Human-Algorithmic Interaction Using a Large Language Model-Augmented Artificial Intelligence Clinical Decision Support System

Niroop Channa Rajashekar*
niroop.rajashekar@yale.edu
Yale School of Medicine
New Haven, Connecticut, United States

Yeo Eun Shin*
joanne.shin@yale.edu
Yale School of Medicine
New Haven, Connecticut, United States

Yuan Pu*
yuan.pu@yale.edu
Yale School of Medicine
New Haven, Connecticut, United States

Sunny Chung
Yale School of Medicine
New Haven, Connecticut, United States
sunny.chung@yale.edu

Kisung You
Mathematics, CUNY Baruch College
New York, New York, United States
kisung.you@baruch.cuny.edu

Mauro Giuffre
Yale School of Medicine
New Haven, Connecticut, United States
mauro.giuffre@yale.edu

Colleen E Chan
Statistics and Data Science, Yale University
New Haven, Connecticut, United States
colleen.chan@yale.edu

Theo Saarinen
Department of Statistics, University of California, Berkeley
Berkeley, California, United States
theo_s@berkeley.edu

Allen Hsiao
Pediatrics, Yale School of Medicine
New Haven, Connecticut, United States
allen.hsiao@yale.edu

Jasjeet Sekhon
Statistics and Data Science, Yale University
New Haven, Connecticut, United States
jasjeet.sekhon@yale.edu

Ambrose H Wong
Yale University
New Haven, Connecticut, United States
ambrose.wong@yale.edu

Leigh V Evans
Emergency Medicine, Yale School of Medicine
New Haven, Connecticut, United States
leigh.evans@yale.edu

Rene F. Kizilcec
Department of Information Science, Cornell University
Ithaca, New York, United States
kizilcec@cornell.edu

Loren Laine
Yale School of Medicine
New Haven, Connecticut, United States
loren.laine@yale.edu

Terika McCall
Department of Biostatistics, Yale School of Public Health
Center for Medical Informatics, Yale School of Medicine
New Haven, Connecticut, United States
terika.mccall@yale.edu

Dennis Shung
Yale School of Medicine
New Haven, Connecticut, United States
dennis.shung@yale.edu

ABSTRACT

Integration of artificial intelligence (AI) into clinical decision support systems (CDSS) poses a socio-technological challenge that is impacted by usability, trust, and human-computer interaction

*These Authors contributed equally to this research.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642024

(HCI). AI-CDSS interventions have shown limited benefit in clinical outcomes, which may be due to insufficient understanding of how health-care providers interact with AI systems. Large language models (LLMs) have the potential to enhance AI-CDSS, but haven't been studied in either simulated or real-world clinical scenarios. We present findings from a randomized controlled trial deploying AI-CDSS for the management of upper gastrointestinal bleeding (UGIB) with and without an LLM interface within realistic clinical simulations for physician and medical student participants. We find evidence that LLM augmentation improves ease-of-use, that LLM-generated responses with citations improve trust, and HCI varies based on clinical expertise. Qualitative themes from interviews suggest the perception of LLM-augmented AI-CDSS as a team-member used to confirm initial clinical intuitions and help evaluate borderline decisions.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**; **Empirical studies in HCI**.

KEYWORDS

Health-Clinical, Machine Learning, Medical: Nursing Homes/Hospitals, Qualitative Methods, Quantitative Methods, Artificial Intelligence, Clinical Decision Support Systems, Workflows, Electronic Health Record

ACM Reference Format:
Niroop Channa Rajashekar, Yeo Eun Shin, Yuan Pu, Sunny Chung, Kisung You, Mauro Giuffre, Colleen E Chan, Theo Saarinen, Allen Hsiao, Jasjeet Sekhon, Ambrose H Wong, Leigh V Evans, Rene F. Kizilcec, Loren Laine, Terika McCall, and Dennis Shung. 2024. Human-Algorithmic Interaction Using a Large Language Model-Augmented Artificial Intelligence Clinical Decision Support System. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642024

1 INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) algorithms have the potential to provide value to clinicians in their already-complex clinical workflows. AI interventions in medicine in randomized controlled trials have shown limited improvement in clinical outcomes, with no clear evaluation of the human-AI interaction [76]. In order to optimize the beneficial effect of these AI interventions and mitigate the potential harms, there is a need to study and involve clinicians as end-users in ML technology to create and iterate systems that improve clinical workflows [117]. ML-based clinical decision support systems can achieve practical clinical relevance when they work seamlessly with existing workflows [116] that utilize the electronic health record (EHR) [85].

Previous efforts at qualitatively characterizing healthcare provider interaction with AI systems has spanned multiple domains, including alert systems [88, 92], risk estimators [10, 12], and image-retrieval [16]. These studies usually explore user interactions as single-user systems between providers and ML models or integrating interactions between patients and ML models [40]. While these single-user interactions may be salient in specific care settings, such as patients in an outpatient clinic, the complexity of modern medicine has led to a transition from receiving care from a single provider to provider teams. This includes cooperation among physicians and medical providers across specialties, training levels, and responsibilities [31]. There is a paucity of studies that evaluate the behavior of provider teams and their user experience with AI clinical decision support systems (AI-CDSS). One challenge to shifting this paradigm is the accessibility of these AI systems to multiple users.

Recently, large language models (LLMs) have emerged as systems with potential to aid clinical decision-making. Recent exploratory studies have assessed LLMs' ability to answer clinical questions [97, 120]. LLMs are accessible to users across a spectrum of expertise, which provides an opportunity to design AI-CDSS that can be used in a team-based setting. LLMs' ability to generate text answers to clinical questions positions it as a potentially useful tool that providers can interact with similarly to how they interact with human team members and experts [95]. Previous work has shown that adoption of ML tools in medicine is more likely when clinicians view the tool as a "partner" to enhance their expertise, this is similar to the role of team members in the clinical team [36]. To our knowledge, there are no studies that evaluate usability of large language model (LLM)-based systems for active clinical workflows in a team-based clinical decision setting. Our study seeks to understand the user patterns that emerge when physicians utilize AI-CDSS and LLMs to make clinical decisions in a live simulated clinical workflow.

We developed a risk-prediction machine learning model trained on data from patients with upper gastrointestinal bleeding (UGIB). We introduce an interactive dashboard for visualization of risk and GutGPT, an LLM trained on gastroenterology guidelines for UGIB [52]. Limiting GutGPT's context to the risk-prediction ML model or UGIB guidelines places bounds on the LLM in an effort to limit hallucinations and response variability. To test the implementation of the dashboard and GutGPT, we designed a randomized controlled trial to determine how physician and medical student teams utilize and interact with these systems in a series of simulated patient encounters. Using UGIB as a disease process is instructive as it is an acute high-stakes, time-constrained clinical problem with a clear value proposition for risk assessment (in our case, assisted with a high-performing AI-CDSS) that necessitates strong teamwork within provider teams for optimal patient care and frequent inter-specialty collaboration. Evidence-based management of UGIB requires considering guidelines authored by professional societies. The guidance from these guidelines can be difficult to parse quickly while applying to unique patient scenarios [6]. LLM integration into GutGPT is designed to incorporate patient data to give answers that apply the guidelines to complicated situations. To our knowledge, this is the first study to deploy a LLM-based CDSS in a clinical simulation to assess usability, trust, and interaction patterns.

Themes were generated from post-simulation interviews, survey data, and query data from GutGPT inputs. We found that LLM output format and integration into the electronic health record (EHR) influences the perception of usability and may affect the adoption of AI-CDSS technology into physicians' workflows. Trust in AI systems for physicians was limited by preconceived notions of AI-CDSS being unreliable or untrustworthy, and improved after increased use of the system with exposure to answers with high-reliability features (e.g. detailed citations). We contribute to the human-computer interaction community (HCI) by delineating barriers to wide-spread adoption of AI-CDSS in physician workflows, providing understanding of factors that influence physician trust in AI, and presenting three design principles for LLM-augmented AI-CDSS.

2 RELATED WORKS

2.1 Upper Gastrointestinal Bleeding

UGIB is one of the most common causes for hospitalization for gastrointestinal disease in the United States, accounting for over 400,000 emergency department (ED) visits a year in the United States [74]. It is a common cause of hospital readmission, morbidity, and mortality [74]. Common etiologies for UGIB include peptic ulcer disease, esophageal varices, and esophagitis, and the diagnosis of UGIB may include obvious symptoms such as bright red blood in gastric contents or stool to insidious symptoms such as fatigue and dark stools [18, 105].

Multiple elements from patient reported clinical history, physical examination, and laboratory measurements that may suggest UGIB. Diagnosis and management of UGIB usually requires interplay between different medical providers and staff in the healthcare environment. For example, patients with UGIB often first present to the ED. They are first seen and assessed by emergency medicine physicians and clinical staff who provide an initial diagnosis andmanagement. If the condition is deemed to be severe enough to require specialist evaluation, the emergency medicine physician initiates communication with specialist gastroenterologists and internal medicine physicians to consider admission to the hospital with urgent endoscopic evaluation [98]. Risk stratification to identify patients who are "very low risk" and can be discharged from the ED is the first key management decision for the provider caring for a patient with UGIB; identifying very-low-risk patients is recommended by national guidelines for the management of patients with acute UGIB [52]. However, it is possible that these patients may require urgent care: high-risk patients with UGIB can clinically deteriorate quickly if they have uncontrolled bleeding and may require hospital-based interventions such as transfusion of red blood cells or interventions to stop bleeding [48]. No existing studies evaluate the implementation of AI-CDSS for UGIB risk assessment. Our paper provides qualitative themes of provider behaviors when interacting with an AI-CDSS for UGIB risk assessment in a simulated environment.

2.2 Clinical Decision Support Systems with and without Artificial Intelligence

Clinical Decision Support Systems (CDSS) have existed in healthcare for decades as an attempt to reduce errors made by medical staff [44, 63]. CDSS are designed to improve healthcare delivery by providing relevant, timely, and useful clinical knowledge to providers that help them to make decisions regarding diagnosis, prognosis, and treatment [72]. The most basic CDSS usually function by matching the characteristics of an individual patient to a computerized clinical knowledge base. Patient-specific assessments or recommendations are made, and subsequently presented to the clinician for a decision [94]. The clinician's role is to combine these evaluations with their own prior knowledge to make the final decision.

Despite only providing simple diagnostic support, early forms of rules-based CDSS (i.e., a treatment that is suggested when a certain part of patient history is flagged) still showed effectiveness in clinical decision support by identifying high-risk patient groups and reducing cases of misdiagnosis [50, 66]. CDSS are able to support many aspects of the healthcare process including disease prevention, screening, diagnosis, treatment, and follow-up [29] while reducing medical costs by minimizing side effects from drug treatments [60, 73]. In the modern era of electronic health records (EHR), CDSS are often integrated into the EHR [68]. However, many clinicians have expressed concerns regarding their trust in CDSS when introduced into their workflows[75].

In the era of increasing data volume and computational capacity, modern CDSS integrate the use of ML and AI in AI-CDSS [61]. AI interventions in healthcare have been studied in randomized controlled trials with a steadily increasing number of Food and Drug Administration (FDA) - approved medical ML applications [76]. AI-CDSS have evolved to provide predictive clinical insights using medical data available across multiple domains, with over 500 clinical prediction models built on EHR data published and 44 published reports of implementation studies [56, 115]. A slight majority of EHR-based AI-CDSS implemented in published studies have demonstrated some improvement in clinical outcomes after implementation [56]. For example, CDSS have showcased the capacity to predict the probability of diabetic complications among individuals with diabetes and guide clinicians with the optimal timing for diagnostic tests [37, 90]. The issue of trust remains challenging for clinician adoption of AI-CDSS. This includes a lack of justification of model predictions [77], which has been partially addressed with the emergence of "Explainable" AI [7]. Explanations from AI models can be categorized as global or local; in offering an explanation of the entire model or single predictions, respectively [4]. Ante-hoc, or inherently explainable methods, are understandable on their own while post-hoc understandability methods communicate information about an output after the model produces the output [7]. Recent advances in explainable AI-CDSS to improve clinician trust have spanned domains of text, graphical, and image explanations, among others. For example, a convolutional neural network to aid glaucoma diagnosis used class activation mapping to generate heat maps for image analysis [24]. An AI-CDSS for identifying women at risk for gestational diabetes mellitus used Shapley additive explanations to graphically represent model features [25].Stakeholder analysis suggests that clinicians prefer AI-CDSS with feature importance and transparency regarding how confident or uncertain the model was in its predictions, and clinicians also indicated that the ML tools had to be tested in real clinical situations so users could grasp their strengths and weaknesses and foster sustainable trust [101]. Our paper provides qualitative and quantitative descriptions of usability for an AI-CDSS with post-hoc explainability methods that contributes towards the understanding of clinician trust when utilizing AI-CDSS.

2.3 Human-Computer Interaction in Artificial Intelligence-Clinical Decision Support Systems

Currently CDSS are implemented in conjunction with clinicians' medical knowledge, intuition, and willingness to incorporate such systems into their decision-making process. Thus, HCI inherently plays a crucial role in the design of CDSS [85]. This is particularly relevant in healthcare, where providers have suffered from the unintended consequences associated with high alert burden in EHR CDSS that are caused due to system design processes that are not physician-centered, such as sepsis alerts [70, 104, 113, 114]. Poorly designed CDSS can lead to nonadherence, high override rates, and "alert fatigue" in which clinicians neglect the alert, thereby reducing their effectiveness and potential benefits [65].

To prevent such adverse effects and maximize CDSS usability, several methodological approaches for usability engineering and cognitive task analysis have been developed [51]. Most notably, heuristic evaluation of medical device interaction and patient safety [118], cognitive factor analysis for GUI evaluation in tele-mental health psychotherapy services [3], the Task, User, Representation and Function (TURF) framework for EHR usability [119], an ethnographic study to create guidelines on designing electronic communicable disease reporting systems [89], and natural language querying to resolve time-event dependencies in clinical information systems [86] are frameworks for exploring different HCI methods to evaluate and develop CDSS.

HCI becomes particularly crucial when it comes to AI-CDSS, as the complexity and lack of usability of sophisticated computational systems like AI may discourage clinician use [93]. Indeed, the difficulty in explaining modern AI-based systems that have a "black-box" nature may also hinder integration into clinicians' workflow [32, 107]. To address these issues, the first step of HCI should be to provide training to users about the inner workings of AI-CDSS and its strengths and weaknesses [17]. The goal of HCI frameworks for developing AI-CDSS is to seamlessly integrate them into pre-existing healthcare information and clinician workflows [67]. Since AI cannot completely emulate physicians' mental models and physicians are unable to access large amounts of data to make conclusions, AI-CDSS should be designed as interactive systems that physicians can use to support their cognitive processes, as part of a human-AI collaboration paradigm [85, 111]. While explainability plays an important role in trust in AI tools, there are other factors that are vital in clinician adoption. Even in studies where the ML tool underlying the CDSS was opaque, trust was increased when adoption was endorsed by colleagues or superiors [35]. Trust is enhanced by reference to resources that are familiar to clinicians; previous research in AI-CDSS has found that clinicians preferred evidence-based explanation of outputs over model features [41]. Physicians' cognitive model for risk stratification and management incorporates information from reputable clinical guidelines. AI-CDSS that delivers guideline-driven advice mimics the role that human teammates play in the medical team [59]. In this paper, we provide a unique perspective by studying AI-CDSS in a team setting, where different team dynamics may affect perceptions towards interactions with AI-CDSS.

2.4 Large Language Models in Artificial Intelligence-Clinical Decision Support Systems

LLMs represent a subset of AI models that excel in diverse natural language understanding and generation tasks since they are autoregressive, with the ability to predict the next word in a given context [79]. These models owe their proficiency to the massive scale of the transformer-based neural networks (with billions of parameters) and extensive training on a vast corpora of text [14]. In the realm of healthcare, LLMs' exceptional natural language processing capabilities render them a powerful tool to be integrated into EHRs, which are vast repositories of patient data that include substantial amounts of unstructured note text. The potential for LLMs has already attracted significant research and commercial attention, with partnerships established between electronic health record vendors and AI companies with cutting-edge LLMs. Examples include the collaboration between Microsoft and Epic on integrating GPT-4-powered services into EHR, as well as the incorporation of Google-designed Med-PaLM 2 healthcare AI chatbot into Meditech [15, 19].

ChatGPT is a famous application of LLMs [71]. Its underlying LLM, the Generative Pre-trained Transformer (GPT), is trained on diverse online text sources to produce human-like responses in versatile conversational interactions [79]. Since its release in November 2022, active investigations into ChatGPT's potential in healthcare have spanned research, practice, and education [57, 87]. ChatGPT's ability to process health-related information from the EHR and ability to interact with users in natural language offers unique opportunities for a wide scope of potential applications in clinical decision support [27]. By comparing GPT-3.5-powered ChatGPT's responses to human medical experts' answers to clinical questions in multiple subspecialties, several studies in general medicine, radiology, and pediatrics have suggested the adequacy of LLMs for providing decision support throughout the pathway of clinical care, from diagnosis to treatment recommendations [42, 80, 81].

However, these studies also reveal limitations in ChatGPT including the opacity of its training data, the phenomenon of hallucinations, and limited model explainability [42, 58, 80, 81]. Recent studies suggest that new LLMs reproduce and amplify human biases [49]. Misuse of ML tools in the healthcare environment can also promote over-reliance on these tools, leading to errors when clinicians delegate verification and safety checks [61]. While certain strategies and frameworks for ChatGPT-based CDSS have been suggested to address these limitations [27], an AI-CDSS that queries reliable clinical guidelines with guardrails could ameliorate many of these complaints while reducing response variability. There is an urgent need for a deeper understanding of user behavior when integrating an LLM in clinical workflows to further develop design principles and usage guidelines for real-world adoption. Our study seeks to elucidate specific patterns of user behavior with LLMs within simulation scenarios.

2.5 Medical Simulation

Medical simulation can be defined as a technique to replace or amplify real experiences with immersive guided interactive experiences to replicate aspects of the real world [28]. Simulations can have enough fidelity with real clinical environments that they can be used to study human factors and behaviors that contribute to the effectiveness of a provider team [38]. In medicine, simulation has a key role in maintaining and promoting patient safety and quality improvement for high-stakes scenarios where provider error could have adverse effects on patient outcomes.

In 1999, the Institute of Medicine released a report on medical errors that revolutionized the approach towards patient safety [47]. The report highlighted simulation as a key driver of healthcare improvement [109]. Simulation allows for improvement both for individual practitioners and for provider teams. On the individual level, simulation centers can help individual medical trainees to practice skills and techniques in safe environment to prepare for situations in which real patients might be at risk, such as learning central line insertion techniques to increase successful insertion rates [9] and lower central line infection rates [8]. On the provider team level, simulation studies have been successful at studying the human factors involved in domains of teamwork [82] and team communication [11]. As medicine has increased in complexity providers increasingly work in teams to provide clinical care for disease management.

Beyond training individuals and provider teams on existing best practices and protocols, simulation centers can also add value in the testing of new medical devices and advanced technology, such as AI/ML. [62, 96] Usability testing for EHR technology, anesthesia machines, and numerous other medical devices is frequently performed in simulation centers [53, 64]. AI and ML products in medicine are considered software as medical devices (SaMD) [103] and require rigorous real-world clinical deployment and evaluation [102]. Simulation environments are underutilized in the development pipeline of these SaMDs to be tested within simulation center environments.

LLMs have rapidly evolving capabilities relevant to clinical applications, and solutions integrating LLMs into AI-CDSS are potential SaMDs that may be integrated into the clinical workflow [43]. Existing partnerships between EHR vendors and LLM companies provide a trajectory for LLMs to be used by providers in routine clinical care [15, 34]. However, clinician skepticism remains a formidable challenge [78]. Among several concerns regarding safety is the potential for hallucinations that result in fabricated citations [21] that may lead to errors when integrated into high-stakes clinical environments. No study to our knowledge has used medical simulation to test LLM-augmented AI-CDSS, which we believe may be useful for developers of AI healthcare systems to facilitate clinical evaluation and safety testing by understanding user needs and behavior.

Our paper demonstrates the feasibility of using a simulation setting to test and to evaluate usability, trust, and human-AICDSS interaction for an EHR-integrated LLM-augmented AI-CDSS.

3 METHODS

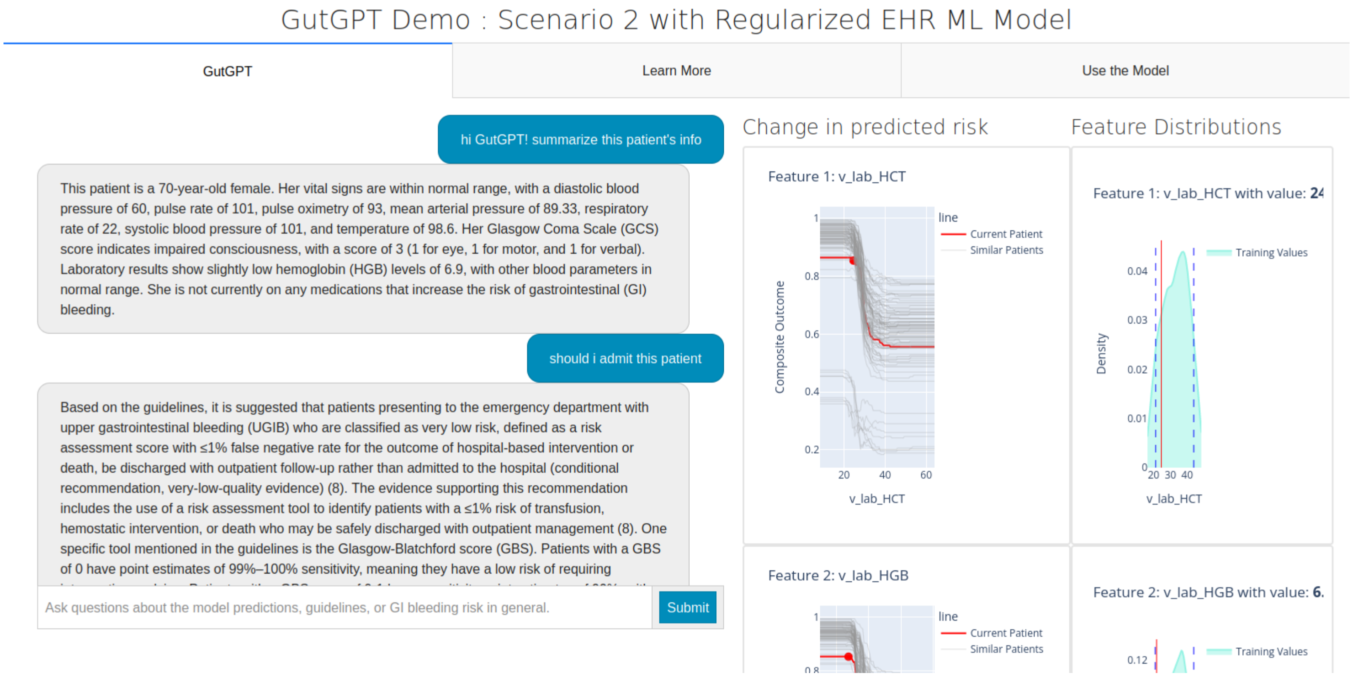
3.1 GutGPT

GutGPT is an in-house CDSS designed and developed to provide a natural language-based interface for two tasks: guideline-based question answering and an interactive dashboard for risk prediction. It is built on top of a high-fidelity ML model validated using an existing clinical dataset. With patient data automatically loaded at launch, GutGPT provides patient-specific predictions of the risk for hospital-based intervention and grounds its reasoning on this information to generate responses to clinicians' questions. Formulation, development, and implementation of dashboard and chatbot tools were performed by a multidisciplinary team. Practicing clinicians in this team directly contributed to the creation of GutGPT and oversaw building the tools from their genesis to experimental trial.

GutGPT's risk-prediction machine learning model was developed using electronic health records (EHR) of patients presenting with signs or symptoms of gastrointestinal bleeding at a large health system. The inputs to the model include demographic data (age and sex), nursing assessments, lab test results, personal medical history, and medication classes in the form of Clinical-Classification-Software codes [1]. We consider a composite binary variable as the outcome, where the value of 1 signifying a high-risk patient that required a hospital-based intervention, such as red blood cell transfusion, intervention to stop bleeding, or 30-day all-cause mortality, and 0 otherwise. Multiple machine learning (ML) and deep learning models were explored, including LASSO regression [100], random forests with honesty [110], gradient boosted trees [23], and feedforward neural networks with 2 and 5 layers [84]. Data pre-processing included dimensionality reduction via LASSO regression to the patients' medical history and medication classes tuned using 10-fold cross validation. Random forests with honesty was applied to the variables with non-zero coefficients, in addition to demographics, nursing assessments, and lab test variables. This final model exhibited the highest true negative rate (TNR) at a true positive rate (TPR) of 99% recommended by national UGIB guidelines as the very low risk threshold [52]. The model had an AUC 0.91 (0.88-0.93) on an internal validation set and 0.92 (0.90-0.95) on an external validation set (data from a different hospital). At the 99% sensitivity threshold, our model exhibited a specificity of 0.46 on the internal validation set and 0.33 on the external validation set, which outperforms existing recommended clinical risk scores.

The interactive dashboard displays risk predictions with interpretability plots for the ML model used within GutGPT. Users can visualize partial dependency plots (PDPs), individual conditional expectation (ICE) plots, and accumulated local effects (ALE) plots for any covariate in the model, assisting their understanding of the effect of selected covariates on the model's predicted risk [69]. The incorporation of these interpretability plots was implemented after an iterative process where a multidisciplinary team including clinicians, data scientists, statisticians, and human factors experts to enhance users' understanding of the ML model's decision-making process, ensuring alignment with their clinical mental model.

Furthermore, users have the ability to modify patient covariate values in real time and observe how the predicted risk of hospital-based intervention or 30-day mortality changes accordingly. The dashboard also provides other information to help contextualize the risk with regards to the general population of patients with acute gastrointestinal bleeding. For example, it reports a patient similarity index, quantifying how similar the queried patient is to patients in the training data. To facilitate population-level understanding, histograms depict the distribution of each variable and highlight the target patient's value relative to all patient values in the training data.


GutGPT Demo : Scenario 2 with Regularized EHR ML Model

hi GutGPT! summarize this patient's info

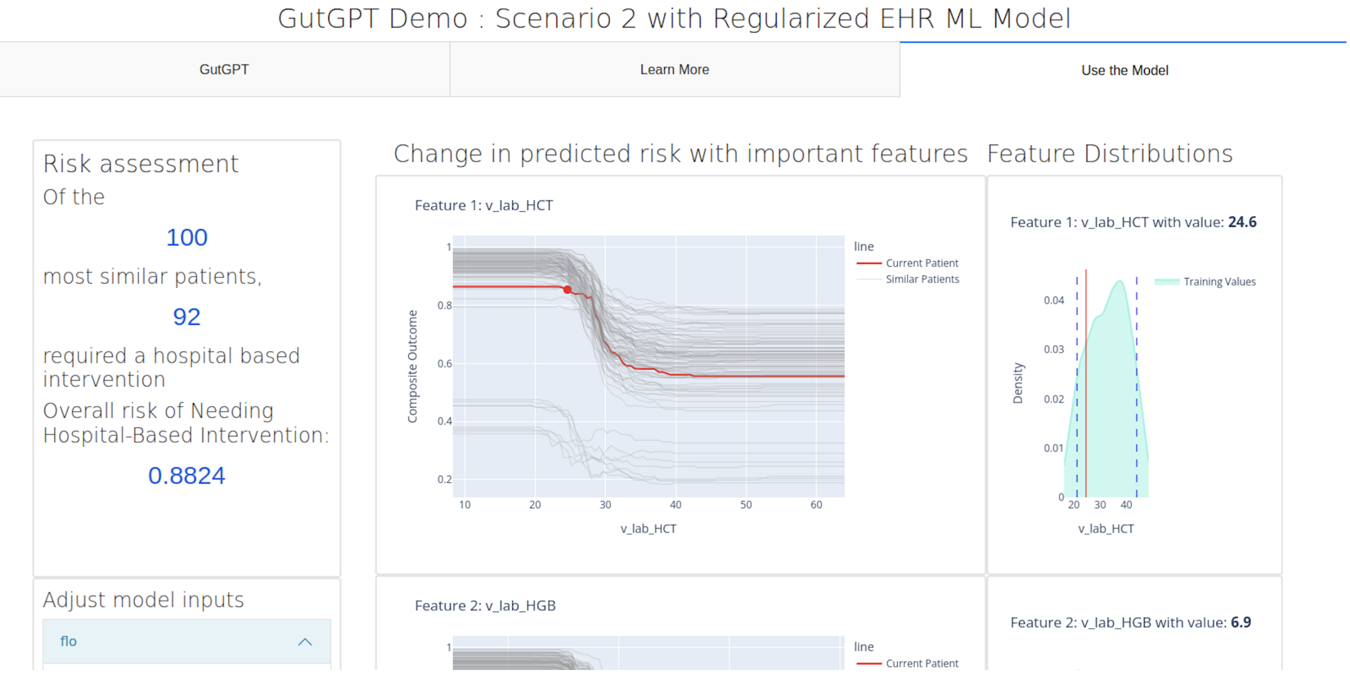
This patient is a 70-year-old female. Her vital signs are within normal range, with a diastolic blood pressure of 60, pulse rate of 101, pulse oximetry of 93, mean arterial pressure of 89.33, respiratory rate of 22, systolic blood pressure of 101, and temperature of 98.6. Her Glasgow Coma Scale (GCS) score indicates impaired consciousness, with a score of 3 (1 for eye, 1 for motor, and 1 for verbal). Laboratory results show slightly low hemoglobin (HGB) levels of 6.9, with other blood parameters in normal range. She is not currently on any medications that increase the risk of gastrointestinal (GI) bleeding.

should i admit this patient

Based on the guidelines, it is suggested that patients presenting to the emergency department with upper gastrointestinal bleeding (UGIB) who are classified as very low risk, defined as a risk assessment score with ≤1% false negative rate for the outcome of hospital-based intervention or death, be discharged with outpatient follow-up rather than admitted to the hospital (conditional recommendation, very-low-quality evidence) (8). The evidence supporting this recommendation includes the use of a risk assessment tool to identify patients with a ≤1% risk of transfusion, hemostatic intervention, or death who may be safely discharged with outpatient management (8). One specific tool mentioned in the guidelines is the Glasgow-Blatchford score (GBS). Patients with a GBS of 0 have point estimates of 99%–100% sensitivity, meaning they have a low risk of requiring

Ask questions about the model predictions, guidelines, or GI bleeding risk in general. Submit

Figure 1: GutGPT chatbot interface. The chat interface on the left adopts a typical conversation-like design. Figures on the right display the patient's vital data and their effect on hospital-based-intervention risks predicted by our model's underlying ML model, in the context of patients in the training database.


GutGPT Demo : Scenario 2 with Regularized EHR ML Model

Risk assessment

Of the 100 most similar patients, 92 required a hospital based intervention

Overall risk of Needing Hospital-Based Intervention: 0.8824

Adjust model inputs

flo

Figure 2: GutGPT dashboard interface. The left column displays on top the hospital-intervention risk for the current patient and has sliders below for the users to calibrate the model by adjusting the patient's vitals, labs, medications, and more (not entirely captured). The same figures in the chatbot interface are displayed on the right.

When a user types a question into the chatbot interface, GutGPT classifies the query as either a question about the predicted risk from the ML model or regarding clinical management from the guideline recommendations. For both types, the structured datafields stored in the EHR are automatically loaded onto GutGPT for individualized prediction. If the query pertains to the risk prediction of GIB, GutGPT retrieves the interactive dashboard, extracts information relevant to the user's query, and provides interpretation of graphically presented information in human language. For example, for a question regarding the predicted risk itself, GutGPT generates a paragraph stating the risk score of a specific patient, with an addendum according to clinical guidelines. It notes that a risk score below the 99% sensitivity threshold should be considered as "very low risk" according to the American College of Gastroenterology (ACG), and "not very low risk" otherwise.

GutGPT also can answer a user's questions regarding clinical management by drawing upon care recommendations for a patient's profile based on the guidelines from the ACG for the management of upper GIB [52]. These guidelines are organized into discrete sections, including pre-endoscopic and endoscopic management, summary of evidence, recommendations, and conclusions. Preprocessing of the sections include separating each section into separate text chunks and converting each chunk into a vector embedding using OpenAI's text embedding model. When a user types a question, the query is converted into a vector embedding and then compared with the vector embeddings of the guideline text sections. This process enables the retrieval of the most relevant sections from the guidelines through a similarity search. The retrieved portions of the guidelines are then integrated into a user's question, along with the patient's EHR data. Instructions on text and reference formatting is also provided in the prompt, which is then supplied to the GPT model to generate a response for the user.

3.2 Participants and Simulation

We recruited 31 participants from various medical education levels. Of those participants, 9 were Emergency Medicine (EM) resident physicians, 6 were internal medicine (IM) resident physicians, and 16 were medical students (MS). They were placed into provider teams of 2-4 participants, for a total of 12 provider teams across the study period. Efforts were made to recruit resident physicians and medical students of all experience levels. IM resident physicians ranged from training levels of post-graduate year 1 (PGY-1) to PGY-3. EM resident physicians ranged from PGY-1 to PGY-4. Medical students ranged from the second year of medical school to fourth year students (including students taking a research year and MD-PhD candidates in the research portion of their degrees).

We sought to include both internal medicine and emergency medicine as these two specialties have frequent contact with patients who have UGIB. Resident physicians of different experience levels were solicited to help identify trends in the experience using the AI system based on training level. Medical students were sought to further diversify the participant pool based on experience - medical students are less likely to be familiar with UGIB management and ACG guidelines than resident physicians. The provider teams performing the simulation activities consisted solely of one training category (EM, IM, or MS). However, within that category, provider teams had varying experience levels. This was done to mimic a typical provider team in clinical environments.

The randomized controlled study is comprised of two arms (see Figure 3). Each provider team was randomized to either the GutGPT arm or dashboard arm separately for the two separate phases of the study (Risk Assessment and Content Assessment). If randomized to the GutGPT arm, a workstation with access to GutGPT, the interactive dashboard, and any internet tool was available to the participants. If randomized to the dashboard arm, the workstation could only access the interactive dashboard and any other internet tool. During the Risk Assessment phase, the participants underwent three risk scenarios in which they decided to admit the simulated patient to the hospital, observe in the ED, or discharge from the ED. During the Content Assessment phase the participants underwent two scenarios which tested their medical management of simulated patients. For all phases, the provider teams were presented with cases of UGIB and the order in which the scenarios were presented was randomized. They interacted with a SimMan full-body mannequin (Laerdal) for the interview and physical exam of the simulated patient. A gastroenterology specialist voiced the patient, and their voice was broadcasted through a speaker in the mannequin. In the simulated clinical environment, the simulated patient's chart was accessible through a workstation that mimicked the electronic medical record - complete with past medical history, laboratory values, and medications. To simulate the clinical team dynamics, the most senior member was assigned to use the dashboard and/or GutGPT interface. The other members occupied the rest of the clinical team involved in data gathering from the mannequin and the EHR. Figure 4 displays the number of sessions where the chatbot feature was accessible by the provider team for each simulation scenario. This study was deemed exempt by a university Institutional Review Board.

3.3 Data Collection

To understand the user interaction pattern of clinicians with AI-CDSS in realistic clinical simulations, we collected three types of data: qualitative interviews, GutGPT chatbot conversations, and quantitative surveys.

3.3.1 Post-simulation Qualitative Interview. We conducted brief one-on-one, semi-structured interviews with each of the participants in separate rooms directly following after participants had finished both the Risk-focused and Care Management scenarios. Each interview lasted between 5-10 minutes. The interviews began with general reflections on participant experiences interacting with the GutGPT during the session. Then, the researchers asked about the participant's willingness to use GutGPT in their clinical decision making process in a real clinical situation, and to elaborate their reasoning. Next, the researcher asked for feedback on the user interface of the Chatbot and the Dashboard features within GutGPT. Each interview session was audio recorded with participants' consent.

3.3.2 Post-trial Quantitative Survey. We administered an adapted version of the System Usability Scale (SUS) [13] to participants immediately after completing the Risk-focused scenarios. We retained four positive items that encompassed similar themes covered in the

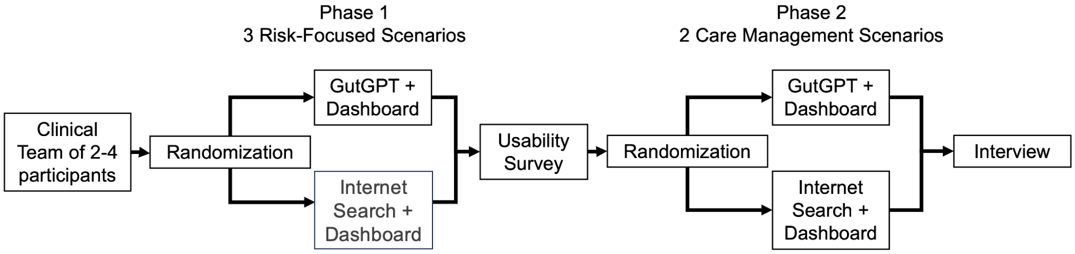


Figure 3: Flowchart depicting the study design. The study is comprised in two phases, with randomization occurring separately at each phase.

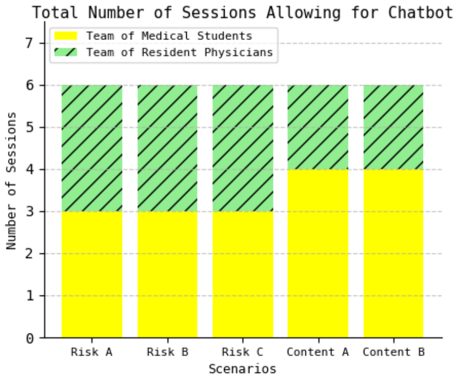


Figure 4: Bar plot displaying number of sessions with the GutGPT chatbot.

original scale (Table 1). The abbreviated SUS survey was appended to an already-long survey administered after the Risk scenarios for a different experiment. We chose the positively-keyed items from the SUS to add to the existing survey, they matched the existing positively-keyed survey items to prevent participant confusion or errors and reduce time taking the survey to avoid an excessively long simulation. We note here that this portion of data collection was added later, resulting in 22 available observations out of the 31 participants.

3.3.3 GutGPT Chatbot Prompting History. All conversations between GutGPT and the participants, including all question inputs ("prompts"; examples see Table 2) from the participants and response outputs from the LLM-augmented chatbot, were automatically recorded in text format with information about their corresponding simulation sessions and scenarios.

3.4 Analysis

3.4.1 Qualitative Analysis of Participant Interviews. Our team of three researchers led the analysis of 31 participant's interview data and regularly discussed emerging themes. We used rapid qualitative analysis methods to effectively extract insights from our data [33]. For the rapid analysis, we created an interview summary template that asked each reviewer to consider initial impressions, system usability, and the role of GutGPT in clinical decision-making. We began by holistically reviewing the transcripts to familiarize ourselves with the data and then delved into paragraph-level understanding. Subsequently, all three researchers compiled significant quotes and observations from the interview summaries onto a shared research board. Through an iterative process, we categorized and organized these notes into common themes and broader feedback categories. This analysis led to the identification of several key insights outlined in this paper: usability in managing various aspects of AI-CDSS and variations in chatbot utilization based on medical specialties and levels of training. Our team's unique interdisciplinary composition, combined expertise in HCI, clinical practice (including specialized knowledge in UGIB), and AI/ML, facilitated a comprehensive understanding of our participants, especially when adhering to a user-centered research framework.

3.4.2 Quantitative Analysis of Post-trial System Usability Scale Responses. In the SUS survey, participants rated the usability of GutGPT for each statement using a 5-point Likert scale, from "strongly disagree" to "strongly agree". We recorded the responses for each sentiment per statement. We quantified participants' average attitude towards each SUS statement by assigning a numerical score to each sentiment category: "strongly disagree" as -2, "disagree" a as -1, "neutral" as 0, "agree" as 1, and "strongly agree" as 2. Following this assignment, the mean score was then zero-centered, hence "neutral"-centered, weighted by the frequency of responses for each sentiment. A positive mean value thus suggested a general agreement with the statement, while a negative mean value indicated disagreement. Additionally, separate calculations of the average attitude were made based on the GutGPT chatbot's accessibility (Figure 5).

It is important to note that we determined a threshold of 85 participants in each arm for the experiment to reach an effect size of the technology acceptance metrics of UTAUT (Unified Theory of Acceptance and Use of Technology [108]) to reach Cohen's $f^2 = 0.1$ with 80% statistical power. While UTAUT data were used in another study in our GutGPT series [20], we adhered to this threshold for the sake of the overall study's coherence. At the time of this manuscript, enrollment for the study has continued. Therefore, the quantitative scores presented herein are primarily indicative of observed trends rather than being conducive to conclusive statistical significance testing.

Table 1: SUS Statements in the Post-trial Survey. Responses were in 5-point Likert scale.

Index	Statement
1	"I thought the system was easy to use."
2	"I would imagine that most people would learn to use this system very quickly."
3	"I found the various functions of this system were well integrated."
4	"I felt confident using the system."

Table 2: Example Question Inputs to the GutGPT Chatbot

Index	Prompt
1	"62 yo M w h/o active etoh cirrhosis, p/w 3 d dark stools, sudden episode of bloody vomit this morning. mentating well. mild tachy, MAP 68."
2	"85-year-old woman with history of hypertension presents with one day of hematemesis in the setting of persistent vomiting for the past two days. Her vitals are 110/70 without other hemodynamic changes, and she takes no medications other than Calcium and amlodipine. What is a relevant differential and should we admit her? "
3	"Hello GutGPT, acting as a consulting gastroenterologist, write a consult note for a 70 y.o. male who presents with chest pain and a week ago with melena, with PMH of heart failure? THe patient is currently taking aspirin, statin and ACE inhibitor. What should the next steps for management be?"
4	"Labs are all normal, I think this pt should be discharge do you agree?"
5	"should i admit this patient"
6	"What is this pt's risk of in-hospital intervention?"
7	"Can you help me calculate the patients GBS score"
8	"what is the next best steps in management for a patient with GBS score of 7"
9	"what is the patient's age"
10	"Do you give both octreotide and vasopressin or one"

3.4.3 *Quantitative Analysis of Chatbot Prompting Pattern.*

We measured the frequency and length of questions asked by the participants when using the GutGPT chatbot, taking into consideration their medical education level and the type of clinical scenarios (risk versus content).

As described in Section 3.2, participants were randomly assigned access to the GutGPT chatbot for both the content and risk scenarios separately. In addition, we conducted simulation sessions with medical student and resident physician teams, with varying numbers of sessions for each group. Hence, to ensure fair comparison, we tallied the total number of sessions allowed for using the chatbot for each medical education level (medical students or resident physicians), type of scenario (risk or content), scenario (A, B, or C for risk scenarios, and A or B for content scenarios), and combination of these conditions, respectively. The question frequency in each situation (e.g., by provider teams of medical students in risk scenario A) was then calculated by dividing the total number of questions typed into the chatbot with the corresponding total number of sessions when chatbot usage was allowed (Figure 6).

Conversely, the average length of questions was straightforwardly defined as the total word count of questions asked in a situation divided by the corresponding number of questions (Figure 7). To maintain simplicity and consistency, a "word" here referred to a continuous string of text between empty spaces. Under such definition, abbreviations such as "yo" (short for "year old") were considered as single words.

Likewise, the insufficient number of participants restricts a robust statistical analysis, making these statistics indicative of trends rather than allowing for definitive statistical significance testing.

3.4.4 *Elucidating Design Principles.*

After completing the initial quantitative and qualitative analysis, the research board created after rapid analysis (section 3.4.1) was re-examined. In conjunction with quotes and sentiments from the qualitative analysis, quantitative results from the SUS and prompt data were analyzed with the goal to extract principles for the effective use of AI-CDSS in clinical care. Qualitative themes and preliminary conclusions from quantitative data were pooled into common themes and insights that constitute the three design principles outlined in 5.4.

4 FINDINGS

First, we explore in Section 4.1 user behavior with the LLM chatbot through quantitative and qualitative analysis of user-generated prompts, user reaction to the generated responses, and how either the LLM chatbot or the interactive dashboard affected the clinical workflow. Then, we focus in Section 4.2 on the effect of clinical context and user characteristics on human-computer interaction with the LLM chatbot, such as the type of clinical task expected, varying levels of prior exposure to AI-CDSS or clinical expertise, and provider team dynamics. Finally, we describe provider concerns specifically pertaining to trust in Section 4.3.

4.1 Usability

4.1.1 System Usability Score. With the score assignment described in 3.4.2, the reported agreement to the SUS statements one through four from our participants could be summarized as follows using the format mean (standard deviation): 0.75 (0.698), 0.7 (0.954), 0.55 (0.921), and 0.35 (1.014), respectively. The medians were consistently near 1 (corresponding to Likert scale response "Agree") for the first three statements and 0.5 for the fourth, while the most frequent score (mode) across all respondents was 1 for each statement (Figure 5). The calculated average, median, and mode were all positive, suggesting with insufficient statistical power that participants appeared to have an overall positive attitude towards our model's usability, regardless of their access to the chatbot feature.

4.1.2 User-Generated Prompts. LLM chatbots like ChatGPT have garnered significant attention from both the public and the media, given their widespread availability online [99]. As a result, many people have first-hand experience with LLMs and have set expectations for LLM performance and functionality. When interacting with GutGPT, many participants drew indirect or direct comparisons with other LLMs they had used previously. Eight participants referenced other LLM chatbots they had used in response to questions about GutGPT's usability. Many cited their familiarity with ChatGPT as a reason for finding GutGPT easy to use, noting similarities between the two. One IM resident physician answered: "I think it was extremely user friendly, I had used ChatGPT before and so it seemed pretty similar to it." An IM resident physician also reported "I immediately knew what to do when I started using it, it's just like ChatGPT." Many of these participants noted that prior experience with similar systems facilitated a smooth transition for users to GutGPT. We found that higher average prompt frequency per scenario (3.9 versus 2.4) and higher average word counts per prompt (15.3 versus 11.0) in content scenarios compared to risk scenarios, with similar frequency and word counts regardless of clinical expertise level.

Unlike the interactive dashboard, the chatbot requires direct user input to produce an output. While this allows for personalized questions, it also means users must craft questions they believe the chatbot can answer. This extra decision-making step proved challenging for many participants, particularly for those unfamiliar with chatbots. An EM resident physician explained: "The hardest part is AI is brand new to everybody, we don't really know the right questions to ask it or what it can and can't do. What is it going to give me appropriate data for… is it going to mislead me because I don't understand it?" However, as participants interacted more with the system as the trial progressed, their comfort grew. "At first I wasn't really sure what it knew and didn't know and how to make sure the questions I asked were the appropriate questions, it got easier as I went" reported an IM resident physician. To make the transition to use easier, several participants recommended a frequently asked questions (FAQ) section or to adopt autocomplete functionality similar to email clients or search engines.

4.1.3 GutGPT Text Responses. Clinicians value clinical decision support systems that are easy to use and deliver desired information quickly and intuitively. Seven of the participants suggested that GutGPT's text output was too lengthy for efficient use. One participant commented, "It puts out large blocks of text at times, especially when citing sources… that takes a while to read through." Time pressure is a significant concern for all physicians, but it's especially pressing for EM physicians who see a large number of patients during their shifts in the emergency department. They must process vast amounts of data and make numerous clinical decisions in short periods of time. An EM resident physician noted about the text output: "I like the response, however I can see myself saying 'this is taking too long to read' on shift, and I don't think I would do it for every patient, I would probably do it for patients I'm a little unsure about." In addition to the volume of text in the typical chatbot responses, the structure of the responses were also emphasized. Three participants pointed out that the information was often presented as a dense paragraph, making it hard to skim or quickly comprehend. They suggested using bullet points or emphasizing key management principles for a clearer presentation, rather than the uniform format of GutGPT's outputs.

4.1.4 Integration into Existing Workflow. Resident physicians and medical students are accustomed to using the EHR to acquire patient information and aid their clinical decision-making. Users' experience with the tool's EHR integration varied based on their usage patterns and trust in the model's incorporation of patient data. Several contrasted this with traditional CDSS data entry. A medical student noted, "I thought it made [GutGPT] very different than existing clinical prediction tools because I don't need to input every detail myself because they are already incorporated in and it makes me more comfortable that I'm able to use such information." Five participants said they were not pleased with the EHR integration of the chatbot. While laboratory and vital data were populated into the chatbot's risk calculations, some participants still took considerable time entering this data via text entry into the chat queries. As a result, several participants indicated during the post-trial interview that this data entry significantly slowed down their interaction with GutGPT and emphasized the need to refine this feature. When prompted to consider using GutGPT in a real clinical situation, a medical student indicates: "I think it could definitely help, I feel like it would probably be dependent on its integration into Epic."

A common refrain from participants dealt with the scope of information the chatbot could access to generate its responses. Clinical workflows in evaluating UGIB by IM and EM physicians usually involve calculating the Glasgow-Blatchford score (GBS), a tool that stratifies patients with suspected UGIB into high or low risk bleeds [55]. High risk bleeds are more likely to require hospital intervention, while patients with low risk bleeds can likely be safely discharged. Physicians and medical students frequently use online medical reference tools such as MDCalc to access the GBS. During the simulations, several participants tried to use GutGPT to score their simulated patients on the GBS. Often they would directly query GutGPT to calculate the GBS for the patient. As GutGPT is an LLM trained on clinical gastroenterology guidelines, it does not have access to clinical calculators. Thus, when questioned about the GBS, GutGPT typically either stated its inability to compute the score or listed the GBS's components without performing the actual calculation. Such responses understandably frustrated participants, three spoke about it during the qualitative interview. The prevailing

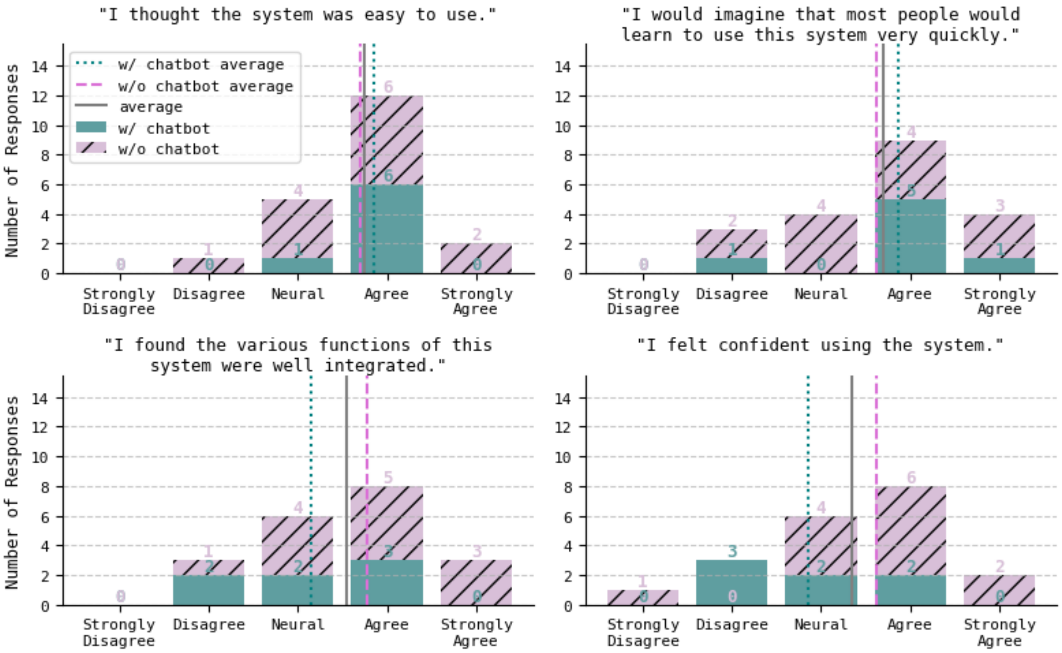


Figure 5: Participant responses to the System Usability Scale survey. Data from participants who had access to the chatbot feature are colored in turquoise, while data from those did not used the chatbot were in purple and hatched. The total height of each bar reflects the combined data.

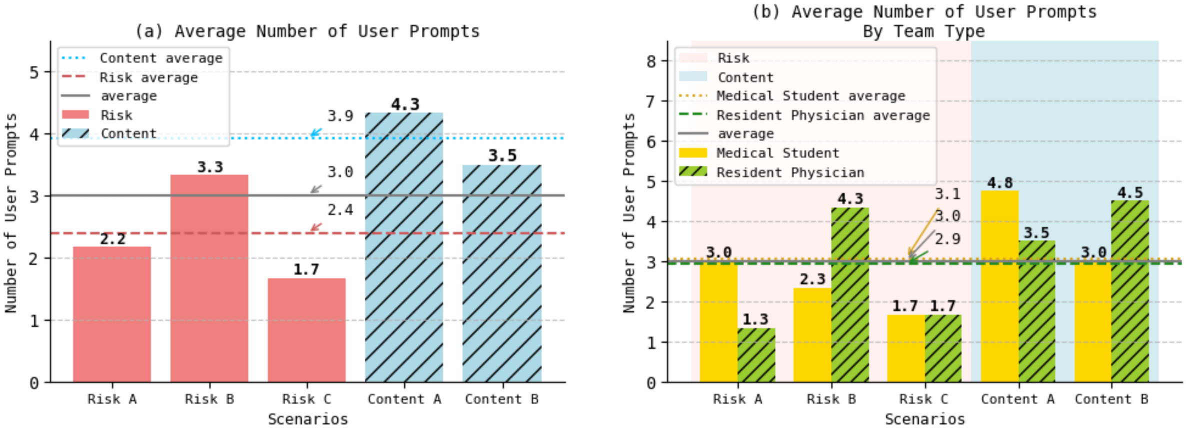


Figure 6: Number of user prompts entered into GutGPT chatbot in each scenario per simulation session. Plot (b) displays the same data in (a) stratified by level of clinical expertise.

suggestion was to integrate GBS calculation within the chatbot. An IM resident physician remarked "it's easier to go onto MDCalc and do [a GBS calculation] rather than using GutGPT if it's going to say 'well this is what GBS means, but we don't actually have the data to pull'." Participants had a preference to access familiar CDSS and anticipated that GutGPT had the ability to access those tools.

4.1.5 Interactive Dashboard Usability. The interactive dashboard displays patient risk scores either numerically or via a graphical representation. One medical student noted their preference for the interactive dashboard over the chatbot interface: "I trust much more when I see the numbers than the words… I have seen other AIs that are text-based and I've personally experienced that they are not working well, so I'm less inclined to trust it." They found it easier to comprehend the explainability of the interactive dashboard which fostered trust in the system. Five other participants did not find the dashboard as intuitive as the chatbot: "a quick glance didn't tell me how to assess [the dashboard], but the [chatbot], I caught on pretty quick what the goal was, how to use it, how to interpret it."

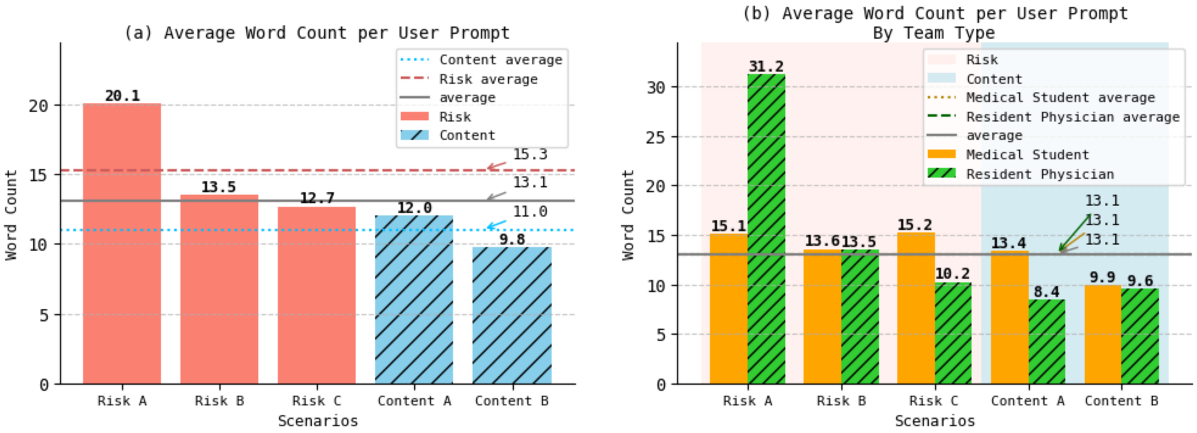


Figure 7: Average user prompt length submitted to the GutGPT chatbot in each scenario. Plot (b) displays the same data in (a) stratified by level of clinical expertise.

A number of participants volunteered answers that indicated that the dashboard was not as straightforward for quick interpretation. An IM resident physician explained, "I couldn't figure out how to use the dashboard in time," a sentiment echoed by others.

Participants found the graphical representation challenging to interpret. An EM resident physician likened it to "looking at the engine of a car that you just bought and you have no idea how it actually runs", highlighting its complexity. The resident physician further reasoned that while potentially beneficial, learning how to use the tool would demand significant time. The intricacy of the graphical representation deterred some from further interaction, believing it wasn't worth the time they could otherwise spend interacting with the scenario, collaborating with teammates, or using other clinical tools. Nine participants reported difficulty in interpreting the graphs.

4.2 Human-Computer Interactions

4.2.1 Clinical Tasks. The qualitative analysis of the GutGPT chatbot conversations reveals a consistent interaction pattern among provider teams of clinicians, irrespective of their clinical expertise level. A provider team typically prompted the chatbot between 1 to 5 times per simulation scenario, with an average of 3 times (Figure 6). The questions had an average length of 13 words (Figure 7).

Notably, when faced with content scenarios, participants asked an average of 1.5 more questions than in risk scenarios. This difference can be possibly attributed to the varying complexity of tasks in these two scenarios. In risk scenarios, the only task for the participants is to decide whether to admit the patient. A brief query like "Should I admit the patient?" or a single search for risk scores could suffice. In contrast, content scenarios require the participants to make a series of management decisions, necessitating reference to medical guidelines and adaptability to changing patient conditions. Ideally, a single query to our LLM chatbot could supply comprehensive guidance, but participants frequently probed further for detailed information or clarification, leading to an increased number of questions to the chatbot.

A related observation is that the participants asked many more questions in Scenario B of the risk assessment phase compared to the other two risk scenarios. Scenario B is a "borderline" case, where the decision to admit or discharge is not as straightforward as in the other cases as the patient's medical data could support either decision. This elevates the case's complexity and requires additional decision support.

4.2.2 Familiarity with the System. Alongside the difference in tasks, the sequence of risk scenarios before content scenarios might contribute to the observed difference in the prompting frequency. As participants became more familiar with GutGPT through risk scenarios, their confidence and willingness to use the system were likely to be higher in the subsequent content scenarios. This might lead to an increase in their interactions with the model, including its chatbot feature.

The reduced average word count per question in content scenarios also suggests a possible influence of familiarity with the system on clinician-chatbot interactions. By inspecting the prompting data, we find that longer questions often unnecessarily repeat patient information embedded in the model, indicating user unfamiliarity. Then, shorter questions in content scenarios could signify improved user understanding and more efficient interactions.

However, we note that this interpretation should be taken with caution, since the conversation between the users and the GutGPT chatbot for each scenario appears to follow a pattern where the users list out the patient's details in the first question and ask follow-up questions without repeating the information. Therefore the lower average word count of questions in the content scenarios could simply be a result of participants asking more follow-up questions in those scenarios.

4.2.3 Level of Clinical Expertise. While quantitative comparisons of participating medical students' and resident physicians' prompt data revealed similar model interaction patterns, qualitative post-simulation interviews indicate that the purposes and experiences of the interactions differed across levels of clinical expertise. Compared to the resident physicians, participating medical students reported more frequently that GutGPT was "super helpful" and could "provide expertise" given their knowledge level. Nine medical students responded in this way, compared to only two resident physicians. Some medical students even positioned GutGPT in leadership roles such as "consultant" and "attending", while resident physicians recognized GutGPT more as a "partner" or "team member" that performs an assistant role in decision-making. One medical student mentioned adopting the model's suggestions even if they conflicted with their own judgment. In contrast, half of resident physicians that performed the leader role during the simulation reported that the recommendations generated by the model did not affect their decisions.

4.2.4 Team Composition and Dynamics. In addition to individual-level factors such as medical expertise, composition and dynamics of provider teams are expected to shape clinician-model interactions as well. Our provider team-based simulation design did not reflect this aspect in the quantitative data, but the post-simulation interviews provided some valuable insights. The roles of the team members in the simulation were designed to reflect real life clinical teams, with senior members in executive decision-making roles (using the dashboard and/or GutGPT) and junior members more responsible for gathering information (e.g. interviewing the mannequin and gathering lab data). That perspective informed some of the participants' responses in the interview. As an IM resident physician placed as a junior member stated: "I was focused on the patient so I didn't get a good look [at the GutGPT system], but it seemed like a useful response from my brief look." Several other IM resident physicians and medical students in the junior member role shared similar comments. Another junior IM resident physician said "I let [the senior team member] deal with the model, and I just worried about the history and physical." Some participants also remarked that their adoption of an AI system depends on team dynamics. A medical student projected that their usage of the model would depend on other provider team members' opinions toward the model: "I think I would be more likely to use it if my attending wanted me to use it." Several participants' opinions diverged when discussing if using GutGPT would be more amenable to good clinical practice in a team or alone. An IM resident physician noted "Typing into the chatbot takes you away from the primary focus of the patient, I would only use it if I'm part of a team for that reason." Others appreciated the input of another source that could function like a team member. Two participants reported their perceived benefits of using the model could depend on the number of people present in the provider team. The context the medical team works in also made a difference for some participants. A medical student said "I think I would feel weird using it in acute situation," indicating the setting a team practices in could influence adoption.

4.3 Trust

Clinicians' trust in the AI-CDSS they interact with plays a large role in whether they decide to adopt the tool into their workflow [32, 107]. In our interviews, several participants discussed their trust of AI-CDSS, some reflecting on their general attitudes about AI-CDSS, and others on their trust of GutGPT after interacting with it first-hand.

While talking about their general attitudes surrounding AI-CDSS, participants voiced concerns about the moral and legal implications of fully adopting these tools in healthcare. Many stated that they "did not believe that clinicians should 100% rely on the model," while citing reasons such as how they were "concerned with liability and responsibility if [they] followed the model and the patient had a bad outcome." Some participants who had prior knowledge of AI systems also reasoned that the AI may output false responses. For instance: "I was concerned that the chatbot will hallucinate, which is particularly bad in medicine." Participants reported they would be less willing to employ LLM-augmented CDSS when there are inaccuracies in the information they output.

Participants also provided direct feedback on their trust levels about GutGPT, based on their experience using the tool during the simulation scenarios. The most commonly cited reason for why participants could not trust the chatbot's outputs was that they did not know what data the chatbot was drawing from. For example, one participant expressed dissatisfaction as the chatbot "did not provide any citations", and another participant said "it would be nice to have hyperlinks of sources and knowledge of where the AI pulled from." This is consistent with [26]'s findings that having fully transparent insight into how an AI generates its output is principal in a clinician's decision to utilize the tool. Further, some participants believed that an AI chatbot could not fully replace a clinician's intuition, and therefore could not be trusted fully. One participant claimed that collecting atmospheric and "emotional" data when entering a patient room is an important part of their workflow, so a limitation of GutGPT was the fact that it did not have such information. We believe that further exploring the implications of these feedback is important for establishing trust in LLM-augmented AI-CDSS in the future.

5 DISCUSSION

Our qualitative and quantitative findings suggest that an LLM-augmented AI-CDSS may increase ease of use in Section 5.1, address challenges with user trust in Section 5.2, and elicit different user patterns based on clinical context and user background in Section 5.3. In Section 5.4 we synthesize our findings into three principles for building LLM-augmented AI-CDSS systems that can meaningfully enhance the work of provider teams in clinical care.

5.1 Large Language Models May Increase Ease of Use for AI-CDSS, but Familiarity Affects User Perceptions

Familiarity is a key aspect of usability, as users are more likely to find recognizable features intuitive. GutGPT was designed with popular AI systems like ChatGPT and text messaging platforms in mind. The participants' opinions on usability sharply differed between the LLM-augmented AI-CDSS (GutGPT) and the interactive dashboard AI-CDSS alone. GutGPT had a recognizable interface that may have contributed to its positive initial reception supported by the interview feedback from many participants as well as the survey results suggesting the perception that systems are easy to use. While the chatbot was seen as immediately intuitive by a large portion of participants, the majority of participants commented that the dashboard was difficult to interpret or not worth taking the time to interpret. However, the qualitative results are more nuanced - despite GutGPT being easier to use than the interactive dashboard, participants remarked on an "activation energy" required to use a chatbot that is not an issue for AI-CDSS without LLM. For example, the interactive dashboard requires no user input for a risk score to be displayed. For GutGPT, there were specific user remarks regarding hesitancy to use due to uncertainty regarding prompt formation that must be overcome. This hesitancy can be especially problematic if the user is unfamiliar with LLMs - reflected by participants who found themselves at a loss on what questions to ask GutGPT. The slight disagreement in the SUS statement "I felt confident using the system" among participants who used the GutGPT chatbot compared to those who used the dashboard alone could be explained by the unfamiliarity with the system as well, given that no one disagreed with the statements that "the system was easy to use" and "most people would learn to use this system very quickly". Uncertainty with use can be problematic in a real clinical situation, as there are many competing demands make clinical care increasingly time-constrained. Typing a query and reading a response in natural language places a further demand on the clinician that could be costly from a time and cognitive-load perspective. Reassuringly, difficulty in prompting faded as participants became familiar with the chatbot and became comfortable working with it. From these observations, the initial approachability was a key factor in allowing participants to experiment with the chatbot and eventually become accustomed to it. However, the dashboard's interface seemed like too steep a challenge to interpret in a short simulated case - and was ignored in many trials.

User familiarity with the interface is not the only experience that matters; in clinical decision-making, familiarity with existing and traditional CDSS can hinder use of AI-CDSS. In UGIB, traditional CDSS is a clinical score, the GBS. Since this CDSS is familiar to providers when caring for patients with UGIB, it was natural that participants reached for this CDSS rather than utilizing the AI-CDSS in our study. This finding reflects similar findings in another usability study of AI-CDSS [111], where the frustration at using the AI-CDSS comes in part from an incomplete understanding of the technical capabilities of AI-CDSS. Interestingly, participants also expressed a desire for the text output from GutGPT to mirror their preferred clinical reference styles. Many desired bullet points or highlighted management steps in the text output, similar to medical reference texts like UptoDate. Possible solutions proposed by users include clear statements of the AI-CDSS capabilities to prevent frustrations that impair usability, as well as the functionality to access traditional CDSS.

The LLM effect on usability for AI-CDSS is consistent with "Unremarkable AI", an idea that stresses unobtrusiveness as crucial to successful adoption of AI-CDSS [116]. If the ideal implementation scenario for an AI-CDSS is one that fits smoothly into the existing workflow of a clinician with little deviation, there should be effort made to craft AI-CDSS that resemble existing tools or applications that clinicians have confidence in navigating. Ideally, AI-CDSS would complement activities that physicians already perform in their jobs. Much of physician responsibility involves data collection, writing clinical notes in the EHR, and deciding which tests and treatments to order. EHR integration is an important factor to access the familiarity that will promote use of an AI-CDSS. EHRs not only contain patient data but also offer clinical calculators, like MDCalc, and clinical pathways to guide diagnostic and treatment choices. During the trials, participants expressed the desire that any AI tool needed to be integrated seamlessly with EHRs. Adequate integration addresses the time pressure and ease of use that many participants alluded to in their answers - an embedded assistant within the EHR that is quickly accessible and helpfully collates relevant patient data. Borrowing from the TURF framework for EHR usability, a system attains acceptable usability when it is easily "learnable" and requires little mental effort to use [119].

5.2 Large Language Models Require Justification with Citations to Promote Trust

We understand trust of algorithmic interfaces as Kizilcec does: "an attitude of confident expectation in an online situation of risk that one's vulnerabilities will not be exploited" [46]. Overall, participants expressed their lack of trust towards LLM-augmented CDSS, and that this lack of trust would deter them from adopting the tool into their workflow. This is consistent with findings by Rousseau *et al.* that trust plays an important role in determining whether or not one is willing to adopt new technologies, particularly involving AI [83].

However, from qualitative interviews we found that one factor that may positively affect trust in the GutGPT responses was the presence of relevant citations, which may indicate the need for transparency regarding the data used to generate the responses. Clinicians are inundated with vast amounts information that they must sift through to make evidence-based diagnostic and management decisions. Clinical guidelines from medical professional societies can be lengthy and difficult to parse for relevant details pertaining to a specific patient. The primary literature from which the guidelines are constructed can be even lengthier and sometimes contradictory. In response to this, clinical reference websites such as UptoDate or ClinicalKey have risen in popularity, offering concise, aggregated information with relevant citations. Many participants reported that GutGPT chat outputs were hard to trust because some of them did not provide citations outlining the source of the information provided. When the chat included citations, participants specifically emphasized how useful they found the response to be. Clear communication about the data used to generate responses from LLM-augmented CDSS is consistent with other studies that found that high-quality labeling leads to higher perceived training data credibility, which in turn enhances users' trust in AI [22]. It is thus imperative to be transparent about the data from which the LLM is generating its responses; when providing a recommendation for clinical management, direct relevant citations should be displayed with every response that is generated. While websites like UptoDate have made evidence-based clinical decision-making easier, they are still general reference materials. They need to be tailored to individual patient scenarios and might not cover unique clinical situations. This represents an opportunity for chat-based AI-CDSS, as information from primary sources can be presented to the clinician in easy-to-understand natural language.

5.3 Human-Computer Interactions Vary By Clinical Tasks and Team Dynamics, But Large Language Model Usage Metrics Are Similar

We found that participants interacted with GutGPT differently based on the clinical task required. Teams using GutGPT in content scenarios submitted more queries to the chatbot than teams in risk scenarios. There were an average of 3.9 queries in content scenarios compared to 2.4 in risk scenarios. This indicates that the chatbot feature was used more heavily when making decisions regarding a care plan, and less utilized for risk assessment. Teams in risk scenarios were asked to determine the risk assessment for the simulated patient - essentially sorting the patient into one of three risk categories. Teams in the content scenarios were asked to stabilize and treat the simulated patient - this is an open-ended situation in which the management options are numerous and unstructured. Choosing the "correct" management decisions requires clinical expertise and a familiarity with UGIB guidelines. GutGPT's guideline-driven recommendations can be perceived as more helpful in these management situations. This differential use is consistent with the paradigm that workflow incorporation of CDSS depends on the needs of human practitioners [94].

We also found differences in how teams interacted with GutGPT according to their level of clinical expertise. Our qualitative interviews suggested that provider teams with more clinical expertise (resident physicians) usually interacted with the AI-CDSS to confirm their own impression or decision, whereas those with less real-world experience (medical students) attributed more expertise to the AI-CDSS and interacted with the system with more deference. From the qualitative interview data, inter-team dynamics contributed to potential use behaviors. The medical team can be a hierarchical structure, group dynamics are often modeled after senior members [106]. Participants assigned as junior team members volunteered that their likeliness of using an AI-CDSS would be increased if those tools are accepted by superiors and peers, indicating that the social expectations of the medical team are an important influence on AI-CDSS adoption and continued use. Division of labor in the medical team also tracks along seniority level, with junior members of the team functioning primarily as data gatherers and reporters while senior members shoulder a larger burden in decision-making, resource allocation, and planning. These roles were reproduced in our simulations, with junior members reporting that they did not occupy their time with familiarizing themselves with the system but instead dove into their roles in interviewing and examining the simulated patient. As our simulated teams approximate real clinical teams, these findings show the importance engaging the key stakeholders in targeting AI-CDSS. In the busy medical team, junior members may find lengthy interactions with an AI-CDSS poorly suited to their role while more senior members might be better situated to devote time and cognitive energy to properly use AI-CDSS. Our findings can be placed in context with existing literature suggesting that interactive technologies are highly dependent on team processes and can influence leadership and team management [30, 54]. These team dynamics are particularly important to consider when deploying technologies such as LLM-augmented AI-CDSS in environments with heterogeneous teams.

Interestingly, we found that provider teams had similar patterns of prompt generation and length across different level of clinical expertise, and suggests a baseline for interactions of on average 3 prompts with 13 words each for provider teams using LLM-augmented AI-CDSS in time-limiting, high-stakes scenarios. This benchmark is particularly valuable because, to our knowledge, we are the first group to measure usage of a LLM-augmented AI-CDSS under real-world clinical simulation conditions.

5.4 Design Principles

Drawing from the user-model interaction insights gleaned from our study, we propose three design principles for AI-CDSS with LLM-augmented interfaces:

5.4.1 Comprehensive Usability Focus. The reported common frustration with dashboard graph interpretation as well as chatbot prompting in our study underscores the necessity of crafting an integrated solution that gives due attention to improving both the usability of the algorithmic output and the LLM-augmented user interface. While enhancing interpretability of algorithm-generated outputs remains crucial for AI-CDSS, equal importance should be placed on providing users with clear guidance on how to interact with and what to expect from new technologies like LLMs. Moreover, a strong design should prioritize seamless integration of these functionalities, an aspect our LLM-augmented AI-CDSS users expressed dissatisfaction with in the SUS survey. Parcipants placed a special emphasis on EHR integration, which is a common refrain from several usability studies with AI-CDSS [112].

5.4.2 Customized Deployment Strategies. As reported in their interviews, participants formed different perceptions of our model's usability and role in clinical decision-making through the same simulation setup, according to their own clinical expertise levels and roles in the workflow. This emphasizes the importance of tailoring the model's deployment strategies to accommodate the varying medical specialties and specific needs of different users within the healthcare ecosystem. As Sendak et al. highlights, stakeholders of varying specialties and expertise should be engaged to provide and iterate feedback of LLM-based AI-CDSS [91].

5.4.3 Understanding and Navigating Team Dynamics. Our study provides preliminary evidence that provider team composition (e.g., in a team or alone) and dynamics (e.g., other team members' perception of the model) exert a complex influence on clinician-CDSS interactions. While further investigations are necessary for a deeper understanding of this topic, design of AI-CDSS should prioritize adaptability and customization to adapt to diverse team compositions. Additionally, strategies such as training in the use of emerging technology may be implemented to ensure effective and harmonious clinician-AI interactions.

6 LIMITATIONS

Medical simulation is primarily designed as an educational exercise to facilitate acquisition of skills by medical trainees in an environment that emulates some of the practical realities of interacting with a patient. However, the simulation environment is an imperfect approximation of a real clinical environment. Use of a simulation mannequin, the lack of distractions, and the abridged time-course of a simulation are examples of factors that prevent medical simulation from achieving strict fidelity with the clinical environment. As a result, medical simulation is a calmer environment than the clinical one, which could encourage AI-CDSS use when the time and social pressures of the real clinical environment might cause trainees to fall back on familiar traditional CDSS. Medical simulation is an environment in which experimentation is welcomed, participants took time to test out and interpret the dashboard and chatbot - luxuries that might not have been afforded to them in the clinical environment. While we provide a quantitative snapshot of potential user patterns of an LLM-augmented AI-CDSS, the interactions were pooled by all members of the provider team and could not depict the individual-level user behaviors (e.g., how model interactions vary based on the user's role in the provider team). Likewise, our study did not capture scenarios in which the clinician interacts with the model independently: some participants touched on this aspect in their qualitative interviews, presenting contradictory views for model usage in such situations. All participants in this study were trainees, and had not yet qualified to practice independently. The majority (76.7%) were under the age of 29, younger than the average independently practicing attending physician. Trainees are in a period of rapid learning of tools and methods that help in clinical care. It would likely be easier for trainees to adopt new technologies in their clinical workflows than more experienced clinicians. Clinicians with a greater amount of experience are more confident in their clinical decision-making and might be less willing to incorporate a new tool into their workflow. Younger people are also more likely to have higher acceptance of AI technology [45].

Another limitation of our study arises from the ongoing and rapid advancement of AI. New models, architectures, and techniques emerge with improvements in their capabilities and performance, so the usability challenges associated with AI-powered systems are likely to shift quickly. One immediate example is that the response latency issue of GutGPT has been mitigated with recent improvements in GPT-3.5-Turbo's inference time.

Lastly, our research plan could be improved. The present design of the study struggled to distinguish between the impacts of increased model familiarity from learning, task complexity, and model performance on our data, especially the prompting pattern. Further exploration of the dashboard's usability through an independent assessment is needed to establish a baseline for better evaluating the value of LLM integration. The truncated SUS survey may limit its comparisons to standards of usability; the fact that the SUS survey was administered only for risk scenarios hindered its ability to reveal users' perception of usability for the whole trial.

7 FUTURE WORK

Our work to evaluate GutGPT and elucidate a more comprehensive understanding about clinicians' attitudes surrounding the AI-CDSS is still ongoing.

We will continue recruiting participants for our usability research to reach the effective size for statistical testing. We will improve our study design to address the limitations described at the end of Section 6. More relevant data such as performance metrics of the LLM component, and time spent for each simulation scenario with or without the chatbot will be collected and assessed in the future trials for a better understanding user-model interactions.

We acknowledge the importance of iterative design in the human-centered approach. We plan to extend our current understanding of user preferences with the following research directions: 1) for usability, we plan to provide guidance on query construction and evaluate its effect on decreasing the initial activation energy that hampers use of the chatbot; 2) for trust, we plan to explore a more active role of an LLM as a team member that listens to and summarizes provider team interactions during the clinical decision process; 3) for user-computer interactions, we plan to customize a workflow that allows individuals to interact with the LLM-augmented AI-CDSS and integrate the user prompts with the provider team interaction with LLM-augmented AI-CDSS.

We plan on updating GutGPT to reflect the three design principles for LLM-augmented AI-CDSS that we proposed in Section 5.4. To improve usability in respect to the LLM-augmented user interface, one potential solution is to provide guidance on writing queries: this can be achieved through query suggestions generated based on commonly asked questions, or "query building blocks" in which clinicians can simply click on components of queries to quickly build their prompt. To better serve users with different level of clinical expertise in UGIB, customizable modes might be developed: clinicians with fewer years of training could be defaulted to model responses with more detailed explanations and more references that provide required expertise, while those from higher training level could choose to receive more concise replies for fact- or decision-checking. To address difficulties in model interactions in special cases such as when clinicians work alone in emergency care, advanced features like real-time speech recognition might be augmented to GutGPT to enable automatic patient-interview summarization that streamline the clinicians' workflow. We plan on implementing these design changes to GutGPT before testing on additional participants, further examining how these changes influence user behavior both at the individual- and team-level.

We also believe that user experience should be studied under conditions that are difficult to achieve in the physical simulation environment. Medical simulations are a valuable training tool that has been found to enhance clinical competence at the undergraduate and postgraduate levels [2]. Even so, medical trainees struggle when transitioning into a real clinical setting [5] due to discrepancies including a static physical environment and lack of environmental distractions present in simulation rooms. Virtual Reality (VR)/Augmented Reality (AR) solution has the potential to improve the levels of realism to enhance learning for simulation studies of LLM-augmented AI-CDSS [2, 39]. Transitioning to VR/AR simulation would allow scalable research for new AI-CDSS like GutGPT, increase the capacity to introduce diversity in medical training (including patient "dummies" of diverse demographics), and increase flexibility in creating and extending simulation environments. We are optimistic about this transition and are interested investigating how it affects future AI-CDSS HCI research.

8 CONCLUSION

In this paper, we sought to extract insights from healthcare providers after they interacted with a LLM-augmented AI-CDSS in simulated clinical scenarios. We present preliminary findings from a randomized controlled trial with 31 participants arranged in provider teams who undergo simulated scenarios of UGIB with an interactive dashboard AI-CDSS with or without an LLM. We find that LLM-augmented AI-CDSS increases ease of use, and that trust can be improved with transparency with supporting evidence of citations in the responses. We found that there appeared to be a baseline utilization pattern of the LLM-augmented AI-CDSS of approximately 3 prompts averaging about 13 words per prompt in each scenario across all participants, though the perception of the LLM-augmented AI-CDSS in human-computer interaction varies by clinical expertise - medical students appreciated the model's expertise while physicians used the model as a check on their intuition. Senior and junior members of the clinical team displayed different behaviors towards AI-CDSS, with greater engagement from senior-level decision-makers. These insights underscore the importance of closely involving healthcare providers in the design and implementation of AI-CDSS. In light of our findings, we propose three fundamental design principles that can guide future refinements of GutGPT and the broader spectrum of AI-CDSS.

ACKNOWLEDGMENTS

We express sincere gratitude to Jeff Hoffman and Tony Polcaro, staff members at Yale Center for Healthcare Simulation, for their invaluable assistance and support throughout our simulation study.

We also extend our thanks to Shinpei Nakamura and Matthew Shen for their technical expertise and support in building GutGPT.

The project described was supported by GI T32 DK007017 from NIH NIDDK. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIDDK or NIH.

REFERENCES

[1] Agency for Healthcare Research and Quality (AHRQ). 2019. Clinical Classifications Software (CCS) for ICD-10-PCS (Beta Version). Online software. https://hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp Accessed November.

[2] Abdulmohsen H. Al-Elq. 2010. Simulation-based medical teaching and learning. *Journal of Family and Community Medicine* 17, 1 (2010), 35–40. https://doi.org/10.4103/1319-1683.68787

[3] Ebrahim Oshni Alvandi, George Van Doorn, and Mark Symmons. 2019. Emotional Awareness and Decision-Making in the Context of Computer-Mediated Psychotherapy. *Journal of Healthcare Informatics Research* 3, 3 (Sept. 2019), 345–370. https://doi.org/10.1007/s41666-019-00050-7

[4] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A. Becker, and Catherine Mooney. 2021. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences* 11, 11 (2021), 5088. https://www.mdpi.com/2076-3417/11/11/5088

[5] Anique Atherley, Diana Dolmans, Wendy Hu, Iman Hegazi, Sonita Alexander, and Pim W Teunissen. 2019. Beyond the struggles: a scoping review on the transition to undergraduate clinical training. *Medical Education* 53, 6 (June 2019), 559–570. https://doi.org/10.1111/medu.13883

[6] Bjarne Austad, Irene Hetlevik, Bente Prytz Mjølstad, and Anne-Sofie Helvik. 2016. Applying clinical guidelines in general practice: a qualitative study of potential complications. *BMC Family Practice* 17, 1 (2016), 92. https://doi.org/10.1186/s12875-016-0490-3

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[8] Jeffrey H. Barsuk, Elaine R. Cohen, Joe Feinglass, William C. McGaghie, and Diane B. Wayne. 2009. Use of Simulation-Based Education to Reduce Catheter-Related Bloodstream Infections. *Archives of Internal Medicine* 169, 15 (2009), 1420. https://doi.org/10.1001/archinternmed.2009.215

[9] Jeffrey H. Barsuk, William C. McGaghie, Elaine R. Cohen, Kevin J. O'Leary, and Diane B. Wayne. 2009. Simulation-based mastery learning reduces complications during central venous catheter insertion in a medical intensive care unit *. *Critical Care Medicine* 37, 10 (2009), 2697–2701. https://journals.lww.com/ccmjournal/fulltext/2009/10000/simulation_based_mastery_learning_reduces.3.aspx

[10] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[11] Y. Bogenstätter, F. Tschan, N. K. Semmer, M. Spychiger, M. Breuer, and S. Marsch. 2009. How accurate is information transmitted to medical professionals joining a medical emergency? A simulator study. *Hum Factors* 51, 2 (2009), 115–25. https://doi.org/10.1177/0018720809336734

[12] Meghan Brennan, Sahil Puri, Tezcan Ozrazgat-Baslanti, Zheng Feng, Matthew Ruppert, Haleh Hashemighouchani, Petar Momcilovic, Xiaolin Li, Daisy Zhe Wang, and Azra Bihorac. 2019. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study. *Surgery* 165, 5 (2019), 1035–1045. https://doi.org/10.1016/j.surg.2019.01.002

[13] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[15] Giles Bruce. 2023. Meditech, Google partner to bring generative AI to EHRs. Retrieved September 12, 2023 from https://www.beckershospitalreview.com/digital-health/meditech-google-partner-to-bring-generative-ai-to-ehrs.html

[16] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[17] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 104:1–104:24. https://doi.org/10.1145/3359206

[18] Mitchell S. Cappell and David Friedel. 2008. Initial Management of Acute Upper Gastrointestinal Bleeding: From Initial Evaluation up to Gastrointestinal Endoscopy. *Medical Clinics of North America* 92, 3 (2008), 491–509. https://doi.org/10.1016/j.mcna.2008.01.005

[19] Microsoft News Center. 2023. Microsoft and Epic expand strategic collaboration with integration of Azure OpenAI Service. Retrieved September 12, 2023 from https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/

[20] Colleen Chan, Kisung You, Sunny Chung, Mauro Giuffrè, Theo Saarinen, Niroop Rajashekar, Yuan Pu, Yeo Eun Shin, Loren Laine, Ambrose Wong, René Kizilcec, Jasjeet Sekhon, and Dennis Shung. 2023. Assessing the Usability of GutGPT: A Simulation Study of an AI Clinical Decision Support System for Gastrointestinal Bleeding Risk. arXiv:2312.10072 [cs.HC]

[21] Anjun Chen and Drake O. Chen. 2023. Accuracy of Chatbots in Citing Journal Articles. *JAMA Network Open* 6, 8 (2023), e2327647–e2327647. https://doi.org/10.1001/jamanetworkopen.2023.27647

[22] Cheng Chen. 2023. Is this AI trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust - CHI '23. https://programs.sigchi.org/chi/2023/program/content/95899

[23] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[24] Omer Deperlioglu, Utku Kose, Deepak Gupta, Ashish Khanna, Fabio Giampaolo, and Giancarlo Fortino. 2022. Explainable framework for Glaucoma diagnosis by image processing and convolutional neural network synergy: Analysis with doctor evaluation. *Future Generation Computer Systems* 129 (2022), 152–169. https://doi.org/10.1016/j.future.2021.11.018

[25] Yuhan Du, Anthony R. Rafferty, Fionnuala M. McAuliffe, Lan Wei, and Catherine Mooney. 2022. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Scientific Reports* 12, 1 (2022), 1170. https://doi.org/10.1038/s41598-022-05112-2

[26] Juan Manuel Durán and Karin Rolanda Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47, 5 (2021), 329–335. https://doi.org/10.1136/medethics-2020-106820 arXiv:https://jme.bmj.com/content/47/5/329.full.pdf

[27] Jannatul Ferdush, Mahbuba Begum, and Sakib Hossain. 2023. ChatGPT and Clinical Decision Support: Scope, Application, and Limitations. *Annals of Biomedical Engineering* (07 2023), 1–6. https://doi.org/10.1007/s10439-023-03329-4

[28] D. M. Gaba. 2004. The future vision of simulation in health care. *Qual Saf Health Care* 13 Suppl 1, Suppl 1 (2004), i2–10. https://doi.org/10.1136/qhc.13.suppl_1.i2

[29] Amit X. Garg, Neill K. J. Adhikari, Heather McDonald, M. Patricia Rosas-Arellano, P. J. Devereaux, Joseph Beyene, Justina Sam, and R. Brian Haynes. 2005. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 293, 10 (March 2005), 1223–1238. https://doi.org/10.1001/jama.293.10.1223

[30] Eleni Georganta, Claudia Peus, and Jasmin Niess. 2023. Interactive technologies through the lens of team effectiveness: an interdisciplinary systematic literature review. *European Journal of Work and Organizational Psychology* 0, 0 (Feb. 2023), 1–16. https://doi.org/10.1080/1359432X.2023.2178904 Publisher: Routledge _eprint: https://doi.org/10.1080/1359432X.2023.2178904.

[31] Megan E Gregory, Ashley M Hughes, Lauren E Benishek, Shirley C Sonesh, Elizabeth H Lazzara, LeChauncy D Woodard, and Eduardo Salas. 2021. Toward the development of the perfect medical team: Critical components for adaptation. *Journal of patient safety* 17, 2 (2021), e47–e70.

[32] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (Dec. 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120 Publisher: American Association for the Advancement of Science.

[33] Alison Hamilton. 2013. Qualitative Methods in Rapid Turn-Around Health Services Research. https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=780

[34] UNC Health and UNC School of Medicine. 2023. UNC Health Works with Epic on Integration of Generative Artificial Intelligence (AI) Tools. https://news.unchealthcare.org/2023/05/unc-health-works-with-epic-on-integration-of-generative-artificial-intelligence-ai-tools/ Accessed on November 2023.

[35] Katharine E Henry, Rachel Kornfield, Anirudh Sridharan, Robert C Linton, Catherine Groh, Tony Wang, Albert Wu, Bilge Mutlu, and Suchi Saria. 2022. Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. *NPJ digital medicine* 5, 1 (2022), 97.

[36] Katharine E. Henry, Rachel Kornfield, Anirudh Sridharan, Robert C. Linton, Catherine Groh, Tony Wang, Albert Wu, Bilge Mutlu, and Suchi Saria. 2022. Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. *npj Digital Medicine* 5, 1 (2022), 97. https://doi.org/10.1038/s41746-022-00597-7

[37] Namki Hong, Heajeong Park, and Yumie Rhee. 2020. Machine learning application in diabetes and endocrine disorders. *The Journal of Korean Diabetes* 21, 3 (2020), 130–139.

[38] S. Hunziker, F. Tschan, N. K. Semmer, M. D. Howell, and S. Marsch. 2010. Human factors in resuscitation: Lessons learned from simulator studies. *J Emerg Trauma Shock* 3, 4 (2010), 389–94. https://doi.org/10.4103/0974-2700.70764

[39] S. Barry Issenberg, William C. McGaghie, Emil R. Petrusa, David Lee Gordon, and Ross J. Scalese. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher* 27, 1 (Jan. 2005), 10–28. https://doi.org/10.1080/01421590500046924

[40] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 659, 14 pages. https://doi.org/10.1145/3411764.3445385

[41] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (, Yokohama, Japan,) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 659, 14 pages. https://doi.org/10.1145/3411764.3445385

[42] Hsu-Ju Kao, Smile Chien, Wen-Chung Wang, Willy Chou, and Julie Chow. 2023. Assessing ChatGPT's capacity for clinical decision support in pediatrics: A comparative study with pediatricians using KIDMAP of Rasch analysis. *Medicine* 102 (06 2023), e34068. https://doi.org/10.1097/MD.0000000000034068

[43] M. Karabacak and K. Margetis. 2023. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus* 15, 5 (2023), e39305. https://doi.org/10.7759/cureus.39305

[44] Rainu Kaushal, Kaveh G. Shojania, and David W. Bates. 2003. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Archives of Internal Medicine* 163, 12 (June 2003), 1409–1416. https://doi.org/10.1001/archinte.163.12.1409

[45] Sage Kelly, Sherrie-Anne Kaye, and Oscar Oviedo-Trespalacios. 2023. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics* 77 (2023), 101925. https://doi.org/10.1016/j.tele.2022.101925

[46] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2390–2395. https://doi.org/10.1145/2858036.2858402

[47] Linda T. Kohn, Janet M. Corrigan, and Molla S. Donaldson. 2000. *To Err Is Human: Building a Safer Health System.* The National Academies Press, Washington, DC. https://doi.org/10.17226/9728

[48] Marin H Kollef, Denise A Canfield, and Gary R Zuckerman. 1995. Triage considerations for patients with acute gastrointestinal hemorrhage admitted to a medical intensive care unit. *Critical care medicine* 23, 6 (1995), 1048–1054.

[49] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender Bias and Stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (<conf-loc>, <city>Delft</city>, <country>Netherlands</country>, </conf-loc>) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. https://doi.org/10.1145/3582269.3615599

[50] Gilad J. Kuperman, Anne Bobb, Thomas H. Payne, Anthony J. Avery, Tejal K. Gandhi, Gerard Burns, David C. Classen, and David W. Bates. 2007. Medication-related Clinical Decision Support in Computerized Provider Order Entry Systems: A Review. *Journal of the American Medical Informatics Association : JAMIA* 14, 1 (2007), 29–40. https://doi.org/10.1197/jamia.M2170

[51] Andre W. Kushniruk and Vimla L. Patel. 2004. Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of Biomedical Informatics* 37, 1 (Feb. 2004), 56–76. https://doi.org/10.1016/j.jbi.2004.01.003

[52] Loren Laine, Alan N Barkun, John R Saltzman, Myriam Martel, and Grigorios I Leontiadis. 2021. ACG clinical guideline: upper gastrointestinal and ulcer bleeding. *Official journal of the American College of Gastroenterology| ACG* 116, 5 (2021), 899–917.

[53] Adam B Landman, Lisa Redden, Pamela Neri, Stephen Poole, Jan Horsky, Ali S Raja, Charles N Pozner, Gordon Schiff, and Eric G Poon. 2013. Using a medical simulation center as an electronic health record usability laboratory. *Journal of the American Medical Informatics Association* 21, 3 (2013), 558–563. https://doi.org/10.1136/amiajnl-2013-002233

[54] Lindsay Larson and Leslie DeChurch. 2020. Leading Teams in the Digital Age: Four Perspectives on Technology and What They Mean for Leading Teams. *The leadership quarterly* 31, 1 (Feb. 2020), 101377. https://doi.org/10.1016/j.leaqua.2019.101377

[55] Stig Borbjerg Laursen, Jane Moller Hansen, and Ove B. Schaffalitzky de Muckadell. 2012. The Glasglow Blatchford Score is the Most Accurate Assessment of Patients with Upper Gastrointestinal Hemorrhage. *Clinical Gastroenterology and Hepatology* 10, 10 (2012), 1130–1135.E1.

[56] Terrence Lee, Neil Shah, Alyssa Haack, and Sally Baxter. 2020. Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review. *Informatics* 7 (07 2020), 25. https://doi.org/10.3390/informatics7030025

[57] Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024. ChatGPT in healthcare: A taxonomy and systematic review. *Computer Methods and Programs in Biomedicine* 245 (2024), 108013. https://doi.org/10.1016/j.cmpb.2024.108013

[58] Zhiqiang Liao, Jian Wang, Zhuozheng Shi, Lintao Lu, and Hitoshi Tabata. 2024. Revolutionary potential of ChatGPT in constructing intelligent clinical decision support systems. *Annals of Biomedical Engineering* 52, 2 (2024), 125–129.

[59] JCHBM Luijten, MJ Westerman, GAP Nieuwenhuijzen, JEW Walraven, MN Sosef, LV Beerepoot, R van Hillegersberg, K Muller, R Hoekstra, JJGHM Bergman, et al. 2022. Team dynamics and clinician's experience influence decision-making during Upper-GI multidisciplinary team meetings: A multiple case study. *Frontiers in Oncology* 12 (2022), 1003506.

[60] Laura Légat, Sven Van Laere, Marc Nyssen, Stephane Steurbaut, Alain G Dupont, and Pieter Cornu. 2018. Clinical Decision Support Systems for Drug Allergy Checking: Systematic Review. *Journal of Medical Internet Research* 20, 9 (Sept. 2018), e258. https://doi.org/10.2196/jmir.8206

[61] Farah Magrabi, Elske Ammenwerth, Jytte Brender McNair, Nicolet F De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, Philip J Scott, Tuulikki Vehko, Zoie Shui-Yee Wong, et al. 2019. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications: a position paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. *Yearbook of medical informatics* 28, 1 (2019), 128.

[62] Jolanta Majer, Milosz J. Jaguszewski, Michael Frass, Marcin Leskiewicz, Jacek Smereka, Jerzy R. Ładny, Oliver Robak, and Łukasz Szarpak. 2019. Does the use of cardiopulmonary resuscitation feedback devices improve the quality of chest compressions performed by doctors? A prospective, randomized, cross-over simulation study. *Cardiology Journal* 26, 5 (2019), 529–535. https://doi.org/10.5603/cj.a2018.0091

[63] Elizabeth Manias, Snezana Kusljic, and Angela Wu. 2020. Interventions to reduce medication errors in adult medical and surgical settings: a systematic review. *Therapeutic Advances in Drug Safety* 11 (Nov. 2020), 2042098620968309. https://doi.org/10.1177/2042098620968309

[64] Allison Y. Strochlic Michael E. Wiklund P.E., Jonathan Kendler. 2016. *Usability Testing of Medical Devices, Second Edition.* CRC Press, Boca Raton, FL, USA. https://doi.org/10.1201/b19082

[65] Kristen Miller, Danielle Mosby, Muge Capan, Rebecca Kowalski, Raj Ratwani, Yaman Noaiseh, Rachel Kraft, Sanford Schwartz, William S Weintraub, and Ryan Arnold. 2017. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. *Journal of the American Medical Informatics Association : JAMIA* 25, 5 (Nov. 2017), 585–592. https://doi.org/10.1093/jamia/ocx118

[66] R A Miller. 1994. Medical diagnostic decision support systems–past, present, and future: a threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association* 1, 1 (1994), 8–27. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC116181/

[67] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* 19, 6 (Nov. 2018), 1236–1246. https://doi.org/10.1093/bib/bbx044

[68] Lorenzo Moja, Koren H. Kwag, Theodore Lytras, Lorenzo Bertizzolo, Linn Brandt, Valentina Pecoraro, Giulio Rigon, Alberto Vaona, Francesca Ruggiero, Massimo Mangia, Alfonso Iorio, Ilkka Kunnamo, and Stefanos Bonovas. 2014. Effectiveness of Computerized Decision Support Systems Linked to Electronic Health Records: A Systematic Review and Meta-Analysis. *American Journal of Public Health* 104, 12 (Dec. 2014), e12–e22. https://doi.org/10.2105/AJPH.2014.302164 Publisher: American Public Health Association.

[69] Christoph Molnar. 2020. *Interpretable machine learning.* Lulu. com.

[70] Daniel R. Murphy, Brian Reis, Dean F. Sittig, and Hardeep Singh. 2012. Notifications received by primary care practitioners in electronic health records: a taxonomy and time analysis. *The American Journal of Medicine* 125, 2 (Feb. 2012), 209.e1–7. https://doi.org/10.1016/j.amjmed.2011.07.029

[71] OpenAI. 2022. Introducing ChatGPT. Retrieved September 5, 2023 from https://openai.com/research/chatgpt

[72] Jerome A Osheroff, Jonathan Teich, Donald Levick, Luis Saldana, Ferdinand Velasco, Dean Sittig, Kendall Rogers, and Robert Jenders. 2012. *Improving outcomes with clinical decision support: an implementer's guide.* Himss Publishing, Chicago, IL.

[73] Jerome A. Osheroff, Jonathan M. Teich, Blackford Middleton, Elaine B. Steen, Adam Wright, and Don E. Detmer. 2007. A Roadmap for National Action on Clinical Decision Support. *Journal of the American Medical Informatics Association : JAMIA* 14, 2 (2007), 141–145. https://doi.org/10.1197/jamia.M2334

[74] Anne F. Peery, Seth D. Crockett, Caitlin C. Murphy, Elizabeth T. Jensen, Hannah P. Kim, Matthew D. Egberg, Jennifer L. Lund, Andrew M. Moon, Virginia Pate, Edward L. Barnes, Courtney L. Schlusser, Todd H. Baron, Nicholas J. Shaheen, and Robert S. Sandler. 2022. Burden and Cost of Gastrointestinal, Liver, and Pancreatic Diseases in the United States: Update 2021. *Gastroenterology* 162, 2 (2022), 621–644. https://doi.org/10.1053/j.gastro.2021.10.017

[75] Haroldas Petkus, Jan Hoogewerf, and Jeremy C Wyatt. 2020. What do senior physicians think about AI and clinical decision support systems: Quantitative and qualitative analysis of data from specialty societies. *Clinical Medicine* 20, 3 (May 2020), 324–328. https://doi.org/10.7861/clinmed.2019-0317

[76] Deborah Plana, Dennis L. Shung, Alyssa A. Grimshaw, Anurag Saraf, Joseph J. Y. Sung, and Benjamin H. Kann. 2022. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. *JAMA Network Open* 5, 9 (2022), e2233946–e2233946. https://doi.org/10.1001/jamanetworkopen.2022.33946

[77] Aaron I F Poon and Joseph J Y Sung. 2021. Opening the black box of AI-Medicine. *Journal of Gastroenterology and Hepatology* 36, 3 (2021), 581–584. https://doi.org/10.1111/jgh.15384

[78] S. V. Praveen and R. Deepika. 2023. Exploring the perspective of infection clinicians on the integration of Large Language Models (LLMs) in clinical practice: A deep learning study in healthcare. *Journal of Infection* 87, 4 (2023), e68–e69. https://doi.org/10.1016/j.jinf.2023.07.011

[79] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[80] Arya Rao, John Kim, Meghana Kamineni, Michael Pang, Winston Lie, Keith J. Dreyer, and Marc D. Succi. 2023. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *Journal of the American College of Radiology* 20, 10 (2023), 990–997. https://doi.org/10.1016/j.jacr.2023.05.003

[81] Arya Rao, Michael Pang, John Kim, Meghana Kamineni, Winston Lie, Anoop K Prasad, Adam Landman, Keith Dreyer, and Marc D Succi. 2023. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *Journal of Medical Internet Research* 25 (2023), e48659.

[82] Michael A. Rosen, Eduardo Salas, Teresa S. Wu, Salvatore Silvestri, Elizabeth H. Lazzara, Rebecca Lyons, Sallie J. Weaver, and Heidi B. King. 2008. Promoting Teamwork: An Event-based Approach to Simulation-based Teamwork Training for Emergency Medicine Residents. *Academic Emergency Medicine* 15, 11 (2008), 1190–1198. https://doi.org/10.1111/j.1553-2712.2008.00180.x

[83] Nikki Rousseau, Elaine McColl, John Newton, Jeremy Grimshaw, and Martin Eccles. 2003. Practice based, longitudinal, qualitative interview study of computerised evidence based guidelines in primary care. *BMJ : British Medical Journal* 326, 7384 (Feb. 2003), 314. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC143528/

[84] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.

[85] Leonardo Rundo, Roberto Pirrone, Salvatore Vitabile, Evis Sala, and Orazio Gambino. 2020. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *Journal of Biomedical Informatics* 108 (Aug. 2020), 103479. https://doi.org/10.1016/j.jbi.2020.103479

[86] Leila Safari and Jon D. Patrick. 2018. Complex analyses on clinical information systems using restricted natural language querying to resolve time-event dependencies. *Journal of Biomedical Informatics* 82 (June 2018), 13–30. https://doi.org/10.1016/j.jbi.2018.04.004

[87] Malik Sallam. 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* 11, 6 (2023), 887. https://doi.org/10.3390/healthcare11060887

[88] Sahil Sandhu, Anthony L Lin, Nathan Brajer, Jessica Sperling, William Ratliff, Armando D Bedoya, Suresh Balu, Cara O'Brien, and Mark P Sendak. 2020. Integrating a Machine Learning System Into Clinical Workflows: Qualitative Study. *J Med Internet Res* 22, 11 (2020), e22421. https://doi.org/10.2196/22421

[89] Arabella Scantlebury, Alison Booth, and Bec Hanley. 2017. Experiences, practices and barriers to accessing health information: A qualitative study. *International Journal of Medical Informatics* 103 (July 2017), 103–108. https://doi.org/10.1016/j.ijmedinf.2017.04.018

[90] Matthew W. Segar, Muthiah Vaduganathan, Kershaw V. Patel, Darren K. McGuire, Javed Butler, Gregg C. Fonarow, Mujeeb Basit, Vaishnavi Kannan, Justin L. Grodin, Brendan Everett, Duwayne Willett, Jarett Berry, and Ambarish Pandey. 2019. Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. *Diabetes Care* 42, 12 (Dec. 2019), 2298–2306. https://doi.org/10.2337/dc19-0587

[91] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 99–109. https://doi.org/10.1145/3351095.3372827

[92] Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, Kelly Kester, Sahil Sandhu, Kristin Corey, Nathan Brajer, Christelle Tan, Anthony Lin, Tres Brown, Susan Engelbosch, Kevin Anstrom, Madeleine Clare Elish, Katherine Heller, Rebecca Donohoe, Jason Theiling, Eric Poon, Suresh Balu, Armando Bedoya, and Cara O'Brien. 2020. Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Med Inform* 8, 7 (2020), e15182. https://doi.org/10.2196/15182

[93] Edward H. Shortliffe and Martin J. Sepúlveda. 2018. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* 320, 21 (Dec. 2018), 2199–2200. https://doi.org/10.1001/jama.2018.17163

[94] Ida Sim, Paul Gorman, Robert A. Greenes, R. Brian Haynes, Bonnie Kaplan, Harold Lehmann, and Paul C. Tang. 2001. Clinical Decision Support Systems for the Practice of Evidence-based Medicine. *Journal of the American Medical Informatics Association* 8, 6 (Nov. 2001), 527–534. https://doi.org/10.1136/jamia.2001.0080527

[95] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[96] J. Smereka, L. Szarpak, M. Czekajlo, A. Abelson, P. Zwolinski, T. Plusa, D. Dunder, M. Dabrowski, Z. Wiesniewska, O. Robak, M. Frass, G. U. Sivrikaya, and K. Ruetzler. 2019. The TrueCPR device in the process of teaching cardiopulmonary resuscitation: A randomized simulation trial. *Medicine (Baltimore)* 98, 27 (2019), e15995. https://doi.org/10.1097/md.0000000000015995

[97] Vera Sorin, Eyal Klang, Miri Sklair-Levy, Israel Cohen, Douglas B. Zippel, Nora Balint Lahat, Eli Konen, and Yiftach Barash. 2023. Large language model (ChatGPT) as a support tool for breast tumor board. *npj Breast Cancer* 9, 1 (2023), 44. https://doi.org/10.1038/s41523-023-00557-8

[98] Pierre-Clément Thiebaud, Youri Yordanov, Jacques-Emmanuel Galimard, Pierre-Alexis Raynal, Sébastien Beaune, Laurent Jacquin, François-Xavier Ageron, Dominique Pateron, and Group the Initiatives de Recherche aux Urgences. 2017. Management of upper gastrointestinal bleeding in emergency departments, from bleeding symptoms to diagnosis: a prospective, multicenter, observational study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 25, 1 (2017), 78. https://doi.org/10.1186/s13049-017-0425-6

[99] H. Holden Thorp. 2023. ChatGPT is fun, but not an author. *Science* 379, 6630 (2023), 313–313. https://doi.org/10.1126/science.adg7879

[100] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58, 1 (1996), 267–288.

[101] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, University of Michigan, Ann Arbor, MI, 359–380.

[102] U.S. Food and Drug Administration. 2017. Software as a Medical Device (SAMD): Clinical Evaluation Guidance for Industry and Food and Drug Administration Staff Food and Drug Administration Center for Devices and Radiological Health. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation

[103] U.S. Food and Drug Administration. 2022. Clinical Decision Support Software - Guidance for Industry and Food and Drug Administration Staff. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

[104] Heleen van der Sijs, Jos Aarts, Arnold Vulto, and Marc Berg. 2006. Overriding of Drug Safety Alerts in Computerized Physician Order Entry. *Journal of the American Medical Informatics Association : JAMIA* 13, 2 (2006), 138–147. https://doi.org/10.1197/jamia.M1809

[105] M. E. van Leerdam. 2008. Epidemiology of acute upper gastrointestinal bleeding. *Best Practice & Research Clinical Gastroenterology* 22, 2 (2008), 209–224. https://doi.org/10.1016/j.bpg.2007.10.011

[106] Meredith Vanstone and Lawrence Grierson. 2022. Thinking about social power and hierarchy in medical education. *Medical Education* 56, 1 (2022), 91–97. https://doi.org/10.1111/medu.14659

[107] Alfredo Vellido. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications* 32, 24 (Dec. 2020), 18069–18083. https://doi.org/10.1007/s00521-019-04051-w

[108] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (September 2003), 425–478. https://ssrn.com/abstract=3375136

[109] Rebecca Voelker. 2009. Medical Simulation Gets Real. *JAMA* 302, 20 (2009), 2190–2192. https://doi.org/10.1001/jama.2009.1677

[110] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[111] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445432

[112] Liuping Wang, Zhan Zhang, Dakuo Wang, Weidan Cao, Xiaomu Zhou, Ping Zhang, Jianxing Liu, Xiangmin Fan, and Feng Tian. 2023. Human-centered design and evaluation of AI-empowered clinical decision support systems: a systematic review. *Frontiers in Computer Science* 5 (2023), 1187299.

[113] Saul N. Weingart, Maria Toth, Daniel Z. Sands, Mark D. Aronson, Roger B. Davis, and Russell S. Phillips. 2003. Physicians' decisions to override computerized drug alerts in primary care. *Archives of Internal Medicine* 163, 21 (Nov. 2003), 2625–2631. https://doi.org/10.1001/archinte.163.21.2625

[114] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. 2021. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine* 181, 8 (2021), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

[115] C. Yang, J. A. Kors, S. Ioannou, L. H. John, A. F. Markus, A. Rekkas, M. A. J. de Ridder, T. M. Seinen, R. D. Williams, and P. R. Rijnbeek. 2022. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. *J Am Med Inform Assoc* 29, 5 (2022), 983–989. https://doi.org/10.1093/jamia/ocac002

[116] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

[117] Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. 2023. Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Trans. Comput.-Hum. Interact.* 30, 2 (2023), Article 33. https://doi.org/10.1145/3582430

[118] Jiajie Zhang, Todd R Johnson, Vimla L Patel, Danielle L Paige, and Tate Kubose. 2003. Using usability heuristics to evaluate patient safety of medical devices. *Journal of Biomedical Informatics* 36, 1 (Feb. 2003), 23–30. https://doi.org/10.1016/S1532-0464(03)00060-1

[119] Jiajie Zhang and Muhammad F. Walji. 2011. TURF: Toward a unified framework of EHR usability. *Journal of Biomedical Informatics* 44, 6 (Dec. 2011), 1056–1067. https://doi.org/10.1016/j.jbi.2011.08.005

[120] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. 2023. Pre-trained Multimodal Large Language Model Enhances Dermatological Diagnosis using SkinGPT-4. *medRxiv* (2023), 2023.06.10.23291127. https://doi.org/10.1101/2023.06.10.23291127