> IS 2024

Informacijska družba

Zbornik 27. mednarodne multikonference

Zvezek K

Information Society

Proceedings of the 27th International Multiconference

Volume K

ChatGPT v zdravstvu

ChatGPT in Medicine

Uredniki > Editors:

Matjaž Gams, Monika Simjanoska Misheva, Stevo Lukić, Franz Wotawa

9. oktober 2024 > Ljubljana, Slovenija / 9 October 2024 > Ljubljana, Slovenia

Zbornik 27. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2024

Zvezek K

Proceedings of the 27th International Multiconference

INFORMATION SOCIETY – IS 2024

Volume K

ChatGPT v zdravstvu
ChatGPT in Medicine

Uredniki / Editors

Matjaž Gams, Monika Simjanoska Misheva, Stevo Lukić, Franz Wotawa

http://is.ijs.si

9. oktober 2024 / 9 October 2024
Ljubljana, Slovenia

Uredniki:

Matjaž Gams
Odsek za inteligentne sisteme, Institut »Jožef Stefan«, Ljubljana, Slovenija

Monika Simjanoska Misheva
Fakulteta za računalništvo in elektrotehniko, Skopje, Makedonija

Stevo Lukić
Nevrološka klinika Niš, Srbija

Franz Wotawa
Inštitut za tehnologijo programske opreme, Gradec, Avstrija

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Lana Zemljak
Oblikovanje naslovnice: Vesna Lasič

Dostop do e-publikacije:
http://library.ijs.si/Stacks/Proceedings/InformationSociety

Ljubljana, oktober 2024

Informacijska družba
ISSN 2630-371X

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 214534915
ISBN 978-961-264-309-6 (PDF)

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2024

Leto 2024 je hkrati udarno in tradicionalno. Že sedaj, še bolj pa v prihodnosti bosta računalništvo, informatika (RI) in umetna inteligenca (UI) igrali ključno vlogo pri oblikovanju napredne in trajnostne družbe. Smo na pragu nove dobe, v kateri generativna umetna inteligenca, kot je ChatGPT, in drugi inovativni pristopi utirajo pot k superinteligenci in singularnosti, ključnim elementom, ki bodo definirali razcvet človeške civilizacije. Naša konferenca je zato hkrati tradicionalna znanstvena, pa tudi povsem akademsko odprta za nove pogumne ideje, inkubator novih pogledov in idej.

Letošnja konferenca ne le da analizira področja RI, temveč prinaša tudi osrednje razprave o perečih temah današnjega časa – ohranjanje okolja, demografski izzivi, zdravstvo in preobrazba družbenih struktur. Razvoj UI ponuja rešitve za skoraj vse izzive, s katerimi se soočamo, kar poudarja pomen sodelovanja med strokovnjaki, raziskovalci in odločevalci, da bi skupaj oblikovali strategije za prihodnost. Zavedamo se, da živimo v času velikih sprememb, kjer je ključno, da s poglobljenim znanjem in inovativnimi pristopi oblikujemo informacijsko družbo, ki bo varna, vključujoča in trajnostna.

Letos smo ponosni, da smo v okviru multikonference združili dvanajst izjemnih konferenc, ki odražajo širino in globino informacijskih ved: CHATMED v zdravstvu, Demografske in družinske analize, Digitalna preobrazba zdravstvene nege, Digitalna vključenost v informacijski družbi – DIGIN 2024, Kognitivna znanost, Konferenca o zdravi dolgoživosti, Legende računalništva in informatike, Mednarodna konferenca o prenosu tehnologij, Miti in resnice o varovanju okolja, Odkrivanje znanja in podatkovna skladišča – SIKDD 2024, Slovenska konferenca o umetni inteligenci, Vzgoja in izobraževanje v RI.

Poleg referatov bodo razprave na okroglih mizah in delavnicah omogočile poglobljeno izmenjavo mnenj, ki bo oblikovala prihodnjo informacijsko družbo. “Legende računalništva in informatike” predstavljajo slovenski “Hall of Fame” za odlične posameznike s tega področja, razširjeni referati, objavljeni v reviji *Informatica* z 48-letno tradicijo odličnosti, in sodelovanje s številnimi akademskimi institucijami in združenji, kot so ACM Slovenija, SLAIS in Inženirska akademija Slovenije, bodo še naprej spodbujali razvoj informacijske družbe. Skupaj bomo gradili temelje za prihodnost, ki bo oblikovana s tehnologijami, osredotočena na človeka in njegove potrebe.

S podelitvijo nagrad, še posebej z nagrado Michie-Turing, se avtonomna RI stroka vsakoletno opredeli do najbolj izstopajočih dosežkov. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejel prof. dr. Borut Žalik. Priznanje za dosežek leta pripada prof. dr. Sašu Džeroskemu za izjemne raziskovalne dosežke. »Informacijsko limono« za najmanj primerno informacijsko tematiko je prejela nabava in razdeljevanjem osebnih računalnikov ministrstva, »informacijsko jagodo« kot najboljšo potezo pa so sprejeli organizatorji tekmovanja ACM Slovenija. Čestitke nagrajencem!

Naša vizija je jasna: prepoznati, izkoristiti in oblikovati priložnosti, ki jih prinaša digitalna preobrazba, ter ustvariti informacijsko družbo, ki bo koristila vsem njenim članom. Vsem sodelujočim se zahvaljujemo za njihov prispevek k tej viziji in se veselimo prihodnjih dosežkov, ki jih bo oblikovala ta konferenca.

Mojca Ciglarič, predsednica programskega odbora

Matjaž Gams, predsednik organizacijskega odbora

PREFACE TO THE MULTICONFERENCE INFORMATION SOCIETY 2024

The year 2024 is both ground-breaking and traditional. Now, and even more so in the future, computer science, informatics (CS/I), and artificial intelligence (AI) will play a crucial role in shaping an advanced and sustainable society. We are on the brink of a new era where generative artificial intelligence, such as ChatGPT, and other innovative approaches are paving the way for superintelligence and singularity—key elements that will define the flourishing of human civilization. Our conference is therefore both a traditional scientific gathering and an academically open incubator for bold new ideas and perspectives.

This year's conference analyzes key CS/I areas and brings forward central discussions on pressing contemporary issues—environmental preservation, demographic challenges, healthcare, and the transformation of social structures. AI development offers solutions to nearly all challenges we face, emphasizing the importance of collaboration between experts, researchers, and policymakers to shape future strategies collectively. We recognize that we live in times of significant change, where it is crucial to build an information society that is safe, inclusive, and sustainable, through deep knowledge and innovative approaches.

This year, we are proud to have brought together twelve exceptional conferences within the multiconference framework, reflecting the breadth and depth of information sciences:

- CHATMED in Healthcare
- Demographic and Family Analyses
- Digital Transformation of Healthcare Nursing
- Digital Inclusion in the Information Society – DIGIN 2024
- Cognitive Science
- Conference on Healthy Longevity
- Legends of Computer Science and Informatics
- International Conference on Technology Transfer
- Myths and Facts on Environmental Protection
- Data Mining and Data Warehouses – SIKDD 2024
- Slovenian Conference on Artificial Intelligence
- Education and Training in CS/IS.

In addition to papers, roundtable discussions and workshops will facilitate in-depth exchanges that will help shape the future information society. The “Legends of Computer Science and Informatics” represents Slovenia’s “Hall of Fame” for outstanding individuals in this field. At the same time, extended papers published in the Informatica journal, with over 48 years of excellence, and collaboration with numerous academic institutions and associations, such as ACM Slovenia, SLAIS, and the Slovenian Academy of Engineering, will continue to foster the development of the information society. Together, we will build the foundation for a future shaped by technology, yet focused on human needs.

The autonomous CS/IS community annually recognizes the most outstanding achievements through the awards ceremony. The Michie-Turing Award for an exceptional lifetime contribution to the development and promotion of the information society was awarded to Prof. Dr. Borut Žalik. The Achievement of the Year Award goes to Prof. Dr. Sašo Džeroski. The "Information Lemon" for the least appropriate information topic was given to the ministry's procurement and distribution of personal computers. At the same time, the "Information Strawberry" for the best initiative was awarded to the organizers of the ACM Slovenia competition. Congratulations to all the award winners!

Our vision is clear: to recognize, seize, and shape the opportunities brought by digital transformation and create an information society that benefits all its members. We thank all participants for their contributions and look forward to this conference's future achievements.

Mojca Ciglarič, Chair of the Program Committee

Matjaž Gams, Chair of the Organizing Committee

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia
Sergio Campos-Cordobes, Spain
Shabnam Farahmand, Finland
Sergio Crovella, Italy

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Mitja Lasič
Blaž Mahnič

Programme Committee

Mojca Ciglarič, chair
Bojan Orel
Franc Solina
Viljan Mahnič
Cene Bavec
Tomaž Kalin
Jozsef Györkös
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Mitja Luštrek
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah
Niko Zimic
Rok Piltaver
Toma Strle
Tine Kolenik
Franci Pivec
Uroš Rajkovič
Borut Batagelj
Tomaž Ogrin
Aleš Ude
Bojan Blažica
Matjaž Kljun
Robert Blatnik
Erik Dovgan
Špela Stres
Anton Gradišek

KAZALO / TABLE OF CONTENTS

Zbornik 27. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2024

Zvezek K

Proceedings of the 27th International Multiconference

INFORMATION SOCIETY – IS 2024

Volume K

ChatGPT v zdravstvu
ChatGPT in Medicine

Uredniki / Editors

Matjaž Gams, Monika Simjanoska Misheva, Stevo Lukić, Franz Wotawa

http://is.ijs.si

9. oktober 2024 / 9 October 2024
Ljubljana, Slovenia

PREDGOVOR

Umetna inteligenca, zlasti generativna umetna inteligenca, kot je ChatGPT, je spremenila pravila igre v številnih panogah. Vendar je njen vpliv v medicini še posebej pomemben, saj gre v zdravstvu za več kot le informacije - gre za človeška življenja. Vključitev umetne inteligence lahko bistveno izboljša izide zdravljenja bolnikov, racionalizira delovne procese in izboljša procese odločanja na celotnem področju medicine.

Vloga umetne inteligence presega pomoč strokovnjakom; neposredno vpliva na oskrbo bolnikov. Z virtualnimi posvetovanji, izobraževanjem bolnikov in preverjanjem simptomov lahko ChatGPT izboljša dostop do zdravstvenega varstva za skupine prebivalstva, ki bi se sicer soočale z ovirami zaradi lokacije ali časovnih omejitev. Poleg tega umetna inteligenca z avtomatizacijo rutinskih opravil zmanjšuje upravno breme zdravnikov, tako da lahko več časa posvetijo oskrbi bolnikov in manj papirologiji. To je lahko ključni dejavnik v boju proti izgorelosti zdravnikov, ki je v sodobnem zdravstvu vse večji problem. Čeprav je potencial ChatGPT v zdravstvu vznemirljiv, moramo obravnavati etične in varnostne izzive, ki jih prinaša. Ključna vprašanja so zagotavljanje zasebnosti pacientov, zmanjšanje pristranskosti algoritmov umetne inteligence in ohranjanje natančnosti zdravniških nasvetov. UI mora podpirati - ne pa nadomestiti - človeško presojo, zlasti pri kritičnih medicinskih odločitvah. Zagotavljanje preglednosti, odgovornosti in zasnove teh sistemov s pristopom, pri katerem je v ospredju varnost, je bistvenega pomena za krepitev zaupanja v te tehnologije.

V prihodnosti se bo vloga ChatGPT in podobnih sistemov umetne inteligence verjetno še povečala. Pravkar začenjamo raziskovati njihove aplikacije v personalizirani medicini, kjer bi umetna inteligenca lahko pomagala prilagoditi zdravljenje posameznim pacientom na podlagi genetskih podatkov, podatkov o okolju in življenjskem slogu. Poleg tega lahko umetna inteligenca z analizo trendov v populacijah prispeva k zgodnjemu odkrivanju bolezni, odkrivanju zdravil in globalnim zdravstvenim pobudam.

Ta konferenca je priložnost, da se poglobimo v najsodobnejše raziskave, nastajajoče aplikacije in etične vidike uporabe ChatGPT v medicinski praksi. Skupaj bomo raziskali sedanje zmogljivosti in prihodnje možnosti, hkrati pa se bomo posvetili izzivom, ki jih prinaša vključevanje umetne inteligence na tako občutljivo in tvegano področje.

Matjaž Gams
Monika Simjanoska Misheva
Stevo Lukić
Franz Wotawa
predsedniki konference ChatGPT v zdravstvu

FOREWORD

AI, and particularly conversational AI like ChatGPT, has been a game-changer in many industries. However, its impact in medicine is particularly significant because healthcare is about more than just information—it's about human lives. The integration of AI has the potential to dramatically improve patient outcomes, streamline workflows, and enhance decision-making processes across the medical field.

The role of AI extends beyond assisting professionals; it directly impacts patient care. Through virtual consultations, patient education, and symptom checks, ChatGPT can enhance access to healthcare for populations who might otherwise face barriers due to location or time constraints. Additionally, by automating routine tasks, AI reduces the administrative burden on clinicians, allowing them to spend more time on patient care and less on paperwork. This can be a crucial factor in combating physician burnout, a growing issue in modern healthcare. While the potential of ChatGPT in healthcare is exciting, we must address the ethical and safety challenges that come with it. Ensuring patient privacy, minimizing bias in AI algorithms, and maintaining the accuracy of medical advice are key concerns. AI should support—not replace—human judgment, particularly in critical medical decisions. Ensuring that these systems are transparent, accountable, and designed with a safety-first approach is essential to building trust in these technologies.

Looking forward, the role of ChatGPT and similar AI systems will likely expand. We are just beginning to explore its applications in personalized medicine, where AI could help tailor treatments to individual patients based on genetic, environmental, and lifestyle data. Additionally, AI can contribute to early detection of diseases, drug discovery, and global health initiatives by analyzing trends across populations.

This conference is an opportunity for us to delve into the cutting-edge research, emerging applications, and ethical considerations surrounding the use of ChatGPT in medical practice. Together, we will explore both the current capabilities and the future possibilities, while also addressing the challenges that come with integrating AI into such a sensitive and high-stakes field.

Matjaž Gams
Monika Simjanoska Misheva
Stevo Lukić
Franz Wotawa
ChatGPT in Medicine chairs

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Matjaž Gams

Monika Simjanoska Misheva

Stevo Lukić

Franz Wotawa

Žiga Kolar

Automatic Reviewing of Conference Papers in Healthcare and Other Sciences Using ChatGPT

Matjaž Gams
Department of Intelligent Systems
Jozef Stefan Institute
Ljubljana Slovenia
matjaz.gams@ijs.si

Luka Bolješić
University of Ljubljana
Faculty of Computer and Information Science
Ljubljana Slovenia
luka.boljesic0@gmail.com

Alex Korlat
University of Ljubljana
Faculty of Computer and Information Science
Ljubljana Slovenia
alex.korlat.ak@gmail.com

Primož Kocuvan
Department of Intelligent Systems
Jozef Stefan Institute
Ljubljana Slovenia
primoz.kocuvan@ijs.si

ABSTRACT

The emergence of artificial intelligence (AI) has significantly impacted various fields, including the peer review process in scientific research. ChatGPT, a large language model developed by OpenAI, has shown potential in automating and enhancing the review of conference healthcare papers. Our HealthReview AI tool can process and analyze large volumes of text rapidly, providing feedback and insights that streamline the peer review process, reduce human workload, and increase efficiency. This paper presents a web application developed using the Flask framework that enables users to upload PDF files containing research papers and utilizes ChatGPT to generate reviews for each paper. The methodology, results and potential implications of this application are discussed, highlighting both the advantages and the challenges of integrating AI into the academic review process.[1]

KEYWORDS

Artificial Intelligence, ChatGPT, Peer Review, Healthcare, Scientific Research, Flask Framework, PDF Processing, Academic Writing, Conference Papers

POVZETEK

Pojav umetne inteligence (UI) je pomembno vplival na različna področja, vključno s postopkom strokovnega pregleda v akademskih in znanstvenih raziskavah. ChatGPT, velik jezikovni model, ki ga je razvil OpenAI, je pokazal potencial za avtomatizacijo in izboljšanje pregleda medicinskih konferenčnih prispevkov. To orodje UI lahko hitro obdela in analizira velike količine besedil ter zagotovi povratne informacije in vpoglede, ki poenostavijo postopek strokovnega pregleda, zmanjšajo delovno obremenitev in povečajo učinkovitost. Ta članek predstavlja spletno aplikacijo HealthReview, razvito s pomočjo ogrodja Flask, ki uporabnikom omogoča nalaganje datotek PDF, ki vsebujejo raziskovalne prispevke, in uporablja ChatGPT za generiranje pregledov za vsak prispevek. Obravnavane so metodologija, rezultati in možni vplivi te aplikacije, ki poudarjajo tako prednosti kot izzive integracije UI v akademski pregledni postopek.

KLJUČNE BESEDE

Umetna inteligenca, ChatGPT, strokovni pregled, zdravstvo, znanstvene raziskave, Flask ogrodje, obdelava PDF, akademsko pisanje, konferenčni prispevki

1 Introduction

The integration of artificial intelligence (AI) across various domains is significantly transforming complex processes, including the peer review of conference papers in healthcare and other sciences. Among AI tools, ChatGPT, developed by OpenAI, stands out for its potential to automate and enhance the review process. Its ability to quickly analyze large volumes of text and provide insightful feedback could streamline peer reviews, reduce human workload, and enhance overall efficiency.

Recent studies highlight the diverse applications of ChatGPT in healthcare education, research, and practice. For example, it has been shown to improve scientific writing, analyze datasets, and aid drug discovery [1, 2, 3]. Additionally, its role in generating paraphrased content and literature reviews indicates the potential to expedite academic tasks, although concerns about originality and accuracy persist [4, 5].

ChatGPT's integration into medical literature reviews has been explored, demonstrating its ability to synthesize medical knowledge, though ethical and accuracy issues require further research [6, 7]. Beyond healthcare, ChatGPT enhances research efficiency across various scientific fields. It effectively generates Boolean queries for systematic reviews and supports rapid literature searches [8]. The AI's potential to streamline peer reviews and address biases, is also evident, though managing issues like bias, plagiarism, and inaccuracies remain crucial to maintaining academic integrity [9, 10, 11]. In our opinion, tools like ChatGPT offer significant opportunities to enhance the peer review process. However, careful deployment is necessary to ensure ethical considerations, accuracy, and the preservation of academic integrity. This paper explores these

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.chtm.3

aspects and presents insights into effectively integrating ChatGPT into the peer review process designed with our system HealthReview. The paper consists of Section 1 Introduction, section 2 Methodology where the system is presented. Results are demonstrated in Section 3, and the paper concludes with a discussion.

2 Methodology

When tested, GPT-4o was already able to reply to all basic questions, e.g. “Is grammar in that text correct”. Therefore, the first task was to create a list of commands to perform the sequence of the review, chosen for the Information Society conference (is.ijs.si). The second task was to fine-tune the process since several output issues were not as desired. The third task was to include additional knowledge, and that was executed by including the Insieme knowledge base as the core medical information. In this way, the HealthReview performance achieved reasonable levels.

The Insieme platform was selected for integration with GPT [12]. This platform features a user-friendly interface that enables users to efficiently access valuable healthcare information from a centralized website, either via manual search or through the Insieme search function. Insieme serves as the successor to the national Electronic and Mobile Health (eHealth) initiative, a project that was characterized by collaboration among 15 partners. Furthermore, the platform's development has been significantly shaped by insights derived from the analysis of various EU healthcare platforms, particularly those that specialize in elderly care. Insieme not only builds upon the foundation laid by these prior initiatives but also aims to set a new standard in the accessibility and usability of healthcare information, thereby providing a more integrated and cohesive experience for its users. Additional medical knowledge therefore comes from the Insieme database that was created from a national electronic and mobile project for smart cities. Including the Insieme platform into the GPS is described elsewhere and is not part of this paper. The HealthReview web application is developed in Python using the Flask framework. Flask is a lightweight web framework for the Python programming language, used for building fast and stable web applications [13]. This application allows users to upload one or several PDF files containing multiple research papers, processes these files to extract the individual papers, and then employs ChatGPT to generate reviews for each paper.

The application is initialized and configured to use an upload folder named 'uploads'. This folder is created if it does not exist to ensure that uploaded files have a designated storage location. To maintain file security, the application restricts uploads to PDF files by specifying an allowed file extension set.The user interface of the application consists of an HTML form rendered by the index route. This form allows users to upload their PDF files through a file input and submit button. Client-side JavaScript enhances user experience by displaying the selected file name and showing a loading indicator upon form submission. Upon file upload via the /submit route, the PDF is securely saved in the upload folder. The file function ensures that only files with a .pdf extension are accepted. Once the file is stored, the papers_strArr function extracts the text of individual papers from the PDF. This function searches for specific markers like "ABSTRACT" and "Author index" to identify the boundaries of each paper. Text is extracted from each page and grouped into separate papers based on these markers. If no papers are detected, an error message is returned. Otherwise, the program proceeds to each extracted paper.

The API Call function facilitates the interaction with the ChatGPT model. It sends the text of each paper to the ChatGPT API, along with a predefined prompt, and retrieves the model's generated review. These reviews are stored in a global dictionary RESULTS with unique identifiers for each paper.

The */paper/<paper_id>* route enables users to view the generated reviews. When a specific paper ID is requested, the corresponding review content is fetched from the RESULTS dictionary and displayed using the paper.html template. Navigation links allow users to browse through the list of papers and view their respective reviews.

HTML templates, including form.html and paper.html, provide the user interface for uploading files, displaying instructions, and viewing individual paper reviews. The form.html template includes the file upload form and instructions for use, while the paper.html template displays individual paper reviews and includes navigation for switching between different papers. This program integrates Flask for web functionality, PDFMiner for PDF text extraction, and OpenAI's ChatGPT-4o for generating paper reviews. This combination offers a streamlined and efficient method for reviewing conference papers, potentially enhancing the speed and consistency of the review process, which is crucial in managing the increasing volume of research submissions. HealthReview accepts input in the form of any language and similarly outputs replies in any language supported by ChatGPT-4o.

3 Results

The developed web application HealthReview was tested using a dataset of conference papers to evaluate its performance and efficiency in generating reviews. An example of the application’s output is shown below.

Title: Anonymous in this paper
Authors: Anonymous
Abstract:
The article focuses on physical inactivity, which increases the risk of chronic diseases, especially cardiovascular diseases, among the adult population as well as among children and

youth. Based on data from the Institute for Health Metrics and Evaluation, the article reports relative risks of cardiovascular disease by age group, gender, and country. In doing so, we focus on the comparison between Slovenia and the world average. Based on the analysis, we conclude that cardiovascular diseases impact people in different age segments; as such they become one of the important risk factors through aging. Also, physical activity importantly influences behavioral, nutrition, and other factors, which contribute to cardiovascular diseases. For that reason, it is very important to systematically analyze and monitor these factors to contribute to better guidelines and measures of policymakers.

Overall evaluation: 1 - Accept with minor changes
(All options: 2 - Accept as is | 1 - Accept with minor changes | 0 - Accept with major changes | -2 - Reject)
The paper provides valuable insights into the relationship between physical inactivity and cardiovascular diseases, particularly in different age segments. However, some minor improvements could elevate the presentation and clarity of the findings.
Overall quality: 7/10.
The scientific quality is solid, but minor methodological limitations and the need for more comprehensive data analysis slightly reduce its impact.
Presentation: 3 – Routine
(All options: 5 - Crystal-clear and particularly attractive | 4 - Clear and attractive | 3 - Routine | 2 - Not very clear and/or attractive | 1 - Unclear and/or particularly unattractive.)
The presentation is well-structured but lacks visual aids and a more engaging writing style.
Language modifications: Minor grammatical mistakes. The paper includes some minor grammatical mistakes that should be fixed before publication.
Proposed modifications: Correct "prerana" to "prehrana" in the "REZULTATI" section. Ensure consistency in the use of terms such as "risk factors" and "prevalence" across sections.
Improve the translation and readability of some sentences for better clarity.
Consider adding graphical representations of data to enhance interpretability.
Suggestion: Use tools like Grammarly or GPTs.
Reviewer's confidence: 4 – High
(All options: 5 - Expert | 4 - High | 3 - Medium | 2 - Low | 1 - None)
Excellence: 3 - Definite but modest methodological novelty and/or good application
(All options: 5 - Important methodological novelty | 4 - Definite methodological novelty | 3 - Definite but modest methodological novelty and/or good application | 2 - Debatable methodological novelty and/or acceptable application | 1 - No methodological novelty and poor application)
Impact: 3 - Used/useful and moderately mature
(All options: 5 - Highly impactful application and/or very widely used, at least moderately mature | 4 - Important application and/or widely used, at least moderately mature | 3 - Used/useful and moderately mature | 2 - Used/useful in limited cases and/or immature | 1 - No practical use)

The HealthReview web application successfully generated reviews for multiple conference papers. Table 1 represents a summary of the evaluations for the conference section. The paper number is a serial number assigned to each paper in the conference dataset. Examples of the meanings of other fields are listed in this section.

4 Discussion

The results of our study demonstrate the promising potential of integrating AI, specifically ChatGPT, into the peer review process for healthcare and medical conference papers. The application that was developed effectively generated insightful reviews, which were evaluated against traditional human-generated reviews for quality and consistency. While the overall performance of the AI-based review system was favorable, several considerations and implications warrant further discussion. First, the ability of ChatGPT to process and analyze large volumes of text rapidly offers a significant advantage in terms of efficiency. This is particularly beneficial in the context of increasing research output and the growing burden on peer reviewers. However, while the AI-generated reviews were generally accurate and aligned with human assessments, there were instances where the feedback provided by ChatGPT lacked depth, particularly in areas requiring domain-specific expertise. This highlights a limitation of current AI technologies, where the ingenuity, flexibility and expertise of human reviewers are still crucial.

It is not clear to what extent the automatic reviewing system applies to all domains, not only healthcare. In any case, when adopting a specific form of review, the instructions as part of the prompt programming should be modified. In practical terms, it should also be noticed that the API GPT call is not free of charge.

Additionally, the application of AI in the peer review process raises ethical concerns, particularly around the potential for bias, the risk of plagiarism, and the integrity of the review process. Although ChatGPT can streamline the review process, these tools must be used as supplements rather than replacements for human reviewers. Maintaining a balance between AI efficiency and human oversight is critical to preserving the integrity and quality of academic peer reviews. Moreover, the reliance on AI for academic tasks necessitates continuous monitoring and updates to the AI models to ensure accuracy, relevance, and fairness. Future developments should focus on enhancing the contextual understanding of AI tools like ChatGPT to better mimic the critical thinking and analytical capabilities of human reviewers.

Table 1: Summary of Reviews

Paper number	Overall evaluation	Overall quality	Presentation	Language (grammatical mistakes)
1	2/4	6/10	2/5	Quite some
2	3/4	7/10	3/5	Minor
3	3/4	7/10	3/5	Some
4	3/4	7/10	3/5	Quite some
5	3/4	7/10	3/5	Some
6	3/4	7/10	3/5	Minor
7	3/4	8/10	3/5	Minor
8	3/4	8/10	3/5	Some
9	3/4	7/10	3/5	Minor
10	3/4	7/10	3/5	Minor
11	3/4	8/10	4/5	Some
12	3/4	7/10	3/5	Minor
13	4/4	9/10	4/5	Minor

In conclusion, the inclusion of HealthReview, i.e. an additional automated review layer introduces several advantages, such as increased objectivity and the potential to generate supplementary suggestions, further enriching the review process. AI tools like ChatGPT offer substantial potential to enhance the peer review process. However, their successful integration requires careful implementation and continuous evaluation to effectively address inherent challenges and ensure that these tools make a meaningful contribution to academic research. In any case, the automatic review by HealthReview or any other review tool should be marked in a way explicitly denoting the source and type of the reviewing tool.

5 Acknowledgments

This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214)

REFERENCES

[1] R. Garg, V. L. Urs, A. A. Agrawal, S. K. Chaudhary, V. Paliwal, and S. K. Kar. 2023. Exploring the role of chatgpt in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promotion Perspectives, 13, 183–191. https://hpp.tbzmed.ac.ir/Article/HPP_19952_20230501034121

[2] P. Y. Lee, H. Salim, A. Abdullah, and C. H. Teo. 2023. Use of chatgpt in medical research and scientific writing. Malaysian Family Physician, 18, 58. https://www.e-mfp.org/ojs3/index.php/MFP/article/view/1213

[3] F. Tustumi, N. Andreollo, and J. E. Aguilar-Nascimento. 2023. Future of language models in healthcare: the role of chatgpt. Arquivos Brasileiros de Cirurgia Digestiva: ABCD, 36. https://doi.org/10.1590/0102-672020230002e1800

[4] Ö. Aydın and E. Karaarslan. 2022. Openai chatgpt generated literature review: digital twin in healthcare. In Emerging Computer Technologies 2. Ö. Aydın, editor. İzmir Akademi Dernegi, 22–31. https://doi.org/10.5281/zenodo.7094305S.

[5] S. Ruksakulpiwat, A. Kumar, and A. Ajibade. 2023. Using chatgpt in medical research: current status and future directions. Journal of Multidisciplinary Healthcare, 16, 1513–1520 https://doi.org/10.2147/JMDH.S419830.

[6] M. Sallam. 2023. Chatgpt utility in healthcare education, research, and prac- tice: systematic review on the promising perspectives and valid concerns. Healthcare, 11. doi: 10.3390/healthcare11060887 D. Solomon, K. Allen, P. Katz, A. Sawalha, and E. Yelin. 2023. Chatgpt, et al... artificial intelligence, authorship, and medical publishing. Arthritis & Rheumatology, 75. doi: 10.1002/art.42497

[7] O. Temsah et al. 2023. Overview of early chatgpt's presence in medical literature: insights from a hybrid literature review by chatgpt and human experts. Cureus, 15. doi: 10.7759/cureus.37281

[8] S. Wang, H. Scells, B. Koopman, and G. Zuccon. 2023. Can chatgpt write a good boolean query for systematic review literature search? In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. doi: 10.1145/3539618.3591703

[9] Biswas, D. Dobaria, and H. L. Cohen. 2023. Chatgpt and the future of journal reviews: a feasibility study. The Yale Journal of Biology and Medicine, 96, 415–420. https://pubmed.ncbi.nlm.nih.gov/37407690/

[10] D. Solomon, K. Allen, P. Katz, A. Sawalha, and E. Yelin. 2023. Chatgpt, et al... artificial intelligence, authorship, and medical publishing. Arthritis & Rheumatology, 75. doi: 10.1002/art.42497

[11] J. Verharen. 2023. Chatgpt identifies gender disparities in scientific peer review. eLife, 12. https://doi.org/10.7554/eLife.88173

[12] The Insieme platform presentation: https://www.youtube.com/watch?v=ozAporFOw64, Accessed 1.9.2024

[13] Flask framework documentation: https://flask.palletsprojects.com/en/latest/ Accessed 1.9.2024

Leveraging Federated Learning for Secure Transfer and Deployment of ML Models in Healthcare*

Zlate Dodevski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of N. Macedonia
zlate.dodevski@students.finki.ukim.mk

Tanja Pavleska
Laboratory for Open Systems and Networks
Jozef Stefan Institute
Ljubljana, Slovenia
atanja@e5.ijs.si

Vladimir Trajkovik
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of N. Macedonia
vladimir.trajkovik@finki.ukim.mk

Abstract

Federated learning (FL) represents a pivotal advancement in applying Machine Learning (ML) in healthcare. It addresses the challenges of data privacy and security by facilitating model transferability across institutions. This paper explores the effective employment of FL to enhance the deployment of large language models (LLMs) in healthcare settings while maintaining stringent privacy standards. Through a detailed examination of the challenges in applying LLMs to the healthcare domain, including privacy, security, regulatory constraints, and training data quality, we present a federated learning architecture tailored for LLMs in healthcare. This architecture outlines the roles and responsibilities of participating entities, providing a framework for secure collaboration. We further analyze privacy-preserving techniques such as differential privacy and secure aggregation in the context of federated LLMs for healthcare, offering insights into their practical implementation.

Our findings suggest that federated learning can significantly enhance the capabilities of LLMs in healthcare while preserving patient privacy. In addition, we also identify persistent challenges in areas such as computational and communicational efficiency, lack of benchmarks and tailored FL aggregation algorithms applied to LLMs, model performance, and ethical concerns in participant selection. By critically evaluating the proposed approach and highlighting its potential benefits and limitations in real-world healthcare settings, this work provides a foundation for future research in secure and privacy-preserving ML deployment in healthcare.

Keywords

Federated Learning, Large Language Models, Data Privacy, Healthcare ML, Privacy Preservation, Model Transferability

1 Introduction

The advancements in hardware and software technologies, hyper-connectivity, and the fourth industrial revolution lead to the creation of mass amounts of health-related data. Machine learning and AI, in general, are the biggest winners from this endless pool of structured and unstructured data, as these technologies thrive on large datasets to identify patterns and make predictions. The novel adaptable, predictive approach to generating insights, decision support, and assistance in tasks that have long been considered solely reserved for human expertise is based on this paradigm's capabilities to recognize patterns from the data without being explicitly programmed.

Conventional machine learning implies that the data owner communicates with a specific central server with significant computational power. The central server consumes data from different sources and applies training techniques and algorithms to devise an effective model. ML requires large amounts of data to satisfy the expectations for the model's performance.

Large-language models (LLMs), as representative of ML advancements, have been a particular point of interest in recent years. They have already proven their applicability and massive potential in multiple fields [30]. LLMs are designed to understand, generate, and interact with human-like text and can understand context, making them suitable for performing a wide range of complex language-related tasks. They are trained in two main phases. First, the model learns general knowledge about language patterns in the pre-training phase. Then, it can be fine-tuned to execute downstream tasks to specialize its expertise in a specific domain [28].

However, like other machine learning models researched and implemented, this paradigm is data-hungry, meaning that it inherently requires massive training data to achieve the expected performance [10]. Thus, LLMs are designed to perform better with an increase in training data volume and computational power.

Various unfortunate scenarios related to the misuse of private and personal data cast a shadow on AI's capabilities, underscoring the growing concerns about data privacy, specifically in the phases when the models are trained. [13]. The year 2016 is particularly significant for two key developments aiming to overcome these challenges. The first is the attempt to regulate personal data collection, processing, and storing by introducing the General Data Protection Regulation (GDPR) in Europe [12]. The second key development was the introduction of Federated Learning (FL) by Google researchers, which provided a groundbreaking scientific approach to addressing data privacy and security concerns in ML [16].

This paper aims to discuss the possibility of satisfying the needs of both data owners and ML experts by leveraging the concept of federated learning. On the one hand, data owners can be supported to participate in collaborative training in a privacy-preserving manner when their data is insufficient to craft a high-

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.chtm.5

performance model, such as LLM. On the other hand, ML experts can develop and advance their approaches by utilizing large volumes of real-life institutional data and access to diverse scenarios, which are essential for building a robust model.

Being aimed at investigating the FL potential for application in the health domain, the insights presented in this work offer support in finding a more robust, secure, and effective use of AI that does not require technical proficiency of the medical experts, ultimately contributing to improved patient care and data protection.

The remainder of this paper is organized as follows: Sections 2 and 3 introduce FL and LLMs, and their relevance to healthcare; section 4 presents our proposed FL architecture for LLMs in healthcare, detailing the system components and their roles; section 5 discusses challenges in implementing LLMs in healthcare using FL; and section 6 explores privacy-preserving techniques for integration with our FL architecture.

2 LLM-based Healthcare Applications

LLMs are usually trained on high-quality public data, but their performance is often limited when tasked with specialized or narrower-spectrum tasks. With specific expertise in mind, such as healthcare, different approaches should be considered to build that corpus. Healthcare institutions can use their local datasets, leading to less efficient LLM, or they can join other collaborative efforts to make high-quality training data facing the inevitable challenges of privacy and regulations.

LLMs are attractive in the healthcare area because of their capability to simplify the interaction with an intelligent system without needing technical expertise. Lack of technological proficiency of medical experts can decrease the adoption of a specific software solution and can be marked as overly complex. The core principles of the LLMs allow users to interact with their domain rules, persistent knowledge, and past experiences without the need to rely on their computer literacy. One of the enormous benefits of utilizing LLMs in potential areas of medical workflows relies on the output, which is in an understandable form of natural language. The ease of use of natural language to provide instructions and ask for decision support bridges the gap between the domain experts and the utilization of an intelligent computer system. Furthermore, much of the data that persists within healthcare institutions is in unstructured formats, such as clinical notes, conversations, diagnoses, prescriptions, and research articles. LLMs are particularly effective at processing and using these natural language texts. In that way, the transparency is increased, and the expert can examine the reasoning behind the answers provided straightforwardly.

In the past several years, we have already witnessed the potential of LLMs in healthcare in many areas, such as interpreting images from a specific medical domain, summarizing reports and medical history, identifying patterns in electronic health records (EHR), and offering support for decision-making processes. The use of natural language can also influence patient engagement processes [16].

Even though exceptional remarks on providing suitable LLMs for healthcare are already in place and the community is verifying the theoretical and conceptual findings, the decision of one healthcare institution to utilize such a system can face a lot of hesitation. Most of the training data for the LLMs comes from publicly available sources, lacking the nuances that bring the real-life data isolated in the infrastructure of a single healthcare institution. Additionally, each institution has many characteristics that make it unique in how it works. Workflows, dictionaries, specific characteristics of the population it serves, or particular domain attributes can result in difficulties for off-the-shelf LLMs in providing the correct output to the instruction given. This calls for the institutions to further tailor and tune the capabilities of the LLM. Although state-of-the-art LLMs allow for such modifications and fine-tunings and making this process feasible, this comes with a heavy involvement and effort by the institution representatives and with extensive computational resources. Finally, even if one institution is capable of making efforts to adapt a generalized LLM for its use, it faces the inevitable obstacle of data insufficiency. In general, a single institution either cannot provide enough data to receive proper, correct output for the downstream task or is incapable of solving instruction of so-called new events or conditions.

A healthcare institution would need support in multiple areas to make the process easier to follow and adopt. As a result, processes related to finding a suitable LLM model, maintaining it, and keeping it up to date should be outsourced to a separate body owning the expertise. To effectively adapt LLMs in the healthcare domain, collaboration among institutions in compliance with the industry regulations should be established to build a rich training corpus.

3 FL Principles Relevant to Healthcare Data Privacy and Security

In healthcare, data is often distributed across multiple institutions, each possessing unique and valuable patient information. Traditional approaches to AI model training require centralizing this data, which poses significant privacy and security risks. Federated learning provides a solution by enabling collaborative model training without exchanging raw data. Instead, each institution trains the model locally and shares only aggregated updates with a central server. This method ensures that sensitive patient data remains within the institution, facilitating the transfer and deployment of AI models across different settings without compromising data security. FL is an iterative process, and each communication round aims to improve the model's performance. A typical FL scenario consists of two main phases in each round: local training of the model done on the participant side and aggregation of updates, which aims to create the most accurate consensus model.

There are three main types of FL based on how the data is distributed across participants. In horizontal federated learning, the datasets share the same feature space but differ in the samples they contain. Vertical federated learning, on the other hand, involves datasets with the same samples but different features. Lastly, federated transfer learning encompasses datasets that vary in both their feature and sample spaces [8].

FL in healthcare is predominantly covered in theoretical research, with many studies exploring its potential, such as for improved personalized treatment and public health monitoring. However, there are real-life applications, such as in radiology, where FL enables collaborative training on medical images like MRIs and X-rays without sharing patient data [23].

In the context of machine learning (ML) applications involving healthcare data, there are three critical vulnerability points that require attention: the data itself, the training of ML models, and the communication and transfer of data. Each area carries specific challenges and risks that must be mitigated to ensure the privacy, security, and efficacy of ML systems in healthcare. Health-related data is inherently complex, with characteristics such as high dimensionality, variance over time, heterogeneity, difficult interoperability, sparsity, and isolation [4]. Protecting the privacy of patients' personal and sensitive health information is crucial. Due to the sensitive nature of healthcare data, security breaches can lead to severe consequences, including identity theft, fraud, and violation of patient confidentiality.

Healthcare data often comes from various sources, such as hospitals, clinics, wearable devices, and electronic health records (EHRs). This data is typically non-independent, identically distributed (non-iid), unbalanced, and fragmented across different systems. Additionally, data may be sparse or isolated, making it challenging to build comprehensive patient profiles or conduct large-scale analyses.

Federated learning offers a promising approach to overcoming these challenges by allowing ML models to be trained across multiple decentralised data sources while keeping data local. This technique improves data privacy and security by not requiring raw data to be transferred to a central location. In an FL environment, each data controller defines its governance processes and privacy policies. This includes setting conditions for data access, training, and validation phases [3, 7, 19].

Communication between institutions, especially in healthcare, must adhere to strict regulatory requirements, such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and GDPR in Europe. Thus, privacy-preserving mechanisms should be implemented "by design" and "by default" to ensure that sensitive patient information is processed securely. A compliant ML system requires secure data transfer mechanisms, consent management, and audit trails. The FL setting can offer advancement in this area by letting institutions keep sensitive information, prevent unnecessary data transfers and processing that could violate regulatory requirements, and minimize the risk of data breaches [27].

Training ML models with healthcare data presents unique challenges, including addressing data bias, limited sample sizes, and ensuring model performance. Healthcare data may be biased due to demographic imbalances, socio-economic factors, or varying levels of care access across populations.

Training of ML models with diverse datasets enhances their generalizability and robustness. By incorporating data from various sources and populations, models can better adapt to new and unforeseen health events, improving their predictive power and reliability. Federated learning, in particular, enables the use of diverse datasets while maintaining privacy, thus improving overall model performance [21, 17].

4 FL Architecture for LLM-Based Healthcare Applications

Figure 1 depicts the three major components of a typical FL architecture. The participants involved in our cross-silo FL setting are the healthcare institutions, the manager (e.g., aggregation server or global server), and the communication-computation layer, which aggregates local updates and orchestrates communication phases in the ecosystem. Each component has its own responsibilities, which are essential for the model to satisfy the preset expectations.

Leveraging FL in utilizing LLMs adds a layer of complexity and implies different approaches based on the level of decentralization that needs to be achieved [29]. FL can help in both the pre-training and fine-tuning phases of LLM, and it is up to the requirements' specific characteristics and the parties involved computational power to choose the right strategy [2, 11]. We will cover the different approaches while examining the three major architecture components.

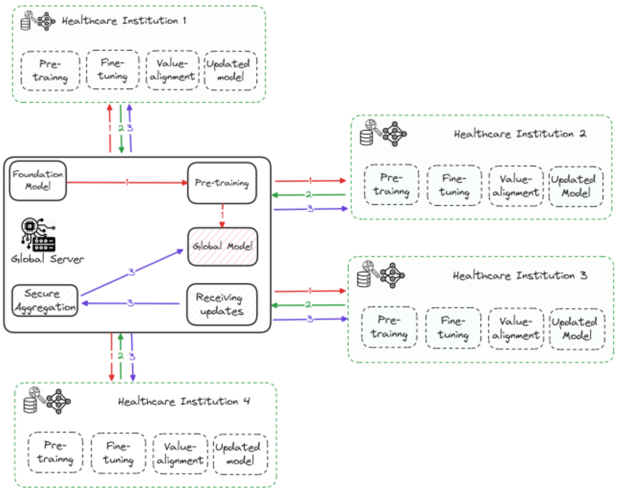


Figure 1: Typical FL Architecture that can be deployed for LLM-based applications

4.1 Global Server

The global server plays a central role, and due to the lack of properly established taxonomy and standards, this component in the literature is also considered as the manager or creator of the whole ecosystem. It is often referred to as the "manager" or "aggregation server" tasked with overseeing the entire collaboration and ensuring its smooth functioning. In healthcare, the manager can be a single healthcare institution that holds a lot of data and wants to leverage the FL setting to collaborate with other institutions, either to complete the missing domains and dimensions (by utilizing Vertical FL) or to enrich and expand the feature set in the same dimension (Horizontal FL). The global server's responsibilities can be broken down into several distinct areas: strategy for choosing a foundation model, strategy for exploiting data distribution and the client selection.

The foundation model, also called the base model, forms the initial point of the LLM training in the FL setting. It represents a starting point for institutions to leverage the pre-training process with their own data or fine-tune it to perform specific downstream tasks since foundation models are usually trained on publicly available datasets. The pre-training process is computationally and time-consuming, meaning that the global server must find the most suitable scenario for satisfying requirements.

There are two main kinds of pre-training models: one based on the BERT model and the other on the GPT model. Both perform differently for different tasks and scenarios [5, 22]. Many attempts are made to use publicly available literature specific to the medical domain and create ready LLM models for usage, such as BioBert, which show superior performance than general pre-trained models [15].

One approach is to select a suitable foundation model based on the options examined before or to initialize FL pre-training, where each party will contribute to the pre-training of the foundation LLM model. The latter approach requires the institutions to have expertise and sufficient computational power in their infrastructure to complete the assignment. Another consideration is the model's size and complexity, which will influence the following steps if not chosen according to the participant's IT infrastructure.

Federated Learning can work differently depending on how data is spread and distributed across entities. In horizontal FL, each institution has data with similar features (for example, multiple hospitals with similar patient data). In vertical FL, institutions have different features for the same set of patients (for example, one entity has clinical data, and another has genetic data). Transfer learning can also be used when the model needs to generalize across different datasets [20]. The global server is responsible for choosing the appropriate strategy based on the data distribution and the desired outcomes.

Client selection in FL refers to choosing which institutions to participate in each training round. The global server must ensure that diverse institutions contribute to model updates without overloading the communication system. Institutions with more data or better computational resources might participate more frequently. Still, the system should be flexible enough to rotate clients or dynamically adjust client participation based on resource availability.

4.2 Participants

The entities participating in this collaboration technique are also tasked with significant responsibilities. In the medical domain, these institutions own huge structured or unstructured datasets and are willing to participate in a distributed training process. Their responsibilities can be broken down into the pre-training process, fine-tuning, value alignment, and strategy for local updates.

If the FL-specific training approach is adopted, as discussed previously in the strategy for choosing the foundation model, then each institution may pre-train the selected model on its data and ensure that the initial model updates sent to the global server are more relevant and valuable. This capability in an FL setting can depend significantly on the institution's computational power, and even though possible and theoretically feasible, it can require a lot of additional expertise for the healthcare institution to engage in this kind of activity.

Fine-tuning is crucial in adopting an LLM in the healthcare domain. The approach of LLM fine-tuning is to make downstream tasks required by the domain be instructed with human feedback [18]. Each institution should provide input-output pairs where instructions are explicitly offered to solve some already defined downstream tasks. These datasets are designed to give the model an idea of what kind of output is expected. The expectation is that the LLM will learn to generalize and can handle novel instructions even though they were not a part of the fine-tuning instruction dataset.

The variety of downstream tasks that LLMs can perform in the healthcare domain is often the critical reason institutions engage with this concept. Based on a benchmark for generalist biomedical AI, some of the most frequently performed medical-relevant tasks suitable for the LLM domain are question answering, visual question answering (for example, based on radiology or pathology images), report summarization and generation, and medical image and medical documents classification [26]. Additionally, relation extraction in combination with named entity recognition can be added to the list of medical-relevant tasks. This is helpful in the medical domain to extract medical terms such as diseases, conditions, procedures, and symptoms from unstructured data and find suitable interpretations and connections in the unstructured data pool.

Thanks to the LLaMA, each institution can make significant attempts to build its domain-specific instruction set and contribute to global instruction tuning in the FL setting. With the FL paradigm, each downstream task can be trained on multiple datasets instead of a single dataset, giving more suitable responses and outputs [25].

In the FL setting, the value-alignment step occurs on the participant's side during local training. Its purpose is to ensure that the model's objectives are aligned with each institution's values and goals. This step is particularly crucial in the medical field, where ethical guidelines and patient care standards are of the utmost importance.

Technically, value-alignment is solved similarly to instruction tuning, with each participant's preference dataset containing combinations of instruction, preferred, and misreferred responses.

In FL, participants typically have far fewer computational resources than centralized cloud servers and fine-tuning all parameters of LLMs can be an obstacle. Parameter-efficient tuning techniques, such as Lora, are used to address this limitation [6]. Instead of updating the entire pre-trained model weights to obtain local updates, participants modify only specific parameters and send them back to the global server for aggregation.

4.3 Communication-Computation Layer

As presented above, the global server is responsible for managing the whole ecosystem, and one of the most complex tasks is related to the communication-computation layer. The global server should manage the aggregation process of local model updates and ensure that the global LLM and updates are securely transmitted across the system.

Choosing the suitable FL algorithm for combining all findings and improvements made by each participant in the form of parameter weights is a step that has attracted many researchers and experts. One of the first and most used algorithms is Federated Averaging (FedAvg), but more sophisticated approaches may be necessary in different scenarios [16]. The model's performance relies significantly on how model updates are aggregated. Even though the only data transmitted through the network in an FL setting are the model and its updates, the communication layer is responsible for ensuring that the transfer is done securely and continuously. The communication layer component must develop a strategy for creating a pipeline from a live data connection to the model and inference to transmitting new model parameters via secure channels to the aggregating server. Size and complexity of the model must be considered as well, since they can cause a bottleneck.

In addition, the communication layer also ensures that the data transfer is seamless and uninterrupted. This component is tasked with developing a robust strategy to create an efficient pipeline, from managing real-time data connections to facilitating model utilization and transmitting updated model parameters securely to the central aggregation server. A key consideration for the communication layer is the size and complexity of the used model. Large models with huge parameter lists can introduce significant bottlenecks during transmission, especially when dealing with limited bandwidth or less powerful devices. As such, the communication layer must be adept at handling these challenges, ensuring that updates are transferred efficiently without compromising the speed or security of the system.

5 Hype, Vision and Challenges

Implementing LLMs in healthcare using FL presents a set of intertwined challenges when viewed through the lenses of privacy and security. There is a foundational challenge between the need for diverse and high-quality data generated by institutions in the specific domain and the importance of protecting sensitive information. FL enables availability and access to a broader spectrum of data sources while maintaining privacy. Still, the inability to directly act upon raw data can impact the convergence of the model and model performance. Data transfer needs in FL, even though minimized to just model updates, still introduce a risk for security attacks. This risk increases with the communication overhead caused by distributing complex and large LLMs.

By introducing a central figure in the architecture in the name of the global aggregation server, the FL setting in LLM opens up a single point of failure in the ecosystem. Adversarial attacks can be performed, compromising model integrity, which could lead to data breaches and incorrect outputs.

FL is still a young and immature topic in the context of LLM. One of the biggest challenges is the lack of benchmarks and comprehensive reviews that can examine the solution's success based on different tasks, architectures, the number of clients, network bandwidth, computational resources, etc. These reviews and benchmarks can further expose security and privacy-preserving issues and initiate proper risk mitigation strategies. Multiple algorithms exist in the literature for aggregating local updates, but no specific algorithm is proposed or adapted for LLMs.

The analysis of the three major components in the previous section pointed out the responsibilities, approaches, and strategies that need to be considered in order to collaboratively design and implement training, and utilize LLM properly. The analysis emphasized that training LLMs in a federated learning setting requires a thoughtful, tailored approach to address the unique challenges. Additionally, there are various approaches to take, depending on factors such as participant resources, data distribution, model size and complexity, data transfer, etc. This section will further examine the challenges of implementing such LLM training in the FL setting. Fine-tuning LLMs in FL is a time-consuming and computationally expensive task [6].

The client selection process, in which the ecosystem manager decides which participants should be involved, can raise many ethical concerns, such as fairness. The purpose of the collaboration is to make the LLM more robust. Still, some participants' data volume and computational power can squeeze out institutions that are not on that level but still can add to the diversity and offer unique cultural, ethical, and contextual values. While FL addresses many privacy concerns by design, it also introduces new security considerations that must be carefully managed. Successfully navigating these challenges requires a detailed approach that balances privacy protection, security enhancement, and the pursuit of practical and robust LLM in healthcare.

6 Privacy-Preserving Techniques

The deployment LLMs in the healthcare field through FL promises advancements in preparing models to react to given domain-specific downstream tasks. The FL can enhance LLMs' effectiveness and proper application while safeguarding patient confidentiality and ensuring regulatory compliance, providing medical professionals greater confidence in adopting these tools.

However, while FL enables collaborative learning without direct data sharing, it's not immune to privacy threats. With this approach, raw data remains local, but the model updates shared during training can still leak information. In addition, LLMs trained with healthcare data could memorize and potentially regenerate sensitive patient information. A privacy breach in this context can cause severe consequences, including exposure to medical history, compromising patient confidentiality, and misuse of sensitive health information [1].

During this collaborative process, the model or its updates could become targets for various attacks. For instance, model inversion attacks performed on the global model might allow the reconstruction of individual patient records. Similarly, membership inference attacks could reveal the presence of specific institutions or patient data in the training, potentially exposing the entire medical history. Malicious participants in the process could poison the model by introducing biases or backdoors, potentially leading to improper results generated by the LLMs [9, 24].

To counter these risks and threats, researchers and practitioners evaluate the effects of several privacy-preserving techniques, such as secure aggregation and differential privacy. Secure aggregation, a cryptographic protocol, allows the central server to observe aggregated results without accessing individual model updates. This approach maintains accuracy but adds significant communication costs. Differential privacy, on the other hand, adds calibrated noise to data or model parameters, offering statistical privacy guarantees. While effective against inference attacks, it may reduce model accuracy and require additional workload in the parameter-tuning process [14]. The choice of privacy-preserving techniques must be made with a thorough understanding of the specific use case, the sensitivity of the data involved, and the potential impacts of privacy breaches. The tailored approach should calibrate the trade-off between model performance and data protection. More robust privacy protection might require limiting the model's access to much-needed data for LLMs to offer a proper answer to a specific task, degrading the model performance and increasing the computational and communicational overhead. As research in this field progresses, finding the right balance between privacy, system performance, and efficiency will be crucial for deploying LLMs in healthcare using FL.

7 Conclusion

This paper has explored the potential of FL in enhancing the deployment of LLMs in healthcare settings. By enabling privacy-preserving collaboration, FL allows healthcare institutions to collectively train and improve LLMs without compromising sensitive patient data. This approach not only addresses fundamental privacy concerns but also enhances model performance by leveraging diverse datasets across institutions, potentially improving the generalizability and robustness of LLMs in healthcare applications. To facilitate the implementation of healthcare LLM with FL, we examined a tailored architectural framework that outlines the roles and responsibilities of participating entities. In addition, challenges and consideration of risks and threats were reviewed, especially in combination with privacy-preserving techniques.

Looking ahead, several areas require further research and development. Optimization of computational and communication efficiency for LLMs, development of standardized benchmarks, establishment of ethical frameworks for participant selection, and exploration of advanced privacy-preserving techniques are crucial for future work.

8 Acknowledgments

This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214).

REFERENCES

[1] [Aouedi, O., Sacco, A., Piamrat, K. and Marchetto, G. 2023. Handling Privacy-Sensitive Medical Data With Federated Learning: Challenges and Future Directions. IEEE Journal of Biomedical and Health Informatics. 27, 2 (Feb. 2023), 790–803. DOI:https://doi.org/10.1109/JBHI.2022.3185673.

[2] Chen, C., Feng, X., Zhou, J., Yin, J. and Zheng, X. 2023. Federated Large Language Model: A Position Paper. arXiv.

[3] Dasaradharami Reddy, K. and Gadekallu, T.R. 2023. A Comprehensive Survey on Federated Learning Techniques for Healthcare Informatics. Computational Intelligence and Neuroscience. 2023, 1 (Jan. 2023), 8393990. DOI:https://doi.org/10.1155/2023/8393990.

[4] Deep learning for healthcare: review, opportunities and challenges - PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6455466/

[5] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[6] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models.

[7] Joshi, M., Pal, A. and Sankarasubbu, M. 2022. Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges. ACM Transactions on Computing for Healthcare. 3, 4 (Oct. 2022), 1–36. DOI:https://doi.org/10.1145/3533708.

[8] Kairouz, P. et al. 2021. Advances and Open Problems in Federated Learning. Foundations and Trends® in Machine Learning. 14, 1–2 (Jun. 2021), 1–210. DOI:https://doi.org/10.1561/2200000083.

[9] Kaissis, G.A., Makowski, M.R., Rückert, D. and Braren, R.F. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence. 2, 6 (Jun. 2020), 305–311. DOI:https://doi.org/10.1038/s42256-020-0186-1.

[10] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv.

[11] Kuang, W., Qian, B., Li, Z., Chen, D., Gao, D., Pan, X., Xie, Y., Li, Y., Ding, B. and Zhou, J. 2024. FederatedScope-LLM: A Comprehensive Package for Fine-tuning Large Language Models in Federated Learning. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (New York, NY, USA, Aug. 2024), 5260–5271.

[12] Kuner, C., Bygrave, L.A., Docksey, C. and Drechsler, L. eds. 2020. The EU General Data Protection Regulation (GDPR): A Commentary. (Feb. 2020).

[13] Laurie, G., Jones, K., Stevens, L. and Dobbs, C. 2015. A Review of Evidence Relating to Harm Resulting from Uses of Health and Biomedical Data.

[14] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X. and He, B. 2023. A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. IEEE Transactions on Knowledge and Data Engineering. 35, 4 (Apr. 2023), 3347–3366. DOI:https://doi.org/10.1109/TKDE.2021.3124599.

[15] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H. and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics. 23, 6 (Nov. 2022), bbac409. DOI:https://doi.org/10.1093/bib/bbac409.

[16] McMahan, B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A. y 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Apr. 2017), 1273–1282.

[17] Nguyen, D.C et al. 2022. Federated Learning for Smart Healthcare: A Survey. ACM Comput. Surv. 55, 3 (Feb. 2022), 60:1-60:37. DOI:https://doi.org/10.1145/3501296.

[18] Ouyang, L. et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 35, (Dec. 2022), 27730–27744.

[19] Pfitzner, B., Steckhan, N. and Arnrich, B. 2021. Federated Learning in a Medical Context: A Systematic Literature Review. ACM Transactions on Internet Technology. 21, 2 (Jun. 2021), 1–31. DOI:https://doi.org/10.1145/3412357.

[20] Prayitno, Shyu, C.-R., Putra, K.T., Chen, H.-C., Tsai, Y.-Y., Hossain, K.S., Jiang, W. and Shae, Z.-Y. 2021. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. Applied Sciences. 11, 23 (Dec. 2021). DOI:https://doi.org/10.3390/app112311191.

[21] Qayyum, A., Qadir, J., Bilal, M. and Al-Fuqaha, A. 2020. Secure and Robust Machine Learning for Healthcare: A Survey. arXiv.

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

[23] Teo, Z.L. et al. 2024. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. Cell Reports Medicine. 5, 2 (Feb. 2024), 101419. DOI:https://doi.org/10.1016/j.xcrm.2024.101419.

[24] Thapa, C. and Camtepe, S. 2021. Precision health data: Requirements, challenges and existing techniques for data security and privacy. Computers in Biology and Medicine. 129, (Feb. 2021), 104130. DOI:https://doi.org/10.1016/j.compbiomed.2020.104130.

[25] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv.

[26] Tu, T. et al. 2024. Towards Generalist Biomedical AI. NEJM AI. 1, 3 (Feb. 2024), AIoa2300138. DOI:https://doi.org/10.1056/AIoa2300138.

[27] Xu, J., Glicksberg, B.S., Su, C., Walker, P., Bian, J. and Wang, F. 2021. Federated Learning for Healthcare Informatics. Journal of Healthcare Informatics Research. 5, 1 (Mar. 2021), 1–19.

[28] Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B. and Hu, X. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. ACM Trans. Knowl. Discov. Data. 18, 6 (Apr. 2024), 160:1-160:32. DOI:https://doi.org/10.1145/3649506.

[29] Ye, R., Wang, W., Chai, J., Li, D., Li, Z., Xu, Y., Du, Y., Wang, Y. and Chen, S. 2024. OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning. (2024). DOI:https://doi.org/10.48550/ARXIV.2402.06954.

[30] Zhou, C. et al. 2023. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. arXiv.

Use and Limitations of ChatGPT in Mental Health Disorders

Jovana Petrović†
Clinic of Psychiatry
University Clinical center
Nis, Serbia
zdravkovicj91@gmail.com

Iva Binić
Clinic of Psychiatry
University Clinical center
Nis, Serbia
ivabinic@gmail.com

Maša Vacev
Clinic of Psychiatry
University Clinical center
Nis, Serbia
vacevmasa@gmail.com

Stevo Lukić†
Clinic of Neurology,
University Clinical centre
Niš, Serbia
slukic@medfak.ni.ac.rs

Abstract / Povzetek

ChatGPT is one of the most advanced and rapidly evolving large language model-based chatbots. It excels in everything from handling simple questions to performing complex medical examinations. While current technology cannot replace the expertise and judgment of skilled psychiatrists, it can assist in early detection of mental problems, patient evaluations, differential diagnoses, psychotherapy and in planning and conducting medical research. Ensuring privacy and adhering to professional, ethical, and legal standards is crucial when processing training data. This is especially important in mental health settings, where disclosing sensitive personal information increases the risk of data misuse and the potential for harmful advice. Current uses of ChatGPT in mental health care are constrained by its design as a general chatbot, rather than a specialized psychiatric tool. Despite this, the model proves useful for handling routine psychiatric and administrative tasks. As GPT technology evolves, it holds significant promise for psychiatry, including integration into diagnostics, psychotherapy, and early detection of mental health issues. To deploy these advancements responsibly and effectively, it is crucial to develop and refine professional ethical standards and practice guidelines.

Keywords / Ključne besede

Keywords mental health disorders, large language models, deep learning, ChatGPT

Introduction

ChatGPT has emerged as one of the most advanced and rapidly evolving large language model-based chatbot systems. Its extensive capabilities, ranging from responding to basic inquiries to performing well in complex medical examinations, have garnered significant attention from the global scientific and research communities, prompting ongoing discourse regarding its potential applications across diverse domain [1]

* All authors contributed equally to this research
† Corresponding authors

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.chtm.6

The discourse surrounding the potential applications of ChatGPT in mental health disorders remains relatively underexplored. This work seeks to offer a quick overview of the current state of ChatGPT implementations within the mental health domain, while also projecting future advancements in digital mental health care through the integration and development of ChatGPT technology.

Current use of ChatGPT for mental health disorders

In managing mental health disorders, human contact is especially crucial compared to other medical fields, as it is accompanied by understanding and empathy. This is why, in the most vital aspects of psychiatric work, human relationships remain central [2]. While ChatGPT's practical applications in the field of mental health are limited because it specializes in language generation, it can still support certain routine tasks within the field. Although evaluations, diagnoses, psychotherapy, and patient assessments are mainly conducted by human therapists, ongoing trials are exploring how ChatGPT might be utilized in mental health services. Several platforms, like ChatBeacon, or Koko, are already available on the market, claiming to offer mental health assistance powered by ChatGPT [3,4]. For instance, Koko is a peer-support platform that provides crowdsourced cognitive therapy. It's experimenting with using GPT-3 to generate bot-written responses to users while they wait for peer responses. Koko is an online mental health intervention that has reached nearly two million people, mostly adolescents. The platform started as a clinical trial at MIT and is based on the concept of crowdsourced cognitive therapy. Users are taught to help each other think more hopefully about the world. Unlike traditional peer support platforms, all interactions on our service are supported and augmented by AI.

Applied to the classification of psychiatric disorders.

Recent advancements in deep learning, the foundational algorithm of GPT, have significantly impacted the field of mental health disorders. This technology has been applied to classify psychiatric disorders using neuroimaging data [5], develop models based on electroencephalograms [6], and utilize a range of patient characteristics for diagnosing and predicting mental disorders [7]. These deep learning models have shown good diagnostic accuracy (AUC 0.74- 0.81) suggesting the possibility of combining genetics and registry data to predict both mental disorder diagnosis and disorder progression in a clinically relevant, cross-diagnostic setting prior to clinical assessment. The main limitation of this model is that it is restricted to learning from historical data and should be continuously assessed and evaluated by trained clinicians and never stand alone in the decision-making

Utilized to alleviate the burdens associated with clinical documentation, communication, and research tasks.

Also, new technologies can assist clinicians by allowing them to focus more on direct patient care and alleviate the high clinical workload and bureaucratic tasks- such as handling admissions and managing paperwork- that have been linked to burnout in earlier research [8]. ChatGPT can assist in processing clinical case transcripts, generating summaries, completing medical record documentation, and efficiently preparing discharge summaries. It can also help facilitate communication between clinicians of different specialties when consultations are needed, aiding in an integrative approach [9].

While current technology cannot replace the expertise and judgment of skilled psychiatrists, it can assist in generating differential diagnoses based on relevant signs and symptoms. ChatGPT is recognized for its substantial potential to assist experts with clinical and laboratory diagnoses, as well as in planning and conducting medical research [10].

Applied in psychotherapy to enhance therapeutic processes

Chatbots can be also beneficial in psychotherapy. The therapist's emotions and the emotional alignment between therapist and client are crucial factors influencing the process and outcomes of therapy [11]. A study conducted during the COVID-19 pandemic showed that technology can offer an effective method, providing at least a first level counseling support structure [12]. This implies that GPT models may potentially develop cognitive empathy over time, making it possible for ChatGPT to achieve a notable level of accuracy in identifying users' emotions [13], nevertheless it is important to make systematic testing to ensure a non-superficial comparison between human and artificial intelligences [14]. Namely, ChatGPT currently lacks the capability to accurately assess personality traits, a task that demands extensive training and expertise from psychiatrists. AI researchers are diligently pursuing technical advancements to improve the precision of personality detection [15].

A study designed to assess the accuracy and appropriateness of psychodynamic formulations generated by ChatGPT found that the model produced suitable results even without additional psychoanalytic information. It also demonstrated a strong ability to generate psychodynamic formulations consistent with various psychoanalytic theories when given appropriate instructions. [16].

The research underscores that ChatGPT is not intended to substitute psychiatrists or psychologists but rather to function as an initial resource and a first line of support for those dealing with mental distress. When used wisely and within appropriate limits, ChatGPT can be an effective tool in supporting mental health services. [17]. However, it is essential to process training data in a manner that ensures privacy protection and adheres to all professional, ethical, and legal standards. Given that individuals may be at increased risk of data misuse when disclosing sensitive personal and family information during mental health treatment [18].

Employed for the prevention and early detection of mental health issues

The role of AI in the prevention and early detection of mental problems can also be very significant. Patients frequently turn to ChatGPT to seek information about their symptoms, possible diagnoses, and treatment options. ChatGPT performs better than Google Search in delivering general medical knowledge but scores lower when it comes to providing medical recommendations [19]. A recent study highlighted early success for an AI model that can detect cognitive distortions in text messages with accuracy comparable to that of clinically trained human raters [20]. GPT's ability to recognize mental health warning signs in routine conversations or text-based telemedicine interactions has the potential to improve early and effective intervention strategies when necessary.

Risks and limitations ChatGPT use for mental health disorders

Despite its strengths and potential, the use of AI technologies in psychiatric clinical practice carries several risks. A significant concern is the phenomenon of "artificial hallucinations" where the conversational model may confidently produce text that is factually incorrect, nonsensical, or misleading [18].

Recent systematic review that included 118 articles identified some limitations regarding the potential of ChatGPT in patient care and medical research, noted that the solutions provided by ChatGPT are often insufficient and contradictory, raising concerns about their originality, privacy, accuracy, and legality [10]. It is well established that ChatGPT may generate inaccurate facts and references when summarizing previous research, and the quality of its responses often hinges on how the questions are phrased [21].

Even with the advanced GPT-4 model, there is still a risk of providing harmful advice. The absence of clinical reasoning and experience in ChatGPT can lead to the omission of important clinical details in patient summaries and medical records. Thus, the most prudent approach is to employ AI systems as supplementary tools for mental health professionals, ensuring they are used under close supervision to uphold the safety and quality of patient care. [22]

Conclusion

The recent introduction of GPT-4 has significantly enhanced the capabilities of the GPT system. Current implementations of ChatGPT within mental health care are limited by its inherent design as a chatbot, rather than as a specialized AI tool specifically tailored for psychiatric use. Nonetheless, this sophisticated language model demonstrates significant utility in addressing various routine psychiatric and administrative functions.

As this technology evolves and advances, we anticipate substantial potential for future applications of GPT technology in psychiatry, including its integration into diagnostic processes, the provision of psychotherapy within clinical environments, and the rapid identification of early warning signs for mental health disorders. Crucially, the development and refinement of professional ethical standards and practice guidelines are imperative for the responsible and effective deployment of these transformative GPT technologies in the mental health sector.

Acknowledgments / Zahvala

The work presented in this paper was partially funded in part by the European Union under Grand 101159214- ChatMED.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. Insert paragraph text here.

References / Literatura

[1] Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. (2023) Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. 89:103770.

[2] Spencer L, Broome M. (2023) The epistemic harms of empathy in phenomenological psychopathology. Phenom Cogn Sci doi: https://doi.org/10.1007/s11097-023-09930-1.

[3] ChatBeacon. Mental health assistant powered by ChatGPT. (2024) URL: https://www.chatbeacon.io/industrychatgpt/mentalhealth (Accessed 24-Aug-2024)

[4] Koko. GPT-3 mental health intervention. (2024) URL: https://gpt3demo.com/apps/koko-ai (Accessed 24-Aug-2024).

[5] Quaak M, van de Mortel L, Thomas RM, van Wingen G. (2021) Deep learning applications for the classification of psychiatric disorders using neuroim- aging data: Systematic review and meta-analysis. Neuroimage. Clin. 30:102584.

[6] de Bardeci M, Ip CT, Olbrich S. (2021) Deep learning applied to electroencephalogram data in mental disorders: A systematic review. Biol. Psychol. 162:108117.

[7] Allesoe RL, Thompson WK, Bybjerg-Grauholm J et al. (2023) Deep learning for crossdiagnostic prediction of mental disorder diagnosis and prognosis using Danish Nationwide register and genetic data. JAMA Psychiatry 80:146-155

[8] Kumar S. (2007) Burnout in psychiatrists. World Psychiatry 6:186-189.

[9] Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. (2023) Using ChatGPT to write patient clinic letters. Lancet Digit. Health 5:e179-e181.

[10] Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. (2023) Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. Health Promot Perspect. 13(3):183-191.

[11] Chui H, Li X, Luk S. (2022) Therapist emotion and emotional change with clients: Effects on perceived empathy and session quality. Psychotherapy (Chic). 59(4):594-605.

[12] Mawani A, Nderu L. (2020) Towards an online empathy assisted counselling web application. EAI Endorsed Transactions on Context-aware Systems and Applications 7(22):167792.

[13] Kosinski M. (2023) Theory of mind may have spontaneously emerged in large language models. ArXiv abs/2302.02083.

[14] Strachan J, Albergo D, Borghini G. et al. (2024) Testing theory of mind in large language models and humans. Nat Hum Behav 8:1285-1295.5

[15] El-Demerdash K, El-Khoribi RA, Ismail Shoman MA, Abdou S. (2022) Deep learning based fusion strategies for personality prediction. Egypt Inform J 23:47-53.

[16] Hwang G, Lee DY, Seol S. et al. (2024). Assessing the potential of ChatGPT for psychodynamic formulations in psychiatry: An exploratory study. Psychiatry Res. 331:115655

[17] Arjanto P, Senduk FFW, Nahdiyah U, Utami MS. (2024) AI and ethics in mental health: exploring the controversy over the use of ChatGPT. J Public Health (Oxf). 46(2):e340-e341.

[18] Wei Y, Guo L, Lian C, Chen J. (2023) ChatGPT: Opportunities, risks and priorities for psychiatry. Asian J Psychiatr. 90:103808

[19] Ayoub NF, Lee YJ, Grimm D, Divi V. (2024) Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. Otolaryngol Head Neck Surg. 170(6):1484-1491.

[20] Tauscher JS, Lybarger K, Ding X et al. (2023) Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. Psychiatr. Serv. 74: 407-410.

[21] The Lancet Digital Health. (2023) ChatGPT: friend or foe? Lancet Digit Health. 5(3):e102.

[22] Jo E, Song S, Kim JH, et al. (2024) Assessing GPT-4’s Performance in delivering medical advice: Comparative analysis with human experts. JMIR Med Educ. 10:e51282.

Testing ChatGPT's Performance on Medical Diagnostic Tasks

Alexander Perko*
Franz Wotawa*
alexander.perko@ist.tugraz.at
wotawa@ist.tugraz.at
Graz University of Technology, Institute of Software Technology
Graz, Austria

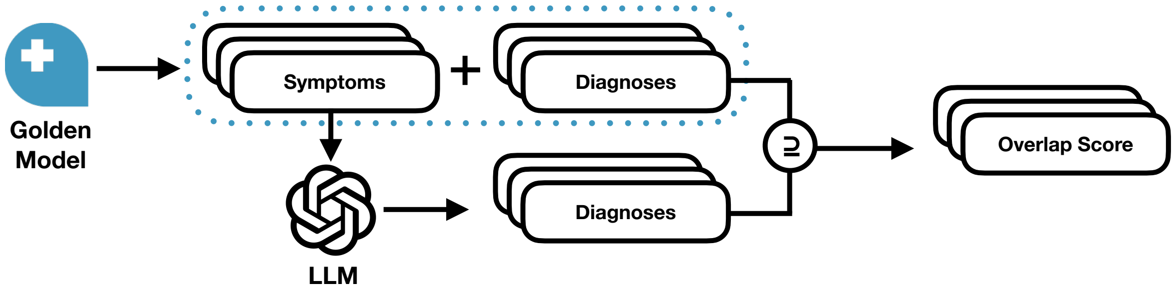


Figure 1: Semi-Automatic Evaluation of an LLM on Medical Diagnostics Using a Medical Expert System as a Golden Model.

Abstract

Large Language Models and chat interfaces like ChatGPT have become increasingly important recently, receiving a lot of attention even from the general public. People use these tools not only to summarize or translate text but also to answer questions, including medical ones. For the latter, giving reliable feedback is of utmost importance, which is hard to assess. Therefore, we focus on validating the feedback of ChatGPT and propose a testing procedure utilizing other medical sources to determine the quality of feedback for more straightforward medical diagnostic tasks. This paper outlines the problem, discusses available sources, and introduces the validation method. Moreover, we present the first results obtained when applying the testing framework to ChatGPT.

Keywords

Large Language Models, ChatGPT, NetDoktor, Testing, Validation

1 Introduction

Large Language Models (LLMs) are omnipresent in today's society, as they are used by a wide audience for a growing number of tasks. This study sheds light on one area of application in particular, which is asking for medical diagnoses. Assessing one's health and medical diagnostics are complex tasks, that fall into the domain of medical experts. However, since the dawn of search engines and medical websites, like NetDoktor [13], people have turned to the internet for getting health advice. Previously, users searching for answers had to consult multiple online resources, compare page contents, and evaluate whether their set of symptoms matched what they found. Nowadays it is seemingly easy to find answers in one spot as LLM-powered chatbots, like ChatGPT [8], are happy to respond with a diagnosis. This, of course, implies much risk of harm or misinterpretation. After all, the very reason many users - being non-experts - turn to chatbots is that they cannot assess symptoms themselves. Clusmann et al. [5] further point out that there is a lack of mechanisms to guarantee that the LLM's output is correct. All of this makes it important to test such systems on a practical level, which is close to the use cases of non-experts. As for its popularity, our evaluation focuses on ChatGPT [8], which is powered by OpenAI's most recent model, GPT-4o [9, 10]. The task of medical diagnostics shares many traits with the natural language processing (NLP) task of question answering (QA). Namely, this task tests for medical knowledge as well as basic reasoning facing medical language. MedQA [6] is a popular benchmark in literature, which is tailored to the medical domain. In recent years, open-domain LLMs such as GPT-3.5 [3], GPT-4 [9], and LLaMA-2 [16] as well as domain-specific LLMs like Med-PaLM 2 [15], Meditron [4] and Med-Gemini [14] have been evaluated on medical QA. The United States Medical Licensing Examination (USMLE) part of MedQA is used particularly often as a performance indicator in this domain. Table 1 shows reported scores of the mentioned LLMs and demonstrates GPT-4's and MedGemini's superiority, with GPT-4 performing marginally worse despite being an open-domain model.

Table 1: LLMs Evaluated on Medical Question Answering. Accuracy Results on the United States Medical Licensing Examination (USMLE) Part of MedQA [6], as Reported in [7, 14, 4, 15].

Model	Domain-Specific	MedQA USMLE
Med-Gemini	Yes	91.1
GPT-4	No	90.2
Med-PaLM 2	Yes	86.5
Meditron	Yes	75.8
LLaMA-2	No	63.8
GPT-3.5	No	60.2

*Both authors contributed equally to this research.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.chtm.7

Alongside ChatGPT's popularity, these results are a major reason why this paper focuses on GPT-4o in particular. This work contributes by introducing a semi-automated validation procedure for medical diagnostics performed with LLMs using an expert system as a golden model (compare to Figure 1). Specifically, we evaluate the performance of ChatGPT powered by GPT-4o with a focus on symptom descriptions in German and compare it to NetDoktor's Symptom-Checker [13], which is curated by medical professionals. Our setup is guided by the following questions regarding prompting ChatGPT:

- Does ChatGPT provide equivalent diagnoses when presented with the same symptoms as NetDoktor?
- Does the output quality - as measured by the overlap - change when asked for a specific amount of "most likely" diagnoses?
- Does the output increase in quality when ChatGPT is queried in English instead of German?

2 Validation Methodology

For the purpose of introducing our methodology, we use myocardial infarction (i.e. heart attack) as a guiding example. According to Statisik Austria's annual report, cardiovascular diseases, which include heart attacks, are the most common cause of death in Austria. The symptoms of a myocardial infarction include:

- Feeling of tightness or constriction
- Feeling of anxiety/panic attacks
- Sudden severe shortness of breath, unconsciousness, or severe dizziness
- Nausea and vomiting
- Blood pressure and pulse drop

These symptoms are now linked to an imaginary person's sex and age to form a persona whom for we want to retrieve diagnoses. Our exemplary set of symptoms shall be linked to an adult man and can be identified by ID 1 in all tables and plots. Besides this exemplary persona, where we first fixed a disease, all other sets of symptoms are picked at random. This can be done due to our assumption of a golden model, which we use as our baseline.

2.1 Golden Model

We use NetDoktor's "Symptom-Checker" [13] as a baseline for our evaluation. Symptom-Checker is a freely accessible, medical expert system for retrieving likely diagnoses corresponding to a person's symptoms. The system can be interacted with via a questionnaire but is only available in German. Parts of the questionnaire are static, such as questions regarding sex, age, and selecting the general area of one's body where symptoms occur most prominently, while others are adapting to the previously asked questions. The dynamically changing questions are always asked expecting an answer from the set: "Yes", "No" and "Skip". According to NetDoktor, the system is continuously validated by medical professionals and is based on the medical database AMBOSS [1] and follows the medical guidelines of professional societies [2]. We assume this expert system to be our golden model, as it comprises curated knowledge of high quality and is fully deterministic. The latter makes it possible, to generate a decision tree from a person's (or persona's) interaction with the system, that is reproducible across multiple calls [1]. Figure 2 shows the tree generated from the interaction of our exemplary persona having a heart attack. The tree is to be read from top to bottom, starting with the first question as the root node. It should be noted, that the very first question "Um wen geht es?" (i.e. "Who is it about?"), was always answered by "jemand anderen" (i.e. "somebody else") for this study. Rectangles represent questions and the ellipses represent the respective possible answers to choose from. The node at the second to last level, which is denoted by "Mögliche Erkrankungen" (i.e. "possible diseases") symbolizes the retrieval of diagnoses from the database, while the leaf nodes on the bottom level signify the results of the query. In this exemplary case, the questions were answered to correspond to the symptoms of a heart attack for demonstration. However, we can also use Symptom-Checker to automatically and randomly traverse the questionnaire's tree-like structure to retrieve sets of symptoms and corresponding diagnoses. This allows for a scaleable framework for comparing other methods against a strong and valid baseline. Sets of symptoms and corresponding "golden" diagnoses are extracted from such a tree as follows: Firstly, for each path from the root node to the bottom level nodes (i.e. the diagnoses), questions-answer-pairs are stored in a JSON data structure. Each full path represents one set of symptoms. Secondly, each set of symptoms is summarized in a textual representation in German taking special care not to lose or add information. This is then translated from German to English. The first rows of Tables 3 and 4 contain the textual descriptions of our example in German and English, respectively. Lastly, the diagnoses provided by the golden model are extracted from the bottom layer (i.e. the leaf nodes) of the tree, which is always a set of three diagnoses. These sets of diagnoses are referred to as NetDoktor diagnoses for the remainder of this paper.

[1]As long as the underlying knowledge base does not change.

2.2 Evaluation Metric

The main evaluation metric used in this work is the overlap of diagnoses as compared to NetDoktor. A set of diagnoses is considered as being good if it contains a large overlap with the golden model diagnoses of NetDoktor. Since the NetDoktor baseline always yields three diagnoses, the highest overlap any other system can achieve is 3/3. Thus, the score ranges from 0/3 to 3/3. We explicitly do not normalize, although we want to compare sets of diagnoses with varying cardinalities. The reason for this is that yielding more diagnoses should not be penalized (as they might be worth considering, as well), and yielding fewer should not lead to a better score automatically.

2.3 Equivalence of Diagnoses

This study compares systems designed for direct interaction with humans. These systems' output is presented to the users in natural language. A key feature of medical language is its interchangeable use of semantically equivalent terminology originating from different languages such as Latin, German, or English. Additionally, when talking to patients, medical personnel often have to use simplified terminology, which includes the use of colloquial synonyms, hypernyms, and hyponyms. Hence, the semantic equivalence of diagnoses must be considered to ensure the comparability of different systems.

- Synonyms are terms, which can be used interchangeably with one another.
- Hypernyms are superordinate or umbrella terms of a term.
- Hyponyms are describing subordinate terms (i.e. more specific) or another term.

Table 2 lists diagnoses that are treated as equivalents for this study.

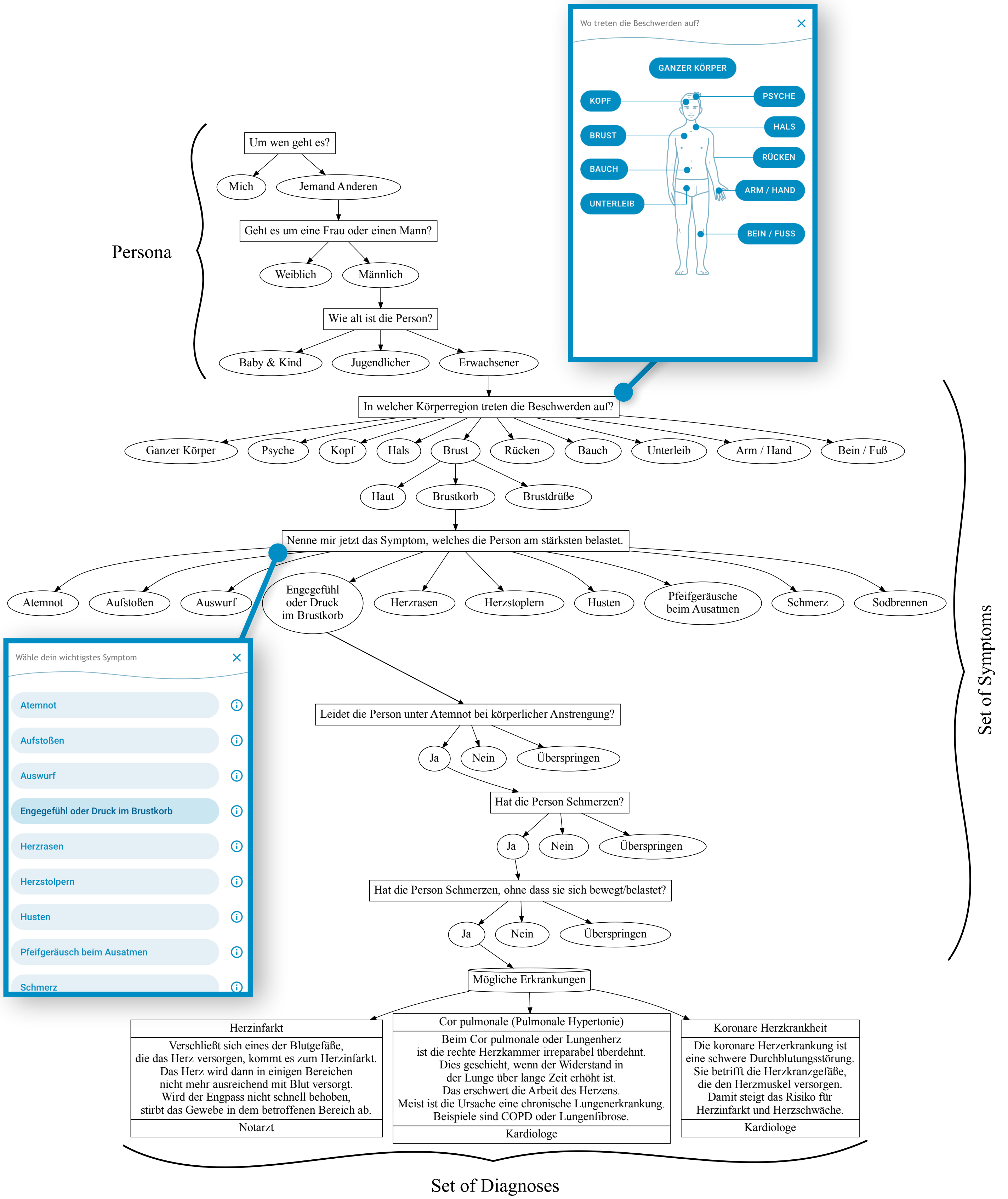


Figure 2: Golden Model: Exemplary Decision Tree Based on NetDoktor's Symptom-Checker Questionnaire [13] Filled-Out for a Persona Having a Heart Attack. Blue Boxes are Screenshots from Symptom-Checker Corresponding to Nodes in the Tree. We Set a Persona and Automatically Extract A) a Set of Symptoms and B) a Set of Diagnoses for Each Path From the Root Node to the Leaf Nodes on the Bottom-Most Level.

Table 2: Equivalent Diagnoses: Synonyms, Hypernyms, Hyponyms & Translations

Diagnosis	Equivalence (as Occurring in ChatGPT Output)
Herzinfarkt	Myokardinfarkt
	Akutes Koronarsyndrom
	Myocardial Infarction
	Heart Attack
Reiter-Syndrom	Reaktive Arthritis
	Morbus Reiter
	Reactive Arthritis
	Reiter's Syndrome
Kawasaki-Syndrom	Kawasaki Disease
	Kawasaki Syndrome
Blinddarmentzündung	Appendizitis
Vorhofflimmern	Herzrhythmusstörungen
Glutenunverträglichkeit	Zöliakie
Bakterielle Pharyngitis	Mild Bacterial Conjunctivitis with Pharyngitis
Krätze	Scabies
Erkältung	Virale Infekte
Pfeiffer-Drüsenfieber	Pfeiffersches Drüsenfieber
	Mononukleose
Blasenentzündung	Zystitis
	Harnwegsinfektion
	Urinary tract infection
Gürtelrose	Herpes Zoster
Mastopathie	Fibrozystische Mastopathie
Lipom	Lipoma

2.4 Sets of Symptoms & Personas

For this evaluation, we retrieved 12 sets of symptoms from NetDoktor - 6 for females and 6 for males, and for each sex, we used all of NetDoktor's 3 age categories (baby/child, adolescent, adult) twice. In addition, we used the exemplary set of symptoms for an adult man having a heart attack, as discussed in the previous section. This yields the 13 sets of symptoms listed in Tables 3 and 4. In Figure 2, the parts of the questionnaire are marked, which correspond to the persona and the set of symptoms respectively. In the following, both terms are used interchangeably.

2.5 Model, Prompts & Diagnose Retrieval

For all of our experiments, we used GPT-4o [9, 10] through ChatGPT [8]. More specifically, we used version GPT-4o-2024-08-06, which has been released in August 2024. We evaluate the same model in German and English and denote this with a trailing "[DE]" for German and "[EN]" for English for the respective results. We extended this convention to our golden model NetDoktor as well. The full list of prompts used can be found in the next section, Section 3. All LLM results were retrieved in a zero-shot methodology, without samples or additional context besides the prompt itself. Every symptom description is sent within a new chat to isolate individual queries. However, we cannot guarantee that we are indeed interacting with a "blank slate" as ChatGPT and GPT-4o are both black boxes and our user profile might interfere with the output.

Table 3: Sets of Symptoms per ID [DE]

ID	Description of Symptoms in German
1	Ein erwachsener Mann verspürt ein Engegefühl im Brustkorb. Er hat Schmerzen, auch wenn er sich nicht bewegt oder belastet. Außerdem leidet er unter Atemnot, wenn er sich anstrengt.
2	Ein Bub hat gerötete Augen und Fieber. Außerdem hat er Schmerzen beim Wasserlassen.
3	Ein kleiner Junge hat starke Bauchschmerzen, die bei Druck schlimmer werden. Es wurde auch festgestellt, dass er allgemein druckempfindlich ist.
4	Ein jugendlicher Junge verspürt ein Engegefühl oder Druck im Brustkorb. Er bemerkt, dass sein Herz sehr schnell schlägt und unregelmäßig arbeitet. Er fühlt sich oft müde und weniger leistungsfähig.
5	Ein jugendlicher Junge hat an Gewicht verloren und leidet unter anhaltender Müdigkeit. Er hat regelmäßig Durchfall, der besonders voluminös und übelriechend ist.
6	Ein erwachsener Mann hat eine Rachenentzündung und bemerkt gerötete Augen. Es gibt jedoch kein Fieber oder geschwollene Lymphknoten.
7	Ein erwachsener Mann hat entzündliche Hautveränderungen am Unterschenkel, die stark jucken, insbesondere nachts. Es wurde kein Zusammenhang mit Allergien festgestellt.
8	Ein kleines Mädchen hat seit einiger Zeit ihren Appetit verloren, fühlt sich ungewöhnlich müde und hat ungewollt an Gewicht verloren. Es wird auch über verminderten Urinfluss berichtet.
9	Ein kleines Mädchen hat Fieber, eine Rachenentzündung und geschwollene Lymphknoten. Sie fühlt sich abgeschlagen und schwitzt besonders nachts stark.
10	Eine jugendliche Mädchen hat Blut im Urin und Schmerzen beim Wasserlassen. Der Harndrang ist häufig, aber es wird nur eine geringe Urinmenge ausgeschieden. Zudem verspürt sie ein Brennen beim Wasserlassen.
11	Ein jugendliches Mädchen klagt über ausstrahlende Schmerzen im Nackenbereich und hat einen Hautausschlag mit kleinen Bläschen.
12	Eine erwachsene Frau hat Spannungsgefühle in der Brust und tastet schmerzlose Knoten. Die Haut ist nicht gerötet.
13	Eine erwachsene Frau verspürt Druckempfindlichkeit im Oberbauch, die Haut wölbt sich vor und die Region ist geschwollen.

3 Experimental Evaluation

Figure 1 depicts our experimental setup: NetDoktor is used as a golden model to automatically derive sets of symptoms and corresponding diagnoses as exemplified in Figure 2. The extracted symptoms are then used as input to the LLM GPT-4o via ChatGPT. ChatGPT diagnoses are then compared to NetDoktor diagnoses to compute an overlap score. Figure 4 gives an overview of our evaluation results. For each set of symptoms, NetDoktor results are shown, followed by four diagnosis strategies utilizing ChatGPT. The grey bars denote the cardinality of every resulting set of diagnoses. Blue overlays are used to show the overlap between NetDoktor diagnoses and ChatGPT diagnoses. These overlays correspond to the values in Table 5, which comprises the occurrences of overlaps in each category from 0/3 to 3/3. In addition to the 13 sets of symptoms, Figure 4 and Table 5 include averages computed over all sets for easier comparison of the prompts/-diagnosis retrieval methods. In the following, you can find the used prompts/methodologies corresponding to the depicted bars:

A **NetDoktor [DE]:** Diagnoses from NetDoktor were retrieved via the Symptom-Checker questionnaire as is documented in Subsection 2.1. This is our golden model and

Table 4: Sets of Symptoms per ID [EN]

ID	Description of Symptoms in English
1	An adult man feels a tightness in his chest. He experiences pain even when he is not moving or exerting himself. Additionally, he suffers from shortness of breath when he exerts himself.
2	A boy has red eyes and a fever. He also has pain when urinating.
3	A little boy has severe abdominal pain, which worsens with pressure. It was also found that he is generally sensitive to pressure.
4	A teenage boy feels a tightness or pressure in his chest. He notices that his heart beats very fast and irregularly. He often feels tired and less capable.
5	A teenage boy has lost weight and suffers from persistent fatigue. He has regular diarrhea that is particularly voluminous and foul-smelling.
6	An adult man has a throat infection and notices red eyes. However, there is no fever or swollen lymph nodes.
7	An adult man has inflammatory skin changes on his lower leg that itch intensely, especially at night. No connection with allergies was found.
8	A little girl has lost her appetite for some time, feels unusually tired, and has unintentionally lost weight. Reduced urine output is also reported.
9	A little girl has a fever, a throat infection, and swollen lymph nodes. She feels weak and sweats heavily, especially at night.
10	A teenage girl has blood in her urine and pain when urinating. The urge to urinate is frequent, but only a small amount of urine is passed. She also feels a burning sensation when urinating.
11	A teenage girl complains of radiating pain in the neck area and has a rash with small blisters.
12	An adult woman has a feeling of tension in her breast and can feel painless lumps. The skin is not reddened.
13	An adult woman feels tenderness in the upper abdomen, the skin bulges, and the area is swollen.

overlaps with its diagnoses are marked in blue, in Figure 4. The questionnaire and results are in German.

B **ChatGPT [DE]:** Ad-hoc query sent to ChatGPT using the symptom descriptions in German from Table 3 as is.

C **ChatGPT [DE] "3 Most Likely":** More elaborate query sent to ChatGPT using the symptom descriptions in German from Table 3, additionally requesting the "3 most likely" diagnoses.

D **ChatGPT [DE] "10 Most Likely":** More elaborate query sent to ChatGPT using the symptom descriptions in German from Table 3, additionally requesting the "10 most likely" diagnoses.

E **ChatGPT [EN]:** Ad-hoc query sent to ChatGPT using the symptom descriptions in English from Table 4 as is.

Out of convenience, the letters introduced in this list are used when referring to a specific prompt in the following paragraphs. The main takeaway from this evaluation is that none of the used prompts achieves a complete overlap of 3/3 with NetDoktor for any of the personas. Prompt B, achieves the highest score, with 0.92/3 i.e. 31%. For our small test set of 13 sets of symptoms, these results constitute from 2/3 for two, and 1/3 for seven sets of symptoms. The same prompting strategy in English, denoted by E, yields worse results, having overlaps of 2/3 for one, and 1/3 for three sets of symptoms. The two prompts giving ChatGPT the task of answering with the "n most likely" diagnoses are equally not performing as well as the simple prompt in German: C achieves 1/3 for 9 sets of symptoms and D yields 2/3 for three and 1/3 for five sets of symptoms. This means that prompt D,

Die Symptome, die der Mann beschreibt, könnten auf verschiedene medizinische Zustände hinweisen, von denen einige ernsthafte medizinische Notfälle darstellen könnten. Die wichtigsten Punkte sind das Engegefühl im Brustkorb, Schmerzen, die unabhängig von Bewegung oder Belastung auftreten, und Atemnot bei Anstrengung.

Einige mögliche Ursachen für diese Symptome sind:

1. **Angina Pectoris**: Dies ist ein Zeichen dafür, dass das Herz nicht genügend Sauerstoff erhält, oft aufgrund einer Verengung oder Blockierung der Herzkranzgefäße. Die Schmerzen und das Engegefühl im Brustkorb können bei körperlicher Anstrengung oder emotionalem Stress auftreten.
2. **Herzinfarkt (Myokardinfarkt)**: Wenn der Schmerz anhaltend ist und nicht nachlässt, könnte dies ein Hinweis auf einen Herzinfarkt sein. Dabei handelt es sich um eine akute Blockierung einer Koronararterie, die die Sauerstoffversorgung des Herzmuskels unterbricht. Dies ist ein medizinischer Notfall.
3. **Lungenembolie**: Dies tritt auf, wenn ein Blutgerinnsel eine Lungenarterie blockiert, was zu Atemnot und Brustschmerzen führt. Auch dies ist ein ernstzunehmender Notfall.
4. **Pleuritis**: Eine Entzündung der Pleura (Lungen- oder Brustfell) kann ähnliche Symptome verursachen, insbesondere wenn die Schmerzen durch tiefe Atemzüge oder Husten verschlimmert werden.
5. **Gastroösophageale Refluxkrankheit (GERD)**: Auch Sodbrennen kann starke Brustschmerzen verursachen, die manchmal mit Herzproblemen verwechselt werden können. Die Symptome können durch das Liegen oder nach einer Mahlzeit verstärkt werden.
6. **Panikattacken**: Psychische Ursachen wie Panikattacken können ebenfalls ähnliche Symptome verursachen, einschließlich Engegefühl im Brustkorb, Schmerzen und Atemnot. Allerdings gibt es hier meist auch andere begleitende Symptome wie starkes Zittern, Schweißausbrüche und ein Gefühl des bevorstehenden Unheils.

Es ist wichtig, dass der Mann so schnell wie möglich medizinische Hilfe sucht, besonders wenn die Symptome plötzlich auftreten, sich verschlimmern oder mit anderen Anzeichen wie Übelkeit, Schwitzen oder Ohnmachtsgefühl einhergehen. Ein Arzt oder das medizinische Personal kann eine genauere Diagnose stellen und die geeignete Behandlung einleiten.

Figure 3: ChatGPT Output for ID 1, Using Prompt B. Overlap with NetDoktor Marked in Blue. Compare to Figure 2.

asking for the "10 most likely" diagnoses is the runner-up with an average of 0.85/3 i.e. 28%. Surprisingly, the simple prompt in English, E, performs poorest, which contradicts our hypothesis of English prompts performing better.

Table 5: Overlaps of Diagnoses with NetDoktor per Prompt

Score	Diagnosis Retrieval Method				
	A	B	C	D	E
3/3	13	0	0	0	0
2/3	0	2	0	3	1
1/3	0	8	9	5	3
0/3	0	3	4	5	9
Avg.	3/3	0.92/3	0.69/3	0.85/3	0.38/3
Avg.[%]	100%	31%	23%	28%	13%

Apart from the overlaps, other interesting observations can be made on closer inspection of the results: ChatGPT seems to rigorously follow the instruction to generate n diagnoses and as such, yields consistently 3 diagnoses for prompt C and 10 diagnoses for prompt D. However, it can be doubted that "most likely" is interpreted in a scientifically backed manner, as ChatGPT often does not include even one of the NetDoktor diagnoses and not once all of them. Equally interesting is the inclusion of the necessity to consult a doctor in one form or the other at the end of every result we received, which is likely due to being "hard-coded" for legal reasons on the part of OpenAI. This can also be seen in Figure 3. Although ChatGPT and GPT-4o are black boxes and LLMs are non-deterministic, we try to document our reported results as well as possible for replication. You can find all of our

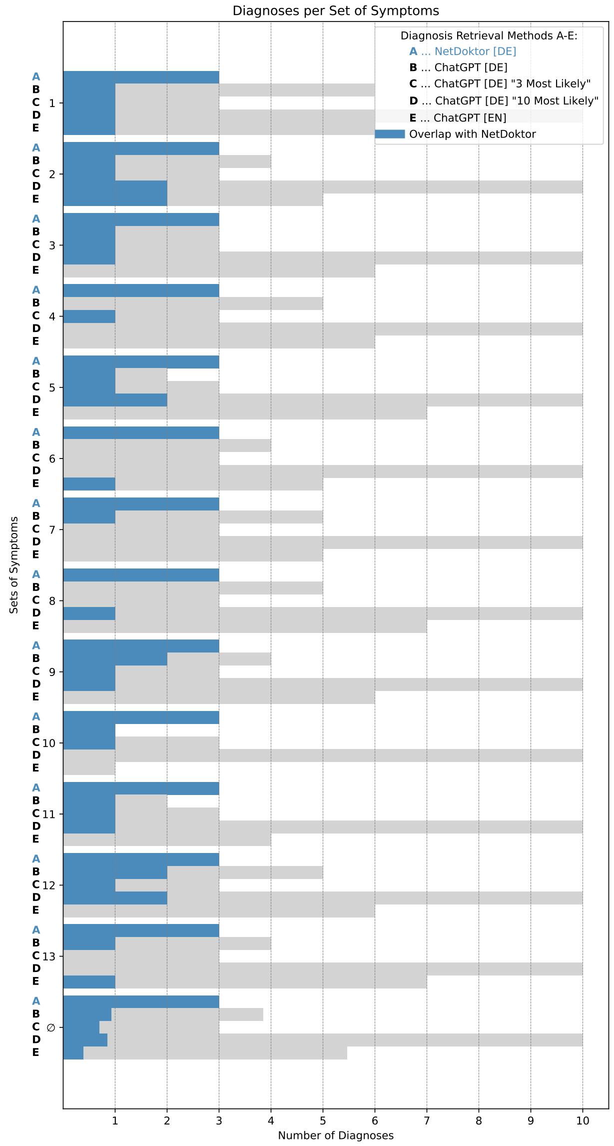


Figure 4: Comparison of Diagnoses for Symptoms Seen in Table 3

experimental results as a replication package under the provided URL [2].

4 Conclusions

In summary, ChatGPT diagnoses only partially match the diagnoses retrieved from our golden model NetDoktor. ChatGPT diagnoses are mostly well-structured and are seemingly valid but fail to include all NetDoktor diagnoses in any of the tested cases. This holds for all tested prompts and across all tested symptoms. The highest overlap, on average, could be achieved with the simplest prompt in German, giving only a description of the symptoms for a persona. When asked for a specific amount of "most likely" diagnoses, ChatGPT always delivered exactly the asked-for number of diagnoses. However, this does not benefit the quality of the output as measured by the overlap metric. Neither does an interaction in English change the output quality for the better. In our tests, ChatGPT always includes a notice to consult a doctor. Human assessment of the diagnoses cannot be fully bypassed by the proposed evaluation methodology. This is due to the immanent presence of semantic equivalence and the necessary medical knowledge to find those equivalences. Although such a task is automatable via LLMs as well, the authors of this paper underline the potential implications for undermining the quality of an evaluation, when fully automated. While our evaluation reports results achieved using ChatGPT and GPT-4o, the proposed methodologies transcend to other LLMs as well. As part of future work, we want to repeat our experiments at a larger scale to achieve representative results. Additionally, we want to consider stability metrics, as seen in [11]. Another interesting direction can be further analysis of the relationship between prompt (engineering) and the retrieval of matching diagnoses as well as their stability. Finally, it would be interesting to compile a corpus of medical symptoms corresponding to diagnoses including named entities and logical abstractions to perform evaluations as seen in [12] on the medical domain.

[2]https://zenodo.org/doi/10.5281/zenodo.13765345

Acknowledgements

The work presented in this paper was partially funded by the European Union under Grant 101159214 – ChatMED. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] AMBOSS GmbH. 2024. Amboss. https://www.amboss.com. Accessed: 2024-09-03. (2024).
[2] AWMF. 2024. Arbeitsgemeinschaft der wissenschaftlichen medizinischen fachgesellschaften (awmf) - leitlinien. https://www.awmf.org/leitlinien. Accessed: 2024-09-03. (2024).
[3] Tom B. Brown et al. 2020. Language models are few-shot learners. arXiv: 2005.14165 [cs.CL]. (2020).
[4] Zeming Chen et al. 2023. Meditron-70b: scaling medical pretraining for large language models. (2023). https://arxiv.org/abs/2311.16079 arXiv: 2311.16079 [cs.CL].
[5] J. Clusmann et al. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3, 141. DOI: https://doi.org/10.1038/s43856-023-00370-1.
[6] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.
[7] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. (2023). https://arxiv.org/abs/2303.13375 arXiv: 2303.13375 [cs.CL].
[8] OpenAI. 2023. ChatGPT. (2023). chat.openai.com/chat.
[9] OpenAI. 2023. GPT-4 technical report. arXiv: 2303.08774 [cs.CL]. (2023).
[10] OpenAI. 2024. Introducing gpt-4o and more tools to chatgpt free users. (May 2024). https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/.
[11] Alexander Perko and Franz Wotawa. 2024. Evaluating openai large language models for generating logical abstractions of technical requirements documents. In *Proceedings to the 24th International Conference on Software Quality, Reliability, and Security (QRS)*. IEEE.
[12] Alexander Perko, Haoran Zhao, and Franz Wotawa. 2023. Optimizing named entity recognition for improving logical formulae abstraction from technical requirements documents. In *Proceedings to the 10th International Conference on Dependable Systems and Their Applications*.
[13] Jens Richter, Hans-Richard Demel, Florian Tiefenböck, Luise Heine, and Martina Feichter. 2024. Symptom-checker. https://www.netdoktor.at/symptom-checker/. Accessed: 2024-09-03. (2024).
[14] Khaled Saab et al. 2024. Capabilities of gemini models in medicine. (2024). https://arxiv.org/abs/2404.18416 arXiv: 2404.18416 [cs.AI].
[15] Karan Singhal et al. 2023. Towards expert-level medical question answering with large language models. (2023). https://arxiv.org/abs/2305.09617 arXiv: 2305.09617 [cs.CL].
[16] Hugo Touvron et al. 2023. Llama 2: open foundation and fine-tuned chat models. *ArXiv*.

HomeDOCtor App: Integrating Medical Knowledge into GPT for Personal Health Counseling

Matic Zadobovšek
matic.zadobovsek@gmail.com
Faculty of Computer and Information Science
Ljubljana, Slovenia

Primož Kocuvan
primoz.kocuvan@ijs.si
Department of Intelligent Systems, Jožef Stefan Institute
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Department of Intelligent Systems, Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

The escalating workload demands on healthcare professionals are leading to systemic overload, resulting in a decline in the efficiency of public health services. This situation necessitates the development of solutions that can alleviate the burden on physicians while ensuring comprehensive patient care. Recent advancements in generative artificial intelligence, particularly in the field of medicine, have demonstrated that large language models (LLMs) can outperform doctors in specific tasks, highlighting their potential as valuable tools for reducing the strain on healthcare providers. This study focuses on the development of the HomeDOCtor application, which integrates additional validated medical knowledge into the GPT-4o LLM. The objective of this application and the enhanced LLM is to offer users reliable access to a medical chatbot capable of providing accurate and timely responses to health-related inquiries. The chatbot's behavior has been meticulously tested and refined in collaboration with a team of physicians. The findings of this research offer insights into the development of such systems and explore their potential application within the Slovenian healthcare system.

Keywords

large language models, GPT-4o, vector embeddings, vector databases, natural language processing, prompt engineering

1 Introduction

There are worldwide growing problems in the health sector due to an ageing population and a shortage of health experts [13, 17]. The field of Natural Language Processing (NLP) has recently seen an increase in the number of LLMs being customised for different domains [2]. Increasingly, we are also seeing their integration into the field of medicine, which is one of the core domains of today's society [14, 22, 11, 20]. Currently, OpenAI's GPT-4o is considered to be the most powerful LLM, which also performs best in the area of health questionnaires and other related tasks [16, 12, 1, 5].

Because of their advanced ability to understand natural language text, there are many potential applications [8, 3]. Patients can talk to the GPT, describe their symptoms in detail, include documents to past treatments and then get a friendly response in seconds. This is particularly useful when the patient is unable to access their personal doctor or simply wants to get feedback before taking further action. LLMs are trained on a huge amount of data, but there is the possibility of hallucination, especially in less researched areas and specific issues [19]. As a result, it makes sense to think about augmenting the existing LLM with verifiable resources that can help to improve it. One of the most important issues regarding the use of LLMs relates to the aspect of secure use of health data. When using commercial LLMs, the data is passed to them and the user should be informed with appropriate disclaimers about the use and processing of the data they have entered. In an ideal situation, we would use our own LLM, running on a local server, and have full control over the implementation, but this raises different issues. Creating your own LLM is a time-consuming and costly process, so our research focused on extending the best LLM currently on the market, GPT-4o.

In our case, we focused on the Slovenian healthcare sector and the adaptation of the LLM to the needs of the average user. The aim was to develop an application (HomeDOCtor) that would allow all Slovenian citizens to get medical help 24/7 without the need to contact a personal doctor. As a result, performance testing is crucial, as even the smallest errors can lead to serious consequences [8].

The core research hypothesis posits that the system, enhanced with additional modules integrated into a GPT, will provide more effective medical advice to the general Slovenian population compared to existing GPTs.

In section 2 we present all the datasets used to extend the LLM. Section 3 systematically describes the approach and technologies used to develop the software solution. Section 4 shows the results of the developed solution and an example of a user conversation with the improved LLM. Section 5 describes how performance was tested and how doctors helped to guide LLM to get the desired outcome. Section 6 presents concluding thoughts and possible improvements.

2 Datasets

To improve the knowledge of the existing ChatGPT-4o, data from verified sources was obtained. The addition of new data allows the GPT to answer questions using its existing knowledge, as well as to address a broader range of questions that require specialized knowledge. In this case, it is about integrating information on Slovenian healthcare [4].

2.1 Insieme Platform

The Insieme platform is the core building block of the dataset used [9]. The platform integrates hand-crafted expert-based knowledge that is accessible to users on all devices and contains basic information on Slovenian healthcare.

The information is organised hierarchically by medical specialty. By clicking (or visiting) on one of the branches, the user is redirected to a sub-page that provides an overview of the diseases and other services that belong to the selected field of medicine.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.chtm.8

Each of the diseases is accompanied by a set of key data that allows the user to get everything they need in one place. This includes:

- basic information about the diagnosis,
- professional help (list of clinics in Slovenia),
- apps (where the user can upload their pictures, ask for opinion, etc.),
- associations in Slovenia,
- articles,
- forums,
- video content, and
- image content.

Figure 1 shows an example of one of the pages illustrating the display of the information for skin cancer.

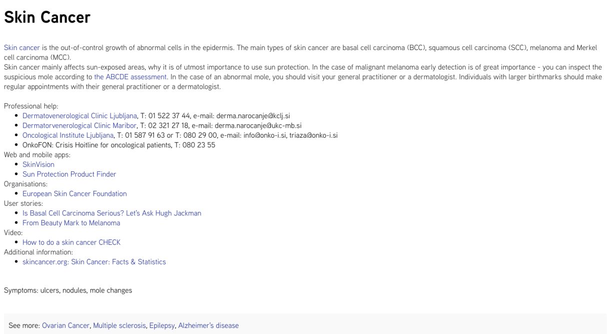
All data have been compiled in collaboration with doctors and are therefore suitable for use as a primary source for the extension of the LLM. The aim is to enable users of the platform to quickly obtain information on Slovenian healthcare that they can understand.

2.2 Other resources

The Med.Over.Net Health Forum was used as one of the two remaining data sources. The forum is divided into different branches of medicine where users can create new conversations and ask questions within them. Each of the sub-forums has an assigned moderator (a doctor) who carefully reviews and answers the questions. All other forum users can also join the conversation. All data from the online forum was extracted using web scraping, whereby all extracted data was stored in a single PDF file.

The second source is the Home Doctor Encyclopaedia (The Home Doctor - Practical Medicine for Every Household), which is in the field of practical home medicine. It contains useful tips on how people can help themselves by applying helpful advice without the need to visit a doctor.

In the future, consideration will be given to adding new resources, in particular a better medical book suitable for Slovenian healthcare. For now, we have focused on the three data sources mentioned above.



Skin Cancer

Skin cancer is the out-of-control growth of abnormal cells in the epidermis. The main types of skin cancer are basal cell carcinoma (BCC), squamous cell carcinoma (SCC), melanoma and Merkel cell carcinoma (MCC).
Skin cancer mainly affects sun-exposed areas, why it is of utmost importance to use sun protection. In the case of malignant melanoma early detection is of great importance - you can inspect the suspicious mole according to the ABCDE assessment. In the case of an abnormal mole, you should visit your general practitioner or a dermatologist. Individuals with larger birthmarks should make regular appointments with their general practitioner or a dermatologist.

Professional help:
- Dermatovenerological Clinic Ljubljana, T: 01 522 37 44, e-mail: derma.narocanje@kclj.si
- Dermatorvenerological Clinic Maribor, T: 02 321 27 18, e-mail: derma.narocanje@ukc-mb.si
- Oncological Institute Ljubljana, T: 01 587 91 63 or T: 080 29 00, e-mail: info@onko-i.si, triaza@onko-i.si
- OnkoFON: Crisis Hoitline for oncological patients, T: 080 23 55

Web and mobile apps:
- SkinVision
- Sun Protection Product Finder

Organisations:
- European Skin Cancer Foundation

User stories:
- Is Basal Cell Carcinoma Serious? Let's Ask Hugh Jackman
- From Beauty Mark to Melanoma

Video:
- How to do a skin cancer CHECK

Additional information:
- skincancer.org: Skin Cancer: Facts & Statistics

Symptoms: ulcers, nodules, mole changes

See more: Ovarian Cancer, Multiple sclerosis, Epilepsy, Alzheimer's disease

Figure 1: Image of the skin cancer disease subpage on the Insieme web platform.

3 Methodology

This section presents the design of a system that extends the LLM GPT-4 with acquired knowledge resources and outlines the architecture of the developed application.

3.1 Fine-tuning

One solution is fine-tuning, which allows a pre-trained LLM to add external data sources to a LLM that already has its own existing knowledge [10]. The idea is that instead of creating an LLM from scratch, an existing LLM is used and further adapted to the user's operational needs. Fine-tuning typically works on the principle of supervised learning, where the LLM needs to be given examples of the input and the expected output. With additional learning, the LLM is able to use the new data. This directly influences the change of parameters (weights) in the model and its performance. The problem arises in that, due to the large amount of parameters in modern LLMs, the whole learning process may be impractical due to the cost of training an outsourced LLM. Additionally, any incorporated data is static, meaning that the learning process must be repeated whenever adding new data. As a result, fine-tuning is not the best solution in cases where we know that the data will change, and we will be constantly adding new data.

3.2 Storing data

The data obtained by web scraping was stored in a vector database, which allows efficient management of the vector data [21].

The raw data obtained (e.g. PDF and Markdown files) are unsuitable for direct use with LLMs as they cannot be directly managed by the LLM. While documents can be added as attachments in the web interfaces of platforms such as ChatGPT (OpenAI's web LLM access), this is only useful in the context of a single conversation. In our case, we would like the LLM to have permanent access to information sources without the need for manual addition of documents by the user.

3.2.1 Vector embeddings. Each of the documents is converted into vector embeddings, which allow the text to be presented in a numerical notation that can be understood by a computer [6]. From this notation, the computer can understand the meaning of the text and determine the meaning between the words in the input. This is useful in many NLP tasks, e.g. search engines, sentiment analysis, recommendation systems and so on. We use vector embeddings to be able to extract information relevant to the corresponding question based on user input.

Modern LLMs have a limitation in the form of a context window, which determines how much text they can process within a single user input. Because of this limitation, we need to be careful what additional information is sent to the LLM, as we could quickly exceed the context window by sending all data sources. E.g. if the user's question is about skin cancer, we only want to get information about that specific disease. As a consequence, we have divided the text into chunks, where each chunk contains data for only one of the diseases. Similar care has been taken with the encyclopaedia and the Med.Over.Net online forum. Meaningful paragraphs have been grouped together to form one chunk of text. Thus, only the key chunk that is most likely to contain the answer to the question asked will be provided to the LLM.

For each of the chunks, a vector embedding has to be created using an appropriate model (e.g. text-embedding-ada-002 from OpenAI). The vector databases (e.g. Redis) have to be used to store the resulting vector embeddings properly.

3.2.2 Vector databases. The vector database allows the storage of unstructured data and fast retrieval due to efficient indexing of the data. In our case, we use them to store individual chunks and their associated vector embeddings [7].

Over all stored chunks, vector search is enabled, which means that the chunks whose vector embeddings most closely match the query vector are returned. The query vector represents the user input, which is converted into vector format. Cosine similarity, Euclidean distance, inner product and other metrics can be used to measure similarity between vectors.

The configuration used to retrieve the chunks can be modified: e.g. specifying the required similarity threshold and the number of chunks retrieved.

3.3 RAG

An alternative solution is Retrieval-Augmented Generation (RAG), where the LLM calls an external database containing all our data when the user provides a question [15]. Relevant data is retrieved from there and passed directly to the LLM, which uses this data in the generation of the answer. In this case, the basic architecture of the LLM used remains completely unchanged, as it accesses a separate building block to retrieve the data. It is a cheaper solution that allows dynamic data extraction. Data can be deleted, modified and newly added to the database at any time.

The Retriever is responsible for retrieving relevant pieces of text from the vector database. The user question is converted into vector embedding, and then the most similar chunks are retrieved to help guide the LLM to the correct answer. A merged query is then created containing the original user question, the extracted chunks of text and any other system instructions given. In the final step, the LLM generation produces an answer based on the query, which is passed to the user.

Compared to fine-tuning, RAG allows the system to change continuously (adding new knowledge sources), makes the operation more understandable (we can check which pieces of text have been passed to the LLM) and reduces the possibility of hallucination (verified pieces of text from the selected domain are added). The style of writing the answers cannot be changed directly, as the parameters of the original model remain completely unchanged, but we can help by writing system instructions. System instructions further guide the operation of the chatbot and play a very important role in achieving a user-friendly behaviour of the system.

All of these features make RAG a suitable choice for enriching the work of an existing LLM with validated information in the chosen field.

The schematic design of the RAG system can be seen in the figure 2 below.

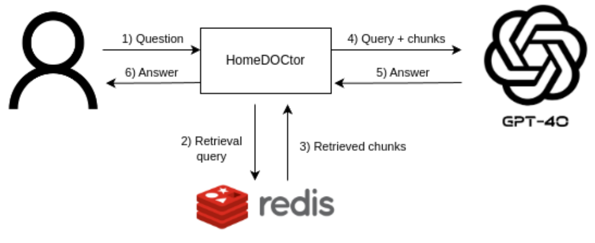


Figure 2: Demonstration of how the RAG system works.

3.4 Application architecture and used technologies

A brief overview of the technologies used to develop the software solution is presented below.

3.4.1 Flutter. The Flutter framework was used to develop the front-end part of the system. The reason for using it is that it allows us to develop applications that work on all platforms by writing unified code in the Dart programming language. It then gets compiled into code for the dedicated platform. This way, the application works on web browsers as well as in the form of a mobile application. Otherwise, we would have to use multiple programming languages to ensure support on multiple devices, which would slow down development.

3.4.2 Redis. The Redis database was used to store all additional knowledge resources. It is a NoSQL database that stores data in key-value format and also works as a vector database. The data is stored in memory, which helps significantly with data retrieval and overall application responsiveness.

3.4.3 LangChain. The LangChain framework offers a number of functionalities that facilitate the development of applications that involve working with LLMs [18]. It allows us to easily connect to a vector database, retrieve documents, call arbitrary LLMs, modify system instructions and so on.

3.4.4 FastAPI. The FastAPI framework allows the creation of RESTful APIs that can be accessed by the HomeDOCtor application. It uses Uvicorn for the server implementation and allows processing of requests received from users. These requests are related to the processing of the received questions and the execution of all the necessary tasks to obtain the appropriate response from the LLM.

3.4.5 Docker. The entire application was packaged in a Docker container to facilitate development and deployment to the server. All the necessary configuration is included in the container, which runs on the system without the need to install the rest of the programming libraries and tools.

3.4.6 Firebase. The Firebase platform offers a number of useful functionalities such as user authentication, application usage analytics, cloud storage and a real-time database (NoSQL cloud database, which in our case is used to store the necessary user data).

4 Results

The HomeDOCtor app allows users to access a chatbot that is constantly available to answer their questions. Users can access the app on their browsers or via a mobile app. The user interface allows users to view the history of the conversation, ask questions, include their own documents as an attachment (e.g. PDF documents of previous exams), get a PDF summary of the conversation and additional information in three side windows. By clicking on the "End session" button at the bottom of the screen, all data is deleted, and the user is redirected to the main page where they can start a new conversation with the chatbot.

Initially, HomeDOCtor determines what kind of conversation is being had. In the case of a simple chat, no additional information is displayed. However, if the system detects that the conversation is about the user's health, additional information is displayed in three side windows. This additional information is dynamically updated according to the progress of the user's conversation with the chatbot. The first window gives a list of possible medical conditions, the second window gives advice on how the user could help himself, and the third window tells whether a visit to the doctor would be necessary. This way, the user can receive a comprehensive overview of the conversation so far and get an opinion that adapts to the course of the conversation.

Figure 3 below shows the user interface of the HomeDOCtor application. We can see the user's conversation with the chatbot in Slovenian and the display of additional information in the three side windows along the right edge. When answering questions, GPT-4o uses its existing knowledge, incorporating as often as possible additional sources of knowledge (especially from the Insieme platform).

Using the app, users get instant help from a chatbot. The chatbot first tries to collect sufficient data before proceeding to further advice. In the course of the treatment, the user is advised on different forms of help (websites, questionnaires, forums, apps, visits to pharmacies, etc.). A visit to the doctor is only advised in urgent cases, when HomeDOCtor considers that the situation is too complicated or further medical treatment is needed.

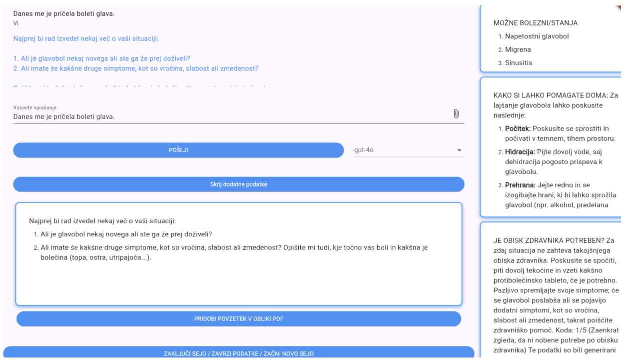


Figure 3: An example of a user talking to the HomeDOCtor chatbot. Additional information is visible on the side.

An example of a conversation from a user who consulted the developed chatbot is shown below. We can see the user's questions and all the corresponding answers.

- *User:* I have a slight headache and a bit of a sore throat today.
- *HomeDOCtor:* I understand. Let's gather some more information to better assist you.
  1. How long have you been experiencing these symptoms?
  2. Do you have any other symptoms, such as fever, cough, or body aches?
- *User:* I've been feeling this way for about two days now. It started with just a mild sore throat, and the headache came yesterday.
- *HomeDOCtor:* Thanks for the information.
  1. How severe are your headache and sore throat on a scale from 1 to 10?
  2. Have you noticed any other symptoms, like nasal congestion, fatigue, or body aches?
- *User:* I'd say my headache is about a 4 out of 10. It's more of a dull ache, not too intense, but definitely noticeable. The sore throat is a bit worse, probably a 6 out of 10.
- *HomeDOCtor:* Thank you for the details.
  1. Have you taken any over-the-counter medications or home remedies for relief? If so, have they helped?
  2. Is there anything that seems to make your symptoms better or worse, such as certain foods, drinks, or activities?
- *User:* I tried drinking some warm tea with honey, which soothed my throat for a while. My headache seems to get a bit worse when I'm looking at screens for too long.
- *HomeDOCtor:* Thank you for the additional information. It sounds like you might be experiencing symptoms of a common cold or another mild infection. Here are a few recommendations:
  1. Rest and Hydration: Ensure you get plenty of rest and stay well-hydrated.
  2. Warm Teas: Continue drinking warm teas with honey to soothe your throat.
  3. Limit Screen Time: Try to reduce your screen time to help with your headache.
  4. Steam Inhalation: Consider steam inhalation to relieve any nasal congestion you might have.
  For additional relief, you can visit a pharmacy for over-the-counter options. You can find more information about medications at https://mediately.co/si and http://www.cbz.si.
  If your symptoms worsen or persist for more than a week, it's advisable to consult a healthcare professional.
- *User:* Thank you for the advice. Where could I get more information?
- *HomeDOCtor:* You're welcome! For more information on common conditions and treatments, you can check these resources:
  1. Med.Over.net: https://med.over.net
  2. Ars Pharme - Zdravstveni nasveti: https://www.arspharmae.com
  3. Triglav Zdravstveni Nasveti: http://www.triglavzdravje.si
  If you need more personalized advice, consider visiting a pharmacist or consulting with a healthcare professional.

5 Evaluation

The full operation of the HomeDOCtor app and chatbot was carried out by the authors and medical experts. After consulting with doctors and identifying the key performance points, iterative and collaborative development followed, gradually introducing changes and improving the performance of the chatbot.

Doctors have contributed significantly to improving the app through rigorous testing, ensuring that the user can get the best possible results to their questions. We have been careful to address situations where a doctor's visit is not necessary. In many cases, non-urgent situations only add to the workload of doctors. In our system, we have tried to avoid this as much as possible by trying to refer the user to other resources that would be good and useful enough for them, and only as a last resort recommending a visit to the doctor. Of course, a visit to the doctor should not always be avoided, so it is crucial that the system is able to correctly identify scenarios where this is really necessary. These are mainly cases where additional diagnostics are needed, which are difficult for the LLM to perform on its own. Doctors have been particularly helpful in the project by further adapting the system instructions, which detail how to deal with all scenarios.

We compared HomeDOCtor with other large language models on the market, but focused the most on comparing it with the regular GPT-4o model, which our system uses for its basic operation. The main differences are in the flow of the conversation, as HomeDOCtor is customised by the system instructions to follow specific scenarios and ask questions that are key for providing further advice. Thus, by simulating a conversation of a user who starts the conversation by listing symptoms, a comparison between the two systems can be made. According to the clinicians' evaluations, HomeDOCtor appeared to be better at asking sub-questions and guiding the user through the treatment. A further important difference can be seen in the display of videos and images, as HomeDOCtor can display them right within the app, whereas GPT-4o creates fictional images or often provides links to non-existent videos. HomeDOCtor also often gives practical advice on how users can help themselves at home and carefully evaluates when it is really necessary to redirect the user to professional help.

Testing of the system has demonstrated the added value of the included knowledge sources, though formal validation tests are still pending. Nevertheless, the scientific hypothesis that superior performance can be achieved compared to the original GPT is increasingly supported by the evidence. Through iterative testing by all participants, the system has now reached a level of maturity suitable for deployment in a production environment. Test users will be able to provide feedback and suggest corrections via a form, which will further refine the chatbot to better meet the needs and preferences of the average user in Slovenia. This will display the data sources used to the test users, who will be able to indicate whether they believe that the relevant documents have been extracted from the vector database.

6 Conclusion

In this study, we have shown how verified information can be incorporated into one of the existing LLMs. It was an extension of GPT-4o, which currently achieves the best results, and has been further adapted to the needs of Slovenian public health in the context of the development of the HomeDOCtor application.

Comprehensive testing has been carried out with a team of doctors, which has helped to ensure that users are provided with relevant and understandable information. HomeDOCtor initially gathers information by asking questions, and only then advises on further action. External resources and self-help methods are prioritised, as the intention is that referral to a doctor should only be made in urgent situations.

Possible improvements are in the use of one of the open source LLMs, which would negate the need for external access to OpenAI's LLMs. Currently, in order to run the HomeDOCtor application, an API call is sent for each question asked, for which a fee is paid to OpenAI. This is also problematic from a data protection point of view as it involves working with confidential user data which should not be passed on (altough you are default opt-out for OpenAI API). Currently, this is taken care of with disclaimers and warnings that the user has to agree to before using the application. At the same time, the use of open source models would bring additional problems, as the appropriate infrastructure would have to be established. Open source models also typically perform worse than e.g. GPT-4o. As a result, we have chosen to use GPT-4o, as we cannot afford worse outcomes due to the criticality of the medical domain.

HomeDOCtor could also be adapted to meet the healthcare needs of countries abroad, but this would require ensuring that the relevant data is obtained in accordance with their national guidelines and security laws. The behaviour of the chatbot could remain largely the same, as GPT-4o can by itself convert between many languages. The key component would thus be to obtain all the necessary country-specific information that we want to make available to users (information on clinics, apps, articles, video content in that language).

In the future, work will be carried out on improving the chatbot, taking into account the opinions of external users, who in this case are also our target group. This research has shown that generative artificial intelligence has a great potential application in the field of medicine and could make a significant contribution to relieving the burden on the healthcare system.

Acknowledgements

We thank members of the Department of Intelligent Systems and medical experts for providing info and testing the system. This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214).

References

[1] Josh Achiam et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
[2] Jan Clusmann et al. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3, 1, 141.
[3] Matjaž Gams and Tine Kolenik. 2021. Relations between electronics, artificial intelligence and information society through information society rules. *Electronics*, 10, 4, 514.
[4] Matjaž Gams, Maj Smerkol, Primož Kocuvan, and Matic Zadobovšek. 2024. Developing a medical chatbot: integrating medical knowledge into gpt for healthcare applications. In *Intelligent Environments 2024: Combined Proceedings of Workshops and Demos & Videos Session*. IOS Press, 88–97.
[5] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9, 1, e45312.
[6] Martin Grohe. 2020. Word2vec, node2vec, graph2vec, x2vec: towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 1–16.
[7] Yikun Han, Chunjiang Liu, and Pengfei Wang. 2023. A comprehensive survey on vector database: storage and retrieval technique, challenge. *arXiv preprint arXiv:2310.11703*.
[8] Yining Huang, Keke Tang, and Meilian Chen. 2024. A comprehensive survey on evaluating large language model applications in the medical industry. *arXiv preprint arXiv:2404.15777*.
[9] Jožef Stefan Institute. 2024. Insieme platform. (Aug. 2024). Retrieved August 26, 2024 from https://ise-emh.eu.
[10] Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62, 101983.
[11] Peter Lee, Carey Goldberg, and Isaac Kohane. 2023. *The AI revolution in medicine: GPT-4 and beyond*. Pearson.
[12] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
[13] World Health Organization et al. 2022. Health and care workforce in europe: time to act.
[14] Carl Preiksaitis and Christian Rose. 2023. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR medical education*, 9, e48785.
[15] Mujahid Ali Quidwai and Alessandro Lagana. 2024. A rag chatbot for precision medicine of multiple myeloma. *medRxiv*, 2024–03.
[16] Nicholas R Rydzewski, Deepak Dinakaran, Shuang G Zhao, Eytan Ruppin, Baris Turkbey, Deborah E Citrin, and Krishnan R Patel. 2024. Comparative evaluation of llms in clinical oncology. *Nejm Ai*, 1, 5, AIoa2300151.
[17] Tony Smith. 1991. Medicine in europe. european health challenges. *BMJ: British Medical Journal*, 303, 6814, 1395.
[18] Oguzhan Topsakal and Tahir Cetin Akinci. 2023. Creating large language model applications utilizing langchain: a primer on developing llm apps fast. In *International Conference on Applied Engineering and Natural Sciences* number 1. Vol. 1, 1050–1056.
[19] Ashish Vaswani. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.
[20] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192, 6, 3197–3200.
[21] Matic Zadobovšek, Primož Kocuvan, and Matjaž Gams. 2023. Vpeljava virtualnega asistenta chatgpt v medicinsko platformo = implementation of a virtual assistant chatgpt into the medical platform. In Bibliografija: str. 31. Institut „Jožef Stefan“, 28–31. http://library.ijs.si/Stacks/Proceedings/InformationSociety/2023/IS2023-Volume_A%20-%20SKUI.pdf.

[22] Peng Zhang and Maged N Kamel Boulos. 2023. Generative ai in medicine and healthcare: promises, opportunities and challenges. *Future Internet*, 15, 9, 286.

Standards for Use of LLM in Medical Diagnosis

Mihailo Svetozarević
Clinic for Neurology
University Clinical Center Niš
Niš, Serbia
mihailo.svetozarevic@gmail.com

Isidora Janković
Center for Radiology
University Clinical Center Niš
Niš, Serbia
isidora_jankovic@yahoo.com

Sonja Janković
Center for Radiology
University Clinical Center Niš
Niš, Serbia
sonjasgirl@gmail.com

Stevo Lukić
Clinic for Neurology
University Clinical Center Niš
Niš, Serbia
srlukic@gmail.com

Abstract

Artificial intelligence, particularly large language models (LLMs), is increasingly being recognized for its potential to revolutionize medical diagnosis by mimicking human cognitive functions in clinical decisionmaking. Despite promising developments, such as the ability to pass medical exams and assist in complex diagnostic processes, LLMs still face significant hurdles, including issues with accuracy, bias, and safety. This paper critically consider evaluation of LLMs performance across various criteria to ensure they meet the required standards for clinical use. Several dimensions of evaluations such as accuracy, calibration, and robustness are used. While LLMs and generative AI more broadly show real potential for healthcare, these tools are not ready yet. The medical community and developers need to develop more rigorous evaluation, analyze across specialties, train on real-world data, and explore more useful types of GenAI beyond current models. But ultimately, we believe these tools can help in improving both physician workload and patient outcomes. We urgently need to set up evaluation loops for LLMs where models are built, implemented, and then continuously evaluated via user feedback.

Keywords

large language models, artificial intelligence, clinical AI implementation, AI in clinical practice, AI safety in healthcare

1 Introduction

Artificial intelligence (AI) by its definition, and in the broadest of terms, represents intelligence exhibited by computer systems. The main goal of AI is to enable computers and machines to mimic human cognitive function. In other words, it aims to simulate human learning, comprehension, problem solving and critical decision making. AI approaches human cognition in two distinct ways, the symbolic and the connectionist approach [1]. The symbolic approach aims to replicate human intelligence by analyzing cognition independent of the biological structure of the central nervous system while the connectionist approach aims to create neural networks that imitate the brains's structure. To realize the potential of AI in healthcare, we believe that the systematic approach to evaluation and benchmarking can get us to a place where AI can be a net positive for health systems.

2 LLM's in Medicine

The rapid advancements in AI, particularly in the realm of large language models (LLM's), have transformed various sectors, including healthcare [2,3]. LLM's and Chat GPT in particular has earned much attention in recent years due to its ability to complete tasks previously considered completable by humans alone as in passing United States Medical Licensing Examination [4]. The ability of LLM's to accurately answer questions, provide advice and even triage patients based on clinical input exceeds that of the everyday person. However, the accuracy of these systems to resolve real world medical issues is yet to exceed that of a fully trained physician. Also, a finite percentage of LLM answers to patients had safety errors, and in one instance the advice given to a patient could have been fatal [5]. In order to avoid this error in the future it is essential to assess these models through rigorous comparative benchmarks. One of the most critical aspects of benchmarking medical LLM's is comparing their performance with existing clinical decision support systems (CDSS) and other AI models. Traditional CDSS, often rule-based or statistical, have been used in healthcare for decades to assist clinicians in making evidence-based decisions. By comparing LLMs to these systems, researchers can determine whether the new models offer significant improvements in accuracy, speed, and comprehensiveness [6]. For example, a comparative benchmark might involve evaluating the diagnostic accuracy of an LLM against a well-established CDSS in predicting outcomes for specific conditions, such as sepsis or diabetes. The LLM's ability to incorporate a broader range of data, including unstructured text from electronic health records (EHRs), could be a key factor in outperforming traditional systems [7]. However, it is also crucial to consider scenarios

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.chtm.9

where traditional systems might still have an edge, particularly in specialized tasks where they have been finely tuned over many years of clinical use [8]. Outside of primary care, radiology is perhaps the medical branch that has been the most upfront and welcoming to the use of new technology [2,3]. The concept of computer-assisted diagnosis (CAD) is well known. AI's provide substantial aid by labeling abnormal or most often borderline exams or simply by quickly excluding negative exams in computed tomographies, X-rays, magnetic resonance images especially in high volume settings like the emergency room where human resources might be less available. AI-driven diagnostic tests have the potential to overcome several current limitations in the clinical approach to patient care [9]. Namely the clinical review, time to diagnosis, diagnostic accuracy and consistency. In tandem with AI, diagnosticians of all medical branches are capable of improving measures of diagnostic accuracy (mainly sensitivity and specificity) as well as minimizing observer variability in specific patient interpretation. This proves most useful in settings where the clinical diagnosis is in question – such as with complex patient presentation or in patients with long histories and various comorbidities. Currently not many prospective studies and randomized trials exist in medical AI application. Most are not prospective, are at high risk of bias and deviate from existing report standards. Data availability is lacking and human cooperator groups are more often small and inadequate. LLM's, in particularly GPT-3, has shown promise in various clinical applications, ranging from creation of patient notes to helping healthcare providers diagnose rare conditions. However, it is important to recognize the inherent limitations of these systems.

3 Standardized Evaluation Framework for Assessing LLM's Clinical Utility for Future Clinical Practice

Medical diagnosis involves a complex process in which a practitioner uses objective data from a clinical exam, as well as data collected from medical tests along with self-described subjective symptoms to conclude the most likely health problem. This kind of approach relies heavily on the synthesis and individual interpretation of a vast amount of information from various sources. These most often include available patient histories, clinical exam data correlated with current medical literature. In this setting LLM's open up new opportunities for enhancing the diagnostic process. In order to better evaluate the LLMs clinical utility a direct comparison must be made between LLMs and human clinicians. This approach is essential to gauge how well AI models can replicate or even enhance the decision-making process of experienced healthcare professionals. Studies often involve presenting both clinicians and LLMs with the same clinical cases and comparing their diagnoses, treatment recommendations, and reasoning [10]. Human clinician benchmarking can reveal important insights into the strengths and limitations of LLMs. For instance, while LLMs might excel at processing and synthesizing vast amounts of data quickly, they may struggle with nuanced cases that require deep contextual understanding or ethical considerations that a human clinician might naturally account for [11]. Furthermore, these benchmarks can highlight areas where LLMs might support clinicians, such as providing second opinions or identifying potential errors in human judgment, rather than replacing them [12]. Randomized controlled trials (RCTs) are considered the gold standard in clinical research for evaluating the efficacy of innovations. Comparative benchmarking of LLMs can also involve assessing how well these models predict or align with outcomes from RCTs. For example, an LLM could be tested on its ability to recommend treatments for stroke prevention based on patient data, and its recommendations could be compared with those validated by RCTs [7]. However, this approach presents a set of challenges, as RCTs often involve highly controlled environments that might not fully capture the complexities of real-world clinical settings. Currently LLMs are most often tested on small datasets acquired for a specific research study or large public benchmark dataset, both of which are usually collected on a limited number of very similar sites with consistent diagnostic techniques. This does not reflect the substantial differences in manufacturer, quality and clinical practices often found in real-world hospitals. As an example, the UK Biobank, a widely employed public imaging benchmark dataset includes brain magnetic resonance images (MRI) for a total of 100,000 patients and more. It restricts image acquisition to four sites each of which has identical equipment in terms of hardware and software and performs regular quality check to ensure the harmonization of data. In contrast most medical centers, including our own in Serbia, extracts data from clinical archives over a period of 20 years which reflects the much more diverse array of available data in everyday settings. Another point of interest is a lack of consensus on which dimensions of evaluation to consider and prioritize for various healthcare tasks. While accuracy is the most often examined dimension when evaluating LLM performance, other dimensions such as fairness, bias and toxicity, robustness, and deployment considerations need to be considered as well [13]. Therefore, while alignment with RCT outcomes is a strong indicator of an LLMs clinical relevance, it is also important to test these models in more varied and less controlled environments to ensure their robustness [11]. Unlike traditional systems or statistical models that remain relatively static once developed, LLMs can be continuously updated and refined. This raises the question of how implement models that are constantly evolving. Development of standardized benchmarks that can be applied across different versions of a model are essential to address this challenge [14]. These benchmarks help identify areas where LLMs can enhance clinical practice and highlight the potential risks or limitations that need to be addressed [6]. By rigorously comparing LLMs against existing systems, human clinicians, and traditional models, we can ensure that these advanced AI systems are integrated into healthcare in a way that maximizes their benefits while minimizing potential harms [10]. In general, there is a lack of consensus on what to consider and prioritize for various healthcare tasks. Several dimensions of evaluations such as accuracy, calibration, and robustness are used [13]. While accuracy is the most often examined when evaluating LLM performance, other aspects such as fairness, bias and toxicity, robustness, and deployment considerations need to be considered as well. A list of possible aspects are presented on Table 1. Comparative benchmarks can guide the development of future AI models. Insights gained from these evaluations can inform model improvements, such as enhancing interpretability, reducing bias, or improving performance on specific tasks. As the field of AI in healthcare continues to evolve, comparative benchmarking will remain a crucial tool for ensuring that new models are both safe and effective for clinical use [8].

Table 1. Comparative benchmarks for evaluation of LLG performances in healthcare (adapted and modified from Bedi et al. 2024)

Dimension of Evaluation	Definition	Metric Examples
Accuracy	Measures how close the LLM output is to the true or expected answer	Human evaluated correctness, ROUGE, MEDCON
Calibration and Uncertainty	Measures how uncertain or underconfident an LLM is about its output for a specific task	Human evaluated uncertainty, calibration error, Platt scaled calibration slope
Robustness	Measures the LLMs resilience against adversarial attacks and perturbations like typos	Human evaluated robustness, exact match on LLM input with intentional typos, F1 on LLM input with intentional use of word synonyms
Factuality	Measures how an LLMs output for a specific task originates from a verifiable and citable source. It is important to note that it is possible for a response to be accurate but factually incorrect if it originates from a hallucinated citation	Human evaluated factual consistency, citation recall, citation precision
Comprehensiveness	Measures how well an LLMs output coherently and concisely addresses all aspects of the task and reference provided	Human evaluated comprehensiveness, fluency, UniEval relevance
Fairness, bias and toxiticy	Measures whether an LLMs output is equitable, impartial, and free from harmful stereotypes or biases, ensuring it does not perpetuate injustice or toxicity across diverse groups	Human evaluated toxicity, counterfactual fairness, performance disparities across race
Deployment considerations	Measures the technical and parametric details of an LLM to generate a desired output	Cost, latency, inference runtime

4 Conclusion

Comparative benchmarking is a critical process in the development and deployment of medical large language models. By comparing LLMs to existing clinical decision support systems, human clinicians, traditional statistical models, and outcomes from randomized controlled trials, we can gain a comprehensive understanding of their strengths, limitations, and potential impact on healthcare. As AI continues to play an increasingly prominent role in medicine, rigorous comparative benchmarks will be essential for ensuring that these models deliver on their promise of improving patient care while adhering to the highest standards of safety and effectiveness.

Acknowledgments

Views and opinions expressed in this paper are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor any other authority can be held responsible for them. All authors contributed equally in the final version of this paper. This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214).

References

[1] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al., 2019. A guide to deep learning in healthcare. Nature Medicine, 25(1), pp.24-29.

[2] Thirunavukarasu, A. J. et al., 2023. Large language models in medicine. Nature Medicine, 29, 1930–1940.;

[3] Thirunavukarasu, A.J., et al., 2023. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: Observational study demonstrating opportunities and limitations in primary care. JMIR Medical Education, 9, p.46599.

[4] Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., et al., 2023. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Medical Education, 9, p.e45312.

[5] Chen, S., Guevara, M., Moningi, S., Hoebers, F., Elhalawani, H., Kann, B.H., Chipidza, F.E., Leeman, J., Aerts, H.J.W.L., Miller, T., Savova, G.K., Gallifant, J., Celi, L.A., Mak, R.H., Lustberg, M., Afshar, M., & Bitterman, D.S., 2024. The effect of using a large language model to respond to patient messages. The Lancet Digital Health, 6(6), pp.e379-e381.

[6] Topol, E. J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.;

[7] Rajkomar, A., Dean, J., Kohane, I., 2019. Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.;

[8] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

[9] Fletcher, E., Burns, A., Wiering, B., Lavu, D., Shephard, E., Hamilton, W., et al., 2023. Workload and workflow implications associated with the use of electronic clinical decision support tools used by health professionals in general practice: A scoping review. BMC Primary Care, 24(1), p.23.

[10] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I., 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 1-9.

[11] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Webster, D. R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410.

[12] Ribeiro, M. T., Singh, S., Guestrin, C., 2016. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144)

[13] Bedi S, Liu Y, Orr-Ewing L. et al. (2024) A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs). MedRvix August 16. 2024.

[14] Goodfellow, I., Shlens, J., & Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572

Using Combinatorial Testing for Prompt Engineering of LLMs in Medicine

Alexander Perko
Iulia Nica
Franz Wotawa
alexander.perko@ist.tugraz.at
inica@ist.tugraz.at
wotawa@ist.tugraz.at
Graz University of Technology, Institute of Software Technology
Graz, Austria

Abstract

Large Language Models (LLMs) like GPT-4o are of growing interest. Interfaces such as ChatGPT invite an ever-growing number of people to ask questions, including health advice, which brings in additional risks for harm. It is well known that tools based on LLMs tend to hallucinate or deliver different answers for the same or similar questions. In both cases, the outcome might be wrong or incomplete, possibly leading to safety issues. In this paper, we investigate the outcome of ChatGPT when we ask similar questions in the medical domain. In particular, we suggest using combinatorial testing to generate variants of questions aimed at identifying wrong or misleading answers. In detail, we discuss the general framework and its parts and present a proof-of-concept utilizing a medical query and ChatGPT.

Keywords

Large Language Models, ChatGPT, Prompt Engineering, Combinatorial Testing, Validation

1 Introduction

The use of LLMs in medicine has been of growing interest. In a recent publication [9], the authors discuss the future of LLMs in medical applications. Although using such a model may lead to improved communication and other advantages, some drawbacks prevent using such models and tools. It is well known that LLMs like ChatGPT [33] have shortcomings like hallucinations [46]. Hallucinations are answers with incorrect claims that do not depend on training data. Such answers, unfortunately, cannot be necessarily identified as wrong and, therefore, might be harmful, especially when dealing with medical questions. There are methods for detecting hallucinations, e.g., see [11]. Hence, verifying and validating tools based on LLMs to ensure a harmless use is of utmost importance.

When using LLMs for queries, the form of the query, i.e., the prompt, is of great importance. Although there has been much work on how to improve writing prompts in various setups, e.g., [28], there is only little scientific work, e.g., [24], providing statistical evidence. However, it is generally agreed that the query's structure has a significant impact on the output of a LLM. Therefore, we need to consider different prompts in any verification and validation procedure.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.chtm.10

In this paper, we contribute to verifying and validating LLMs focusing on the impact of prompts. In particular, we introduce and discuss a methodology based on combinatorial testing [21] for generating various versions of prompts for medical queries. We rely on testing because it is a methodology to straighten the view on finding interactions with a system under test (SUT) that leads to unexpected behavior. Hence, in testing, we want to provide interactions that make a system fail. Combinatorial testing is a test methodology that has proven to be effective in finding test cases, i.e., inputs given to a system, to provoke a failing behavior. In particular, combinatorial testing focuses on interacting parameter values that reveal faults. In previous work, Kuhn and colleagues [22] showed that strictly less than 7 interacting parameters must be considered for many applications.

Besides its effectiveness, combinatorial testing is a good testing methodology for LLMs that consider prompts. For the latter, we need different combinations of textual fragments to show differences in the outcome. Combinatorial testing provides such combinations and also avoids leading to a combinatorial explosion of potential prompts when restricting the number of considered fragment interactions.

We organize the paper as follows: We first introduce the foundations. For this purpose, we discuss related research on testing LLMs, and introduce the basic concepts behind combinatorial testing. Afterward, we introduce the general testing methodology for generating different prompts focusing on the medical domain. In addition, we illustrate the use of the methodology considering one particular medical query. Finally, we conclude the paper.

2 Related Research

In the past several years, considerable efforts have been made to evaluate LLMs. The first indicator is the wide variety of *benchmarks* which have emerged in order to test and compare their performance on various tasks. In [6], the authors compile a selection of 46 popular benchmarks. Among them, we can differentiate between benchmarks used for general language tasks, like Chatbot Arena [7], MT-Bench [49], HELM [25], or MMLU [13] and domain-specific benchmarks, like MATH [14], concentrated on assessing reasoning and problem-solving capabilities of AI models in mathematics, APPS [15] for evaluating code generation, or MultiMedQA [41] with focus on medical examinations, medical research, and consumer healthcare questions.

Further on, depending on the human involvement in the evaluation process, there are two common methods: *human evaluation* and *automatic evaluation*. Human evaluation becomes a natural choice in many non-standard cases, where the automated evaluation metrics are either not suitable or insufficient. For example, in [25], the evaluators analyze summarization and disinformation scenarios, while in [2] analogical reasoning tasks. Also, Ziems et al. [50] used the annotations from researchers for generation. Although it can provide more comprehensive and accurate feedback than the automatic evaluation, the human evaluation may suffer from a high variance and instability due to cultural and individual biases. On the other hand, the automatic variant benefits from higher standardization. LLM-EVAL [26], for instance, is a unified multidimensional automatic evaluation method for open-domain conversations with LLMs. Jain et al. [18] proposed a self-supervised evaluation framework, and also PandaLM [45] obtained reproducible and automated language model assessment by training an LLM that acts as the 'judge' to assess different models. For more details on the specific key metrics and factors for both evaluation types, we refer the interested reader to [6].

In addition, *domain-specific evaluation* is also critical, as LLMs are often used in specific areas, such as healthcare or finance, that have specific requirements for the models. In the beginning, most evaluation research has been focused on natural language tasks. A popular direction here is, for instance, the sentiment analysis task, which analyzes and interprets the text to identify the emotional inclination. Further on, in the medical field, the application of LLM has recently gained significant attention. According to the literature ([6], [5]), most LLMs evaluations in the medical field focus on medical queries. ChatGPT, in particular, generated relatively accurate information for various medical queries from genetics [10], biomedicine [17], radiation oncology physics [16]. Furthermore, several studies have evaluated the performance and feasibility of ChatGPT in the medical education field. In [31], GPT-3.5 [4] and GPT-4 [34] models were evaluated in terms of their understanding of surgical clinical information and their potential impact on surgical education and training. These findings demonstrate that LLMs can be successfully used in clinical education, still further efforts may be needed to overcome limitations. For more details concerning the evaluation of other applications like social science, natural science, engineering agent applications, education, search and recommendation, and personality testing, we refer the reader to [6].

Another interesting taxonomy from [6] groups the encountered testing approaches into three directions: (1) from the objective calculation (benchmarking) to human-in-the-loop testing, (2) from static to crowd-sourcing test sets and (3) from unified to challenging test sets. While unified settings involve test sets with no preference for any specific task, challenging settings create test sets for specific tasks. Tools like DeepTest[43] use seeds to generate input modifications for testing, CheckList [38] builds test sets based on templates, whereas AdaFilter [36] constructs tests adversarially. Furthermore, despite the growing number of academic projects designed for prompting LLMs [19, 27, 3], just a few of them support systematic evaluation of textual responses [48, 1]. ChainForge [1] is a visual toolkit that offers on-demand hypothesis testing of the behavior of text-generating LLMs on open-domain tasks.

To our knowledge, the use of combinatorial testing (CT), in particular for the testing of LLMs, is reported in a single paper [12]. Based on a given original sentence, the authors derive new sentences by replacing words with synonyms according to a combinatorial test set. Assuming that the semantics of the original sentence are preserved in the derived sentences, a test oracle is created based on existing annotations. In the experimental evaluation from [12], the authors apply generated pairwise sentence test sets from the BoolQ benchmark set [8] against two LLMs (T5 [37] and LLaMa [44]). The results indicate that the accuracy of the responses remains roughly equivalent to those provided for the original test set.

3 Combinatorial Testing

Combinatorial testing aims to generate test cases by considering a system's input model. The input model comprises a set of parameters (or variables) $\{x_1, \ldots, x_n\}$ and a not necessarily different domain d_i for each parameter x_i. The domain itself is a finite set of values a parameter can take. A test case is a n-tuple specifying a value $v_i \in d_i$ for every parameter $x_i \in \{x_1, \ldots, x_n\}$. A test suite is a set of test cases. Usually, we write a test suite as a table where the columns are the parameters, and the rows have their corresponding values.

Given an input model, a complete test suite comprises a row for each possible value-parameter combination. Obviously, the upper bound of rows is of order $O(D^n)$ where D is the maximum size of all domains d_i, $i = 1, \ldots, n$, i.e., $D = \max_{i=1,\ldots,n}(d_i)$. Hence, computing a complete test suite is not feasible for software or systems comprising a larger number of input parameters. Moreover, applying all test cases is not feasible because the system's behavior must also be evaluated. In combinatorial testing, we do not have a test oracle. The focus is only on input generation. Hence, such a test oracle must be added to classify a test case as passing or failing, i.e., indicating whether a test case leads to a correct or wrong output, respectively. It is worth noting that such a test oracle can be automated, and we will discuss this when showing our application for validating LLMs considering medical queries.

Combinatorial testing avoids computing all possible test cases. The idea behind this is to consider not all parameter combinations but only those combinations of values for a fixed number k (smaller than n) of parameters. Hence, a combinatorial test suite covers all combinations of values for any subset of parameters of size k, which is usually substantially smaller. Such a test suite is said to be of strength k or to be a k-wise test suite. If k is 2, then the test suite is a pairwise test suite, and we speak about pairwise testing. Note that in practice, pairwise testing is not good enough (see [22, 23]). For more information on combinatorial testing and its foundations, we refer the interested reader to [30, 21]. There are many algorithms available, including ACTS [47], for computing combinatorial test suites for arbitrary input models and strengths. It is also worth mentioning that combinatorial testing has been successfully used in many application domains, including autonomous driving [20] and security testing [40].

In the following, we illustrate combinatorial testing using a small example. In this example, we assume four parameters a, b, c, d, all of them only taking values from the Boolean domain $\{T, F\}$ standing for true and false. A pairwise combinatorial test suite for this input model comprises 6 test cases:

	a	b	c	d
1	T	T	F	F
2	T	F	T	T
3	F	T	T	F
4	F	F	F	T
5	F	T	F	T
6	T	F	F	F

For any combination of two parameters, e.g., a and c, this table comprises all possible combinations of values. Rows 1, 2, 3, and 4 already cover all four combinations for these two parameters. For parameters b and d, rows 1, 2, 5, and 6 are required to cover all value combinations. It can be easily checked that this holds also for any other pair of parameters. Note that pairwise testing in this case only requires 6 test cases. Considering all combinations, we would have $2^4 = 16$ test cases. For the remainder of this paper, as we introduce domains extending beyond boolean values, we will use indices when referring to parameter values. For $\{T, F\}$, the indices would be 0 and 1, respectively, and the first row of the table above would be represented as $[0, 0, 1, 1]$.

4 Validation Methodology

Figure 1 gives a high-level overview of our proposed validation methodology. The remainder of this section follows the numbers shown in Figure 1 and discusses the individual elements of our validation pipeline.

The domain of our combinatorial prompt generation pipeline can be seen in Table 1, where parameters are components of a prompt and values are (sub-)phrases. Our prototypical set of parameters comprises a) symptom presentation, which is an introductory sub-phrase to the prompt, b) diagnostic focus, which sets the horizon for which kind of diagnoses are expected, c) an additional hint to consider context information such as age, and d) constraints on how the output should be formulated. Each parameter can assume an indexed value from the given set, and every set of values includes an empty string, which is denoted by "-".

Table 1: Domain: Prompt Components and Values by Index

Parameter (i.e. Prompt Component)	IDX	Value
Symptom	0	-
Presentation	1	list of symptoms
"Given the	2	symptoms
following..."	3	high-level overview of symptoms
Diagnostic	0	-
Focus	1	a probable diagnosis
	2	a differential diagnosis
	3	an emergency diagnosis
	4	the three most likely diagnoses
	5	the ten most likely diagnoses
Contextual	0	-
Information	1	based on patient's age and gender
Constraints	0	-
"the diagnosis	1	be concise
should..."	2	be detailed including explanations
	3	have less than 100 words

In our prototype implementation, we use pairwise testing for all parameters seen in Table 1. Note that we treat the sets of symptoms separately from the rest of the input parameters: Symptoms are excluded from the pairwise combination as we rather want to combine each set of symptoms with all pairwise combinations of the other parameters. This results in the listed 24 pairwise combinations per set of symptoms. Finally, the output of our combinatorial testing pipeline is textual prompts in natural language that act as test cases for evaluating an LLM. The resulting test suite from pairwise combinations for our exemplary set of symptoms can be seen in the first two columns of Table 3. For our preliminary evaluation, we used GPT-4o [34, 35] exclusively. The model can be accessed either via ChatGPT [33] or the OpenAI API [32]. Having both a chatbot interface and a programmable API enables rapid prompt prototyping as well as executing larger test suites with the same underlying model. That said, it would be easy to swap out the SUT and test another LLM using the same validation methodology.

We use the expert system "Symptom-Checker" [39], which is curated by medical professionals and freely accessible via NetDoktor [29]. Given a set of symptoms, we can automatically retrieve diagnoses by traversing a decision tree and answering yes-no-questions. In addition to "yes" and "no" there is also the option to skip a question when the inquired information is not available. Further questions are asking for age, sex, and main symptoms, as well as the body part that is influenced the most. We compute a score by comparing results (i.e., diagnoses) from our SUT ChatGPT with our Golden Model NetDoktor. This is done semi-automatically by first retrieving the diagnoses in the same format and then assessing their overlap. As NetDoktor always yields three diagnoses, our score ranges from 0/3 (no overlap) to 3/3 (complete overlap). In cases where the SUT yields more or less than three diagnoses, we do not normalize the score. This means that a result comprising only one diagnosis cannot achieve a complete overlap with NetDoktor and thus cannot achieve a better score than 1/3. In turn, a result comprising 10 diagnoses cannot achieve a higher score than 3/3. It must be noted that duplicate diagnoses are only counted once, and semantic equivalence is considered when comparing diagnoses.

5 Medical Use Case

For our use case, we assume the downstream task of retrieving diagnoses from an LLM based on a given set of symptoms. As mentioned earlier in this paper, hallucination is a severe problem when using LLMs. Especially in the medical domain, it is of utmost importance that systems are tested and validated in a structured way, as faulty output might have dire consequences ranging from misinformation to taking the wrong medication. However, LLMs cannot be tested exhaustively due to their non-deterministic nature and other factors, such as computational and monetary costs. Even testing a restricted domain, such as prompt formulation, given a limited set of input parameters, may lead to a combinatorial explosion when testing all possible combinations. For instance, compare the 24 pairwise combinations, as described above, to all possible 192 combinations from the values in Table 1 used with reasonably sized corpus of pathologies, such as DDXPlus [42]: This corpus comprises 134,530 samples (i.e. sets of symptoms) from the real world, which yields 25,829,760 possible test cases as compared to 3,228,720 for pairwise testing. Taking into account cost, we look at 206,134.40$ versus 25,766.80$. Table 2 gives a breakdown of the number of tokens for prompts as well as their cost, assuming the current pricing of GPT-4o, which is 5.00$ and 15.00$ per million input and output tokens, respectively. It has to be mentioned that the input length can be fully controlled, whereas the length of the output can only be guided by the prompt. In this case, we always assumed 475 output tokens, which was the average in our preliminary experiments. The average number of input tokens was 171.1, consisting of 19.1, which are derived from combinations of values seen in Table 1, and 152 coming from our exemplary set of symptoms, discussed in the following Subsection 5.1.

5.1 Example

To further explain our proposed validation methodology, we use the following exemplary description of symptoms:

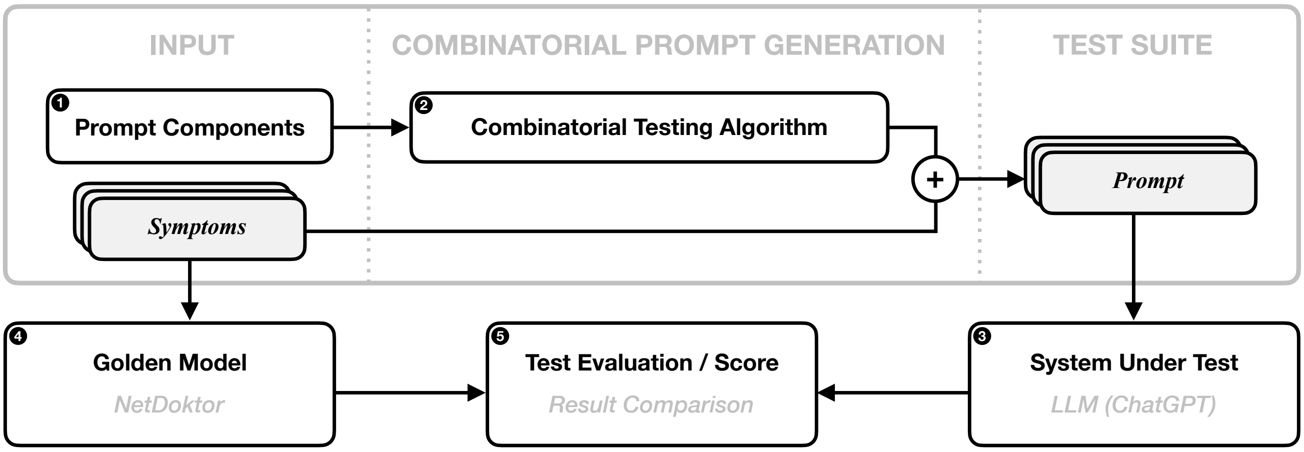


Figure 1: Basic Architecture of Our Validation Methodology.

Table 2: Cost per Size of Medical Corpus

Corpus Size	1	10	100	...	DDXPlus
All Combinations					
Combinations	192	1,920	19,200	...	25,829,760
Input Tokens	0.03M	0.33M	3.29M	...	4,419M
Input Cost [$]	0.16	1.64	16.43	...	22,097.36
Output Tokens	0.09M	0.91M	9.12M	...	12,269M
Output Cost [$]	1.37	13.68	136.80	...	184,037.04
Total Cost [$]	1.53	15.32	153.23	...	206,134.40
Pairwise Combinations					
Combinations	24	240	2,400	...	3,228,720
Input Tokens	0.004M	0.04M	0.41M	...	552M
Input Cost [$]	0.02	0.21	2.05	...	2,762.17
Output Tokens	0.01M	0.11M	1.14M	...	1,533M
Output Cost [$]	0.17	1.71	17.10	...	23,004.63
Total Cost [$]	0.19	1.92	19.15	...	25,766.80

An adult woman is experiencing symptoms in the breast gland area. Her most troubling symptom is fluid discharge, and she can feel a firm, painless lump.

This set of symptom shall be seen as a sample from a corpus of medical pathologies (i.e. sets of symptoms). We use it to test different prompting strategies by combining the values from our domain, seen in Table 1. From this singular sample, we can generate 24 test cases as per our methodology. These test cases are then used to evaluate our SUT based on the output of our golden model. For this set of symptoms, our golden model diagnoses are:

- Breast cancer
- Cyst in the breast
- Mastopathy

Table 3 shows the test result of all 24 test cases. It can be easily spotted that test case 13 was the only prompt achieving a complete overlap with the NetDoktor diagnoses. When fully written-out, prompt 13 corresponding to the combination $[3, 5, 1, 0]$ was:

Given the following high-level overview of symptoms, provide the ten most likely diagnoses based on the patient's age and gender.
An adult woman is experiencing symptoms in the breast gland area. Her most troubling symptom is fluid discharge, and she can feel a firm, painless lump.

Other than the overlap score, there are severe differences in the output depending on the used prompt. Figure 2 illustrates those differences underlined by textual metrics, such as the number of words, while Figure 3 highlights the conciseness of the result as measured by the ratio between the number of words and the number of diagnoses. Most notably, the constraints to asking the LLM to provide concise diagnoses or limiting the number of words to 100 reduce the length drastically. As can be seen in Figure 2, the prompts 3, 4, 7, 8, 11, 12, 16, 17, 20, 22, 23, and 24 all yielded results with less than 200 words. However, prompt 7 exceeds the posed 100 word limit. Furthermore, none of these prompts fully overlapped. When comparing the results for conciseness in particular, Figure 3 shows that the ratio between the number of words and the number of diagnoses is less than 50 for all prompts querying the LLM to provide concise responses (i.e. 3, 4, 16, 17, 23, 24), whereas it is above 60 for all and above 100 for all but one of the prompts asking for a detailed response (i.e. 5, 6, 9, 10, 19).

In an effort to make our work as transparent and reproducible as possible, we provide all prompts and responses of our preliminary study as a replication package [1].

6 Conclusion

This paper highlights the importance of a structured and rigorous validation methodology for LLMs in the medical domain, particularly focusing on prompt engineering. The proposed validation pipeline makes use of pairwise combinatorial testing to systematically evaluate the responses of LLMs like ChatGPT to medical queries. The methodology generates test cases given sets of symptoms and combinations of prompt components. Combinatorial testing ensures that a wide range of prompt variations is tested per set of symptoms without causing a combinatorial

[1] https://zenodo.org/doi/10.5281/zenodo.13765131

Table 3: Overlaps of Diagnoses with Golden Model per Combination for One Exemplary Set of Symptoms. Each Combination Corresponds to One Prompt and Is Denoted by a Code Representing the Indices of the Assumed Values per Prompt Component, as Seen in Table 1. "Mast." stands for " "Mastopathy".

Test Suite		Golden Model Overlap			Score
ID	Combination	Cancer	Cyst	Mast.	
1	[0, 0, 0, 0]	✓	✓		2/3
2	[1, 1, 1, 0]	✓			1/3
3	[2, 2, 1, 1]	✓			2/3
4	[3, 3, 0, 1]	✓			1/3
5	[3, 4, 1, 2]	✓			1/3
6	[2, 5, 0, 2]	✓	✓		2/3
7	[1, 5, 0, 3]	✓	✓		2/3
8	[0, 4, 1, 3]	✓			1/3
9	[0, 3, 1, 2]	✓	✓		2/3
10	[1, 2, 0, 2]	✓	✓		2/3
11	[2, 0, 1, 3]	✓	✓		2/3
12	[3, 1, 0, 3]	✓	✓		2/3
13	**[3, 5, 1, 0]**	✓	✓	✓	**3/3**
14	[2, 4, 0, 0]		✓		1/3
15	[1, 3, 0, 0]	✓			1/3
16	[0, 1, 0, 1]	✓	✓		2/3
17	[1, 0, 0, 1]	✓	✓		2/3
18	[0, 2, 0, 0]	✓	✓		2/3
19	[2, 1, 0, 2]	✓			1/3
20	[3, 2, 0, 3]	✓			2/3
21	[3, 0, 0, 2]	✓	✓		2/3
22	[2, 3, 0, 3]	✓			2/3
23	[0, 5, 0, 1]	✓	✓		2/3
24	[1, 4, 0, 1]	✓			1/3

explosion. Doing so is more efficient and reduces costs as compared to testing all possible combinations of prompt parameters, which is especially important when evaluating LLMs on large medical corpora. The proposed validation pipeline implements semi-automated scoring based on a "golden model", which provides diagnoses curated by medical professionals. In a preliminary study, we demonstrate severe differences in output for prompt variations given the same set of symptoms. Out of 24 test cases, only one achieved a full overlap with our golden model when using GPT-4o. Once more, this highlights the dependence on well-formulated prompts and a need for thorough testing strategies, especially in critical domains like medicine.

Acknowledgements

The work presented in this paper was partially funded by the European Union under Grant 101159214 – ChatMED. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24). ACM, (May 2024), 1–18. DOI: 10.1145/3613904.3642016.

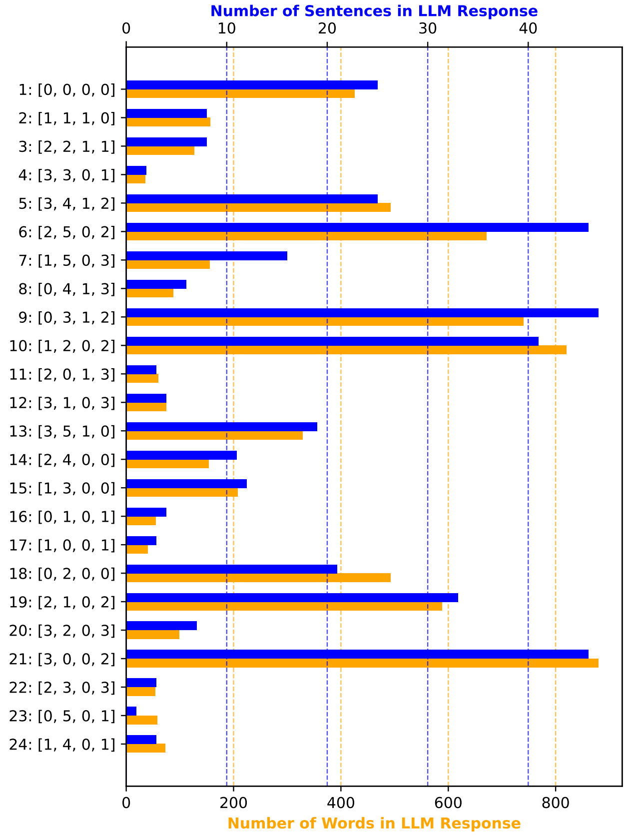


Figure 2: Textual Metrics of LLM Responses: Number of Sentences & Number of Words per Prompt

[2] Yejin Bang et al. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. (2023). https://arxiv.org/abs/2302.04023 arXiv: 2302.04023 [cs.CL].

[3] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: text-to-image generation through interactive prompt exploration with large language models. (2023). https://arxiv.org/abs/2304.09337 arXiv: 2304.09337 [cs.HC].

[4] Tom B. Brown et al. 2020. Language models are few-shot learners. arXiv: 2005.14165 [cs.CL]. (2020).

[5] David Cecchini, Arshaan Nazir, Kalyan Chakravarthy, and Veysel Kocaman. 2024. Holistic evaluation of large language models: assessing robustness, accuracy, and toxicity for real-world applications. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*. Anaelia Ovalle, Kai-Wei Chang, Yang Trista Cao, Ninareh Mehrabi, Jieyu Zhao, Aram Galstyan, Jwala Dhamala, Anoop Kumar, and Rahul Gupta, editors. Association for Computational Linguistics, Mexico City, Mexico, (June 2024), 109–117. DOI: 10.18653/v1/2024.trustnlp-1.11.

[6] Yupeng Chang et al. 2023. A survey on evaluation of large language models. (2023). https://arxiv.org/abs/2307.03109 arXiv: 2307.03109 [cs.CL].

[7] Wei-Lin Chiang et al. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. (2024). arXiv: 2403.04132 [cs.AI].

[8] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: exploring the surprising difficulty of natural yes/no questions. (2019). https://arxiv.org/abs/1905.10044 arXiv: 1905.10044 [cs.CL].

[9] J. Clusmann et al. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3, 141. DOI: https://doi.org/10.1038/s43856-023-00370-1.

[10] Dat Duong and Benjamin D. Solomon. 2023. Analysis of large-language model versus human performance for genetics questions. *medRxiv*. eprint: https://www.medrxiv.org/content/early/2023/01/28/2023.01.27.23285115.full.pdf. DOI: 10.1101/2023.01.27.23285115.

[11] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630. https://doi.org/10.1038/s41586-024-07421-0.

[12] 2023. *Applying pairwise combinatorial testing to large language model testing*. Springer, (Sept. 2023), 247–256. ISBN: 978-3-031-43239-2. DOI: 10.1007/978-3-031-43240-8_16.

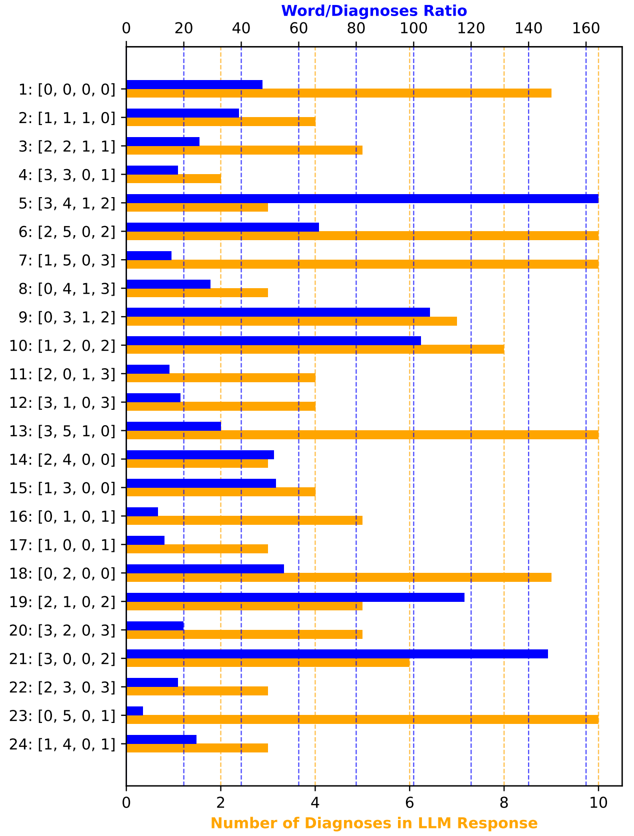


Figure 3: Conciseness of LLM Responses: Number of Diagnoses & Ratio between Number of Words and Number of Diagnoses per Prompt

[13] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. (2021). https://arxiv.org/abs/2009.03300 arXiv: 2009.03300 [cs.CY].

[14] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. (2021). https://arxiv.org/abs/2103.03874 arXiv: 2103.03874 [cs.LG].

[15] Dan Hendrycks et al. 2021. Measuring coding challenge competence with apps. (2021). https://arxiv.org/abs/2105.09938 arXiv: 2105.09938 [cs.SE].

[16] Jason Holmes et al. 2023. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology*, 13, (July 2023). DOI: 10.3389/fonc.2023.1219326.

[17] Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of chatgpt on biomedical tasks: a zero-shot comparison with fine-tuned generative transformers. (2023). https://arxiv.org/abs/2306.04504 arXiv: 2306.04504 [cs.CL].

[18] Neel Jain, Khalid Saifullah, Yuxin Wen, John Kirchenbauer, Manli Shu, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Bring your own data! self-supervised evaluation for large language models. (2023). https://arxiv.org/abs/2306.13651 arXiv: 2306.13651 [cs.CL].

[19] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. Promptmaker: prompt-based prototyping with large language models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI EA '22) Article 35. Association for Computing Machinery, New Orleans, LA, USA, 8 pages. ISBN: 9781450391566. DOI: 10.1145/3491101.3503564.

[20] Florian Klück, Yihao Li, Jianbo Tao, and Franz Wotawa. 2023. An empirical comparison of combinatorial testing and search-based testing in the context of automated and autonomous driving systems. *Information and Software Technology*, 160, 107225. DOI: https://doi.org/10.1016/j.infsof.2023.107225.

[21] D.R. Kuhn, R.N. Kacker, and Y. Lei. 2013. *Introduction to Combinatorial Testing. Chapman & Hall/CRC Innovations in Software Engineering and Software Development Series*. Taylor & Francis.

[22] D.R. Kuhn, R.N. Kacker, Y. Lei, and J. Hunter. 2009. Combinatorial software testing. *Computer*, (Aug. 2009), 94–96.

[23] Rick Kuhn, Yu Lei, and Raghu Kacker. 2008. Practical combinatorial testing: beyond pairwise. *IT Professional*, 10, 3, 19–23.

[24] Boniphace Kutela, Kelvin Msechu, Norris Novat, Emmanuel Kidando, and Angela Kitali. 2023. Uncovering the influence of chatgpt's prompts on scientific writings using machine learning-based text mining approaches. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.4385895.

[25] Percy Liang et al. 2023. Holistic evaluation of language models. (2023). https://arxiv.org/abs/2211.09110 arXiv: 2211.09110 [cs.CL].

[26] Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: unified multi-dimensional automatic evaluation for open-domain conversations with large language models. (2023). https://arxiv.org/abs/2305.13711 arXiv: 2305.13711 [cs.CL].

[27] Aditi Mishra, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2023. Promptaid: prompt exploration, perturbation, testing and iteration using visual analytics for large language models. (2023). https://arxiv.org/abs/2304.01964 arXiv: 2304.01964 [cs.HC].

[28] M. Nazari and G. Saadi. 2024. Developing effective prompts to improve communication with chatgpt: a formula for higher education stakeholders. *Discov Educ*, 3. https://doi.org/10.1007/s44217-024-00122-w.

[29] 2024. Netdoktor. https://www.netdoktor.at. Accessed: 2024-09-03. (2024).

[30] Changhai Nie and Hareton Leung. 2011. A survey of combinatorial testing. *ACM Comput. Surv.*, 43, 2, Article 11, (Feb. 2011), 11:1–11:29.

[31] Namkee Oh, Gyu-Seong Choi, and Woo Yong Lee. 2023. Chatgpt goes to the operating room: evaluating gpt-4 performance and its potential in surgical education and training in the era of large language models. *Annals of Surgical Treatment and Research*, 104, (Apr. 2023), 269. DOI: 10.4174/astr.2023.104.5.269.

[32] OpenAI. 2023. API Reference. (2023). platform.openai.com/docs/api-reference.

[33] OpenAI. 2023. ChatGPT. (2023). chat.openai.com/chat.

[34] OpenAI. 2023. GPT-4 technical report. arXiv: 2303.08774 [cs.CL]. (2023).

[35] OpenAI. 2024. Introducing gpt-4o and more tools to chatgpt free users. (May 2024). https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/.

[36] Jason Phang, Angelica Chen, William Huang, and Samuel R. Bowman. 2021. Adversarially constructed evaluation sets are more challenging, but may not be fair. (2021). https://arxiv.org/abs/2111.08181 arXiv: 2111.08181 [cs.CL].

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21, 140, 1–67. http://jmlr.org/papers/v21/20-074.html.

[38] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: behavioral testing of nlp models with checklist. (2020). https://arxiv.org/abs/2005.04118 arXiv: 2005.04118 [cs.CL].

[39] Jens Richter, Hans-Richard Demel, Florian Tiefenböck, Luise Heine, and Martina Feichter. 2024. Symptom-checker. https://www.netdoktor.at/symptom-checker/. Accessed: 2024-09-03. (2024).

[40] Dimitris E. Simos, Josip Bozic, Bernhard Garn, Manuel Leithner, Feng Duan, Kristoffer Kleine, Yu Lei, and Franz Wotawa. 2019. Testing TLS using planning-based combinatorial methods and execution framework. *Software Quality Journal*, 27. https://doi.org/10.1007/s11219-018-9412-z.

[41] Karan Singhal et al. 2022. Large language models encode clinical knowledge. (2022). https://arxiv.org/abs/2212.13138 arXiv: 2212.13138 [cs.CL].

[42] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A New Dataset For Automatic Medical Diagnosis. (2022). https://arxiv.org/abs/2205.09148 arXiv: 2205.09148 [cs.CL].

[43] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. 2018. DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars. (2018). https://arxiv.org/abs/1708.08559 arXiv: 1708.08559 [cs.SE].

[44] Hugo Touvron et al. 2023. LLaMA: Open and Efficient Foundation Language Models. (2023). https://arxiv.org/abs/2302.13971 arXiv: 2302.13971 [cs.CL].

[45] Yidong Wang et al. 2024. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. (2024). https://arxiv.org/abs/2306.05087 arXiv: 2306.05087 [cs.CL].

[46] Yijun Xiao and William Yang Wang. [n. d.] On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. DOI: 10.18653/v1/2021.eacl-main.236.

[47] Linbin Yu, Yu Lei, R.N. Kacker, and D.R. Kuhn. 2013. ACTS: A Combinatorial Test Generation Tool. In *Software Testing, Verification and Validation (ICST), 2013 IEEE Sixth International Conference on*, 370–375.

[48] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23) Article 437. Association for Computing Machinery, Hamburg, Germany, 21 pages. ISBN: 9781450394215. DOI: 10.1145/3544548.3581388.

[49] Lianmin Zheng et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=uccHPGDlao.

[50] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? (2024). https://arxiv.org/abs/2305.03514 arXiv: 2305.03514 [cs.CL].

Meeting Cultural and Linguistic Demands to Accommodate Fine-Tuned LLMs to Local Medical Customs and Patient Communication

Gordana Petrovska Dojchinovska[1],Monika Simjanovska Misheva[1], Kostadin Mishev[1]
[1]Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering,
Skopje, N. Macedonia
gordana.petrovska.dojchinovska@students.finki.ukim.mk
{monika.simjanoska, kostadin.mishev}@finki.ukim.mk

ABSTRACT

Integrating advanced open-source large language models (LLMs), such as LLaMA and GatorTron, into healthcare offers a novel approach to enhancing communication between physicians and patients. This paper provides a comprehensive review of the potential of these models to improve patient-provider interactions, focusing on their ability to process and generate human-like language in real-time clinical settings. The review outlines the methodology used to evaluate LLMs, which includes a detailed comparison based on qualitative factors such as linguistic adaptability, cultural sensitivity, and context-awareness, alongside quantitative metrics such as accuracy rates, error margins, and patient satisfaction scores from clinical studies. Key ethical considerations are explored, particularly concerning data privacy, patient consent, and accountability. The paper delves into how adopting specific ethical frameworks or guidelines can help mitigate risks associated with bias, misinformation, and patient autonomy. Additionally, the potential for LLMs to perpetuate biases or cultural misunderstandings is discussed, emphasizing the need for fine-tuning these models to align with clinical guidelines and patient needs across different geographies and medical customs. While the paper acknowledges the gaps in current research, such as the emotional sensitivity of LLMs and their ability to understand nuanced patient concerns, it also proposes structured future research directions. This includes the development of LLMs that are more contextually aware, emotionally intelligent, and capable of operating in diverse healthcare settings. By synthesizing current studies and real-world applications, this paper aims to provide a transparent, reproducible framework for evaluating the effectiveness of fine-tuned LLMs in transforming healthcare communication, ultimately improving patient outcomes and satisfaction.

KEYWORDS

large language models, healthcare, patient-doctor communication, ethics in AI, cultural sensitivity

1 INTRODUCTION

Effective communication has always been at the forefront of successful interaction between physicians and patients. It has already been purported that integrating cultural and linguistic competency into healthcare policies, provider training, and patient care strategies will improve the quality of care for diverse populations[5]. Conversely, language barriers, cultural misunderstandings, and a lack of cultural awareness among healthcare providers can lead to miscommunication, misdiagnosis, medication errors, and other safety risks[11]. Large Language Models (LLMs) have shown immense potential in various fields due to their proficiency in generating and understanding natural language. In healthcare, models such as LLaMA and GatorTron present an exciting frontier for improving communication between physicians and patients. These models can process and generate human-like language, which could address significant challenges in clinical communication, such as linguistic, cultural, or emotional barriers. However, significant challenges remain, particularly regarding ethical implications, technical hurdles, local customs, linguistic demands, and the critical aspect of patient-doctor communication. This review examines current applications of LLMs in healthcare, the associated challenges, ethical concerns, and potential gaps, incorporating a range of recent research studies in the field from 2023 and early 2024.[1]

1.1 Purpose and Scope

This paper aims to critically examine the potential of integrating LLMs into healthcare settings, emphasizing their role in enhancing communication. The scope of this review extends to a comparative evaluation of LLaMA, GPT-4, L2M3, and GatorTron, focusing on their potential impact, ethical considerations, and future research needs. The goal is to provide healthcare practitioners, researchers, and developers with a transparent and reproducible framework for adopting LLMs.

2 METHODOLOGY

This review systematically analyzes the current applications of LLMs in healthcare by conducting a structured comparison of models such as GPT-4, Llama, GatorTron, and L2M3. The evaluation criteria include both qualitative and quantitative metrics, focusing on model accuracy, error rates, patient satisfaction scores, and the ability to generate culturally sensitive and clinically appropriate responses. The sources for this review consist of peer-reviewed studies published between 2023 and 2024, covering applications of LLMs in real-time clinical settings, mental health, and patient-doctor communication. Each model's performance is assessed based

.

[1]https://doi.org/10.70314/is.2024.chtm.11

on specific clinical tasks, adaptability to local medical customs, and ethical considerations. This methodology ensures transparency and reproducibility, allowing future researchers to replicate and validate the findings.

3 CURRENT APPLICATION OF LLMS IN HEALTHCARE

LLMs are being adopted for various tasks within healthcare, with their most prominent use in medical communication, clinical decision support, and facilitating patient-doctor interactions. Conversational agents like ChatGPT assist patients in understanding medical terminology and provide immediate answers to healthcare-related queries, often enhancing patient engagement with care providers. LLMs also bridge communication gaps between patients and healthcare providers, improving clarity and comprehension in patient education[25]. In patient-doctor communication, LLMs serve as intermediaries to help patients articulate their symptoms and concerns more clearly. This enhances clinical encounters, as doctors can quickly grasp patient issues without the noise of miscommunication. They can improve the precision of patient-doctor dialogue, particularly when addressing complex conditions and explaining treatment plans[21]. Similarly, mental health applications have seen LLMs being used to summarize counseling sessions, allowing healthcare providers to focus on therapeutic interventions rather than administrative tasks[1]. Beyond patient communication, LLMs assist clinicians with diagnoses and medical research. GPT-4 and LLaMA are increasingly used as clinical assistants, offering support in diagnosis and treatment planning[29]. Large language models enhance decision-making in surgical care by answering patients' questions, thereby improving pre-surgical communication and reducing patient anxiety[18]. Figure **??** shows the areas of the medical domain where LLMs are currently being applied.

Figure 1: Applications of large language models in healthcare

4 IMPROVING PATIENT-DOCTOR COMMUNICATION

Effective communication between patients and doctors is a cornerstone of healthcare. Miscommunication can lead to misunderstanding, reduced adherence to treatment plans, and dissatisfaction with care. LLMs offer a promising solution to enhance communication by simplifying medical information into layman's terms. The role of LLMs in improving patient-doctor interactions is significant, as they ensure the translation of medical jargon into easily understandable language during consultations, ultimately leading to more informed decision-making and patient compliance[19]. Recent research explores how generative AI tools, including LLMs, have been applied to reduce misunderstandings in clinical services. They show that patients often feel overwhelmed by medical language, and LLMs can mitigate this issue by acting as interpreters, promoting clearer communication between patients and their healthcare providers[28]. A pictorial demonstration of the ways large language models can improve the communication between patients and their providers is given in Figure 2.

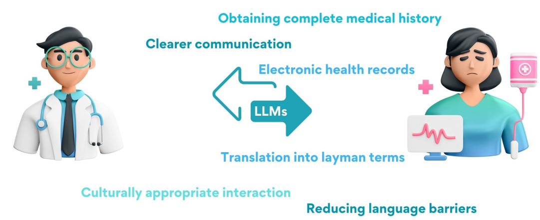


Figure 2: Patient-doctor communication facilitated by the use of LLMs

There is also research that addresses how LLMs can cater to multilingual settings, facilitating better communication in regions with diverse linguistic backgrounds. This ensures that patients from low-resource, multilingual regions receive equitable healthcare without language barriers compromising the patient-doctor relationship[3]. However, the risks of over-reliance on LLMs in direct patient communication must also be acknowledged. Some research argues that cultural sensitivity and local medical customs play a crucial role in healthcare communication. LLMs that fail to adapt to these factors may risk undermining trust between patients and healthcare providers. For instance, culturally inappropriate language or failure to understand local medical norms could weaken the patient-doctor bond and compromise care[11]. The diagram in Figure 3 shows a basic workflow of the integration of LLMs in healthcare.

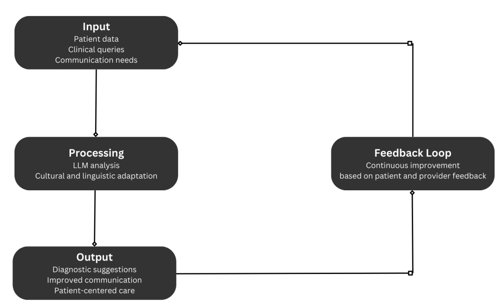


Figure 3: Workflow of LLM integration in healthcare

5 COMPARATIVE ANALYSIS OF THE MOST PROMINENT LLMS IN HEALTHCARE

Our review would be incomplete if we failed to mention the most prominent large language models that have been fine-tuned for

Model Name	Primary Applications	Cultural and Linguistic Adaptations	Current Limitations	Future Research Needs
GPT-4	Diagnostic support, patient communication	Multilingual capabilities, cultural sensitivity	Bias, accuracy issues	Domain-specific adaptations, ethical frameworks
L2M3	Health equity, multilingual support	Designed for multiple languages, cultural context integration	Limited domain-specific fine-tuning	Research on impact in low-resource regions
LlamaCare	Healthcare knowledge sharing	Tailored for healthcare terminology, multilingual support	May lack emotional sensitivity	Enhanced cultural adaptation, emotional sensitivity
GatorTron	Clinical decision-making, patient interaction	Adapted for diverse clinical contexts	Potential bias, limited contextual awareness	Improving contextual understanding and bias reduction
Additional Models (e.g., Me-LLaMA)	Specific clinical domains, patient interaction	Customizable for local languages and cultures	Challenges with generalizability	Further development for specialized clinical needs

Table 1: Comparison of LLMs in Healthcare Applications

applications in the medical domain. Therefore, this section will highlight the strong points and then perform a comparative analysis between GPT-4, LlamaCare, GatorTron, and L2M3. A graphic display of these models and their most prominent features in the medical domain can be found in Figure 4, while Table 1 below summarizes the models' strong points and current areas that can be improved, as well as areas where research can be focused in the future.

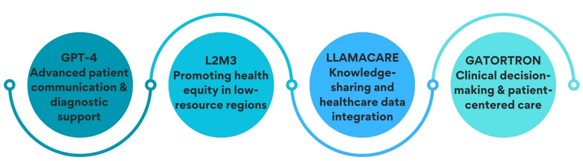


Figure 4: Overview of the key LLMs that are applied in healthcare

5.1 GPT-4

General Application in Medicine. As one of the most versatile LLMs, GPT-4 has been widely studied for its role in clinical decision-making, medical education, and patient engagement. GPT-4's broad applications, from summarizing medical records to assisting with diagnosis and patient interaction. However, GPT-4's lack of specialized medical training means it faces challenges when compared to models like LlamaCare and GatorTron[23].

Patient-Doctor Communication. GPT-4 has also demonstrated potential in improving patient communication by generating empathetic, human-like responses. The model can generate emotionally intelligent text, potentially enhancing patient trust in AI-generated advice[17]. Despite this, GPT-4 struggles with more complex medical queries where detailed clinical context is needed[24].

5.2 L2M3

Multilingual Healthcare Accessibility. One of L2M3's greatest strengths is its support for multilingual healthcare environments. L2M3 is a model designed specifically for low-resource healthcare settings where language barriers and cultural diversity pose significant challenges. By offering support for multiple languages, L2M3 has the potential to increase access to healthcare for marginalized communities globally[8].

Cultural Sensitivity. L2M3 goes beyond just linguistic adaptability by embedding cultural knowledge into its model[23]. This enables L2M3 to provide more contextually appropriate advice that aligns with local medical customs. For instance, it can recommend treatments or healthcare guidelines that resonate with the cultural practices of the patient's region, something that more generalized models like GPT-4 may fail to do effectively.

Consideration for Health Equity. L2M3 plays a pivotal role in reducing healthcare disparities by making culturally and linguistically appropriate care accessible in underdeveloped and diverse regions. It is particularly adept at filling gaps left by monolingual or culturally neutral models, such as GPT-4[23].

5.3 LlamaCare

Knowledge Sharing in Healthcare. LlamaCare was developed to facilitate knowledge sharing among healthcare professionals. Unlike GPT-4, which is designed for general applications, LlamaCare is fine-tuned specifically for healthcare, giving it an edge in clinical decision support[29]. LlamaCare's training on specialized medical datasets makes it particularly useful for knowledge-intensive tasks such as diagnosing complex conditions or synthesizing information from clinical trials[12].

Fine-tuned for Medical Data. LlamaCare's ability to provide accurate, context-specific information gives it a significant advantage over more general models. Its precision comes from training on large volumes of healthcare-specific datasets, allowing it to outperform models like GPT-4 when it comes to specialized clinical decision-making[26].

Patient Communication. Although LlamaCare's primary role is to assist healthcare providers, it can also be used to improve patient communication by offering detailed and reliable medical information. However, unlike GPT-4, which is more conversational, LlamaCare's focus remains on delivering precise medical knowledge rather than generating empathetic dialogue[4].

5.4 GatorTron

Focus on Medical Records. GatorTron is being specifically designed for the U.S. healthcare system, where it excels in processing vast amounts of electronic health records (EHRs). Its ability to rapidly synthesize and analyze patient data has made it an essential tool for improving diagnostic accuracy and reducing administrative burden[2]. GatorTron outperforms models like GPT-4 in this regard, owing to its fine-tuning on clinical records[7].

Enhancing Clinical Workflows. One of GatorTron's key strengths is its ability to streamline clinical workflows by accurately summarizing medical histories, diagnoses, and treatment plans.[14]. GatorTron enhances clinical efficiency by reducing the amount of time doctors spend on administrative tasks like reviewing patient charts, allowing them to focus more on patient care.[16].

Adaptation to U.S. Healthcare. GatorTron's design is optimized for the U.S. healthcare system, making it particularly effective in this context. However, this focus on English-language records means it may not be as adaptable in multilingual or international healthcare environments. GatorTron's success illustrates the need for more research on how to adapt models to diverse healthcare systems[27].

5.5 Comparative Analysis

Local Medical Customs and Linguistic Demands. When comparing L2M3 and GatorTron, L2M3's strength in multilingual settings is highlighted, which makes it ideal for global healthcare applications. GatorTron, by contrast, is more specialized for English-speaking environments. GPT-4 offers broader utility but lacks the cultural and linguistic specificity of L2M3 and the clinical precision of GatorTron and LlamaCare[23].

Patient Communication. GPT-4 and LlamaCare both demonstrate potential in patient-doctor communication, but they serve different purposes. GPT-4 excels in generating empathetic responses, while LlamaCare offers more medically precise information. L2M3's focus on multilingual communication gives it an edge in culturally diverse settings, but GatorTron's role remains more data-focused[17][9][26].

Customization and Accuracy. Both GatorTron and LlamaCare excel in accuracy due to their fine-tuning on medical data[30]. GPT-4, while highly versatile, does not have the same level of specialization[23]. L2M3, on the other hand, is a standout for global healthcare, particularly in low-resource regions[3].

6 CHALLENGES

6.1 Technical Challenges of Implementing Medical LLMs

Despite promising applications, adapting LLMs to specific medical tasks presents technical challenges. A major issue is the need for domain-specific training data. The models need considerable fine-tuning for clinical natural language processing (NLP) tasks, making zero-shot learning an emerging solution[20]. When it comes to the application of zero-shot learning in preventive healthcare, it has an ability to provide accurate decision support even in niche medical contexts[13]. In addition, fine-tuning LLMs to accommodate local medical customs and linguistic variations is critical for effective patient communication across diverse healthcare settings. That is why multilingual models like L2M3, which cater to low-resource regions where local dialects and cultural practices significantly influence healthcare delivery, have such importance[3]. Without incorporating local linguistic demands and medical traditions, LLMs risk misinterpretation, leading to misdiagnosis or poor patient outcomes.

6.2 Ethical Implications of LLMs in Healthcare

The ethical implications of deploying LLMs in healthcare extend beyond technical challenges, touching on critical issues such as data privacy, patient consent, and accountability. are substantial, particularly in terms of biases, transparency, and patient autonomy. While current studies briefly mention these concerns, this review delves deeper into specific ethical frameworks that should guide the development and deployment of LLMs. One such framework is the "Data Protection by Design" principle, which emphasizes the need to integrate privacy safeguards into AI systems from the outset. In addition, patient consent must be a core element when implementing LLMs in clinical settings, ensuring that patients are fully aware of how their data is used and stored. Accountability in AI decision-making also needs to be addressed, particularly in high-stakes scenarios like diagnostics and treatment planning, where errors can have life-altering consequences. Finally, the risk of bias and perpetuating healthcare inequities through LLMs necessitates stricter ethical oversight, with transparent mechanisms for identifying and mitigating bias. Racial and ethnic biases in GPT-4 were explored for medical diagnosis and triage, uncovering disparities in the model's responses. These concerns underscore the need for ethical oversight when using LLMs in culturally diverse settings[10]. Furthermore, LLMs could reinforce healthcare inequities if not properly adapted to the specific needs and practices of various cultures[6]. Additionally, LLMs must account for linguistic diversity in global healthcare contexts. For example, healthcare systems in multilingual nations, such as India, require models that can function across multiple languages while understanding the nuances of local medical customs. There is an innate link between culture, language, and patient safety, making linguistic competency critical for patient-centered care[11]. The LlamaCare model provides a framework for sharing healthcare knowledge across diverse linguistic groups, showing how LLMs can facilitate cross-cultural knowledge sharing in healthcare[22]. Figure 5 below focuses on the challenges, as opposed to the benefits, of applying LLMs in the healthcare domain.

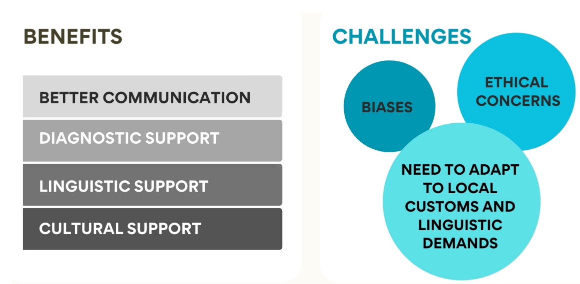


Figure 5: The benefits and challenges of applying large language models in the medical domain

7 EVALUATING THE EFFECTIVENESS OF LLMS IN HEALTHCARE

The evaluation of LLMs in healthcare requires a combination of qualitative and quantitative metrics to ensure a balanced assessment of their performance. In addition to the qualitative analysis of their language generation capabilities, this review incorporates quantitative metrics such as accuracy rates, error rates in specific tasks (e.g., diagnosis or patient communication), and patient satisfaction scores derived from recent studies. These metrics provide a robust framework[9] for comparing LLMs across various clinical applications. For example, models like GPT-4 and LlamaCare have been evaluated for their diagnostic accuracy, while GatorTron has shown promise in reducing administrative burdens through improved EHR management. Quantitative evidence, such as the accuracy rates of LLMs in diagnosing rare conditions or the error rates in patient-facing applications, strengthens the argument for their continued refinement and deployment in healthcare.

Evaluating LLMs in medical applications, focusing on adaptability to local linguistic demands, accuracy, and transparency requires a detailed framework[9]. There are METRICS in place that serve as a tool for standardizing LLM evaluations, ensuring they are tailored to local healthcare systems and relevant cultural needs[19]. In multilingual and multicultural settings, the evaluation of LLMs should prioritize their ability to maintain cultural and linguistic appropriateness. The L2M3 model, designed for regions with limited healthcare resources, demonstrates how multilingual and culturally aware LLMs can improve healthcare outcomes in underserved communities[3].

8 GAPS IN THE EXISTING RESEARCH

As LLMs become more prevalent in clinical environments, the focus on improving patient-doctor communication and addressing cultural and linguistic considerations will be crucial. Overreliance on LLMs without adequate human oversight, particularly in sensitive clinical areas like diagnostics and treatment planning is dangerous[29]. Additionally, there is a need for stringent ethical guidelines to ensure that LLMs do not perpetuate healthcare inequities[6]. In the papers that were evaluated for this review, we have identified the most notable gap in the limited focus on multicultural and linguistic diversity. There is a need for more comprehensive studies and models specifically designed for non-Western and linguistically diverse healthcare settings. LLMs must be fine-tuned to local medical practices and languages to truly serve global healthcare needs. The second gap identified is the bias that exists in LLM healthcare recommendations[10][6]. There is a lack of clear, actionable frameworks for systematically identifying and reducing bias in LLMs. Addressing this will require a multidisciplinary approach combining AI ethics, clinical expertise, and sociocultural considerations. The third gap is the emotional sensitivity in patient-doctor communication[1][21][18]. Finally, there is the danger of over-reliance on LLMs for decision support[6][29]. Still, the field is quite promising, so exploring future directions for research is worthwhile.

9 FUTURE DIRECTIONS FOR RESEARCH

The potential of LLMs in healthcare remains largely untapped, with significant gaps in current research that need addressing. One critical area for future investigation is the development of LLMs that are emotionally sensitive and capable of handling high-stakes, emotionally charged patient interactions. For example, models could be designed to recognize emotional cues in patient language and adjust their responses to provide more empathetic care. Another vital research direction is improving the contextual awareness of LLMs, particularly in culturally diverse settings where understanding local customs, languages, and medical practices is essential. Furthermore, future research should explore ways to integrate LLMs with other healthcare technologies, such as electronic health records (EHRs), to streamline clinical workflows while ensuring data privacy. Finally, the creation of robust, standardized evaluation frameworks[19][9] will be crucial in assessing the long-term effectiveness of LLMs in healthcare, focusing on patient outcomes, model accuracy, and ethical and regulatory compliance[6][15].

10 CONCLUSION

LLMs hold immense potential to revolutionize healthcare by enhancing diagnostic support, improving patient-doctor communication, and facilitating equitable care. Their success, however, hinges on addressing challenges such as local medical customs, linguistic diversity, and ethical concerns. Proper adaptation of LLMs can significantly enhance patient-centered care by making communication between patients and healthcare providers more culturally sensitive and linguistically appropriate. Prioritizing inclusivity and transparency is essential for improving healthcare outcomes and equity. Future research must focus on addressing cultural and linguistic diversity, improving emotional sensitivity in patient-doctor interactions, mitigating biases, and establishing ethical and legal frameworks for AI in healthcare. Specialized research is also needed to tailor LLMs for specific clinical domains, such as mental health and surgery, to ensure these tools are safe, reliable, and contextually aware.

11 ACKNOWLEDGEMENT

This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214).

Funded by
the European Union

REFERENCES

[1] Prottay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K Gopinath, Vijay Krishnan, Swati Kedia Gupta, Koushik Sinha Deb, and Tanmoy Chakraborty. Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study. *JMIR Mental Health*, 11:e57306, 2024.

[2] Xi Chen. Large language models in the medical field: Principles and applications. *International Journal of Computer Science and Information Technology*, 2(3):219–224, 2024.

[3] Agasthya Gangavarapu. Introducing l2m3, a multilingual medical large language model to advance health equity in low-resource regions, 2024.

[4] Marius Geantă, Daniel Bădescu, Narcis Chirca, Ovidiu Cătălin Nechita, Cosmin George Radu, Stefan Rascu, Daniel Rădăvoi, Cristian Sima, Cristian Toma, and Viorel Jinga. The potential impact of large language models on doctor–patient communication: A case study in prostate cancer. In *Healthcare*, volume 12, page 1548. MDPI, 2024.

[5] Tawara D Goode, M Claire Dunne, and Suzanne Bronheim. *The evidence base for cultural and linguistic competency in health care*. Commonwealth Fund New York, NY, 2006.

[6] Joschka Haltaufderheide and Robert Ranisch. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ Digital Medicine*, 7(1):183, 2024.

Indeks avtorjev / Author index

ChatGPT v zdravstvu

ChatGPT in Medicine

Uredniki > Editors:

Matjaž Gams, Monika Simjanoska Misheva, Stevo Lukić, Franz Wotawa