, 2018.

[80] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: adapting large language models can outperform human experts; 2023. *arXiv preprint arXiv:2309.07430*, 2023.

[81] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

[82] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.

[83] Hongwei Wang, Hongming Zhang, and Dong Yu. On the dimensionality of sentence embeddings. *arXiv preprint arXiv:2310.15285*, 2023.

[84] Wikipedia contributors. Cohen's kappa. https://en.wikipedia.org/wiki/Cohen%27s_kappa, 2025. Accessed: 2025-08-15.

[85] World Health Organization. Depression: let's talk says WHO, as depression tops list of causes of ill health, 2017. Accessed: 2025-06-13.

[86] Li-Tzy Wu, He Zhu, and Udi E Ghitza. Multicomorbidity of chronic diseases and substance use disorders and their association with hospitalization: Results from electronic health records data. *Drug and alcohol dependence*, 192:316–323, 2018.

[87] Thomas R Yackel and Peter J Embi. Unintended errors with ehr-based result management: a case series. *Journal of the American Medical Informatics Association*, 17(1):104–107, 2010.

[88] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

[89] Chang Ho Yoon, Sean Bartlett, Nicole Stoesser, Koen B Pouwels, Nicola Jones, Derrick W Crook, Tim EA Peto, A Sarah Walker, and David W Eyre. Mortality risks associated with empirical antibiotic activity in escherichia coli bacteraemia: an analysis of electronic health records. *Journal of Antimicrobial Chemotherapy*, 77(9):2536–2545, 2022.

[90] Kevin Yuan, Chang Ho Yoon, Qingze Gu, Henry Munby, A Sarah Walker, Tingting Zhu, and David W Eyre. Transformers and large language models are efficient feature extractors for electronic health record studies. *Communications Medicine*, 5(1):83, 2025.