Unseen but Influential: The role of friction in General Practitioners Experience with an LLM-Based Diagnostic Feature in EHR Systems

Elise Hallaert[1,*], Jean Albrecq[1], Juliette Barnich[1] and Bruno Dumas[1]

[1]Namur Digital Institute (NaDI), University of Namur, Namur

Abstract

Large language models are increasingly integrated into medical software, with the promise of improved efficiency, reduced diagnostic time, and lower costs. At the same time, these systems raise concerns about their influence on clinicians' reasoning, trust, and responsibility. To examine the role of interface design of this kind of application, this study adopts a human-centered approach to examine how general practitioners experience an LLM-based diagnostic feature embedded in an electronic health record interface. We designed a prototype inspired by existing clinical software and implemented two conditions: a control interface and a frictional variant including a pop-up prompt encouraging critical appraisal of the diagnostic suggestions. This prompt was intended to introduce friction by momentarily increasing cognitive engagement. We pre-generated diagnostic hypotheses along with short explanations of their underlying rationale with GPT-4 (without task-specific fine-tuning), which were delivered through a Wizard-of-Oz setup. Ten general practitioners interacted with both interface conditions in a randomized order across two randomly assigned clinical cases. We collected behavioral data, screen and face recordings, structured observations, and semi-structured interviews. We administered the UEQ on the control interface to assess baseline user experience and ensure that usability did not confound the results. Participants did not report a noticeable difference between the two interface conditions and did not explicitly identify the pop-up as friction. However, we observed a subtle difference in how participants revised or engaged with the proposed diagnoses across conditions. It suggests that friction may influence decision-making without being consciously recognized. Participants raised concerns about automation bias, transparency, data use, and the need to preserve clinical judgment.

Keywords

Design Friction, Large Language Model, Medicine, User Experience

1. Introduction

Large Language Models (LLMs) are increasingly integrated into medical software. Recent advances, including models such as GPT-4 [1], Med-PaLM 2, BioMedLM, Amie, suggest that LLMs may assist in clinical reasoning [2], exploit patient data [3], and suggest potential diagnoses [4, 2]. These capabilities make them attractive for integration into primary care settings [5, 6]. This trend reflects a move toward using Artificial intelligence (AI) to support physicians in their practice, and the capabilities are still explored, including in general practice [7, 8].

However, integrating LLMs into clinical decision-making raises critical concerns. Diagnostic accuracy remains essential, and errors—wether originating from the model or from human reliance—carry real risks for patients [9].

Beyond accuracy, these systems may shape clinicians' reasoning, bias decision-making, or reduce critical reflection [9, 10]. They also raise broader questions related to trust, responsibility, and transparency in healthcare [11]. Ethical concerns such as accountability, trust, and transparency are also central to discussions surrounding LLMs in healthcare [11].To this day, many advances have been made in

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden

*Corresponding author

elise.hallaert@unamur.be (E. Hallaert); jean.albrecq@unamur.be (J. Albrecq); juliette.barnich@unamur.be (J. Barnich); bruno.dumas@unamur.be (B. Dumas)

0009-0009-7555-7069 (E. Hallaert); 0009-0002-2038-3267 (J. Albrecq); 0009-0009-0876-2281 (J. Barnich); 0000-0001-5302-4303 (B. Dumas)

developing software and improving performance that were rarely followed by user-centered approaches [12].

From a Human computer interaction (HCI) perspective, interface design plays a central role in mediating these effects. In particular, design friction—defined as interface features that deliberately slow down interaction to encourage reflection—has been explored as a way to support more mindful and responsible use of digital systems [13, 14]. While such approaches have shown promise in other domains, their role in clinical decision-support systems remains underexplored.

In this context, this paper adopts a human-centered approach to examine how General practitioners (GPs) experience an LLM-based diagnostic feature embedded in an Electronic health record (EHR) interface. We investigate whether introducing friction through a simple interface intervention—a pop-up prompt encouraging critical appraisal—affects how practitioners engage with diagnostic suggestions, and whether such effects are consciously perceived.

We report on a mixed-methods study in which GPs interacted with both a control interface and a frictional variant of an EHR software. Our findings reveal a dissociation between perception and behavior: participants did not explicitly recognize the friction, yet their interactions with the diagnostic suggestions differed across conditions. In particular, participants occasionally revised their initial judgment in favor of the LLM's suggestions, pointing to a potential form of automation bias. Together, these results suggest that friction may influence decision-making without being consciously recognized, raising important implications for the design of clinical AI systems. They also highlight the need to carefully account for bias in clinical contexts, whether or not such interface interventions are present.

Research questions are:

RQ1 How do GPs perceive and evaluate LLM-generated diagnostic suggestions in EHR systems?

RQ2 How does the introduction of design friction influence general practitioners' engagement with and response to LLM-generated diagnostic suggestions?

The remainder of this paper is structured as follows. We first review related work on AI in healthcare and design friction in HCI. We then describe our study design and methodology. Next, we present the results, focusing on the relationship between perceived and observed effects of friction. Finally, we discuss the implications for the design of clinical AI systems and outline directions for future work.

2. Related Work

The integration of AI systems into healthcare has sparked growing interest due to its potential to support clinical decision-making [5]. In Primary Health Care (Primary health care (PHC)), LLMs are used for a range of applications, including medical consultation and health management, health promotion and prevention, diagnosis, triage, mental health support, and medical training [15]. Research has increasingly examined both the benefits of these systems, such as reduced diagnostic time, improved workflow efficiency, and lower operational costs [7, 3, 16], and their limitations. In this regard, Lee et al. [12] reviewed LLM evaluation methods in the medical field and found that models are most often assessed in internal medicine and general practice on completeness, agreement with experts, appropriateness, and reproducibility. They emphasize the need for more comprehensive evaluation approaches that go beyond accuracy alone. Similarly, Gerlich [17] cautions that increased reliance on AI tools is associated with reduced critical thinking skills, mediated by cognitive offloading.

Limitations also include the difficulties related to transparency and interpretability, hallucinations in the generated content, security and legal concerns, biases of the model, and human factors [18, 15]. Human factors refer to the patient's choice to be treated with the help of such tools and the impact these have on the healthcare practitioners. Findings about them show that while most have already used LLMs and appreciated the experience, some expressed concerns about their autonomy and performance, but indicated confidence in improved future interactions [19].

Regarding friction, it has historically been minimized in User experience (UX) design, where smooth and efficient interactions are typically prioritized. Deliberate friction is still an emerging concept within the field of HCI. Researchers explore friction as an intentional design strategy to encourage reflection.

However, this approach remains largely unexplored in the context of medical software, especially the effectiveness of such frictional interventions in clinical contexts, and whether their effects are consciously perceived by practitioners. To our knowledge, no studies in the medical domain have explicitly framed their approach as “design friction”.

Benedetti and Mauri [13] position friction not as a usability flaw, but as a method for reflection and intentional slowing in design interventions. While smooth interfaces can lead to opaque interactions that discourage mindful engagement, introducing frictional elements can prompt users to pause and potentially avoid such unreflective behavior. In this context, friction can be seen as a potential mechanism to counteract automation bias by encouraging more deliberate engagement with system outputs. This theoretical groundwork has yet to be applied or empirically validated in healthcare technology. Nevertheless, a few studies introduced frictional elements in contexts sharing characteristics with clinical systems, such as Decision Support Systems in medical and law domains [20] and demonstrated the feasibility of reducing automation bias and increasing diagnostic accuracy.

3. Materials and Methods

3.1. Materials

ChatGPT (OpenAI, GPT-4) was used to simulate the LLM-assisted diagnostic support. Two generated suggestions were presented to participants during the study. Both suggestions were designed to replicate realistic medical consultation scenarios based on standardized patient data and formatted into bullet points, including age, symptoms, treatments, medical history, and clinical exam results. To ensure these cases’ clinical relevance and consistency, all prompts and resulting diagnostic outputs were reviewed and coded in advance in collaboration with a general practitioner to ensure pertinence, realism and alignment with general practice standards.

The scenarios were presented within a custom-built prototype interface of an electronic health records’ support EHR software based on the Subjective objective assessment plan (SOAP) note style inspired by existing clinical decision support systems on the market. The prototype was reviewed for clinical relevance. While minimal in terms of features, it mirrored essential workflows of current EHR-integrated solutions, with an integration of LLM-generated content (see Figure 1). The frictional prototype generated a warning pop-up window before displaying the diagnostic, followed by a question asking whether the practitioner agreed with the diagnostic. All other features were the same in the control prototype.

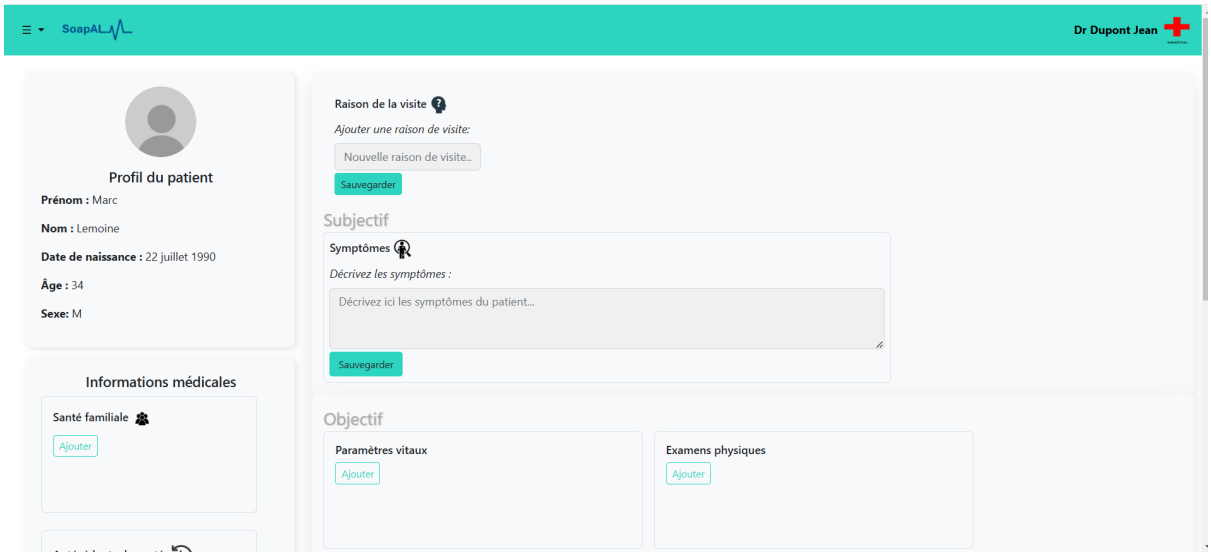

Figure 1: Prototype software main interface. It shows a patient profile structured with SOAP notes, including LLM-generated diagnostic suggestions in the plan section.

During the tests, the software Tobii Pro Lab (v. 25.7.1400) supported the pupil analysis in addition to an observation sheet and an audio recording of each session.

To evaluate the user experience of the prototype, participants completed the User experience questionnaire (UEQ) [21] at the end of the session. The UEQ is a standardized tool designed to measure user perception of the system, enabling benchmark comparison and internal consistency checks.

3.2. Study Design

Ten GPs were selected through convenience sampling. Prior research indicates that five to six participants are typically sufficient for exploratory studies [22]. All participants provided informed consent and received a short briefing about the study, including the role of the AI system beforehand. Each participant interacted with both versions of the interface to complete the diagnostic tasks using the prototype (one with friction, one without), presented in a randomized order. Scenarios were randomly assigned to each condition and simulated realistic clinical cases, requiring practitioners to put themselves in the context of a medical consultation. The diagnostic suggestions presented during the study were generated using a Wizard of Oz approach. In this method, participants interacted with what appeared to be an autonomous AI system while the responses were already planned in the prototype. At the start of each task, GPs received a printed clinical scenario and were prompted to think aloud and to fill in the corresponding fields in the prototype interface. Once completed, the diagnostic suggestion could be generated by clicking a button. The frictional prototype included a warning text about their responsibility and prompted participants to respond to a fixed question: whether they agreed or disagreed with the proposed diagnosis. The non-frictional one created the diagnostic immediately after clicking on the button. In both prototypes, they could modify the diagnosis and were asked to establish a treatment plan based on their scenario evaluation. This aimed to recreate a realistic diagnostic workflow while allowing us to hear and observe their reasoning and interaction with the tool under frictional and non-frictional conditions. Each participant completed both experimental conditions in a within-subjects design, allowing for direct comparison across tasks.

The video recording captured participants' pupil dilatation and focus areas while interacting with the interface. Researchers filled out the structured observation sheet in real time. After completing the tasks, participants completed the User Experience Questionnaire (UEQ) and conducted a semi-structured interview with the two experimenters to explore their impressions and expectations.

To limit bias, the order of test conditions and scenarios was randomized. Clinical relevance was validated by a referring GPs, while the Wizard of Oz method avoided variability in AI output. All experiments were conducted by the same experimenters. It was not mentioned beforehand that the generated diagnosis was not based on the participants' input but was already pre-programmed in the system according to the scenario used.

3.3. Data Collection and Methodology

A combination of quantitative and qualitative metrics have been used:

- Behavioral metrics: pupil dilation data using Tobii Pro Lab, as an indicator of cognitive load.
- Observational data: boolean indicating whether the participant read warnings, accepted the diagnosis or made changes. Observational cues were recorded via the structured observation sheet.
- Subjective evaluation: participants' user experience was assessed using the UEQ, which provided scores across six dimensions.
- Qualitative feedback: interviews and open-ended questionnaire responses were collected for qualitative analysis.

Interviews and open-ended responses were coded inductively, following the principles of thematic analysis. Responses were categorized into usability, trust, diagnostic confidence, and perception of friction. UEQ responses were analyzed using descriptive statistics. Pupil dilation was statistically analyzed

to compare cognitive load across conditions, and the baseline correction was made by subtracting the minimum value detected during the diagnostic reading. Task-evoked pupillary responses have been shown to reliably reflect variations in mental effort and task difficulty [23]. The Tobii Pro Lab software computed pre-processing.

This study involving human participants was reviewed and approved by the University's ethical committee under the approval reference number 215145. All procedures involving human participants followed the ethical standards of the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Before participation, all subjects received detailed information about the study and provided written informed consent. Participants were informed of their right to withdraw at any time without consequence. Informed consent was also obtained for anonymized data and quotations for publication purposes.

4. Results

4.1. Participants

Ten GPs (60% male, 40% female) took part in the study. The participants were distributed across age groups as follows: 10% were aged 20–29, 40% were 30–39, 20% were 40–49, 10% were 50–59, and 20% were 60–69 (N = 10). All participants were volunteers and provided informed consent before participation.

4.2. User Experience

One questionnaire response was excluded from the dataset due to excessive (n=3) inconsistencies in the scale responses, following the UEQ quality criteria. The internal consistency of each scale was assessed using the Guttman lambda-2 coefficient for the small sample size. The coefficient needs to be higher than 0.7 to be considered sufficiently consistent. This assessment is presented in Table 1.

Descriptive statistics for each UEQ scale are presented in Table 2. These values reflect participants' perceptions of the software's usability and design qualities.

Table 1
UEQ Scales: Lambda-2

UEQ Scale	Lambda-2
Attractiveness	0.85
Perspicuity	0.73
Efficiency	0.61
Dependability	0.55
Stimulation	0.80
Novelty	0.79

Table 2
UEQ Scales: Mean and Variance

UEQ Scale	Mean	Variance	Benchmarking
Attractiveness	0.80	1.217	Below average
Perspicuity	1.38	1.180	Above average
Efficiency	1.35	0.937	Above Average
Dependability	1.23	1.024	Above Average
Stimulation	0.48	1.483	Bad
Novelty	-0.30	1.285	Bad

4.3. Pupil Diameter

The pupil diameter change descriptive statistics as an indicator of cognitive load is described in Table 3. An illustrative boxplot is shown in Figure 2.

Table 3
Descriptive statistics results for pupil diameter across conditions Frictional (F) and Non-frictional (NF) (N = 8)

Metric	F	NF
Mean	1.78	0.48
Standard deviation	2.39	0.77
Minimum	0.06	0.07
Maximum	5.40	2.46

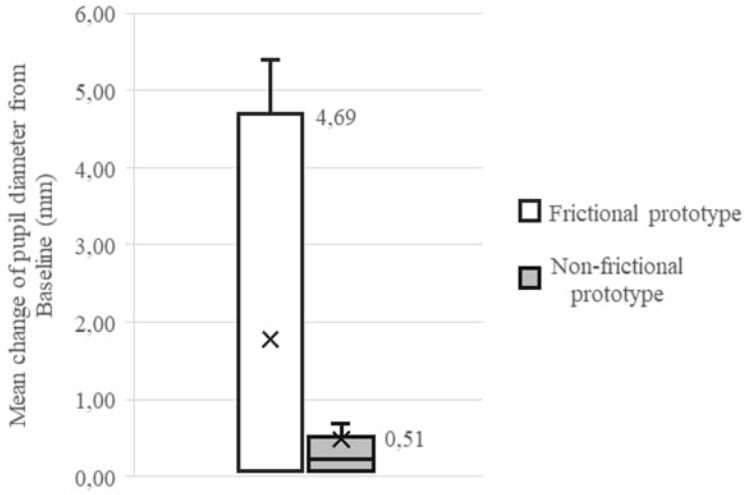

Figure 2: Mean pupil diameter change from baseline (in mm) for F and NF conditions. Error bars represent standard deviation. It shows a greater dilation and variability in the frictional condition (mean 1.78 mm) compared to the non-frictional condition (mean 0.48 mm).

Participants exhibited a larger mean change of dilatation from baseline in the frictional condition (mean = 1.78 mm) compared to the non-frictional condition (mean = 0.48 mm), with a p-value of 0.11, indicating that the difference has no statistical significance. Error bars represent the standard deviation, indicating variability across participants.

4.4. Observational Data

Of the users who modified the diagnostic with the frictional design, three also modified the non-frictional design (n = 3), while four did not (n = 4). No users modified the non-frictional design only (n = 0), and three made no modifications to either version (n = 3). A McNemar test found a light statistically significant effect of frictional design on modification decisions since this p-value is smaller than 0.05, (χ^2 (1) = 4.00, p = .0455).

4.5. Qualitative Analysis

The subjective analysis highlighted distinct themes that emerged through open, axial, and selective coding of subjective data. All users (n=10) expressed critical views towards the LLM suggestions, most of them (n=8) said they already used a LLM before. Some users (n=4) felt that the diagnosis was too focused on hospital situations. Even when they partially agreed with the recommendation, participants (n=3) hesitated to change the LLM suggestion and chose not to modify it with both prototypes. Users

highlighted the importance of understanding the patient's context. Medical decision-making seems rooted in their reasoning, experience, intuition, empathy, and familiarity with each patient's unique situation. Participants (n=6) mentioned the practical benefits of LLM, especially its potential for saving time, helping with simple medical cases, or offering cognitive support for clinical reasoning. Practitioners know the limitations and risks of LLM and raised concerns about overprescription, ethical issues such as blindly following LLM recommendations or overrelying on it, and technical barriers to adopting these tools in everyday medical practice. Despite these concerns, nearly all participants (n=9) were interested in LLM. Half of them (n=5) were open to trying these tools if this could save time. Three users mentioned that they would prefer using LLM for simple medical cases, two users highlighted that LLM should support clinical reasoning, rather than focus on data collection, as some existing tools currently do. Two users reported that the output appeared to be overly influenced by the input data and insufficiently reflective of other information that physicians can obtain through their clinical experience. Four users spontaneously reported that they believed the system to be a real software application. In addition to the standard UEQ evaluation, users completed a supplementary Likert-scale questionnaire to assess the perceived influence of LLM and their trust in its suggestions. Results are summarized in Table 4.

Table 4
User perceptions of LLM influence and trust based on Likert-scale ratings

Measure	Mean	Standard deviation	Interpretation
Influence of LLM on reasoning	2.64	1.46	Low to moderate influence
Trust in LLM diagnosis	3.11	0.62	Moderate trust, consistent answers

5. Discussion

User experience remained stable across conditions, suggesting that the introduction of friction did not negatively impact usability in the short term. This is a key consideration for the integration of such interventions in clinical systems, where efficiency is critical.

Interestingly, during post-test interviews, no participants stated they perceived a difference between the F and NF versions of the interface. Despite the absence of explicit recognition, friction may still have influenced participants' interaction with the system. This highlights a dissociation between perceived and actual effects of interface design: users may not identify friction, yet it can still shape their engagement and decision-making. This observation aligns with previous research suggesting that friction is not an obstacle but a potential reflective trigger [13]. The effect appears more qualitative than behavioral as the pupil diameter change suggests an increased cognitive load, but cannot be generalized.

The most meaningful findings emerged from the qualitative analyses. Half of the participants emphasized that their role as GPs depends heavily on critical thinking and responsible decision-making. Some chose to accept a diagnosis generated by the LLM with which they disagreed. This tension highlights a known risk of AI in healthcare: automation bias, where clinicians defer to system output even when it conflicts with their judgment. While performance and potential bias are still assessed [7], the results of this study align with the results concerning autonomy concerns elicited from practitioners [19]. In two notable cases, participants engaged in retroactive reasoning to reconcile the LLM's diagnosis with their own, post-rationalizing a disagreement. This could be a form of cognitive dissonance reduction previously observed in human-AI interaction [24]. This concern was expressed by participants (n=2); users may suppress valid doubts rather than challenge the system, which raises ethical concerns about underexplained LLM outputs.

One participant expressed discomfort at the idea of asking patients for consent. This balance between legal requirements and seamless integration is still an ongoing debate [25]. Results of this study show that data collection should not be the primary concern of EHR software. This differs from previous results; professionals predicted changes to human interactions while respondents of this study predicted

more changes to HCI [19]. Rather than detailed justifications, three participants wanted lists of possible diagnoses and confidence levels. They wanted autonomy in consulting the generated content rather than receiving recommendations. These results support the trend towards preference for the assistive role rather than a directive one for all but one participant (n=9), as well as a communication role for providing the "name" of the patient's condition to reassure them, as mentioned by one participant.

Another consideration is pre-filled text, employed in the diagnostic suggestion shown to participants. Pre-filled or default content has been identified in HCI literature as a potential dark pattern [26]. It may unintentionally reinforce automation bias or lead to passive acceptance of suggestions. This is critical when users are under time pressure or high cognitive load and highlights the importance of carefully designing AI systems to avoid reducing practitioner autonomy.

5.1. Limitations

This study has several limitations that should be acknowledged. First, the sample size was relatively small (n = 10), limiting the statistical power and internal consistency of the results for quantitative measures. Additionally, the short duration of the tasks did not capture the long-term effects or fatigue-related impacts of repeated use in real-world clinical settings. Although the Wizard of Oz setup for the diagnostic was critical to allow consistent output from the AI system to avoid bias, that may differ from how practitioners interact with a real-time, integrated system, limiting the ecological validity of the findings.

Despite these limitations, the study design includes strengths that mitigate some threats to validity. Constant lighting conditions were maintained to minimize external influence on eye-tracking measures such as pupil dilation. Furthermore, by-task segmentation allowed us to compare specific user interactions under controlled conditions, reducing noise in observational data.

5.2. Recommendations for design

Based on our findings, we propose the following recommendations for designing clinical software integrating LLM-generated diagnostic support:

- **Avoid default answers:** Pre-filled diagnoses can bias decision-making. Present suggestions with confidence levels to encourage critical reflection.
- **Introduce light friction:** Subtle interaction constraints can support clinician autonomy without significantly impacting usability.
- **Promote reasoning over compliance:** Encourage users to avoid or reject LLM outputs. Interfaces should frame the AI as assistive, not authoritative.
- **Address ethical concerns:** Ensure transparency regarding data use and consent. Allow users to manage the recording or use of encoded data according to regulations.

6. Conclusion and future works

This study explored how GPs interact with a diagnostic support tool powered by a LLM, focusing on UX and the potential role of frictional design. While results do not show a significant difference in perceived usability between the frictional and non-frictional prototypes, qualitative and observational data revealed insights into how practitioners engage with LLM-generated suggestions.

These insights reveal that integrating LLMs into primary care cannot be addressed solely through performance metrics or UX scores. The subtle influences on reasoning, autonomy, and professional identity are central to adoption and ethical deployment. This study contributes valuable qualitative evidence to the growing discourse on LLM-human collaboration in medicine and points toward design strategies; friction may serve for critical reasoning without impairing short-term usability.

Participants expressed concerns about autonomy and responsibility. This highlights the importance of designing interfaces that support reflection and preserve professional evaluation. Even when users

disagreed with the diagnosis, some accepted it, indicating possible cognitive and ethical risks associated with unexplained or default LLM outputs.

Future research should involve larger studies to enhance statistical reliability and investigate long-term use of LLM-based tools, including alternative friction mechanisms, alternative result presentations, ethical considerations, and integration in clinical workflows. Another important question concerns whether LLMs could lead to homogenization of outputs, changing and shaping professional practice, and should be explored in future work [27].

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] OpenAI, Introducing ChatGPT, https://openai.com/index/chatgpt/, 2024.

[2] T. Tu, M. Schaekermann, A. Palepu, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, Y. Cheng, E. Vedadi, N. Tomasev, S. Azizi, K. Singhal, L. Hou, A. Webson, K. Kulkarni, S. S. Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, A. Karthikesalingam, V. Natarajan, Towards conversational diagnostic artificial intelligence, Nature (2025) 1–9. doi:10.1038/s41586-025-08866-7.

[3] C. Pinto, J. Faria, L. Macedo, An Active Learning-Based Medical Diagnosis System, in: G. Marreiros, B. Martins, A. Paiva, B. Ribeiro, A. Sardinha (Eds.), Progress in Artificial Intelligence, Springer International Publishing, Cham, 2022, pp. 207–218. doi:10.1007/978-3-031-16474-3_18.

[4] J. A. Fries, L. Weber, N. Seelam, et al., BigBIO: A Framework for Data-Centric Biomedical Natural Language Processing, 2022. doi:10.48550/arXiv.2206.15076. arXiv:2206.15076.

[5] A. Andrew, Potential applications and implications of large language models in primary care, Family Medicine and Community Health 12 (2024) e002602. doi:10.1136/fmch-2023-002602.

[6] X. Chen, J. Xiang, S. Lu, Y. Liu, M. He, D. Shi, Evaluating large language models and agents in healthcare: Key challenges in clinical applications, Intelligent Medicine (2025). doi:10.1016/j.imed.2025.03.002.

[7] E. Goh, R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, E. Horvitz, D. Yang, A. Milstein, A. P. J. Olson, A. Rodman, J. H. Chen, Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial, JAMA Network Open 7 (2024) e2440969. doi:10.1001/jamanetworkopen.2024.40969.

[8] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, E. Horvitz, Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine, 2023. doi:10.48550/arXiv.2311.16452. arXiv:2311.16452.

[9] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–21. doi:10.1145/3449287.

[10] Y. Xu, W. Gao, Y. Wang, X. Shan, Y.-S. Lin, Enhancing user experience and trust in advanced LLM-based conversational agents, Computing and Artificial Intelligence 2 (2024) 1467–1467. doi:10.59400/cai.v2i2.1467.

[11] R. M. Wachter, E. Brynjolfsson, Will generative artificial intelligence deliver on its promise in health care?, JAMA 331 (2024) 65–69. doi:10.1001/jama.2023.25054.

[12] J. Lee, S. Park, J. Shin, B. Cho, Analyzing evaluation methods for large language models in the medical field: A scoping review, BMC Medical Informatics and Decision Making 24 (2024) 366. doi:10.1186/s12911-024-02709-7.

[13] A. Benedetti, M. Mauri, Design for friction. An inquiry to position friction as a method for reflection in design interventions., Convergences - Journal of Research and Arts Education 16 (2023). doi:10.53681/c1514225187514391s.31.139.

[14] A. L. Cox, S. J. Gould, M. E. Cecchinato, I. Iacovides, I. Renfree, Design Frictions for Mindful Interactions: The Case for Microboundaries, Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (2016) 1389–1397. doi:10.1145/2851581.2892410.

[15] H. Qin, Y. Tong, Opportunities and Challenges for Large Language Models in Primary Health Care, Journal of Primary Care & Community Health 16 (2025) 21501319241312571. doi:10.1177/21501319241312571.

[16] K. Shan, M. A. Patel, M. McCreary, T. G. Punnen, F. Villalobos, L. M. Tardo, L. A. Horton, P. V. Sguigna, K. M. Blackburn, S. B. Munoz, K. W. Burgess, T. M. Moog, A. D. Smith, D. T. Okuda, Faster and better than a physician?: Assessing diagnostic proficiency of ChatGPT in misdiagnosed individuals with neuromyelitis optica spectrum disorder, Journal of the Neurological Sciences 468 (2025) 123360. doi:10.1016/j.jns.2024.123360.

[17] M. Gerlich, Ai tools in society: Impacts on cognitive offloading and the future of critical thinking, Societies 15 (2025). URL: https://www.mdpi.com/2075-4698/15/1/6. doi:10.3390/soc15010006.

[18] H. Coelho, On developing ethical ai, Progress in Artificial Intelligence. EPIA 2022 13566 (2022) 512–521. doi:https://doi.org/10.1007/978-3-031-16474-3_42.

[19] J. Sumner, Y. Wang, S. Y. Tan, E. H. H. Chew, A. W. Yip, Perspectives and Experiences With Large Language Models in Health Care: Survey Study, Journal of Medical Internet Research 27 (2025) e67383. doi:10.2196/67383.

[20] N. A. J. Cornelissen, R. J. M. van Eerdt, H. K. Schraffenberger, W. F. G. Haselager, Reflection machines: Increasing meaningful human control over Decision Support Systems, Ethics and Information Technology 24 (2022) 19. doi:10.1007/s10676-022-09645-y.

[21] R. Hartson, P. Pyla, Chapter 20 - Prototyping, in: R. Hartson, P. Pyla (Eds.), The UX Book (Second Edition), Morgan Kaufmann, Boston, 2019, pp. 405–432. doi:10.1016/B978-0-12-805342-3.00020-5.

[22] R. A. Virzi, Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?, Human Factors 34 (1992) 457–468. doi:10.1177/001872089203400407, publisher: SAGE Publications Inc.

[23] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, I. Krejtz, Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze, PloS One 13 (2018) e0203629. doi:10.1371/journal.pone.0203629.

[24] C. E. Seran, M. J. T. Tan, H. Abdul Karim, N. AlDahoul, A conceptual exploration of generative AI-induced cognitive dissonance and its emergence in university-level academic writing, Frontiers in Artificial Intelligence 8 (2025). doi:10.3389/frai.2025.1573368, publisher: Frontiers.

[25] M. Marques, A. Almeida, H. Pereira, The Medicine Revolution Through Artificial Intelligence: Ethical Challenges of Machine Learning Algorithms in Decision-Making, Cureus 16 (2024) e69405. doi:10.7759/cureus.69405.

[26] P. W. S. Newall, What is sludge? Comparing Sunstein's definition to others', Behavioural Public Policy 7 (2023) 851–857. doi:10.1017/bpp.2022.12.

[27] R. Bommasani, K. A. Creel, A. Kumar, D. Jurafsky, P. Liang, Picking on the same person: Does algorithmic monoculture lead to outcome homogenization?, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022, pp. 3663–3678. doi:10.48550/arXiv.2211.13972.