Physician in the Loop Design of Interactive Agents

Maxime Griot[1,2][0009−0000−8180−4737], Jean Vanderdonckt[2][0000−0003−3275−3333], Demet Yuksel[1,2], and Coralie Hemptinne[1,2][0000−0003−0717−3804]

[1] Université catholique de Louvain, Place des Doyens 1, 1348 Louvain-la-Neuve, Belgium
firstname.lastname@uclouvain.be
https://uclouvain.be

[2] Cliniques Universitaires Saint-Luc, Avenue Hippocrate 10, 1200 Brussels, Belgium
firstname.lastname@saintluc.uclouvain.be
https://saintluc.be

Abstract. Medicine poses a unique challenge due to the complexity and range of tasks required to treat patients. We propose an approach to reduce the gap between computer scientists and medical doctors to design relevant interactive agents to assist physicians in their daily work. We demonstrate that an interactive one-hour meeting is enough to equip physicians with enough knowledge and understanding of interactive agents to propose relevant and feasible ideas. Physicians also demonstrated a real interest in designing such systems to assist them in their work.

Keywords: Physician in the loop · Transversal design · Interactive agents · large language models

1 Background

The recent advancements in generative Artificial Intelligence (AI) brought by the transformers architecture and large volumes of text on natural language processing allowed for new applications of machine learning for medicine[15]. Ranging from named entity recognition[16] to performing medical diagnosis[6, 12, 13] and replacing nurses in specific tasks[4], these applications show promise to assist with the increasing workload of healthcare workers. However, a major concern regarding these advances is the relative lack of involvement of physicians. Published work, both in academic settings and commercial solutions either do not involve physicians at all or only involve them to evaluate the end result[14]. Similarly, studies performed by medical doctors to evaluate the relevance and quality of these models in their specialized field usually do not involve computer scientists[5, 7, 18]. This disconnect between the two fields is further widened by the scientific publication process, where computer scientists publish in computer science journals and conferences and medical doctors publish in medical journals and conferences. Adoption and reception of generative artificial intelligence in medicine has been slow and met with resistance by physicians in part because the performance claims made by computer scientists do not translate to clinically relevant performance[3].

The lack of interactions and overlap between these two fields constitutes a major obstacle to creating relevant, safe and adapted interactive agents for medicine. In this work, we detail our approach to develop interactive agents by including both computer scientists and medical doctors in the entire process.

2 Process

2.1 Selection

To allow for a constructive design process, we contacted department leads to discuss their interest in the field of AI, their available resources and interest in being involved in a work group. This initial contact was only used to gauge interest, we shared a two page overview of what generative AI is and our goal of co-creating systems with physicians. We started with department leads who already participate in meetings and projects related to the Electronic Health Record at our hospital and both accepted. We limited the number of departments involved to two, this allowed the available resources to stay focused on the specific needs of these departments ; Oncology and Geriatrics. Limiting the number of participants also reduces the risks of disagreements between departments.

Department leads were tasked with giving a brief introduction to the physicians in their respective departments regarding the department's involvement in this on-going project and studies. Both groups were composed of an average of 10 physicians depending on their availability.

2.2 Knowledge

One of the challenges with co-creation of systems is the mismatch of knowledge between involved parties. We identified key gaps (capabilities and limitations of large language models, and agents) to fill to allow physicians to think about potential uses and applications of the available technology. We organized interactive presentations with each departments to explain the current state of AI in medicine and then involve them in the interactive design of a small agent.

Current State The introduction of the presentation showed existing uses of AI in medicine, such as nodule detection in radiology[10]. We then acknowledged the challenges faced by physicians with the increase in clerical work[17] after which we briefly mentioned ChatGPT as the current state-of-the-art large language model.

Intuitive Agent Design Instead of explaining the concept of agents based on large language models, we used an interactive conversation to guide the physicians into intuitively creating a small agent themselves through prompting.

The demonstration started with a standard chat user interface with Mixtral 8x7B[9] as the model and a vector database seeded with UpToDate paragraphs[11, 2] on which we ask a preset question "A 54 year old male patient presents to the emergency department with abdominal pain, I suspect gallstones, describe the best next step". The model then correctly proposed to perform an ultra-sound.

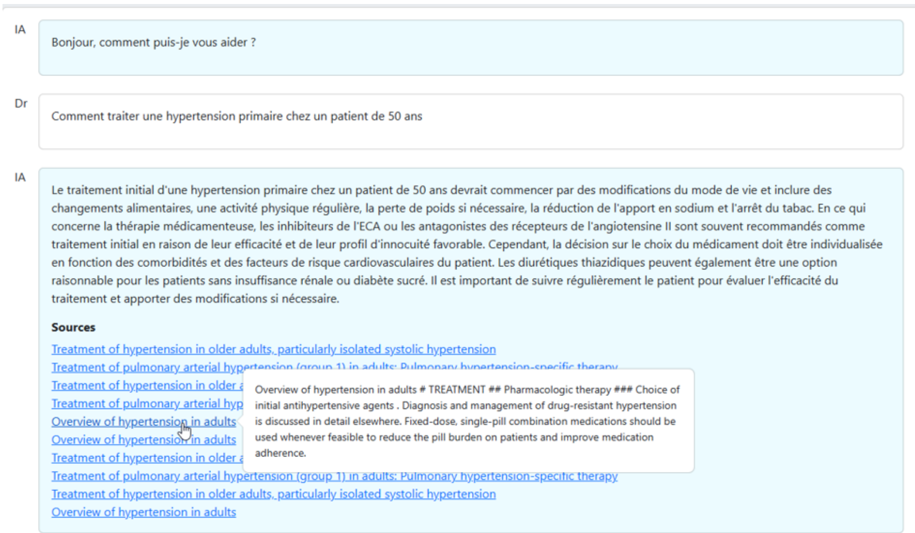

Fig. 1. Chat user interface used to prompt the system with Retrieval Augmented Generation[8] using UpToDate[11].

We used the model's answer to open a discussion and ask if there is anything else we should be doing, to which physicians made multiple suggestions such as:

- Asking specific questions that will help with the diagnosis (i.e. "Did you eat a meal with a lot of fat before the pain appeared?", "Do you have a history of gallstones?")

- Checking past clinical documentation and imagery results to determine if there is a history of similar complaints.

We selected one of the suggestions and proposed that the model could do this for us and modified the prompt to add a tool that the model can use.

Prompt To help you answer the physician's question, you have access to the "Ask" tool, you can use this to ask questions to the patient, use it by writing Ask("question here"), for example Ask("How are you doing?").

Given this prompt the model suggested some questions, demonstrating the use of a tool. In both groups, participants started proposing new tools and asked questions regarding the limitations of tools such as the number of tools and their complexity. Adding some of the proposed tools demonstrated multi-tool usage and allowed the introduction of Retrieval Augmented Generation through the use of a reference database to look up the appropriate guidelines.

Limitations After the co-creation of a small agent we address safety by prompting the model with a simple prompt "Describe the procedure to perform a blood gas analysis". The model used did not mention Allen's test[1], a critical step to ensure the safety of the procedure and suggested that a tourniquet be used to make the artery more visible. This demonstrated that models may not know critical information and can generate incorrect information even for basic medical knowledge.

We also demonstrate cases of hallucinations by prompting the model multiple times with a high temperature setting to increase the likelihood of this happening during the demonstration.

2.3 Design

After the presentation we asked the physicians to take a few days to write down a list of useful applications they believe an agent system could solve. We proposed a common template to simplify collection and asked them to rate the priority of each proposition using a 3 point system. We also asked them to share with us a list of references they use in their practice to use for the Retrieval Augmented Generation.

Table 1. Instructions used to normalize physician design propositions.

Type	Description	Example
Template	As a X, I would like Y, in order to Z.	As a physician, I want the system to look up drug interactions, in order to increase the safety of my prescriptions.
Priority	+++ need ++ want + would like	+++

Given their propositions, we evaluated the feasibility and estimated the time required to achieve a proof-of-concept. We then used the combination of the priority assigned by physicians and our capacity to develop them to chose one proposition per department.

We proposed a retrospective study design for each of the selected propositions that were reviewed and adapted by each department before submission to the ethics committee of our hospital.

3 Results

Our approach of involving physicians in the design was well received by all participants and generated much enthusiasm about the potential contributions of AI in their daily practice. Both departments made many suggestions some of a which are included in Table 2.

Physicians also proposed multiple sources of knowledge to provide models with accurate and updated information some of which are included in Table 3.

Aside from the initial goal of the meetings to introduce the technology and equip physicians with enough understanding to propose applications, both meetings ended with open discussions between all participants on the impact AI will have on clinical practice. We noticed that relatively basic functions such as suggesting relevant questions to ask patients caused physicians to be both impressed and worried about the future of medicine. This led to discussions about the future of medical practice and the role of physicians in an AI driven hospital.

Table 2. Example uses suggested by physicians along with their priorities.

Department	Use	Priority
As an oncologist, I would like AI to	help me write my consultation notes, to reduce my clerical work	+++
	produce instructions for patients at the end of the visit	+++
	inform me if a patient can be included in a study	+++
	compare the patient with similar patients to help me decide between treatment and paliative care	++
	help me estimate life expectancy without cancer	++
	tell me if the patient needs nutritional support based on the treatment and weight changes	++
As a geriatrician, I would like AI to	extract the problem list, medical and surgical history and recent treatment from the healthcare network	+++
	help me optimize treatment based on the current start/stopp guidelines	+++
	notify me of documented events since the last visit	++
	anticipate post-operative complications	+

Table 3. Selected sources of knowledge proposed by physicians.

Department	Source	Kind
Geriatrics	UpToDate	Guidelines
	Gériatrie pour le praticien, Elsevier Masson	Textbook
	European Geriatrics Medicine	Journal
	Journal of the American Geriatric Society	Journal
	Geriatric Depression Scale (GDS-15)	Score
	Stopp and Start v3	Pharmacology Guidelines
	Neuro-Geriatrics. A Clinical Manual. Springer	Textbook
Oncology	European Society For Medical Oncology	Guidelines
	American Society of Clinical Oncology	Guidelines
	National Comprehensive Cancer Network	Guidelines

4 Conclusion

Through this approach of co-creation and design of interactive agents to assist physicians in their daily work, we demonstrate that involving physicians from the beginning does not require extensive effort. We show that a single one hour meeting is enough to equip physicians with enough understanding to make relevant and feasible design suggestions. These meetings also sparked many discussions on the limitations, safety concerns and how patients will benefit from the use of these new tools. The development of a tool requested by physicians contributes to a long term support and their assistance with time consuming tasks such as writing study protocols and manual evaluation of results.

Our work demonstrates that involvement of domain specialists from the beginning is beneficial to ensure relevance and better collaboration on the creation of interactive agents and that bridging the knowledge gap to collaborate does not require many hours of education.

This work also demonstrates that AI generates existential questions for physicians, we believe it is important to allow for such discussions and to explain that current models can only perform simple but time consuming tasks and are limited by their capabilities, hallucinate, need supervision and to be wary of commercial claims.

Acknowledgments. This study was funded by the Fondation Saint-Luc (grant number 467E).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allen, E.V.: Thrombo-Angiitis Obliterans. Bulletin of the New York Academy of Medicine **18**(3), 165–189 (Mar 1942), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933752/

2. ChromaDB: ChromaDB the AI-native open-source embedding database (2023), https://www.trychroma.com
3. Heston, T.F., Lewis, L.M.: ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. PLOS ONE **19**(4), e0301854 (Apr 2024). https://doi.org/10.1371/journal.pone.0301854, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0301854, publisher: Public Library of Science
4. Hippocratic: Hippocratic AI (2024), https://www.hippocraticai.com/
5. Humar, P., Asaad, M., Bengur, F.B., Nguyen, V.: ChatGPT Is Equivalent to First-Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-Service Examination. Aesthetic Surgery Journal **43**(12), NP1085–NP1089 (Nov 2023). https://doi.org/10.1093/asj/sjad130
6. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams (Sep 2020), http://arxiv.org/abs/2009.13081, arXiv:2009.13081 [cs] version: 1
7. JQ, J., AS, D.: ChatGPT for healthcare providers and patients: Practical implications within dermatology. Journal of the American Academy of Dermatology **89**(4), 870–871 (Oct 2023). https://doi.org/10.1016/j.jaad.2023.05.081, https://pubmed.ncbi.nlm.nih.gov/37315798/, place: United States Publisher: School of Medicine, University of California San Francisco, San Francisco, California; Department of Dermatology, University of California San Francisco, San Francisco, California. Electronic address: joy.jin@ucsf.edu.; Department of Dermatology, University of California San Francisco, San Francisco, California.
8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
9. Mistral: mistralai/Mixtral-8x7B-v0.1 · Hugging Face (2024), https://huggingface.co/mistralai/Mixtral-8x7B-v0.1
10. Nam, J.G., Hwang, E.J., Kim, J., Park, N., Lee, E.H., Kim, H.J., Nam, M., Lee, J.H., Park, C.M., Goo, J.M.: AI Improves Nodule Detection on Chest Radiographs in a Health Screening Population: A Randomized Controlled Trial. Radiology **307**(2), e221894 (Apr 2023). https://doi.org/10.1148/radiol.221894, https://pubs.rsna.org/doi/full/10.1148/radiol.221894, publisher: Radiological Society of North America
11. N.V., W.K.: UpToDate: Industry-leading clinical decision support (2023), https://www.wolterskluwer.com/en/solutions/uptodate
12. OpenAI: GPT-4 Technical Report (Mar 2023). https://doi.org/10.48550/arXiv.2303.08774, http://arxiv.org/abs/2303.08774, arXiv:2303.08774 [cs]
13. Rydzewski, N.R., Dinakaran, D., Zhao, S.G., Ruppin, E., Turkbey, B., Citrin, D.E., Patel, K.R.: Comparative Evaluation of LLMs in Clinical Oncology. NEJM AI **1**(5), AIoa2300151 (Apr 2024). https://doi.org/10.1056/AIoa2300151, https://ai.nejm.org/doi/full/10.1056/AIoa2300151, publisher: Massachusetts Medical Society
14. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., Chaves, J.Z., Hu, S.Y., Schaekermann, M.,

Kamath, A., Cheng, Y., Barrett, D.G.T., Cheung, C., Mustafa, B., Palepu, A., McDuff, D., Hou, L., Golany, T., Liu, L., Alayrac, J.b., Houlsby, N., Tomasev, N., Freyberg, J., Lau, C., Kemp, J., Lai, J., Azizi, S., Kanada, K., Man, S., Kulkarni, K., Sun, R., Shakeri, S., He, L., Caine, B., Webson, A., Latysheva, N., Johnson, M., Mansfield, P., Lu, J., Rivlin, E., Anderson, J., Green, B., Wong, R., Krause, J., Shlens, J., Dominowska, E., Eslami, S.M.A., Chou, K., Cui, C., Vinyals, O., Kavukcuoglu, K., Manyika, J., Dean, J., Hassabis, D., Matias, Y., Webster, D., Barral, J., Corrado, G., Semturs, C., Mahdavi, S.S., Gottweis, J., Karthikesalingam, A., Natarajan, V.: Capabilities of Gemini Models in Medicine (May 2024). https://doi.org/10.48550/arXiv.2404.18416, http://arxiv.org/abs/2404.18416, arXiv:2404.18416 [cs]

15. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (Aug 2023). https://doi.org/10.1038/s41586-023-06291-2, https://www.nature.com/articles/s41586-023-06291-2, number: 7972 Publisher: Nature Publishing Group
16. Sousa, H., Mario Jorge, A., Pasquali, A., Santos, C., Lopes, M.: A Biomedical Entity Extraction Pipeline for Oncology Health Records in Portuguese - Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing pp. 950–956 (2023). https://doi.org/10.1145/3555776.3578577, https://doi.org/10.1145/3555776.3578577, place: New York, NY, USA Publisher: Association for Computing Machinery Section: 0
17. Starren, J.B., Tierney, W.M., Williams, M.S., Tang, P., Weir, C., Koppel, R., Payne, P., Hripcsak, G., Detmer, D.E.: A retrospective look at the predictions and recommendations from the 2009 AMIA policy meeting: did we see EHR-related clinician burnout coming? Journal of the American Medical Informatics Association **28**(5), 948–954 (May 2021). https://doi.org/10.1093/jamia/ocaa320, https://doi.org/10.1093/jamia/ocaa320
18. Suchman, K., Garg, S., Trindade, A.J.: Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. The American Journal of Gastroenterology **118**(12), 2280–2282 (Dec 2023). https://doi.org/10.14309/ajg.0000000000002320