Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 1319–1328
https://doi.org/10.54941/ahfe1005699

Minimizing Chat-Bot Risks in Telemedical Applications: A Semi-Rule-Based System Approach for Large Language Model (LLM) Interactions in an Outpatient Setting

Simon Stock[1], Marius Gerdes[1], Florian Mazura[2], Markus Schinle[3], Jonathan Helmond[1], and Wilhelm Stork[1]

[1]Karlsruhe Institute of Technology, Germany
[2]FZI Forschungszentrum Informatik, Germany
[3]Offenburg University of Applied Sciences, Germany

ABSTRACT

We present a patient-centric system integrating Large Language Models (LLMs) into medical applications, focusing on a diverse set of use cases. An initial use case for symptom reporting was explored using natural language, addressing the limitations of traditional questionnaires. This collected data can be analysed by healthcare professionals during visits. Designed within the EU's Medical Device Regulation (MDR), our system incorporates a semi-rule-based approach to guide conversations, ensuring control over the LLM's outputs. With modular architecture and open standards like FHIR, our system supports personalized medicine and future advancements in AI-driven healthcare tools.

Keywords: Large language models (LLMs), Telemedical applications, Risk minimization, Systems engineering, Human-AI interaction

INTRODUCTION

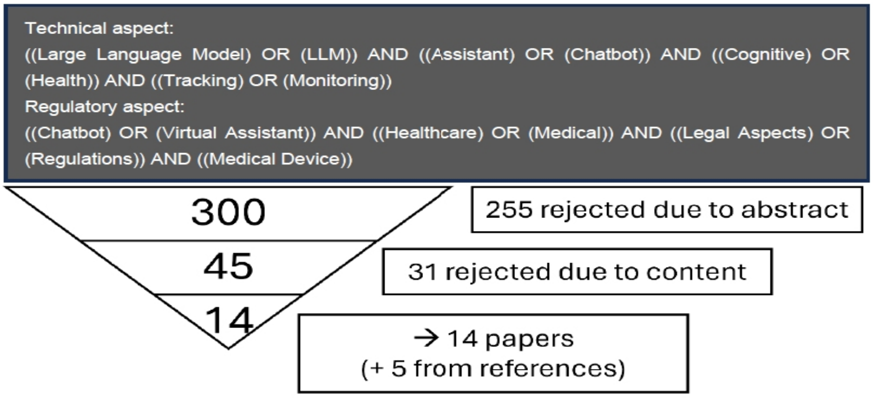
In recent years, telemedicine has experienced rapid growth, driven by advancements in digital technologies and the increasing demand for more accessible healthcare services. A critical component for the future of automated healthcare is the integration of conversational AI systems, particularly Large Language Models (LLMs), into health applications. LLMs, owing to their proficiency in processing and generating natural language, represent promising tools to enhance healthcare efficiency, especially in outpatient settings. However, the implementation of these models also introduces several risks. These include the potential for incorrect medical advice, concerns over patient privacy, and the challenge of limited regulatory oversight. This paper examines the application of a semi-rule-based system designed to mitigate these risks. The goal is to balance the flexibility and adaptability of LLMs with the safety and precision necessary for medical applications.

Figure 1: Logo of the METIS project.

This research was conducted within the METIS transfer-research project at the Karlsruhe Institute of Technology (KIT). METIS is a digital platform (Schinle et al., 2023) designed to support patients with neurodegenerative diseases through a companion app, complemented by a web interface for healthcare providers to manage treatment plans and interventions. A distinctive feature of the METIS project is its startup-like approach, which addresses not only the technical possibilities but also the practical challenges faced by real-world med-tech startups (Gerdes et al., 2024). As part of this project, the potential use of Large Language Models (LLMs) was evaluated. However, initial assessments revealed significant concerns regarding their accuracy, particularly in terms of the classification required for medical devices. Additionally, the vast range of potential use cases would make comprehensive testing nearly impossible. Therefore, this work introduces a semi-rule-based system that incorporates LLM technology while ensuring compliance with medical device regulations.

BACKGROUND AND RELATED WORK

For this study, we adopted the QUA^{3}CK development process proposed by (Becker et al., 2019). Building on this framework, a structured literature review was conducted. Two distinct search strings were designed, as the intersection of the initial search terms did not yield valuable results. The search strings used were:



Technical aspect:
((Large Language Model) OR (LLM)) AND ((Assistant) OR (Chatbot)) AND ((Cognitive) OR (Health)) AND ((Tracking) OR (Monitoring))
Regulatory aspect:
((Chatbot) OR (Virtual Assistant)) AND ((Healthcare) OR (Medical)) AND ((Legal Aspects) OR (Regulations)) AND ((Medical Device))

Figure 2: Resulting papers from the structured literature review. Five papers were added from references within the selected 14 papers.

Initially, we reviewed 300 papers that matched the search criteria at the abstract level to identify those potentially relevant (see Figure 2 for an overview). Of these, 45 papers were identified as likely to provide valuable insights for this research. The next step involved a more detailed examination of these 45 papers, though not a full reading. Following this process, 14 papers were selected as directly valuable and were then thoroughly reviewed to extract pertinent information and knowledge.

In addition, through the analysis of key references in the reviewed literature, five more papers were identified as relevant, contributing further insights into the research topics.

Table 1. Taxonomy and related work to this work.

ID	Paper	Strucutre of a LLM Chatbot System	LLM Models and Techniques	Prompt Engineering	Access Health Data	Accuracy	Safety
1	Abbasian et al., 2023	X			X		
2	Montagna et al., 2023	X					
3	Usman Hadi et al., 2023		X				
4	Wei et al., 2023		X	X			
5	Subramonyam et al., 2023		X				
6	Ahmadi and Fox, 2023		X				
7	Maia et al., 2023	X			X	X	
8	Zhang et al., 2020			X			
9	Lei et al., 2021				X		
10	Hauglid et al., 2023						
11	Sarkar et al., 2023						
12	He et al., 2023		X			X	X
13	Jo et al., 2023		X	X			
14	Min et al., 2022		X				
15	Banerjee et al., 2023					X	
16	Thirunavukarasu et al., 2023		X			X	X
17	Bobrow et al., 1977			X			
18	Jurafsky et al., 2000					X	
19	Vaghefi et al., 2023		X				
0	This paper	X		X	X	X	X

As highlighted by (Jo et al., 2023) and (Lei et al., 2021), Large Language Models (LLMs) should function as supporting tools within the broader healthcare service framework. While LLMs can enhance and streamline healthcare processes, the indispensable roles of healthcare professionals, family members, and caregivers remain paramount.

Our research indicates that, particularly in recent years, LLMs—such as GPT-based models—have seen increasing adoption in healthcare for tasks like patient interaction and symptom tracking. Their ability to manage complex linguistic tasks and provide real-time communication makes them well-suited for telemedicine applications. However, deploying LLMs in sensitive environments like outpatient healthcare introduces several risks. These risks include providing incorrect responses, misinterpreting medical symptoms, and failing to offer context-sensitive advice, which are critical shortcomings in patient safety.

Regulatory frameworks, such as the EU's Medical Device Regulation (MDR), stipulate that AI systems in healthcare must adhere to strict standards to ensure safety and reliability. Yet, there remains a lack of focused research on the regulatory requirements for LLMs in medical devices, particularly concerning safety. This gap was only broadly covered by (He et al., 2023) and (Thirunavukarasu et al., 2023). Current literature predominantly addresses the prototyping and initial testing phases of LLMs, without thoroughly exploring the legal and regulatory aspects. As a result, a significant gap persists in understanding how to fully integrate LLMs into healthcare while ensuring compliance with safety and regulatory standards.

CURRENT LIMITATIONS OF LLMS

LLMs present unique risks in telemedicine due to their reliance on extensive datasets, which may not always align with the specific, context-sensitive requirements of healthcare. This discrepancy raises concerns about the accuracy and safety of medical interactions facilitated by these models. The key limitations of LLMs are as follows: **Non-deterministic behavior**: LLMs exhibit non-determinism due to the vast number of potential input variations, which can lead to inconsistent outputs. **Outdated information**: LLMs may provide outdated information, as their training data does not always include the most recent medical advancements. **Accuracy issues**: LLMs can lack accuracy, especially if there are gaps in the underlying knowledge. Furthermore, they may misinterpret context, leading to incorrect or irrelevant advice. **Coherence challenges**: LLMs lack true understanding and can generate fabricated or misleading information, commonly referred to as “hallucinations.” **Transparency concerns**: The process by which LLMs generate answers is often opaque, making it unclear which data sources were used or how conclusions were derived. **Ethical risks**: LLMs may produce responses that are potentially harmful, offensive, or discriminatory, presenting significant ethical concerns. **Resource-intensive**: The training and deployment of LLMs require considerable time and computational resources, making them costly to develop and maintain. Addressing these limitations is critical for the safe and effective integration of LLMs in healthcare. Ideally, solutions should be implemented either through the chatbot design or by the chatbot engine provider to mitigate these risks.

CONCEPT AND IMPLEMENTATION

A hybrid approach, combining rule-based mechanisms with LLM interactions, offers a potential solution to mitigate the risks associated with using AI in healthcare. In this system, different techniques are employed, utilizing rule-based algorithms to manage critical tasks and ensuring safety and compliance with medical standards. To maintain the system’s intended use, every incoming message is classified according to specific use cases. Any classified use case that is not supported or is safety critical is redirected from the LLM cascade to a rule-based algorithm. To prevent misinterpretation and increase accuracy, critical decisions - such as confirming medical information - are handled by deterministic algorithms triggered by patient input, bypassing LLMs entirely. Additionally, to maintain coherence and reduce the risk of hallucinations, each system response includes a summary of the information understood by the LLM. These measures significantly reduce risks and improve patient safety. The semi-rule-based system operates by first assessing the context of each patient interaction. For high-risk medical queries, such as diagnosis or medication advice, the system relies on predefined responses based on clinical guidelines. LLMs are utilized for lower-risk interactions, such as answering general questions, small talk or scheduling appointments. A key example is the use case of symptom tracking, though the system can accommodate many other non-critical use cases. The general interaction pattern for symptom tracking is as follows (compare Figure 3).

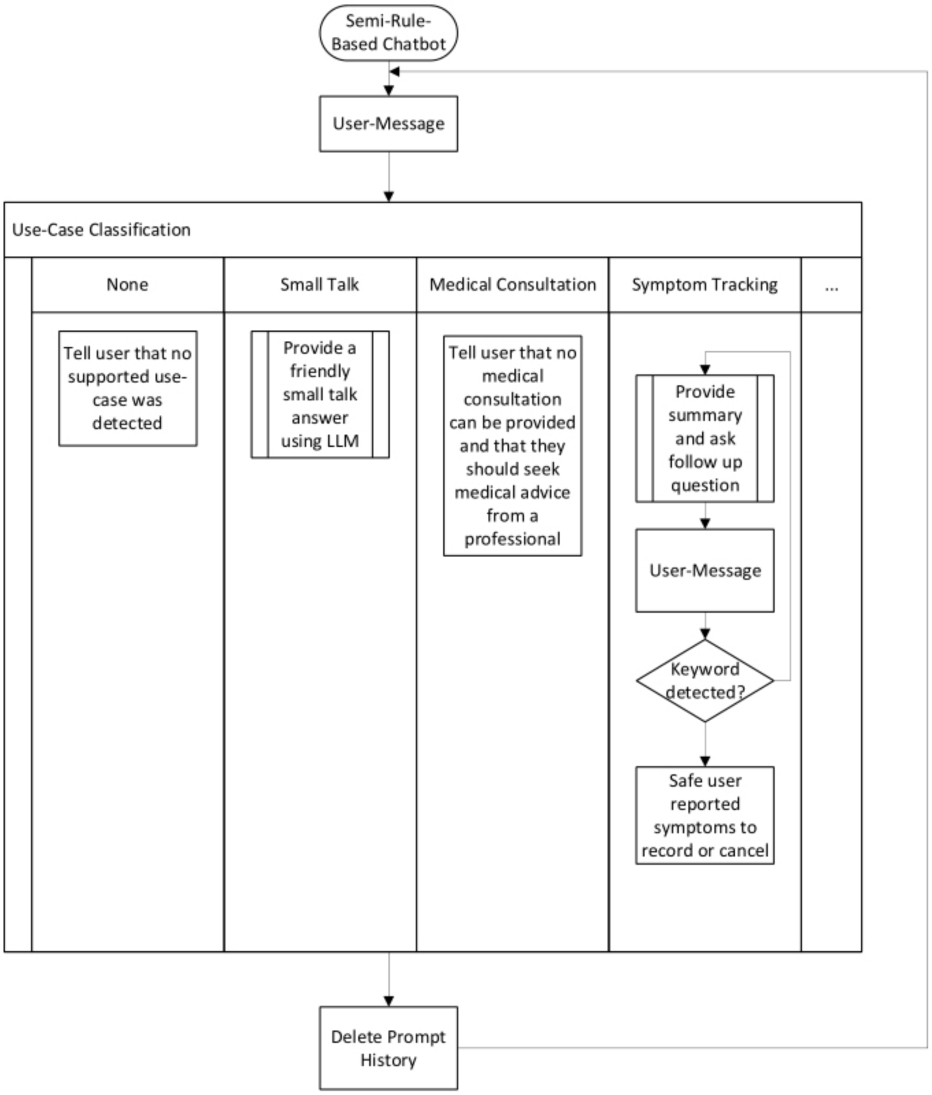

Semi-Rule-Based Chatbot
User-Message
Use-Case Classification
None
Small Talk
Medical Consultation
Symptom Tracking
...
Tell user that no supported use-case was detected
Provide a friendly small talk answer using LLM
Tell user that no medical consultation can be provided and that they should seek medical advice from a professional
Provide summary and ask follow up question
User-Message
Keyword detected?
Safe user reported symptoms to record or cancel
Delete Prompt History

Figure 3: Flowchart illustrating the semi-rule-based chatbot system approach. The process begins with the classification of the initial user message, based on predefined use cases. Depending on the classification, different actions are triggered. High-risk use cases are handled by rule-based algorithms to ensure safety, while lower-risk interactions are processed by the LLM system. This division of tasks ensures that critical decisions are made reliably, while LLMs manage less sensitive tasks.

The user reports a problem → The chatbot asks a clarifying question → The user responds → The chatbot summarizes the information and asks another question.

This cycle continues until sufficient data is collected, with the chatbot summarizing and saving the information to the patient's record. The system message guiding the chatbot instructs asking one question at a time and avoiding any medical advice. Iterative testing has shown that the label "You are a Medical Device Class IIa," improves accuracy of the instructions significantly. Early attempts to have the LLM determine when symptom tracking should conclude were unsuccessful, as the LLM inconsistently classified conversation endings. As a result, user input (using keywords like "CONFIRM" or "CANCEL") is now required to finalize the symptom tracking. The chatbot then uses the confirmed summary of the patient's symptoms based on the conversation to save it to the METIS system's FHIR database. Ultimately, a physician can then access this data on the METIS Web-interface at the time of patient visit.

Testing revealed that combining multiple tasks in a single LLM prompt (such as generating responses and summaries together) resulted in increased errors. To enhance performance, these tasks were separated into distinct processes. The final implementation of the system modules is presented in Figure 4. The chat backend was developed using the Matrix framework. According to matrix.org, "Matrix defines a set of open APIs for decentralized communication, suitable for securely publishing, persisting, and subscribing to data over a global open federation of servers with no single point of control." Matrix is an open standard for communication and instant messaging, and it as well serves as the backbone for the TI-Messenger of Gematik, the communication app of the German healthcare system. The core component of this architecture is the Semi-Rule-Based Chatbot Server, which acts as the intermediary connecting users to the system described in Figure 3. For medical use cases that involve accessing patient data (either reading or writing), the server establishes a secure connection with the FHIR database storage. An important design consideration was ensuring that the LLM interface is modular and easily interchangeable. In our trials, we connected the system to GPT-3.5 and GPT-4.0 via a Microsoft Azure subscription. However, due to privacy and data protection concerns, particularly in relation to GDPR and MDR compliance, a final implementation intended for patient use or deployment within a healthcare system must evaluate whether such external services are compliant. Alternatively, self-hosted solutions, such as those built on LLaMA-based networks, may offer a more feasible option for maintaining data privacy and security.

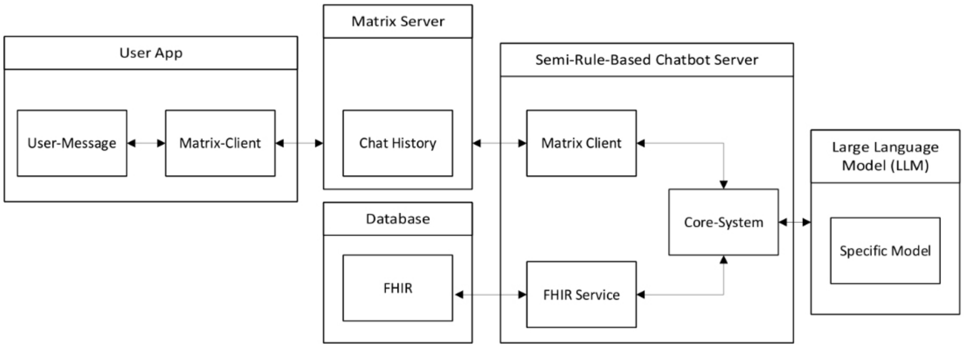


Figure 4: System architecture of the overall implementation. The individual boxes represent different implemented entities, such as the semi-rule-based chatbot server, matrix framework, and fhir database.

REGULATORY ASPECTS OF IMPLEMENTATION

Class IIa devices face stricter requirements than Class I, as outlined in Annex IX of the Medical Device Regulation (MDR). Chapter I mandates a quality management system, while Chapter III requires administrative provisions for 10 years. Key MDR requirements include ensuring the device performs as intended, is effective, minimizes risks, and complies with the General Data Protection Regulation (GDPR). For software, additional requirements focus on repeatability, reliability, and state-of-the-art development practices that consider lifecycle, risk, and information security management. The MDR is harmonized with ISO 13485, which provides best practices for compliance. Additionally, the Artificial Intelligence Act (AIA), effective August 2024, classifies AI systems in Class IIa or higher as "high risk." However, the full impact of the AIA on medical devices is still unclear, and there is no consensus among Notified Bodies on uniform requirements for AI-enabled devices. In our system the symptom tracking use case collects patient health data, focusing solely on data collection without offering diagnoses or medical advice. This separation from symptom monitoring reduces patient risk, as only data collection accuracy needs validation. Given that LLMs cannot yet meet regulatory standards for medical interpretation, the chatbot must be carefully designed to avoid providing unsolicited medical advice. A typical test case involves the system interacting with patients in outpatient settings. Predefined rule-based responses or use-case specific LLM-prompts ensure safety. Our designed prompt was the following:

> You are an intend classification for a chat. Say what the current use case is. They can be: {use_cases}. Answer like 'use case: small talk'
>
> This are the explanations of the use cases: {use_cases_explanations}
>
> This is the chat history: {chat_history}
>
> In case no use case fits, answer with none.
>
> use case:

This hybrid approach as illustrated in Figure 3 allows the system to maintain patient safety while enhancing user experience through more natural, conversational interactions using LLMs and use-cases.

RESULTS

In-house testing of the semi-rule-based system showed significant improvements in medical output accuracy and regulatory compliance compared to LLM-only systems. Usability testing indicated high user satisfaction, especially in conversational interactions. However, limitations were noted when handling unforeseen medical situations not covered by predefined rules. While the system mitigates many risks, ongoing human oversight is still necessary to ensure adherence to medical standards.

During the design phase, we focused on creating an LLM-based solution that complies with medical device regulations while optimizing performance. Various design concepts were explored to balance safety and LLM benefits. The most promising strategies were based on four principles: transparency, user control, deterministic algorithms, and state classification to keep the chatbot within its intended use.

Development followed a modular approach, leveraging Langchain for its scalability and adaptability to future LLM advancements. Integration with healthcare platforms like Metis was enabled using open standards like FHIR and Matrix, ensuring interoperability. Libraries such as Langchain and matrix-nio were selected to minimize third-party risks and ensure patient safety.

For deployment, Docker provided portability and scalability. Prompt development was an iterative process, refining LLM behavior to optimize performance and minimize compliance risks.

CONCLUSION AND FUTURE WORK

This paper presented a semi-rule-based system that combines deterministic rule-based approaches with LLMs to minimize risks in telemedicine applications. The system offers a promising solution for deploying AI in outpatient care. Future work will focus on expanding predefined rules and improving the system's ability to handle complex medical scenarios autonomously. Additionally, a user study in clinical or pre-clinical settings will be conducted to evaluate user satisfaction and system safety.

The evaluation highlighted areas for improvement. Enhancing the chatbot with a persona (e.g., name or face) could boost user engagement. Implementing a keyword correction system would reduce errors from misspellings, improving accuracy. To improve instruction-following in long conversations, summarizing chat history could help the chatbot focus on relevant information, though some details may be lost. Emergency detection features could also be integrated, especially for identifying critical symptoms. For elderly users, a text-to-voice feature or integration with a robot system (Stock et al., 2023) could enhance accessibility, and expanding language support with, automatic detection would allow the chatbot to adapt to the user's language. Finally, adding graphical elements like a "CONFIRM" button could improve user-friendliness across platforms.

REFERENCES

Arun James Thirunavukarasu et al. "Large language models in medicine". In: *Nature Medicine* 29.8 (Aug. 2023). Number: 8 Publisher: Nature Publishing Group, pp. 1930–1940. issn: 1546-170X. doi: 10.1038/s41591-023-02448-8.

Becker, Jürgen, et al. "The QUA^{3}CK Machine Learning Development Process and the Laboratory for Applied Machine Learning Approaches (LAMA)." Symposium Artificial Intelligence for Science, Industry and Society (AISIS 2019), Mexiko-Stadt, Mexiko, 20.10. 2019–25.12. 2019. 2019.

Dan Jurafsky, Daniel Jurafsky, and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Upper Saddle River, NJ: Prentice Hall, 2000. 934 pp. isbn: 978-0-13-095069-7 978-0-13-122798-9.

Daniel G. Bobrow et al. "GUS, a frame-driven dialog system". In: *Artificial Intelligence* 8.2 (Apr. 1, 1977), pp. 155–173. issn: 0004-3702. doi: 10.1016/0004-3702(77)90018-2.

Debarag Banerjee et al. *Benchmarking LLM powered Chatbots: Methods and Metrics*. Aug. 8, 2023. arXiv: 2308.04624[cs].

Eva Maia, Pedro Vieira, and Isabel Praça. "Empowering Preventive Care with GECA Chatbot". In: *Healthcare* 11.18 (Jan. 2023). Number: 18 Publisher: Multidisciplinary Digital Publishing Institute, p. 2532. issn: 2227–9032. doi: 10.3390/healthcare11182532.

Gerdes, Marius, et al. "Digital Healthcare Applications for Preventing Risk Factors in Dementia: Requirements Engineering in the German Healthcare System." *Alzheimer's Association International Conference*. ALZ, 2024.

Hariharan Subramonyam et al. *Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces*. Sept. 25, 2023. doi: 10.48550/arXiv. 2309.14459.

Hannah Lei et al. *COVID-19 Smart Chatbot Prototype for Patient Monitoring*. Aug. 12, 2021. doi: 10.48550/arXiv.2103.06816. arXiv: 2103.06816[cs].

Jingwen Zhang et al. "Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet: Viewpoint". In: *Journal of Medical Internet Research* 22.9 (Sept. 30, 2020). JMIR Publications Inc., Toronto, Canada, e22845. doi: 10.2196/22845.

Jing Wei et al. *Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data*. Sept. 22, 2023. doi: 10.48550/arXiv.2301.

Kai He et al. *A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics*. Oct. 9, 2023. doi: 10.48550/arXiv.2310.05694.

Mahyar Abbasian et al. *Conversational Health Agents: A Personalized LLM-Powered Agent Framework*. Oct. 20, 2023. doi: 10.48550/arXiv.2310.02374.

Muhammad Usman Hadi et al. *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*. Nov. 16, 2023. doi: 10.36227/techrxiv.23589741.v4.

Mathias Karlsen Hauglid and Tobias Mahler. "Doctor Chatbot: The EUs Regulatory Prescription for Generative Medical AI". In: *Oslo Law Review* 10.1 (June 30, 2023). Publisher: Universitetsforlaget, pp. 1–23. doi: 10.18261/olr. 10.1.1.

Sara Montagna et al. “Data Decentralisation of LLM-Based Chatbot Systems in Chronic Disease Self-Management”. In: *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. GoodIT ‘23. New York, NY, USA: Association for Computing Machinery, Sept. 6, 2023, pp. 205– 212. isbn: 9798400701160. doi: 10.1145/3582515.3609536.

Sareh Ahmadi and Edward A. Fox. *AI Chatbot for Generating Episodic Future Thinking (EFT) Cue Texts for Health*. Nov. 6, 2023. doi: 10.48550/arXiv.

Schinle, Markus, et al. “Model-Driven Dementia Prevention and Intervention Platform.” *Caring is Sharing–Exploiting the Value in Data for Health and Innovation*. IOS Press, 2023. 937–941.

Surjodeep Sarkar et al. *Towards Explainable and Safe Conversational Agents for Mental Health: A Survey*. Apr. 25, 2023. doi: 10.48550/arXiv.2304.13191.

Eunkyung Jo et al. “Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention”. In: *Proceedings of the 2023 CHI Conference on Human Fac- tors in Computing Systems*. CHI ‘23: CHI Conference on Human Factors in Computing Systems. Hamburg Germany.

Sewon Min et al. *Rethinking the Role of Demonstrations: What Makes In- Context Learning Work?* Oct. 20, 2022. arXiv: 2202.12837[cs].

Saeid Ashraf Vaghefi et al. *chatClimate: Grounding Conversational AI in Climate Science*. Apr. 28, 2023. doi: 10.48550/arXiv.2304.05510.

Stock, Simon, et al. “AI, Robotics, and Clinical Research for Innovative Dementia Interventions: A Japanese-German Collaboration.” *International German-Japanese Workshop on Digital Dementia Intervention and Prevention Strategies in the Age of AI (2023), Karlsruhe, Deutschland, 26.06. 2023–27.06.2023.*