Generative AI in the Hospital: A Participatory Assessment of Healthcare Needs and Challenges

Anonymous Author(s)
Affiliation
Address
email

Abstract

Although Large Language Models (LLMs) have shown promising performance in several medical applications, their deployment in the medical domain poses unique challenges of ethical, regulatory, and technical nature. In this study, we employ a systematic participatory approach to investigate the needs and expectations regarding clinical applications of LLMs at a major European university hospital. Having identified potential LLMs use-cases in collaboration with stakeholders, we assess the current feasibility of these use-cases focusing on pressing challenges such as regulatory frameworks, data protection regulation, bias, hallucinations, and deployment constraints. This work provides a framework for a participatory approach to identifying institutional needs with respect to introducing advanced technologies into clinical practice, and a realistic analysis of the technology readiness level of LLMs for medical applications, highlighting the issues that need to be overcome to enable safe, ethical and regulatory compliant use of medical LLMs.

1 Introduction

Large Language Models (LLMs) have significantly improved the capabilities in processing large amounts of unstructured text, information retrieval, text summarization, and assistive tasks. Because of these recent advancements, there is an increasing expectation on the use of LLMs in the clinical domain—where unstructured text is a common data format—to overcome pressing healthcare issues such as lack of qualified clinical professionals, increasing pressure on healthcare due to an aging society, alarm fatigue, and a ever-growing administrative burden.

Indeed, LLMs have shown promising performance on several healthcare tasks including medical writing and editing, medical education, answering medical questions and analysing electronic health records (Thirunavukarasu et al., 2023; Michael G. Madden, 2023; Umeton et al., 2024; HyoJe Jung, 2024; Kirk Lower, 2023)

Although the potential applications are numerous, there is a need to prioritize the most impactful, beneficial, and feasible use cases. LLMs in the medical domain present several challenges, which include the ethical and regulatory obligation to protect patient privacy, the risk of indirect patient harm in case of model hallucinations, errors or biases, the need for model interpretability, regulatory challenges with respect to medical device regulations, challenges related to local deployment and long-term costs, and intellectual property, among others.

In this study, we use a systematic participatory approach to evaluate the emerging needs and expectations of a healthcare institution regarding medical LLMs. Having identified top-priority institutional use-cases, we provide an assessment of technical, ethical and regulatory feasibility, highlighting existing challenges and blockers. Using this approach, we identify potential use-cases, and elucidate relevant evaluation criteria to rank them. Finally, we provide an assessment of the current technology

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

readiness level of medical LLMs, and highlight the efforts needed from the scientific community, healthcare institutions and regulatory bodies, to allow LLMs-based technologies to address healthcare needs.

2 Methods

Setting This study was performed at a large university hospital in Europe. The intended implementation of LLMs applications at the hospital required sponsorship from the medical direction, legal department, and IT.

Working group A working group (WG) of institutional stakeholders was created to discuss needs, opportunities and strategic priorities.

Specifically, the working group included clinical representatives of 11 departments across the hospital, representatives of nursing and medical directions, patients representatives, legal experts and representatives from the information technology (IT) department, the clinical informatics unit and the biomedical data science center of the institution. External advisors from 2 other other hospitals in the same country and a technical university from the same country were invited to be part of the group, for a total of 32 people.

Internal processes Monthly discussions were held among WG members over a six-month period. The initial meeting focused on discussing the vision, brainstorming institutional needs, and identifying potential use-cases where LLMs could add value. In the second meeting, we established key evaluation criteria to score the use-cases. Evaluation criteria were mostly based on transversal goals such as potential impact on patient engagement, quality of care, reduced administrative workload, and technical feasibility.

Structured survey Following the identification of the use-cases and their evaluation criteria, we designed and distributed a structured survey among the WG members (see Appendix for details), allowing them to vote for the five most relevant use-cases from the predefined list and rank them according to the identified evaluation criteria. The survey was developed in Qualtrics Version, [06.2024].

Rankings of the use-cases for each evaluation criterion, based on the survey results, were further discussed in subsequent WG meetings until a final agreement was reached. Also relative weight (importance) of each evaluation criteria was discussed.

Final ranking of top-priority use-cases was obtained using a weighted sum $R_i = \sum_{k=1}^{m} w_k \cdot R_{i,k}$, where $R_{i,k}$ is the rank of the i-th use-case according to the k-th evaluation criterion, w_k is the weight of the k-th evaluation criterion, and m is the total number of evaluation criteria.

3 Results

Vision and key evaluation criteria Based on the outcomes of the WG, we identified five major areas of improvement: quality of care, patient engagement and satisfaction, administrative burden, research and digital innovation, and continuous education. We then identified specific use-cases for LLMs applications that target these areas, see complete list in Table 1.

The key evaluation criteria that we have identified, along with an agreement on their relative weight, are: institutional impact [high weight], impact on patient management and quality of care [medium weight], impact on patient satisfaction and engagement [medium weight], impact can be measured with KPIs [high weight], intended use as software as medical device [medium weight], safety defined as perceived absence of risk of harm or misinformation [high weight], and whether a commercial solution is available [low weight].

Expectations for the clinical LLM applications Sixteen members of the WG, corresponding to one representative per service or division, participated in the survey.

Table 1: Use-cases for medical LLMs ranked by healthcare stakeholders

Use-cases	Areas					Ranking	
	QC	**RI**	**PE**	**Adm**	**Edu**	**Popularity**	**Weighted**
Assistive chatbot for complex cases	•	•				1	1
Summarization of discharge letters into patient adapted language	•		•			4	2
Automatic generation of discharge letters				•		3	3
Automatic analysis of patients feedback and complaints	•		•			5	4
Smart retrieval of guidelines	•				•	5	5
Entry notes and medical records summarization	•			•		5	6
Smart summarization of current literature					•	5	7
Chatbot for education of healthcare professionals					•	2	8
Smart search over patients records	•	•				6	–
Chatbot to increase health literacy in patients			•			6	–
Clinical trial matching		•				7	–
Automatic generation of imaging reports	•			•		7	–

QC: Quality of care, **RI**: Research & innovation, **PE**: Patient engagement & satisfaction, **Adm**: Administrative burden, **Edu**: Education

Ranking of medical LLMs use-cases by popularity is shown in Table 1. *Assistive chatbot for complex cases*, *chatbot for education of medical professionals* and *automatic generation of discharge letters* were the most voted use-cases with 10, 8 and 7 votes respectively.

The survey results provided preliminary rankings of the use-cases for each evaluation criterion (see Appendix for additional results), which we then confirmed or modified at open discussions among the WG members. For simplicity, the discussion was limited to the use-cases that were selected by at least 5 respondents, in our case, corresponding to use-cases with popularity ranking 1 to 5. *Assistive chatbot for complex cases* was identified as the top priority use-case for the institution, followed by *summarization of discharge letters into patient adapted language* and *automatic generation of discharge letters*. The *assistive chatbot for complex cases* ranked highest in institutional impact, workload, quality of care, and measurable impact, but was second to last in risk of harm. *Summarizing discharge letters into patient-adapted language* ranked highest for patient satisfaction and engagement, while *summarizing current literature* was identified as the safest use-case. *Chatbot for education of healthcare professionals* ranked last in terms of perceived safety, followed by *assistive chatbot for complex cases* and *automatic generation of discharge letters*. Overall ranking, obtained with the above formula, and final ranking for each evaluation criteria after discussion with the WG is shown in Fig. 1.

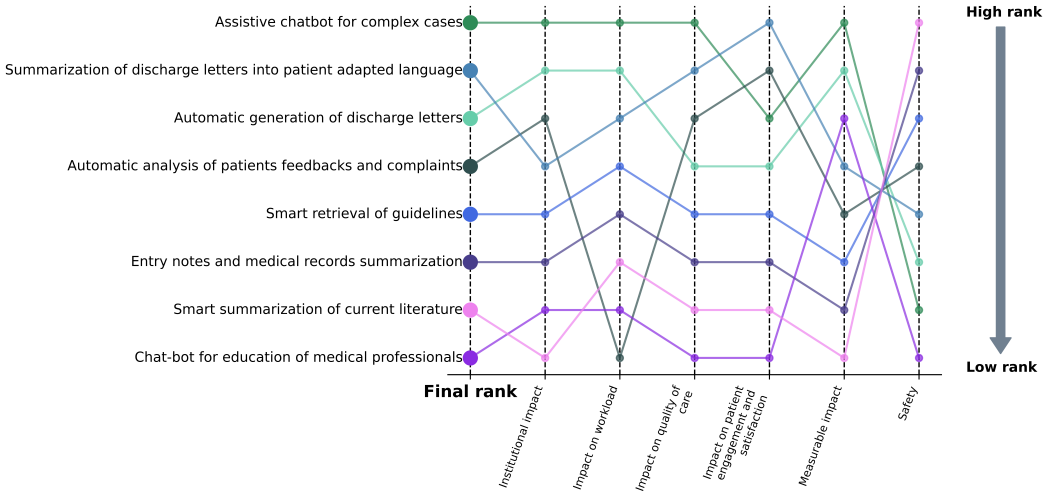

Figure 1: Use-cases final ranking and evaluation criteria.

4 Discussion

In this study we present a systematic approach to assess needs and expectations of a European healthcare institution regarding medical LLMs. The results of this study indicate that institutional stakeholders consider LLMs as a valuable tool to improve differential diagnosis, assisting with administrative tasks, improving patient engagement and monitoring quality of care across the institution. Although preliminary studies suggest that LLMs may reach sufficient performance to answer these needs, performance alone does not ensure successful clinical implementation. When dealing with the strategic implementation of LLMs in a healthcare institution, it is important to ponder the remaining challenges that affect LLMs development, deployment and implementation. In the following section, we reflect on these challenges.

Regulation of medical uses Under the United States (US) Food and Drugs Administration (FDA) regulations, European Law on Medical Devices Regulations (MDR), and the recently approved European Union (EU) AI Act, general purpose LLMs would not automatically be classified as medical devices (Meskó, 2023). It is the intended use that dictates the regulatory framework. As such, LLMs, or general-purpose LLMs, that are developed, fine-tuned or modified in order to serve a more specific medical purpose might be treated as medical devices. Classification as medical device triggers a series of requirements that may be challenging to meet for LLMs. For example, regulatory requirements apply to the entire development lifecycle of the device, not only to the phase where the model is adapted to medical purpose, posing a challenge for models that derive from general-purpose LLM. For these models, the development process most likely did not adhere to stringent medical device regulations. Even if it did, the resulting documentation may not be publicly accessible.

Moreover, current risk assessment approaches may be inadequate for LLMs. For example, in the case of an *assistive chatbot for complex cases*—identified as a top priority use-case in our survey—several questions arise: How can we conduct a comprehensive risk assessment considering that the questions that future users may ask are potentially infinite? what role does the context in which the tool is used play in relation to its safety? Should we consider user-training as the main risk mitigation strategy for medical LLMs? Finally, in case of a LLMs-driven adverse event, would *for-cause* auditing be possible considering LLMs limited explainability? To date, these remain open questions, and proposed risk mitigation strategies focus on extensive user training and consequent user liability. In practice, starting from the assumption that users of medical LLMs tools would be capable of identifying subtle issues in LLMs output. This framework could be particularly problematic for applications such as *Chatbot for education of healthcare professionals* or *Chatbot to increase health literacy in patients*, where users may have limited abilities to question LLMs-generated content. Regarding clinical support applications, clinical investigations will play a crucial role in LLMs risk assessment, but to the best of our knowledge, no example exists to date.

The regulatory framework for continual fine-tuning, or retraining, is another crucial point. In January 2021, the FDA took a step toward addressing this issue by publishing its action plan to facilitate AI-/ML innovation (FDA). The proposition was to regulate AI-powered software as medical device (SaMD) throughout their lifespan, introducing the so-called *"predetermined change control plan"*. The guiding principles for the predetermined change control plans for ML-enabled medical devices were later published in October 2023. With this action the FDA aimed at aligning the speed of regulatory certification with the rapid release of new highly-performant ML models, including LLMs. The EU followed a similar approach in the newly released AI Act (EU), which states that new certification is not deemed for changes in algorithms or algorithms performance that were pre-determined by the manufacturer and pre-assessed at the time of relevant conformity assessment. It remains unclear how this process will work for LLMs as further re-training or fine-tuning typically involve new, larger datasets and a significantly different architecture. Continuous re-training may be required for many of the discussed use-cases, given the continuous advancements of medical care, new guidelines and scientific discoveries.

Hallucinations and omissions In their outputs, LLMs tend to both provide contextually plausible but factually incorrect information, a phenomenon known as hallucinations, as well as potentially omit crucial pieces of information (Ji et al., 2023). Depending on the use-case, these could pose critical issues to the viability of solving a problem with LLMs. For instance, even though a *chatbot for medical professionals education* was deemed a useful tool in our survey, the working group members agreed that it poses significant risk of harm due to hallucinations and omission, thus potentially

misleading inexperienced medical clinicians. Solving these issues is an open problem, and there is no consensus on whether the solution can exist at all, with hallucinations or omissions possibly being an inherent feature of the way the current generation of LLMs are produced (Bender et al., 2021). The pragmatic solutions often require additional application-specific infrastructure on top of LLMs, such as retrieval-augmented generation (Lewis et al., 2020) or precise tooling (Schick et al., 2024). The reliability and safety of medical use-cases where these are an issue will thus depend on the reliability of infrastructure for preventing hallucinations or incorrect omissions.

Bias In certain ways, LLMs can be akin to parrots (Bender et al., 2021), regurgitating low-quality information obtained from their training data, oftentimes containing text from web sites such as reddit and Wikipedia. It should come as no surprise that LLMs could propagate outdated or generally harmful medical beliefs rooted in pseudo-science, conspiracy theories, racism or sexism (see, e.g., Omiye et al., 2023), and incorporate these beliefs in downstream use-cases, e.g., in the recommendations given in the *chatbot* use-cases, be it for *medical education* or *assistance with complex clinical cases*, or *entry notes summarization*. Like with hallucinations, it is unclear whether this issue is solvable with the current generation of LLMs, but pragmatic deployment requires rigorous evaluation in terms of bias (Gallegos et al., 2024).

Infrastructure and data protection Although some medical institutions are able to partner with external LLM training and inference providers (Umeton et al., 2024), this way of incorporating LLMs into clinical practice comes with issues. First, in the case of commercial providers, it leads to the political issue of "opaque commercial interests" (Toma et al., 2023) guiding healthcare solutions. Second, for use-cases involving patient data, e.g., *summarization of discharge letters,* the usage of external infrastructure is complicated by the ethical and legal data protection responsibilities, especially in jurisdictions with strict data protection regulation such as the EU. Indeed, in such use-cases, personal data would have to be transferred outside of the hospital's computational infrastructure either for model training or inference. According to General Data Protection Regulation (GDPR), e.g., the institutional requirements of our hospital, the data has to be sufficiently de-identified in such transfers. In the absence of highly reliable automated tools for deidentifying clinical text, however, every piece of the deidentified text might need to undergo manual inspection for missed personal identifiers. For instance, to release the CheXpert Plus dataset (Chambon et al., 2024) of annotated chest X-ray images, up to 30 human annotators had to review more than $850{,}000$ text fragments. Thus, in the absence of tools for ensuring reliable deidentification, the usage of external infrastructure might require costly manual inspection of all of the patient-related data transferred to external providers.

On-premise infastructure costs As we detailed previously, in jurisdictions with strict data protection regulation, outsourcing LLMs to external infrastructure is challenging for use-cases involving patient data. If we want to use LLMs for such use-cases, another option would be developing internal computational infrastructure on the medical institution's premises. At the same time, the state-of-the-art LLMs with multiple billions of parameters are notoriously costly not only to train, but even to infer from. For instance, a 65B parameter model might require four NVidia A100 GPUs each with 80GB for inference only (Samsi et al., 2023). At the time of writing, such a setup would cost at least $60{,}000$ USD, in addition to recurring energy and personnel costs associated with keeping the infrastructure running. This leaves the question: Is the cost-benefit ratio worth it? Any LLM use-case should be rigorously evaluated in terms of the benefit it brings in terms of healthcare outcomes, education, or relieving administrative burden, against the infrastructure and associated costs.

Leakage of private information Even if a model has been trained on-premise, LLMs can leak privacy-sensitive information contained in their training data (Carlini et al., 2023). This means that LLMs themselves as well as their outputs could contain personal information. In use-cases such as the *assistive chatbot for complex medical cases*, this means that if the model was trained on internal patient data, patient information could be extracted by malicious users, or could be leaked even in non-malicious chatbot interactions. There exists a formal theory of ensuring privacy in statistical and machine learning called *differential privacy*, which enables to obtain privacy guarantees when models are trained on private data, preventing the leakage scenarios mentioned before. Recent work on fine-tuning language models with differential privacy (Yu et al., 2021) have shown promising results, even though normally differential privacy significantly reduces the utility of models. It is still, however, an open question, whether it is possible to obtain a meaningful level of privacy while retaining acceptable performance in high-risk downstream tasks such as assistive medical chatbots.

5 Conclusions

Implementation of LLMs in a hospital setting should take into consideration institutional needs, impact on processes, quality of care and patient satisfaction, costs and return on investment, and the risks of harm. In this work we detailed the entire participatory decision process in which we elucidated the needs, wants, and questions in collaboration with various stakeholders at a practicing medical institution. Our survey showed that healthcare stakeholders see assistive AI-based decision-support applications as most promising, especially for the management of complex cases. However, it remains unclear how such applications can be certified, and how to address key performance issues such as hallucinations and bias. Applications aimed at reducing administrative burden or monitoring quality of care, which also seemed promising to the stakeholders, may have higher likelihood of success. These use-cases are less likely to fall under medical device regulations, thus not requiring certification, and at the same time rely on human validation, thus mitigating the risk of incorrect or incomplete output. On the downside, they may require a computationally powerful infrastructure, leading to significant long-term costs.

Overall, administrative use-cases seem most promising for the rapid and successful implementation of generative AI in hospitals, while other use-cases, focused on quality of care and education, still present some specific critical challenges. Despite their limitations, LLMs hold significant potential, and rather than blocking or limiting their development, we should work as a community to tackle these existing problems. To ensure continuous innovation in healthcare, addressing these open, multidisciplinary issues, together, is of utmost importance.

References

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

John G. Laffey Michael G. Madden, Bairbre A. McNicholas. Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. *Intensive Care Medicine*, 49:1018–1020, 2023.

Renato Umeton, Anne Kwok, Rahul Maurya, Domenic Leco, Naomi Lenane, Jennifer Willcox, Gregory A Abel, Mary Tolikas, and Jason M Johnson. Gpt-4 in a cancer center—institute-wide deployment challenges and lessons learned. *NEJM AI*, 1(4):AIcs2300191, 2024.

Heejung Choi Hyeram Seo Minkyoung Kim JiYe Han Gaeun Kee Seohyun Park Soyoung Ko Byeolhee Kim Suyeon Kim Tae Joon Jun Young-Hak Kim HyoJe Jung, Yunha Kim. Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients. *arXiv preprint arXiv:2404.05144*, 2024.

Bryan Lim Nimish Seth Kirk Lower, Ishith Seth. Chatgpt-4: Transforming medical education and addressing clinical exposure challenges in the post-pandemic era. *Indian Journal of Orthopaedics*, 2023.

Topol E.J. Meskó, B. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *npj Digit. Med.*, 2023.

FDA. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. URL https://www.fda.gov/media/145022/download?attachment.

EU. Eu artificial intelligence act. URL https://artificialintelligenceact.eu/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195, 2023.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.

Augustin Toma, Senthujan Senkaiahliyan, Patrick R Lawler, Barry Rubin, and Bo Wang. Generative ai could revolutionize health care—but not if control is ceded to big tech. *Nature*, 624(7990): 36–38, 2023.

GDPR. General data protection regulation. URL https://gdpr-info.eu/.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*, 2024.

Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE, 2023.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.