Position: Participatory Assessment of Large Language Model Applications in an Academic Medical Center

Giorgia Carra[1], **Bogdan Kulynych**[1,2], **François Bastardot**[1], **Daniel Kaufmann**[1], **Noémie Boillat-Blanco**[1], **and Jean Louis Raisaro**[1,2]

[1]Lausanne University Hospital, Switzerland
[2]University of Lausanne, Switzerland

Abstract

Although Large Language Models (LLMs) have shown promising performance in healthcare-related 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 Lausanne University Hospital, an academic medical center in Switzerland. Having identified potential LLM use-cases in collaboration with thirty stakeholders, including clinical staff across 11 departments as well nursing and patient representatives, we assess the current feasibility of these use-cases taking into account the regulatory frameworks, data protection regulation, bias, hallucinations, and deployment constraints. This study provides a framework for a participatory approach to identifying institutional needs with respect to introducing advanced technologies into healthcare practice, and a realistic analysis of the technology readiness level of LLMs for medical applications, highlighting the issues that would need to be overcome LLMs in healthcare to be ethical, and regulatory compliant.

1 Introduction

Large Language Models (LLMs) represent a significant step forward in computer systems' ability to process and work with text-based information. The healthcare field, which routinely handles a large amount of written documentation, sees LLMs as a useful tool to help overcome pressing issues (Thirunavukarasu et al., 2023; Clusmann et al., 2023) such as lack of qualified clinical professionals, alarm fatigue, an ever-growing administrative burden, and increasing pressure due to an aging society. Indeed, LLMs have shown promising performance in several healthcare tasks including medical writing, editing, summarization (Michael G. Madden, 2023; Umeton et al., 2024; Wu et al., 2024).

In this study, we use a systematic participatory approach to evaluate the emerging needs and expectations regarding LLM applications in a European academic medical center—Lausanne University Hospital in Switzerland—with a representative pool of thirty healthcare professionals across different roles and hospital departments, and discuss technical, ethical, and regulatory feasibility of these use-cases. Our analysis shows what technical advances, institutional changes, and regulatory updates would be needed to implement these systems effectively.

2 Methods

We organized a working group (WG) of institutional stakeholders to discuss the needs, opportunities and strategic priorities regarding LLMs. The working group included 30 participants from across the

2nd Workshop on Regulatable ML at NeurIPS 2024.

Table 1: **Average ranking of LLM applications by a working group of healthcare professionals at the Lausanne University Hospital.** We show the ranks according to six criteria regarding the impact of the applications and their perceived safety, along with the total (summed) ranks. Lower numbers are better.

	Impact Criteria						
Application	**Inst**	**QC**	**PE**	**WL**	**ME**	**Saf**	**Total**
Automatic tool for classifying incident reports	5	2	3	6	2	1	19
Summarization of documentation into patient-adapted language	4	5	2	3	3	2	19
Automatic generation of draft discharge letters	1	7	6	1	1	6	22
Assistive chatbot for clinical decision-making	2	1	5	4	4	7	23
Entry notes and patient history summarization	3	6	4	2	5	4	24
Analysis of patients' feedback and complaints	7	4	1	5	6	3	26
Chatbot for medical education, guidelines retrieval, and literature summarization	6	3	7	7	7	5	35

Inst: Institutional, **QC**: Quality of care, **PE**: Patient engagement, **WL**: Workload, **ME**: Measurability, **Saf**: Safety

hospital: clinical staff from 11 departments, nursing and medical directors, patient advocates, legal experts, IT staff, clinical informaticians, and representativies of the Biomedical Data Science Center of the hospital. We held regular monthly discussions among the members over a six-month period.

Preliminary discussions The initial meeting focused on discussing the vision, brainstorming institutional needs, and identifying potential applications 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.

Phase I: Survey Following the identification of a broad list of use-cases and their evaluation criteria, we designed and distributed a structured survey among the WG members. The survey aimed to assess the use-cases which seemed most promising to the WG members, as well as obtain the initial assessments of the application in terms of the identified criteria. Based on the results of the survey, we identified seven most voted applications, which we studied further in the next phase. See Appendix A for details.

Phase II: Participatory workshop We held a participatory workshop in which the members of the WG organizing team presented in detail each of the seven use-cases identified as most popular in the survey, and facilitated a discussion among the members on various aspects of the use-cases. After the discussions, we conducted an anonymous survey in which the members ranked each of the applications according to the criteria established previously.

3 Results

Next, we present the outcomes of the WG activities: the identified framework for assessing LLM applications, and the ranking of LLM applications. Then, we comment on the technical and regulatory feasibility of these applications.

3.1 Working Group Outcomes

Vision and key evaluation criteria Based on the outcomes of the WG, we identified five major areas of improvement at the hospital: *quality of care, patient engagement and satisfaction, administrative burden, research and digital innovation, and continuous education.* Through discussions, brainstorming, and literature review, we then identified specific use-cases for LLMs applications that target these areas. See the complete list in Appendix A.

Moreover, during the discussions, the WG identified the following key evaluation criteria for applications: *institutional impact*, *impact on patient management and quality of care*, *impact on patient satisfaction and engagement*, *whether the impact is measurable*, *whether the intended use is as software as medical device*, and *safety*, defined as whether the risk of harm is low.

Application ranking After two phases of application selection, the WG ranked the final seven use-cases. We present the average ranks, both across the evaluation criteria, and in total, in Table 1. We obtained two use-cases with the highest average ranking: First, the development of automatic tools for classifying and analyzing incident reports, e.g., cases of falls or incorrect drug dosage. Second, the summarization of discharge documentation into patient-adapted language. Both applications are of administrative nature, are deemed as easy to measure in terms of their impact, have benefits to patient engagement, and are assessed as being the least risky among the final application roster.

Interestingly, the intermediate results from phase I showed the use-case of the assistive chatbot for clinical decision-making as both the most popular and best-ranked (see Appendix A). In phase II, however, after a detailed group discussion of all applications, it scored lower in the final anonymous survey, being ranked the worst in terms of safety, yet having the widest institutional impact.

3.2 Technical and Regulatory Feasibility

Regulation of medical uses Under the United States 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 and Topol, 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 clinical decision-making*—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 the *chatbot for medical education*, 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, with some investigations already ongoing (Andreoletti et al., 2024).

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 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 clinical decision-making*, 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 clinical decision-making*, 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* (Dwork et al., 2014), 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., 2022) 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.

4 Discussion

Our findings reveal a practical reality: although healthcare professionals see multiple promising applications for large language models in their practice, regulatory and infrastructural constraints significantly limit immediate implementation options. The preference for administrative applications, particularly discharge note generation and summarization, suggests a pragmatic path forward that balances regulatory compliance with feasibility.

Two key limitations of our study warrant discussion. First, the conclusions of our participatory methodology, which are limited to the specific context of Lausanne University Hospital, may not generalize across healthcare institutions. That said, we do hypothesize that our findings could largely generalize to similar medium-sized academic medical centers in Western Europe. Moreover, our study provides a framework for institutions to examine their unique needs and constraints through structured stakeholder engagement. Second, despite achieving representation across departments and roles, our participant pool likely shows selection bias, as the self-selected nature of participation in the working group means we may have captured perspectives from professionals who are inherently more enthusiastic about language models in healthcare.

Despite these limitations, our findings highlight how successful implementation of generative models in healthcare settings requires careful navigation of practical constraints, particularly around medical device regulation and data protection.

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.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 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.

Haotian Wu, Paul Boulenger, Antonin Faure, Berta Céspedes, Farouk Boukil, Nastasia Morel, Zeming Chen, and Antoine Bosselut. Epfl-make at "discharge me!": An llm system for automatically generating discharge summaries of clinical electronic health record. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 696–711, 2024.

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

mattia Andreoletti, Berkay Senkalfa, and Alessandro Blasimme. Ongoing and planned randomized controlled trials of ai in medicine: An analysis of clinicaltrials. gov registration data. *medRxiv*, pages 2024–07, 2024.

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.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

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. In *International Conference on Learning Representations*, 2022.

Table 2: Use-cases considered by the working group in Phase I (survey).

Applications	Areas QC	RI	PE	Adm	Edu	Ranking Popularity
Assistive chatbot for clinical decision-making	•	•				1
Chatbot for medical education					•	2
Automatic generation of draft discharge letters				•		3
Summarization of discharge letters into patient-adapted language	•		•			4
Analysis of patients' feedback and complaints	•		•			5
Smart retrieval of guidelines	•				•	5
Entry notes and medical records summarization	•			•		5
Smart summarization of current literature					•	5
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

A Phase I: Survey

Methods In the survey, we allowing the participants to vote for the five most relevant use-cases from the pre-defined list and rank them according to the identified evaluation criteria. The survey was developed in Qualtrics Version, [06.2024].

Results Sixteen members of the WG, corresponding to one representative per hospital service, participated in the survey.

We show the ranking of LLM use-cases by popularity in Table 2. *Assistive chatbot for clinical decision-making*, *chatbot for medical education* and *automatic generation of discharge letters* were the most voted use-cases with 10, 8 and 7 votes respectively.

To scope the WG activities, for the next phase of the assessment, we only proceeded with the use-cases which were voted by at least five WG members. These are the top 8 use-cases in Table 2. To further simplify the discussion, we merged the following applications: smart retrieval of guidelines, chatbot for medical education, and smart summarization of current literature, resulting in six use-cases. During the following discussions, WG members proposed another use-case of classification and analysis of incident reports, which was agreed to be added to the final list of applications for the next phase, resulting in seven total use-cases.