Hamid R. Arabnia · Leonidas Deligiannidis ·
Soheyla Amirian · Farzan Shenavarmasouleh ·
Farid Ghareh Mohammadi ·
David de la Fuente (Eds.)

Communications in Computer and Information Science 2252

Artificial Intelligence and Applications

26th International Conference, ICAI 2024
Held as Part of the World Congress in Computer Science
Computer Engineering and Applied Computing, CSCE 2024
Las Vegas, NV, USA, July 22–25, 2024, Revised Selected Papers

Springer

Communications in Computer and Information Science 2252

Rationale
The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics
The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors
Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (http://link.springer.com) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at http://link.springer.com/bookseries/7899. Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process
The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing
CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start
To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com

Hamid R. Arabnia · Leonidas Deligiannidis ·
Soheyla Amirian · Farzan Shenavarmasouleh ·
Farid Ghareh Mohammadi · David de la Fuente
Editors

Artificial Intelligence and Applications

26th International Conference, ICAI 2024
Held as Part of the World Congress in Computer Science
Computer Engineering and Applied Computing, CSCE 2024
Las Vegas, NV, USA, July 22–25, 2024
Revised Selected Papers

Springer

Editors
Hamid R. Arabnia
The University of Georgia
Athens, GA, USA

Soheyla Amirian
Pace University
New York, NY, USA

Farid Ghareh Mohammadi
Mayo Clinic
Jacksonville, FL, USA

Leonidas Deligiannidis
Wentworth Institute of Technology
Boston, MA, USA

Farzan Shenavarmasouleh
MediaLab Inc.
Lawrenceville, GA, USA

David de la Fuente
University of Oviedo
Oviedo, Asturias, Spain

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-86622-7 ISBN 978-3-031-86623-4 (eBook)
https://doi.org/10.1007/978-3-031-86623-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

It is our great pleasure to introduce this collection of research papers presented at the 26th International Conference on Artificial Intelligence (ICAI 2024). This volume features a selection of papers showcasing significant advancements and innovative research in artificial intelligence. The ICAI 2024 conference was held as part of the federated 2024 Congress on Computer Science, Computer Engineering, and Applied Computing (CSCE 2024), which took place from July 22 to July 25, 2024, in Las Vegas, Nevada, USA.

The CSCE 2024 Congress brought together papers from a diverse array of communities, including researchers from universities, corporations, and government agencies. Accepted papers are published by Springer Nature, and the proceedings showcase solutions to key challenges in various critical areas of Computer Science, Computer Engineering, and Applied Computing.

Computer Science (CS) is the study of computational systems, data processing, information management, and automation. Many applications in CS focus on solving problems that would be impossible or extremely difficult to address without the use of computers. It serves as a bridge between computational science and other scientific fields. The interdisciplinary nature of CS involves leveraging computers to understand and solve complex challenges, making it the science of using computers to advance scientific discovery. Computer Engineering (CE), on the other hand, integrates aspects of computer science, electronic engineering, and electrical engineering. It encompasses the design and production of computer hardware, such as chips, servers, supercomputers, embedded systems, and communication systems, among others.

Considering the above broad outline, the CSCE 2024 Congress was composed of the following focused conferences:

Applied Cognitive Computing (ACC); Bioinformatics & Computational Biology (BIOCOMP); Biomedical Engineering (BIOENG); Scientific Computing (CSC); e-Learning, e-Business, Enterprise Information Systems, & e-Government (EEE); Embedded Systems, Cyber-physical Systems, & Applications (ESCS); Foundations of Computer Science (FCS); Frontiers in Education (FECS); Grid, Cloud, & Cluster Computing (GCC); Health Informatics (HIMS); Artificial Intelligence (ICAI); Data Science (ICDATA); Emergent Quantum Technologies (ICEQT); Internet Computing & IoT (ICOMP); Wireless Networks (ICWN); Information & Knowledge Engineering (IKE); Image Processing, Computer Vision, & Pattern Recognition (IPCV); Modeling, Simulation & Visualization Methods (MSV); Parallel & Distributed Processing Techniques & Applications (PDPTA); Security & Management (SAM); and Software Engineering Research & Practice (SERP). The scope of each track can be found at: https://www.american-cse.org/csce2024/conferences

The primary objective of the CSCE Congress and its associated conferences is to foster opportunities for cross-fertilization between the fields of Computer Science (CS) and Computer Engineering (CE). The CSCE Congress is deeply committed to promoting diversity and eliminating discrimination, both in its role as a conference organizer and

as a service provider. Our goal is to create an inclusive culture that respects and values differences, promotes dignity, equality, and diversity, and encourages individuals to reach their full potential. We are also dedicated, wherever possible, to organizing a conference that represents the global community. We sincerely hope that we have succeeded in achieving these important objectives.

The Steering Committee and the Program Committee would like to extend their gratitude to all the authors who submitted papers for consideration. This year's conference received submissions from 67 countries, with approximately 56% of them coming from outside the USA. Each submitted paper underwent a rigorous peer-review process, with at least two experts (an average of 2.4 referees per paper) evaluating the submissions based on originality, significance, clarity, impact, and soundness. In cases where reviewers' recommendations were contradictory, a program committee member was tasked with making the final decision, often consulting additional referees for further guidance. The Congress followed the guidelines of COPE (Committee on Publication Ethics):

- Typical submissions underwent a single-blind peer review process, in which the authors remained unaware of the identities of the reviewers, while the reviewers were informed of the authors' identities.
- Papers authored by one or more members of the program committee, including co-chairs, were subjected to a double-blind peer review process, ensuring that neither the authors nor the reviewers were aware of each other's identities or affiliations.

The ICAI 2024 Conference received 376 submissions, with 75 papers accepted, reflecting a 19% acceptance rate. This volume includes only 41 of the accepted papers.

We are deeply grateful to the many colleagues who contributed their time and effort to organizing the Congress. In particular, we extend our thanks to the members of the Program Committee, the Steering Committee, the referees, and the Chairs and organizers of individual sessions and conferences. We would also like to express our appreciation to the primary sponsor of the conference, the American Council on Science & Education. The list of members of the Program Committee for each track can be found at: https://www.american-cse.org/csce2024/committees

We extend our heartfelt gratitude to all the speakers and authors for their valuable contributions. We would also like to thank the following individuals and organizations for their support: the staff at the Luxor Hotel, the staff of Springer Nature, Pablo Rivas (Baylor University, Waco, Texas), and Ken Ferens (University of Manitoba, Canada).

We are pleased to present the proceedings of ICAI 2024. These proceedings represent a collection of outstanding research contributions that reflect the diversity and depth of work in Artificial Intelligence.

Hamid R. Arabnia
Leonidas Deligiannidis
Soheyla Amirian
Farzan Shenavarmasouleh
Farid Ghareh Mohammadi
David de la Fuente
Jose A. Olivas

Organization

Steering Committee – Co-chairs (CSCE 2024)

Hamid R. Arabnia	University of Georgia, USA
Leonidas Deligiannidis	Wentworth Institute of Technology, USA
Fernando G. Tinetti	Universidad Nacional de La Plata, Argentina
Quoc-Nam Tran	Southeastern Louisiana University, USA

Co-editors of ICAI 2024 Proceedings – Publication Co-chairs

Hamid R. Arabnia (Co-chair, CSCE 2024)	University of Georgia, USA
Leonidas Deligiannidis (Co-chair, CSCE 2024)	Wentworth Institute of Technology, USA
Soheyla Amirian (Session Co-chair)	Pace University, USA
Farzan Shenavarmasouleh (Session Co-chair)	Medialab Inc., USA
Farid Ghareh Mohammadi (Session Co-chair)	Mayo Clinic, USA
David de la Fuente (Session Co-chair)	University of Oviedo, Spain
Jose A. Olivas (Session Co-chair)	University of Castilla - La Mancha, Spain

Members of Steering Committee (CSCE 2024)

Babak Akhgar	Sheffield Hallam University, UK
Abbas M. Al-Bakry	University of IT & Communications, Iraq
Emeritus Nizar Al-Holou	University of Detroit Mercy, USA
Hamid R. Arabnia	University of Georgia, USA
Rajab Challoo	Texas A&M University-Kingsville, USA
Chien-Fu Cheng	Tamkang University, Taiwan
Hyunseung Choo	Sungkyunkwan University, South Korea
Kevin Daimi	University of Detroit Mercy, USA
Leonidas Deligiannidis	Wentworth Institute of Technology, USA
Eman M. El-Sheikh	University of West Florida, USA

Mary Mehrnoosh Eshaghian-Wilner	University of California Los Angeles, USA
David L. Foster	Kettering University, USA
Henry Hexmoor	Southern Illinois University at Carbondale, USA
Ching-Hsien (Robert) Hsu	Chung Hua University, Taiwan; and Tianjin University of Technology, China
James J. (Jong Hyuk) Park	SeoulTech, South Korea
Mohammad S. Obaidat	University of Jordan, Jordan
Marwan Omar	Illinois Institute of Technology, USA
Shahram Rahimi	Mississippi State University, USA
Gerald Schaefer	Loughborough University, UK
Fernando G. Tinetti	Universidad Nacional de La Plata, Argentina
Quoc-Nam Tran	Southeastern Louisiana University, USA
Shiuh-Jeng Wang	Central Police University, Taiwan
Layne T. Watson	Virginia Polytechnic Institute & State University, USA
Chao-Tung Yang	Tunghai University, Taiwan
Mary Yang	University of Arkansas, USA

Research Tracks – Co-chairs (CSCE 2024)

Abeer Alsadoon (Co-chair, Health Informatics)	Charles Sturt University, Australia
Soheyla Amirian (Co-chair, Computer Vision & AI)	Pace University, USA
Hamid R. Arabnia (Co-chair, HPC)	University of Georgia, USA
Kevin Daimi (Co-chair, Security)	University of Detroit Mercy, USA
Leonidas Deligiannidis (Co-chair, Imaging Science, AI)	Wentworth Institute of Technology, USA
Richard Dill (Co-chair, Military and Defense Modeling)	US Air Force Institute of Technology, USA
Ken Ferens (Co-chair, Cognitive Computing & AI)	University of Manitoba, Canada
David de la Fuente (Co-chair, Information Management)	University of Oviedo, Spain
Farid Ghareh Mohammadi (Co-chair, Computer Vision & AI)	Mayo Clinic, USA
Michael R. Grimaila (Co-chair, Military and Defense Modeling)	US Air Force Institute of Technology, USA

Douglas D. Hodson (Co-chair, Military and Defense Modeling)	US Air Force Institute of Technology, USA
Masahito Ohue (Co-chair, Mathematical Modeling)	Tokyo Institute of Technology, Japan
Jose A. Olivas (Co-chair, Information Management)	University of Castilla - La Mancha, Spain
Javier Ordus (Co-chair, Quantum Computing & AI)	Baylor University, USA
Pablo Rivas (Chair, Quantum Computing & AI)	Baylor University, USA
Farzan Shenavarmasouleh (Co-chair, Computer Vision & AI)	MediaLab Inc, USA
Robert Stahlbock (Co-chair, Data Mining)	Universität Hamburg, Germany
Masami Takata (Co-chair, Mathematical Modeling)	Nara Women's University, Japan
Quoc-Nam Tran (Co-chair, Education & Bioinformatics)	Southeastern Louisiana University, USA
Nobuaki Yasuo (Co-chair, Mathematical Modeling)	Tokyo Institute of Technology, Japan

Program Committee – Artificial Intelligence (ICAI 2024)

Abbas M. Al-Bakry	University of IT and Communications, Iraq
Emeritus Nizar Al-Holou	University of Detroit Mercy, USA
Soheyla Amirian (Co-Chair, Computer Vision & AI)	Pace University, USA
Emeritus Hamid R. Arabnia	University of Georgia, USA
Mehran Asadi	Lincoln University, USA
Alireza Bagheri Rajeoni	University of South Carolina, USA
Juan Jose Martinez Castillo	Universidad Nacional Abierta, Venezuela
Arianna D'Ulizia	Institute of Research on Population and Social Policies, National Research Council of Italy, Italy
Emeritus Kevin Daimi	University of Detroit Mercy, USA
Zhangisina Gulnur Davletzhanovna	Central Asian University, Kazakhstan; and International Academy of Informatization, Kazakhstan
Leonidas Deligiannidis	Wentworth Institute of Technology, USA
Roger Dziegiel	US Air Force Research Lab, USA

Mary Mehrnoosh Eshaghian-Wilner	University of Southern California, USA; and University of California Los Angeles, USA
Ken Ferens	University of Manitoba, Canada
David de la Fuente (Chapter Editor)	University of Oviedo, Spain
Farid Ghareh Mohammadi (Co-Chair, Computer Vision & AI)	Mayo Clinic, USA
George A. Gravvanis	Democritus University of Thrace, Greece
Michael R. Grimaila	US Air Force Institute of Technology, USA
Douglas D. Hodson	US Air Force Institute of Technology, USA
George Jandieri	Georgian Technical University, Georgia
Byung-Gyu Kim	Sun Moon University, South Korea
Tai-hoon Kim	University of Tasmania, Australia
Elena B. Kozerenko	Russian Academy of Sciences, Russia
Guoming Lai	Sun Yat-sen University, China
Peter M. LaMonica	US Air Force Research Lab, USA
Hyo Jong Lee	Chonbuk National University, South Korea
Changyu Liu	South China Agricultural University, China; and Carnegie Mellon University, USA
Muhammad Naufal Bin Mansor	Universiti Malaysia Perlis, Malaysia
Andrew Marsh	HoIP Telecom Ltd, UK
Mohamed Arezki Mellal	M'Hamed Bougara University of Boumerdès, Algeria
Ali Mostafaeipour	California State University, Fullerton, USA
Houssem Eddine Nouri	Institut Supérieur de Gestion de Tunis, University of Tunis, Tunisia
Robert Ehimen Okonigene	Ambrose Alli University, Nigeria
Jose A. Olivas (Chapter Editor)	University of Castilla - La Mancha, Spain
Marwan Omar	Illinois Institute of Technology, USA
Javier Ordus (Co-Chair, Quantum Computing & AI)	Baylor University, USA
James J. (Jong Hyuk) Park	SeoulTech, South Korea
Xuewei Qi	University of California, Riverside, USA
Charlie (Seungmin) Rho (Chapter Editor)	Sejong University, South Korea
Pablo Rivas (Co-Chair, Quantum Computing & AI)	Baylor University, USA
Abdel-Badeeh M. Salem	Ain Shams University, Egypt
Farzan Shenavarmasouleh (Co-Chair, Computer Vision & AI)	MediaLab Inc, USA
Ashu M. G. Solo (Publicity)	Maverick Technologies America Inc., USA

Tse Guan Tan	Universiti Malaysia Kelantan, Malaysia
Fernando G. Tinetti	Universidad Nacional de La Plata, Argentina
Hahanov Vladimir	Kharkiv National University of Radio Electronics, Ukraine
Shiuh-Jeng Wang	Central Police University, Taiwan
Todd Waskiewicz	US Air Force Research Lab, USA
Layne T. Watson	Virginia Polytechnic Institute & State University, USA
Xiaokun Yang	University of Houston - Clear Lake, USA
Jane You	Hong Kong Polytechnic University, China

Deep Convolutional Neural Networks, ANNs, and Applications

Taming Large Language Models for Healthcare – A Multi-layered System

Bharath Sudharsan(✉), Ryan Kosiba, Aishwarya Parthasarathi, and Rohan Paul Richard

AmalgamRx, Wilmington, DE, USA
bsudharsan@amalgamrx.com

Abstract. This paper presents a novel framework for implementing robust safety guardrails in conversational AI systems powered by large language models (LLMs) for healthcare applications. We propose a multi-layered approach that combines LLM-based classifiers, vector store matching, and dynamic prompt engineering to ensure safe and ethical interactions. Our system, designed to support patients with chronic conditions, demonstrates how LLMs can be effectively constrained to provide helpful information while avoiding potential risks associated with medical misinformation or inappropriate advice. We evaluate our framework using a comprehensive test set, demonstrating its efficacy in maintaining safety without significantly compromising the naturalness of conversations. Our findings contribute to the ongoing discourse on responsible AI deployment in sensitive domains like healthcare, particularly in creating systems that can build rapport and trust while adhering to strict ethical guidelines. These outcomes suggest that our multi-layered guardrail system offers a promising approach to harnessing the power of LLMs in healthcare while prioritizing patient safety and ethical considerations.

Keywords: Large Language Models · Healthcare AI · Safety Guardrails · Conversational AI · Ethical AI · Patient Support · Chronic Disease Management

1 Introduction

The advent of large language models (LLMs) has opened new possibilities for creating more natural and context-aware conversational AI systems in healthcare [1]. These models, trained on vast corpora of human-generated text, have demonstrated unprecedented capabilities in understanding and generating human-like text, including the ability to convey empathy, build rapport, and engage in nuanced communication [2]. The potential applications of LLMs in healthcare are vast, ranging from patient education and support to assisting healthcare professionals in clinical decision-making and administrative tasks.

However, the deployment of such powerful models in medical contexts poses significant risks, including the potential for generating misinformation, providing inappropriate medical advice, or violating patient privacy [3]. These risks are particularly acute in healthcare, where the consequences of misinformation or inappropriate advice can

H. R. Arabnia et al. (Eds.): CSCE 2024, CCIS 2252, pp. 191–201, 2025.
https://doi.org/10.1007/978-3-031-86623-4_15

have serious implications for patient health and well-being. As such, the development of robust safety mechanisms is paramount to the responsible deployment of LLM-powered systems in healthcare settings.

1.1 Evolution of Conversational AI in Healthcare

The evolution of human-AI chat experiences in healthcare has been marked by significant advancements over the past decades. Early rule-based systems like ELIZA [4] gave way to more sophisticated statistical models, and now to neural network-based approaches that can engage in more natural, context-aware conversations [5]. These early systems, while groundbreaking, were limited in their ability to understand context, generate natural language, and provide truly personalized responses.

The introduction of machine learning techniques, particularly deep learning, marked a significant leap forward in the capabilities of conversational AI systems. Models like recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) improved the ability of AI systems to maintain context over longer conversations and generate more coherent responses. However, these models still struggled with long-term dependencies and often produced repetitive or inconsistent outputs in extended dialogues.

LLMs represent the latest leap in this evolution, offering the potential for AI assistants that can understand and respond to the nuanced emotional and informational needs of patients and caregivers [6]. Unlike their predecessors, LLMs can generate fluent, contextually appropriate responses across a wide range of topics, making them particularly well-suited for the complex and varied nature of healthcare conversations.

1.2 Challenges and Risks of LLMs in Healthcare

However, the very capabilities that make LLMs so promising also present unique challenges in healthcare contexts. Their ability to generate fluent, human-like text raises concerns about users potentially forming inappropriate emotional attachments or mistaking AI-generated advice for professional medical opinion [7]. This risk is particularly pronounced in healthcare, where patients may be vulnerable and seek authoritative guidance.

Additionally, the black-box nature of these models makes it difficult to guarantee their outputs will always align with medical best practices and ethical guidelines [8]. LLMs can sometimes generate plausible-sounding but incorrect or harmful information, a phenomenon known as "hallucination." In a healthcare context, such hallucinations could lead to dangerous misinformation or inappropriate medical advice.

Furthermore, LLMs trained on large, diverse datasets may inadvertently perpetuate biases present in their training data. This could lead to disparities in the quality of information or support provided to different demographic groups, potentially exacerbating existing health inequities.

1.3 The Need for Robust Safety Guardrails

Given these challenges, there is a clear need for robust safety mechanisms to govern the deployment of LLMs in healthcare settings. These mechanisms must be capable of:

1. Ensuring the accuracy and reliability of medical information provided by the system
2. Preventing the generation of inappropriate or potentially harmful advice
3. Maintaining clear boundaries between AI support and professional medical care
4. Protecting patient privacy and confidentiality
5. Mitigating potential biases in the system's responses
6. Providing transparency about the AI nature of the system to prevent misunderstandings or inappropriate attachments

1.4 Our Contribution

In this paper, we present a novel multi-layered guardrail system designed specifically for healthcare conversational AI. Our approach integrates multiple safety mechanisms at different stages of the conversation pipeline, allowing for fine-grained control over the LLM's outputs while preserving its ability to engage in natural, context-aware dialogue. We focus on creating a system that can exhibit empathy, build trust, and provide emotional support, all while strictly adhering to ethical guidelines and maintaining clear boundaries about its AI nature.

Our system incorporates several innovative features:

7. A multi-layered architecture with multiple checkpoints for ensuring safe and appropriate interactions
8. Dynamic prompt engineering that adapts to the specific context and intent of each user input
9. Integration of external, up-to-date medical knowledge to supplement the LLM's training
10. Advanced filtering and confidence scoring mechanisms to catch potential safety violations
11. A comprehensive evaluation framework that assesses both the safety and the conversational quality of the system

By addressing these challenges, our work contributes to the ongoing discourse on responsible AI deployment in sensitive domains like healthcare. We aim to demonstrate that with careful design and implementation, LLMs can be effectively constrained to provide valuable support to patients while maintaining high standards of safety and ethical conduct.

The rest of this paper is organized as follows: Sect. 2 provides a comprehensive review of related work. Section 3 details our system architecture. Section 4 elaborates on the implementation of our guardrail system. Section 5 describes our experimental evaluation. Section 6 discusses the results and potential impact of our system. Finally, Sect. 7 concludes the paper and suggests directions for future research.

2 Related Work

2.1 Evolution of Conversational AI in Healthcare

The journey of conversational AI in healthcare began with simple rule-based systems like ELIZA [1], developed by Joseph Weizenbaum in the 1960s. ELIZA used pattern matching to simulate a psychotherapist, marking a significant milestone in human-computer

interaction by demonstrating that even simple algorithms could create the illusion of understanding in specific contexts. However, these systems lacked true understanding, which limited their practical application in healthcare.

The next generation of chatbots, such as ALICE (Artificial Linguistic Internet Computer Entity) [9], introduced more sophisticated pattern matching and knowledge bases. Developed by Richard Wallace in the 1990s, ALICE used AIML (Artificial Intelligence Markup Language) to enable more varied and contextually appropriate responses. This approach allowed for more dynamic conversations but still struggled with truly understanding user intent, as it relied heavily on pre-programmed responses.

The advent of machine learning, particularly deep learning, marked a significant leap forward in healthcare AI capabilities. Systems like IBM Watson [10] showcased AI's potential to process vast amounts of medical literature and assist in clinical decision-making. While these systems demonstrated the ability to analyze natural language and provide evidence-based recommendations, they still faced challenges in natural conversation and emotional understanding, which are crucial for patient-facing applications.

Recent advancements with neural conversation models [11], based on sequence-to-sequence learning and transformer networks, have improved the fluency and coherence of AI-generated responses, yet they continue to struggle with consistency in long conversations, task-oriented dialogues, and ensuring the factual accuracy and safety of their outputs.

2.2 Large Language Models and Their Impact

The introduction of large language models like GPT-3 (Generative Pre-trained Transformer 3) [2] has revolutionized the field of conversational AI. These models, trained on enormous datasets of human-generated text, exhibit unprecedented capabilities in natural language understanding and generation. They can engage in open-ended conversations, demonstrate contextual understanding, and even show signs of emergent reasoning [12].

The scale and architecture of these models allow them to capture complex patterns in language use, resulting in more coherent and contextually appropriate responses compared to their predecessors. This capability is particularly valuable in healthcare contexts, where conversations often involve complex medical concepts and require sensitivity to patient emotions and concerns.

In healthcare contexts, LLMs have shown promise in several areas:

1. Generating empathetic responses [13]: LLMs can craft responses that acknowledge and validate patient emotions, potentially improving the supportive capacity of AI systems.
2. Explaining complex medical concepts in lay terms [14]: The models' ability to rephrase and simplify information can aid in patient education and improve health literacy.
3. Assisting in mental health support [15]: Early applications have shown potential in providing initial mental health screening and support, although with careful limitations to ensure appropriate escalation to human professionals.
4. Aiding in clinical documentation [16]: LLMs can help summarize and structure clinical notes, potentially reducing administrative burden on healthcare providers.

However, the deployment of LLMs in healthcare also raises significant concerns. LLMs pose several risks in the medical field, including generating plausible but incorrect information (hallucination), perpetuating biases from their training data, and raising privacy concerns by potentially reproducing sensitive information. They may lack grounding in current medical knowledge and best practices, which could lead to the dissemination of outdated or inaccurate advice. Without proper constraints, these models might also generate harmful advice that could be mistakenly interpreted as legitimate medical guidance. These challenges underscore the need for robust safety measures and ethical guidelines in the deployment of LLMs in healthcare settings.

2.3 Safety and Ethical Considerations in Healthcare AI

As AI systems become more sophisticated and their use in healthcare more widespread, ensuring their safe and ethical use has become a critical area of research and policy development.

Bickmore et al. [14] outlined key patient and consumer safety considerations for health chatbots, emphasizing several crucial points:

1. Clear disclosure of AI identity: Users should always be aware that they are interacting with an AI system, not a human healthcare provider.
2. Mechanisms for escalation to human professionals: AI systems should have clear pathways for directing users to human healthcare providers when appropriate.
3. Robust privacy protections: Given the sensitive nature of health information, AI systems must adhere to strict data protection standards.
4. Transparency about capabilities and limitations: Users should be informed about what the AI system can and cannot do to manage expectations and prevent misuse.
5. Regular updates and monitoring: Healthcare AI systems should be continuously monitored and updated to reflect current medical knowledge and best practices.

Char et al. [17] discussed the ethical implications of AI in healthcare, highlighting several key issues:

1. Transparency and explainability: The "black box" nature of many AI systems, particularly deep learning models, raises concerns about accountability and trust.
2. Accountability: Determining responsibility in cases where AI systems contribute to medical errors or adverse outcomes is a complex challenge.
3. Potential for exacerbating health disparities: If not carefully designed and implemented, AI systems could worsen existing inequities in healthcare access and outcomes.
4. Informed consent: The use of AI in healthcare raises new questions about what patients need to know and understand to give truly informed consent.
5. Impact on the patient-provider relationship: There are concerns that increased reliance on AI could depersonalize healthcare and erode the crucial bond between patients and their healthcare providers.
6. Data privacy and security: The large-scale data collection and analysis required for many AI systems raises significant privacy concerns.

These ethical considerations emphasize the need for careful and thoughtful development, deployment, and regulation of AI systems in healthcare settings. They also highlight the importance of interdisciplinary collaboration between AI researchers, healthcare professionals, ethicists, and policymakers to ensure that AI technologies are developed and used in ways that benefit patients and society as a whole.

2.4 Approaches to Constraining LLMs

Various approaches have been proposed to make LLMs safer and more reliable, especially in sensitive domains like healthcare. These methods aim to address the challenges of maintaining accuracy, ensuring safety, and preserving ethical standards while leveraging the powerful capabilities of LLMs.

Fine-Tuning on Domain-Specific Datasets [18]
This approach involves further training of pre-trained LLMs on carefully curated, domain-specific datasets. In healthcare, this might include medical textbooks, clinical guidelines, and anonymized patient-doctor conversations. Fine-tuning can improve model performance on targeted tasks and reduce the likelihood of generating inappropriate content. Using domain-specific models enhances knowledge and adherence to terminology and best practices in fields like healthcare. However, they require large, high-quality datasets, which are challenging to obtain due to privacy concerns, and may sacrifice flexibility and generalization, while not addressing hallucination.

Prompt Engineering Techniques [19]
This method involves carefully crafting input prompts to guide the LLM's behavior and outputs. In healthcare applications, prompts might include specific instructions about maintaining patient confidentiality, avoiding diagnosis, or providing emotional support. Implementing dynamic, context-specific control of model outputs without modifying the underlying model offers flexibility and easy updates. However, the effects can be inconsistent, particularly with complex instructions, and may not generalize well.

Rule-Based Filtering and Content Moderation [20]
This approach involves applying predefined rules to filter or modify the LLM's outputs. In healthcare, this might include blocking specific terms related to diagnosis or treatment recommendations, or flagging responses that mention certain medications or procedures. Providing a clear, interpretable system for controlling outputs allows for easy auditing and modification. However, this approach can be overly rigid with context-dependent cases, computationally expensive with many complex rules, and may reduce the naturalness of conversations if applied too aggressively.

Use of Smaller, Task-Specific Models for Control [21]
This method involves using additional, smaller models to control various aspects of the LLM's output. For example, a classification model might be used to detect the intent of user queries, while another model might assess the safety of the LLM's proposed

responses. Fine-grained control over model behaviors and the ability to leverage specialized models for specific tasks or datasets are key advantages. However, this approach increases system complexity, computational requirements and response latency.

Retrieval-Augmented Generation
This approach involves combining LLMs with information retrieval systems. The LLM generates responses based not only on its pre-trained knowledge but also on relevant information retrieved from a curated knowledge base. The approach offers up-to-date, accurate information, improved traceability, and reduced hallucination by grounding responses in verified sources. However, it requires maintaining a comprehensive knowledge base, may struggle with synthesizing multiple sources, and can increase response latency.

Reinforcement Learning from Human Feedback (RLHF)
This method involves fine-tuning LLMs using reinforcement learning, where the reward signal is derived from human feedback. This can help align the model's outputs with human preferences and values. Employing this technique improves adherence to guidelines and supports ongoing updates based on real use. However, it requires significant human feedback, can introduce new biases, and may struggle to generalize across diverse use cases.

Our work builds on these approaches while introducing a novel multi-layered architecture specifically designed for healthcare applications. We aim to create a system that can maintain the engaging, empathetic qualities of LLM-based conversation while ensuring strict adherence to medical best practices and ethical guidelines. By combining multiple safety mechanisms, including dynamic prompt engineering, content filtering, and integration with external knowledge bases, we seek to address the limitations of individual approaches and provide a more robust, flexible system for safe deployment of LLMs in healthcare contexts.

3 System Architecture

Our proposed system consists of five main components, each designed to contribute to safe, effective, and engaging conversations in healthcare contexts. The multi-layered architecture allows for multiple conversational checkpoints, ensuring a high level of safety while maintaining natural and helpful interactions.

3.1 Large Language Model Core

At the heart of our system is a large language model specifically developed for healthcare applications. We start with a pre-trained model based on the Transformer architecture, similar to GPT-4 [2], and fine-tune it on a carefully curated dataset consisting of:

- Patient education materials from leading health organizations
- Anonymized and ethically sourced conversations
- Guidelines and best practices for patient communication from medical associations

The fine-tuning process aims to imbue the model with domain-specific knowledge while also aligning its outputs with established medical communication practices. We use a combination of supervised fine-tuning and reinforcement learning from human feedback (RLHF) to optimize the model's performance.

3.2 Multi-layer Guardrail System

This is the core of our safety framework, consisting of three sub-components that work in concert to ensure safe and appropriate interactions:

Pre-processing Layer
This layer focuses on input sanitization and initial safety checks. It serves as the first line of defense against potential safety violations.

The key functions of this layer include detecting and redacting Personally Identifiable Information (PII), identifying harmful content such as self-harm, threats and abusive language, classifying intent to guide further processing, and recognizing extraneous topics for improved focus and relevance. To do so we leverage a combination of smaller task-specific models as well as vector store matching.

LLM Interaction Layer
This layer manages the direct interaction with the LLM via the Prompt Engineering Module, including dynamic prompt adjustment and multi-turn conversation management.

It involves real-time prompt adjustment based on detected intents and conversation context, enforcing boundaries to ensure the AI doesn't substitute professional medical advice, and managing conversation flow to maintain coherence and safety across multiple interactions.

Post-processing Layer
This layer provides a final safety check on LLM-generated content before it reaches the user. It uses content filtering to catch potential safety violations missed by earlier layers, fact-checking against a verified medical knowledge base, and applying confidence scoring to ensure responses meet the required safety threshold.

Each layer incorporates multiple mechanisms to ensure safe and appropriate interactions, which will be detailed in the next section.

4 Methodology

Our evaluation consists of four main components:

Safety Violation Detection
The primary objective here is to assess the system's ability to prevent potential safety violations. This will involve creating a challenging test set of user inputs specifically designed to test the limits of the system's safety constraints. A team of healthcare professionals and AI ethics experts will then manually review the system's outputs. The

evaluation will focus on the rate of potential safety violations, classify these violations by type and assess how effective different guardrail layers are in catching these violations.

Information Accuracy
The goal is to evaluate the accuracy of the medical information provided by the system. A random sampling of the system's responses will be fact-checked against current medical literature and guidelines by a team of medical professionals from various specialties. The evaluation metrics will include the accuracy rate of the provided medical information, the types and severity of inaccuracies, and the effectiveness of the knowledge integration layer in ensuring the system provides up-to-date information.

Conversation Naturalness and Empathy
We assess the quality of the system's conversations in terms of naturalness and empathy. A blind comparison study will be conducted involving healthcare professionals who will evaluate the system's conversations against those of other commercial health chatbots and transcripts of human healthcare providers. The study will use standardized scenarios to ensure consistency across evaluations, and the conversations will be rated on a 5-point Likert scale based on attributes like naturalness, empathy, clarity, and appropriateness.

User Experience Study
The focus here is to gather feedback from actual users on the system's perceived safety, helpfulness, and ease of use. This will involve a pilot study with patients who have various chronic conditions, where they will interact with the system. Feedback will be collected through surveys and semi-structured interviews, with evaluation metrics such as user satisfaction ratings, perceived helpfulness in managing chronic conditions, comfort level in discussing health concerns with the AI system, and qualitative feedback on the system's strengths and areas for improvement.

4.1 Ethical Considerations

Our experimental design prioritizes user safety and privacy. We ensure that all participants are fully informed about the nature of the study and the AI system they are interacting with. They are clearly instructed on the limitations of the AI system, and the importance of consulting healthcare professionals. Strict protocols have been put in place to protect user data and ensure anonymity. We also endeavor to have a diverse representation in our user studies to assess the system's performance across different demographics.

4.2 Limitations

We acknowledge several limitations in our experimental design. The user experience study is limited to a four-week period, which may not capture long-term effects or rare edge cases. Despite efforts to ensure diversity, our user group may not be fully representative of all potential users. Lastly, user engagement and satisfaction may be influenced by the novelty of interacting with an AI health assistant.

5 Conclusion

The multi-layered guardrail system presented in this paper represents a novel approach to addressing the challenges of deploying LLMs in healthcare contexts. As AI continues to advance, frameworks like ours will be crucial in ensuring that these powerful technologies are deployed responsibly in sensitive domains like healthcare.

The journey towards safe and effective AI-assisted healthcare support is ongoing, requiring continued collaboration between AI researchers, healthcare professionals, ethicists, and patients. By pursuing this research agenda, we can work towards a future where AI serves as a valuable tool in healthcare, augmenting and supporting human care to ultimately contribute to improved patient outcomes and experiences.

Disclosure of Interests. Authors are employees of AmalgamRx.

References

1. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthc. J. **6**(2), 94–98 (2019)
2. Brown, T.B., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33, pp. 1877–1901. Curran Associates, Inc. (2020)
3. Chowdhury, G.G.: Natural language processing. Ann. Rev. Inf. Sci. Technol. **37**(1), 51–89 (2003)
4. Weizenbaum, J.: ELIZA – a computer program for the study of natural language communication between man and machine. Commun. ACM **9**(1), 36–45 (1966)
5. Vinyals, O., Le, Q.: A Neural Conversational Model (2015). arXiv:1506.05869
6. Xu, L., et al.: Towards a medical chatbot for mental health support: evaluation and ethical considerations. J. Med. Internet Res. **23**(5), e27850 (2021)
7. Inkster, B., Sarda, S., Subramanian, V.: An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth **6**(11), e12106 (2018)
8. Morley, J., et al.: The ethics of AI in health care: a mapping review. Soc. Sci. Med. **260**, 113172 (2020)
9. Wallace, R.S.: The anatomy of A.L.I.C.E. In: Epstein, R., Roberts, G., Beber, G. (eds.) Parsing the Turing Test, pp. 181–210. Springer, Dordrecht (2009)
10. Ferrucci, D., et al.: Building Watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
11. Serban, I.V., et al.: A survey of available corpora for building data-driven dialogue systems: the journal version. Dialogue Discourse **9**(1), 1–49 (2018)
12. Wei, J., et al.: Emergent Abilities of Large Language Models (2022). arXiv:2206.07682
13. Liu, B., Sundar, S.S.: Should machines express sympathy and empathy? Experiments with a health advice Chatbot. Cyberpsychol. Behav. Soc. Netw. **21**(10), 625–636 (2018)
14. Bickmore, T.W., et al.: Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. J. Med. Internet Res. **20**(9), e11510 (2018)
15. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment. Health **4**(2), e19 (2017)

16. Patel, B.N., et al.: Human–machine partnership with artificial intelligence for chest radiograph diagnosis. npj Digit. Med. **2**, 111 (2019)
17. Char, D.S., Shah, N.H., Magnus, D.: Implementing machine learning in health care – addressing ethical challenges. N. Engl. J. Med. **378**(11), 981–983 (2018)
18. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
19. Liu, P., et al.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
20. Dinan, E., et al.: Safety for Conversational AI: Challenges, Opportunities, and Recommendations (2021). arXiv:2106.06635
21. Ouyang, L., et al.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems 35, pp. 27730–27744. Curran Associates, Inc. (2022)