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][0009−0009−0954−9781], 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

Recent advancements in generative Artificial Intelligence (AI), driven by the transformer architecture and the availability of large-scale text corpora, have opened new avenues for machine learning applications in medicine [16]. These applications span a wide range, from named entity recognition [17] to medical diagnosis [6, 13, 14] and even the potential replacement of nurses in specific tasks [4]. Such developments offer promising solutions to address the increasing workload of healthcare professionals. However, a significant concern surrounding these advancements is the relative lack of physician involvement in their development and implementation.

Published work, both in academic settings and commercial solutions, often either excludes physicians entirely or involves them only in evaluating the end result [15]. Conversely, studies conducted by medical doctors to assess the relevance and quality of these models in specialized fields frequently lack input from computer scientists [5, 7, 19]. This disconnect between the two fields is further exacerbated by the siloed nature of scientific publication, where computer scientists and medical doctors publish in their respective disciplinary journals and conferences, with limited cross-pollination of ideas.

The adoption and reception of generative AI in medicine have been slow and met with resistance from physicians, partly because the performance claims made by computer scientists often fail to translate into clinically relevant outcomes [3]. This misalignment between technological capabilities and clinical needs underscores the importance of interdisciplinary collaboration in developing AI solutions for healthcare.

The lack of interaction 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 developing interactive agents by including both computer scientists and medical doctors throughout the entire process, from conception to implementation and evaluation.

2 Engineering

2.1 Selection

Involving the correct actors in a co-engineering task is the first major step. In this work, we aimed to involve both software engineers and physicians, with or without prior exposure to systems design. To ensure a smooth design process, involving curious and available physicians was key.

To recruit physicians, we contacted department leads to discuss their interest in AI, available resources, and willingness to participate in a work group. This initial contact was used to gauge interest; we shared a two-page overview of generative AI and our goal of co-creating systems with physicians.

We prioritized department leads who already contribute to clinical informatics through their involvement in Electronic Health Record projects at our hospital. The first two department leads contacted agreed to participate in the work group. We limited the number of departments involved to two (Oncology and Geriatrics), allowing resources to remain focused on these departments' specific needs. Limiting the number of participants also reduces the risk of interdepartmental disagreements.

Department leads were tasked with briefly introducing the ongoing project and studies to physicians in their respective departments. Both groups comprised an average of 10 physicians, depending on availability.

2.2 Knowledge

One challenge in system co-creation is the knowledge mismatch between involved parties. Software engineers understand systems design and technological limitations but lack knowledge of patient treatment processes. Physicians can identify key aspects of their daily routine that could be improved but may struggle to formulate solutions. Additionally, both groups alone are unable to make proper risk assessments due to their respective knowledge limitations.

A conventional approach would involve implementing metrics to track time-consuming tasks, shadowing physicians, and conducting interviews to find use cases. Our approach attempts to reduce knowledge gaps between the two groups, allowing for open discussion and enabling physicians to identify relevant and feasible use cases.

To facilitate such discussions, we addressed three key categories that we believed would provide sufficient information for physicians to brainstorm ideas:

1. Understanding basic capabilities of large language models through zero-shot prompts
2. Limitations of these capabilities, including hallucinations
3. Basic agents, extending large language models with external APIs that models can use to enrich user requests

We organized interactive meetings with each department to explain the current state of AI in medicine and involve them in the interactive design of a small agent. The entire co-design process is presented in Figure 1.

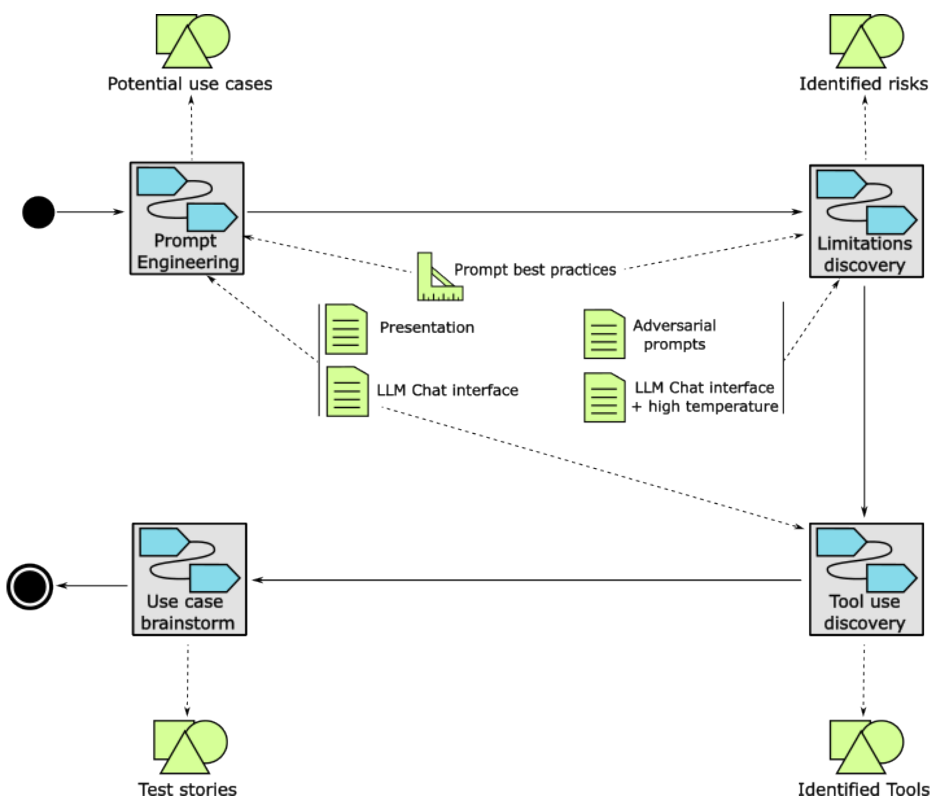

Fig. 1. Overview of the co-engineering process with outcomes for each step using the Software & Systems Process Engineering Meta-Model (SPEM) [12].

The demonstration relied on a custom chat user interface using Mixtral 8x7B[9] as the model and the ChromaDB vector database seeded with UpToDate paragraphs[11, 2].

Prompt Engineering Meetings began with a short introduction to existing AI applications in medicine, primarily addressing classification models in imagery, such as nodule detection in radiology [10]. We acknowledged challenges faced by physicians due to increased clerical work [18] and briefly mentioned ChatGPT as the current state-of-the-art large language model.

The interactive session aimed to teach fundamentals of large language models and agents through an intuitive process. This task helped physicians identify both use cases and potential solutions using existing tools that large language models could utilize.

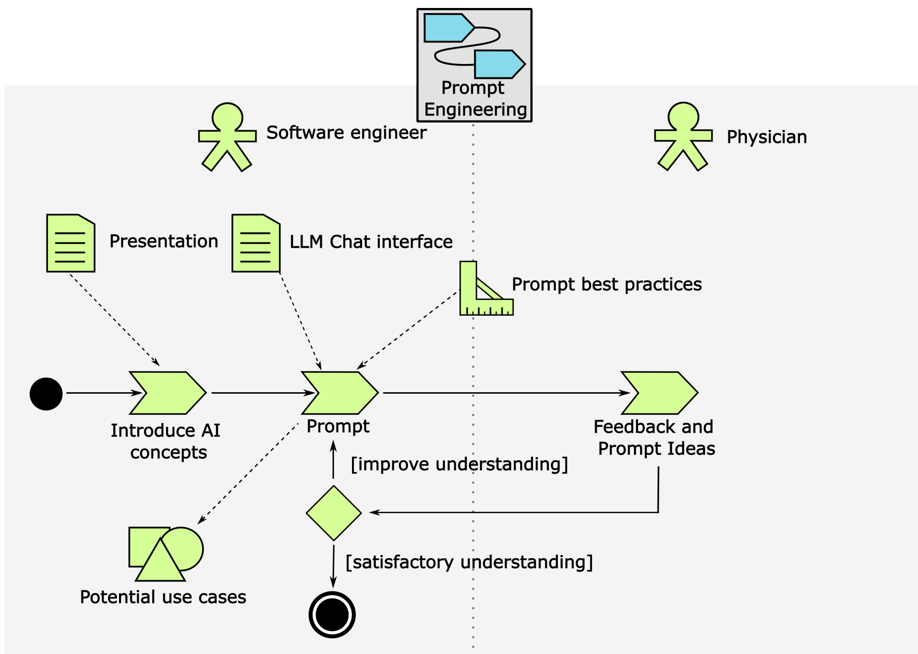

Fig. 2. Definition of the prompt engineering process using SPEM.

Through the process detailed in Figure 2, we prompted an LLM with a simple medical question and obtained feedback and prompt ideas from physicians. This iterative process familiarized physicians with the chat interface and improved their understanding of prompting best practices. Basic use cases were formulated by physicians during this process. After a few iterations, when physicians wanted to move on or lacked new ideas, we proceeded to the next step.

Limitations Discovery A major limitation of LLMs is their ability to be confidently incorrect, commonly called hallucinations, which can be particularly dangerous in critical domains such as medicine where mistakes can lead to adverse patient outcomes.

We demonstrated these limitations through two main approaches:

Omission of Critical Information and Incorrect Suggestions After the co-creation of a small agent, we addressed safety concerns by prompting the model with a simple task: "Describe the procedure to perform a blood gas analysis." The model's response revealed significant limitations:

- The model failed to mention Allen's test [1], a critical step to ensure the safety of the procedure.
- It incorrectly suggested using a tourniquet to make the artery more visible, which is not a standard practice for arterial blood gas sampling and could potentially be harmful.

This demonstration highlighted that models may lack critical information and can generate incorrect instructions even for basic medical procedures, underscoring the potential risks of relying solely on AI-generated medical advice.

Inducing Hallucinations We also demonstrated cases of hallucinations by prompting the model multiple times with a high temperature setting. This approach increased the likelihood of generating inconsistent or factually incorrect responses during the demonstration, allowing physicians to observe firsthand the unreliability that can occur when models operate beyond their knowledge boundaries.

These demonstrations were crucial in illustrating to physicians that while LLMs can be powerful tools, they are not infallible and require careful oversight, especially in medical contexts. The complete process of limitations discovery is illustrated in Figure 3.

Similar to the prompt engineering process, we allowed for iterative changes and discussions. We spent time explaining hallucinations and addressed secondary questions such as detection and prevention. After explaining, we allowed additional feedback and let physicians test more prompts until they reached a satisfactory understanding of the implications of hallucinations. This process allowed physicians to raise concerns regarding the risks in medicine caused by these limitations.

The limitations discovery phase was instrumental in fostering a critical perspective among the participating physicians. It highlighted the necessity for human expertise in verifying and contextualizing AI-generated information, especially in high-stakes medical scenarios. This understanding formed a crucial foundation for the subsequent stages of our co-creation process, ensuring that the developed tools would be used responsibly and with appropriate safeguards.

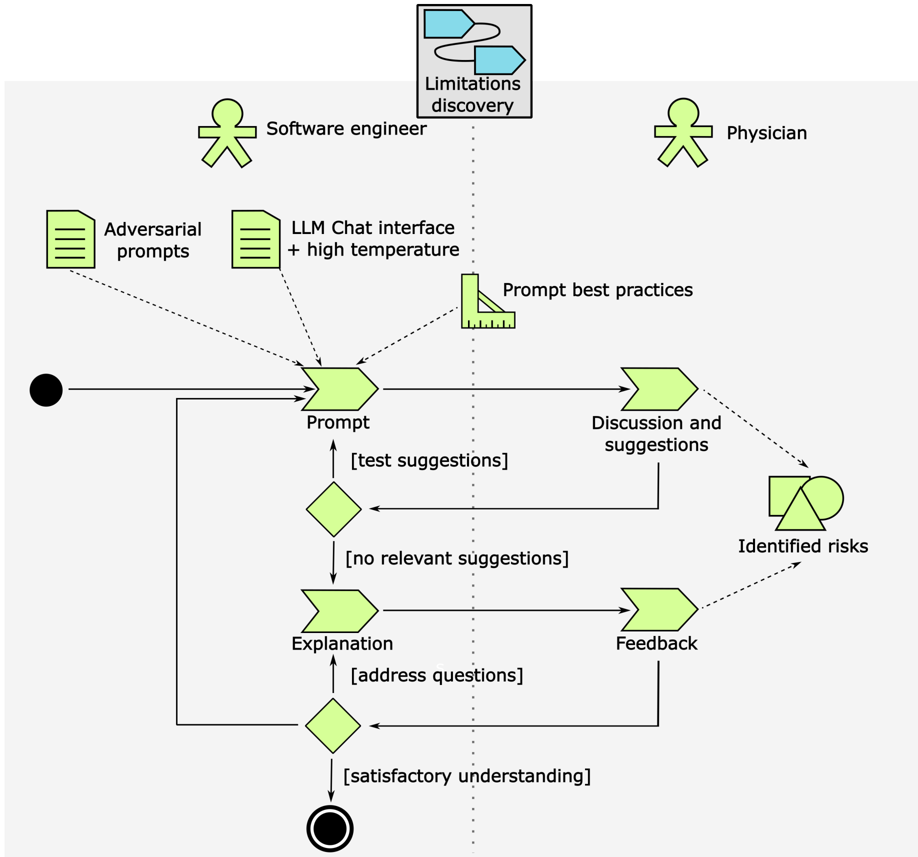

Fig. 3. Definition of the limitations discovery process using SPEM.

Tool Use Discovery With basic chat usage and limitations covered, we introduced the concept of tools—functions that can be used by the LLM to enrich user requests. We implemented an automatic Retrieval Augmented Generation (RAG) system [8] in our chat interface that retrieves documents related to the prompt from UpToDate [11]. The process of tool discovery is illustrated in Figure 4.

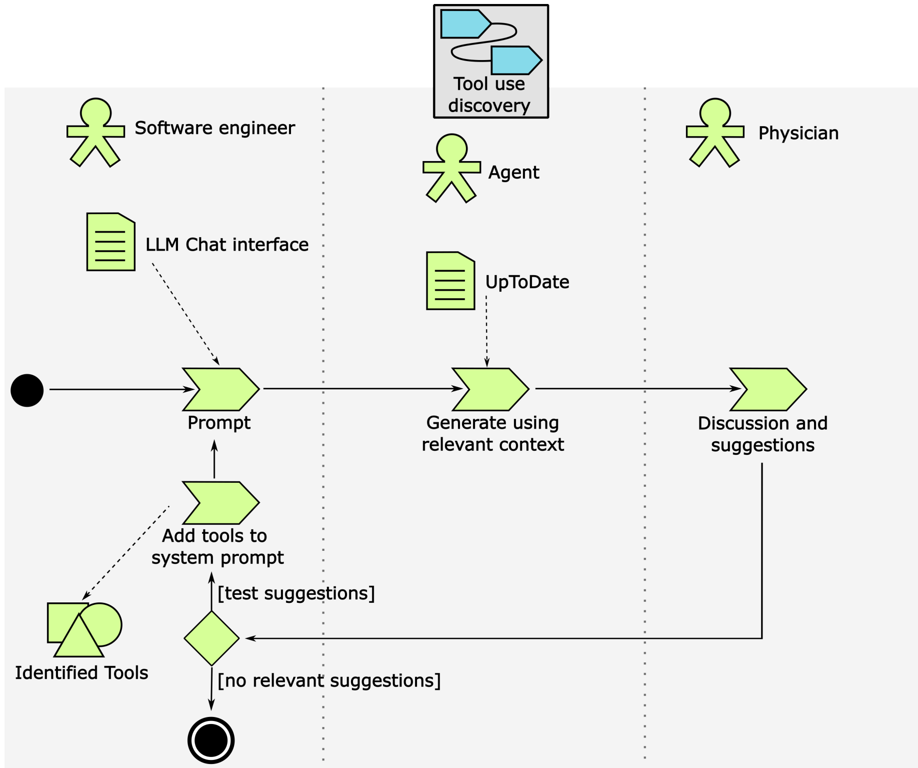

Fig. 4. Definition of the tools discovery process using SPEM.

To demonstrate the RAG tool use, we proposed a basic medical vignette:

Vignette

A 54-year-old male patient presents to the emergency department with abdominal pain. I suspect gallstones. Describe the best next step.

This vignette allowed the model to retrieve information related to the approach to abdominal pain and gallstones diagnosis. The RAG system enriched the LLM's response with relevant medical information from trusted sources.

We used the model's answer to initiate a discussion with the physicians, asking if there was anything else that should be considered. This led to multiple suggestions from the physicians, including:

- Asking specific diagnostic questions (e.g., "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 imaging results to determine if there is a history of similar complaints

Building on these suggestions, we proposed that the model could perform these tasks using additional tools. We modified the prompt to introduce a new tool:

> **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 by writing Ask("question here"). For example: Ask("How are you doing?").

Given this prompt, the model demonstrated the use of the tool by suggesting relevant questions. This demonstration sparked engagement from the participants, who began proposing new tools and inquiring about tool limitations, such as the number of tools that could be used and their complexity.

We further expanded the demonstration by implementing some of the proposed tools, showcasing multi-tool usage. This allowed us to introduce more advanced concepts, including the use of Retrieval Augmented Generation through a reference database tool for looking up appropriate guidelines.

Our user interface was designed to enhance the accessibility and transparency of the retrieved information. As shown in Figure 5, we added a "Sources" section under the LLM response. This section comprised links to the relevant sources of information. Additionally, we implemented a feature where hovering the mouse over a link would display the content of the extract used by the RAG system. This design choice allows users to read the relevant extract without opening a new tab and reading the entire article, thus streamlining the verification process and enhancing the user experience.

This interactive demonstration allowed physicians to understand how the LLM could leverage external, up-to-date medical knowledge and additional tools to provide more accurate and context-aware responses. It also highlighted the importance of transparency in AI-assisted medical decision-making, as users could easily verify the sources and specific information used by the system.

The tools discovery process, as depicted in Figure 4, and the practical demonstration using the sample interface shown in Figure 5, provided physicians with a comprehensive understanding of how AI tools can be integrated into their decision-making process while maintaining transparency and reliability. Furthermore, the collaborative nature of this process fostered engagement and creativity among the participants, leading to the exploration of more advanced and tailored AI tools for medical applications.

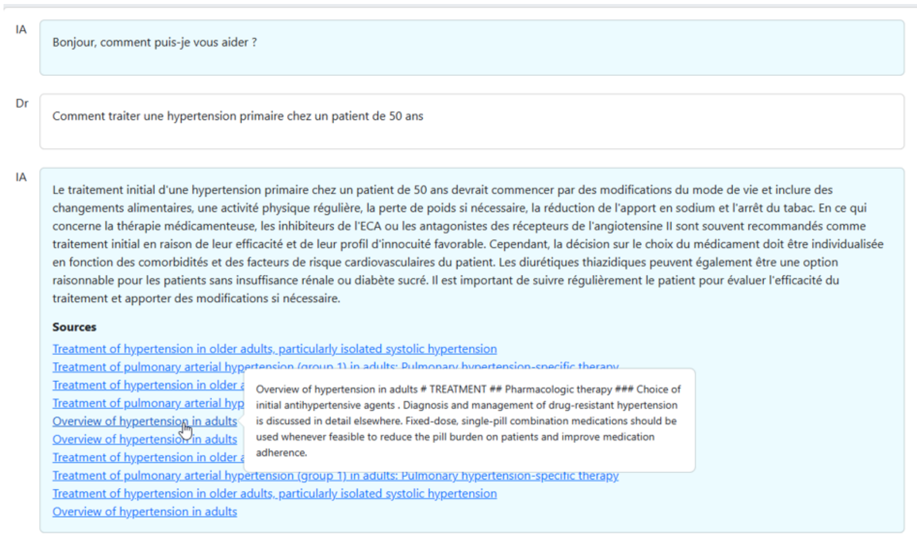

Fig. 5. Sample user interface demonstrating the Retrieval Augmented Generation system, including the LLM response, Sources section, and hover functionality for viewing source extracts.

Brainstorm and Proposition Development Following the interactive meetings, we initiated a structured brainstorming process to capture and prioritize potential applications of agent systems in clinical practice as illustrated in Figure 6. This process was designed to bridge the gap between technological possibilities and genuine clinical needs, ensuring that our development efforts would yield practically relevant AI tools. Our approach consisted of several key phases:

Idea Generation and Prioritization We engaged physicians in a creative yet structured ideation process. Participants were given several days to reflect on their clinical experiences and compile a list of potential AI applications that could address challenges in their practice. To ensure consistency and facilitate analysis, we provided a standardized template for idea submission and a prioritization system, as illustrated in Table 1.

Table 1. Instructions for normalizing 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	+++

This user story format ("As a X, I would like Y, in order to Z") encouraged physicians to articulate not just the desired functionality but also the intended outcome and the specific user role. The priority system (+++ need, ++ want, + would like) allowed participants to indicate the perceived importance of each idea, providing valuable context for the subsequent selection process.

Resource Identification for Knowledge Enhancement Concurrent with the idea generation, we asked physicians to compile a list of references they frequently consult in their practice. This step was crucial for two reasons:

1. It provided insights into the knowledge sources that inform clinical decision-making in different specialties.
2. It supplied valuable input for enhancing our Retrieval Augmented Generation (RAG) system, ensuring that the AI agent could access and utilize relevant, up-to-date medical knowledge.

Proposition Refinement and Technical Evaluation Upon receiving the physicians' ideas, our interdisciplinary team undertook a thorough refinement and evaluation process:

1. **User Story Transformation:** We converted the raw propositions into formal user stories, aligning them with established software development practices. This step helped clarify requirements and facilitated communication between clinical and technical team members.
2. **Feasibility Assessment:** Each proposal underwent a rigorous technical feasibility evaluation. We considered factors such as data availability, integration challenges, and potential ethical implications.
3. **Development Estimation:** We estimated the time and resources required to develop a proof-of-concept for each idea. This step was crucial for balancing ambition with practical constraints.

Selection and Study Design The final phase involved a balanced approach to project selection and study design:

1. **Proposition Selection:** We selected one proposition per department, carefully balancing the priorities assigned by physicians with our development capacity and the potential for innovation.
2. **Retrospective Study Design:** For each selected proposition, we designed a retrospective study to evaluate the potential impact and effectiveness of the proposed agent system. This approach allowed us to assess the value of the AI tool using existing clinical data before proceeding to prospective trials.
3. **Collaborative Refinement:** We engaged in iterative discussions with the respective departments to review and refine the study designs, ensuring they addressed relevant clinical questions and adhered to rigorous methodological standards.

4. **Ethical Approval:** The finalized study proposals were submitted to the hospital's ethics committee for review and approval, ensuring compliance with ethical standards and patient data protection regulations.

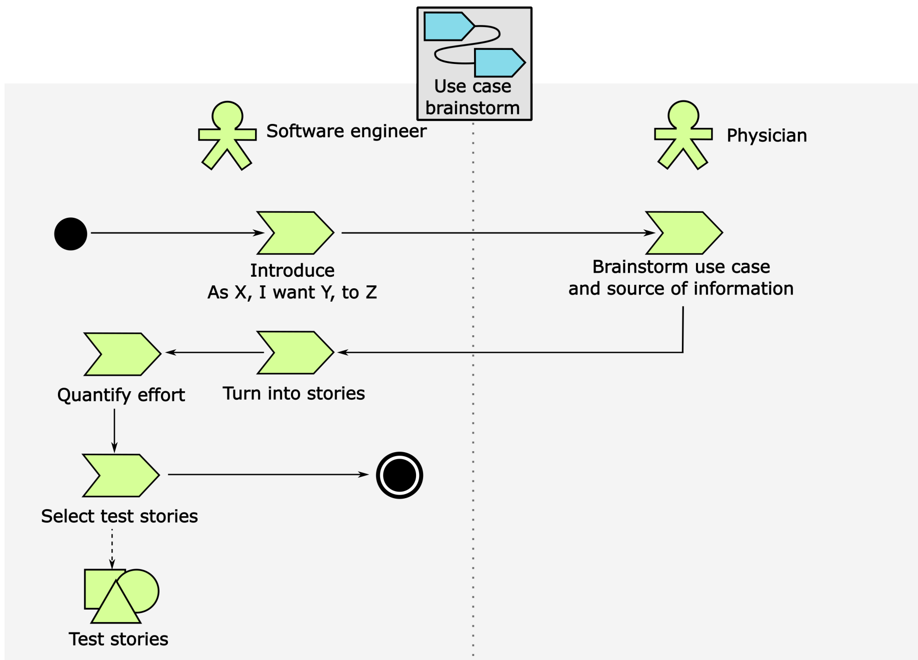

Fig. 6. Definition of the brainstorming and proposition development process using SPEM.

This structured approach to brainstorming and development ensured that the selected projects were not only technically feasible but also aligned with genuine clinical needs and priorities. By combining physician insights with rigorous technical and ethical evaluations, we laid a solid foundation for the subsequent stages of our co-creation methodology. This process promoted the development of AI tools that are both innovative and practically relevant to medical practice, while also setting the stage for meaningful evaluation of their potential impact.

3 Results

Our approach of involving physicians in the design process yielded significant insights and generated considerable enthusiasm about the potential contributions of AI in clinical practice. The collaborative sessions not only achieved their primary goals but also sparked broader discussions about the future of medicine.

3.1 Physician Engagement and Propositions

The participatory design approach was well-received by all participants across both departments. Physicians demonstrated high levels of engagement, offering numerous suggestions for potential AI applications in their daily practice. Table 2 presents a selection of these suggestions, along with their assigned priorities.

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 palliative 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	+

The suggestions revealed a strong interest in AI applications that could streamline administrative tasks, enhance decision-making processes, and improve patient care. Notably, both oncologists and geriatricians prioritized AI assistance in documentation and guideline adherence, indicating common pain points across specialties.

3.2 Knowledge Sources for AI Enhancement

In addition to use case suggestions, physicians proposed multiple sources of knowledge to provide models with accurate and up-to-date information. Table 3 presents a selection of these proposed sources, categorized by department and type of resource.

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
	START and STOPP 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

The proposed knowledge sources encompassed a wide range of materials, including clinical guidelines, textbooks, journals, and assessment tools. This diverse set of resources reflects the complex nature of medical decision-making and highlights the importance of incorporating varied, authoritative sources into AI systems for healthcare.

3.3 Emergent Discussions and Reflections

While the primary goal of the meetings was to introduce the technology and equip physicians with sufficient understanding to propose applications, the sessions evolved beyond this initial scope. Both meetings concluded with open discussions among all participants about the broader impact of AI on clinical practice.

Key observations from these discussions include:

- **Reaction to Basic AI Functions:** Physicians expressed both impressment and concern when presented with relatively basic AI functions, such as suggesting relevant questions to ask patients. This dual reaction underscores the potential of AI to significantly alter clinical workflows.
- **Future of Medical Practice:** The demonstrations sparked conversations about the future landscape of medicine and the evolving role of physicians in an AI-driven healthcare environment.
- **Ethical Considerations:** Participants raised questions about the ethical implications of AI in healthcare, including issues of patient privacy, decision-making autonomy, and the potential for AI to exacerbate or mitigate healthcare disparities.
- **Integration Challenges:** Discussions touched upon the practical challenges of integrating AI tools into existing clinical workflows and electronic health record systems.
- **Legal Implications:** Concerns were raised about physicians' liability when following AI-generated recommendations, highlighting the need for clear guidelines on integrating AI into clinical decision-making.

These emergent discussions revealed a nuanced perspective among physicians regarding AI in healthcare. While they recognized its potential to enhance efficiency and decision-making, they also expressed a keen awareness of the need for careful implementation and ongoing evaluation of AI tools in clinical settings.

3.4 Implications for Future Development

The results of these collaborative sessions have several important implications for the future development of AI in healthcare:

1. **Tailored Solutions:** The diversity of suggested applications underscores the need for AI solutions tailored to specific clinical contexts and specialties.
2. **Knowledge Integration:** The range of proposed knowledge sources highlights the importance of developing robust methods for integrating diverse, authoritative medical information into AI systems.
3. **Ongoing Engagement:** The positive reception and rich discussions suggest that ongoing engagement with healthcare professionals should be a cornerstone of AI development in medicine.

These findings provide a strong foundation for the next phases of our research, informing both the technical development of AI tools and the broader strategies for their integration into clinical practice.

4 Conclusion

This study demonstrates the efficacy and importance of involving physicians in the co-creation and design of interactive AI agents for medical applications. Our findings indicate that a single one-hour meeting is sufficient to equip physicians with the necessary understanding to make relevant and feasible design suggestions. This approach not only generates valuable input for AI development but also sparks crucial discussions on limitations, safety concerns, and potential patient benefits.

Key outcomes of our methodology include:

- **Efficient Knowledge Transfer:** Physicians quickly gained enough understanding of AI capabilities to contribute meaningfully to the design process.
- **Relevant Design Suggestions:** The involvement of domain specialists from the outset ensured the relevance and practicality of proposed AI tools.
- **Interdisciplinary Dialogue:** The process fostered important discussions on ethical considerations, implementation challenges, and the future of medical practice.
- **Long-term Engagement:** By developing tools requested by physicians, we laid the groundwork for sustained support and collaboration in addressing time-consuming tasks.

Furthermore, our work reveals that AI integration in healthcare prompts existential questions for physicians. It is crucial to facilitate these discussions and to contextualize the current capabilities of AI models. We emphasize that present AI systems are best suited for simple yet time-consuming tasks, require supervision, and have limitations including the potential for hallucinations.

In conclusion, this study underscores the value of bridging the knowledge gap between AI developers and medical professionals. By fostering collaboration from the inception of AI tools, we can ensure their relevance, safety, and effective integration into clinical practice. This approach not only enhances the quality of AI applications in healthcare but also prepares the medical community for the evolving landscape of AI-assisted medicine.

By continuing to prioritize collaborative approaches in AI development for healthcare, we can work towards creating tools that truly enhance medical practice while addressing the complex ethical and practical challenges that arise in this rapidly evolving field.

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. OMG: Software & Systems Process Engineering Meta-Model Specification (2008), http://www.omg.org/spec/SPEM/2.0/PDF
13. 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]
14. 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
15. 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]
16. 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
17. 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
18. 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
19. 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