

Symptom reporting quality is reduced when individuals believe to interact with AI instead of human physicians

Moritz Reis, PhD^{1*†}; Florian Reis, MD^{2†}; Yeun Joon Kim, PhD^{3,4}; Aylin Demir, MD⁵;
Jess Lim, MD⁶; Matthias I. Gröschel, MD, PhD⁷; Sebastian D. Boie, PhD²;
Wilfried Kunde, PhD¹

¹ Institute of Psychology, University of Wuerzburg, Wuerzburg, Germany

² Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

³ Judge Business School, University of Cambridge, Cambridge, UK

⁴ Institute of Metabolic Science, School of Clinical Medicine, University of Cambridge, Cambridge, UK

⁵ Department of Pneumology, Heckeshorn Lung Clinic, Helios Klinikum Emil von Behring, Berlin, Germany

⁶ Department of Neurology, Vivantes Hospital Neukölln, Berlin, Germany

⁷ Department of Infectious Diseases, Respiratory and Critical Care Medicine, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[†] Both authors contributed equally.

* Corresponding author. **Email:** moritz.reis@uni-wuerzburg.de

Author Contributions: MR and FR designed the study, did investigations and analyzed data; MR, FR, YJK, AD, JL, MIG, SDB and WK were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Competing Interest Statement: FR and SDB report having employment contracts with Pfizer outside the submitted work. No other disclosures are reported.

Key words: Artificial intelligence; Human-AI interaction; healthcare; bias; medical advice

Abstract

Artificial intelligence (AI) systems demonstrate remarkable potential in healthcare. However, their integration into clinical practice depends on overcoming barriers to user acceptance. While previous research has identified reservations regarding AI-generated medical advice, earlier stages of human-AI interaction, such as how symptoms are communicated, remain largely unexplored. In a preregistered between-subjects experiment ($N = 500$), participants were randomly assigned to provide simulated symptom reports for common medical conditions to either an AI chatbot or a human physician. Report quality was assessed using Large Language Model-based suitability metrics, which were validated by physicians. Participants who believed to interact with an AI (vs. physician) provided lower-quality symptom reports. Our findings indicate a bias at the input-stage of seeking medical advice in digital settings. This outcome could compromise the performance of AI systems in real-world applications, regardless of their actual capacity.

Main text

Healthcare is undergoing a digital transformation characterized by the rapid integration of artificial intelligence (AI) and telemedicine platforms into primary care and triage settings (1-2). Central to both traditional and novel digital care settings are patients' self-reported symptoms, which remain a primary source of information. Accurate and comprehensive symptom descriptions are crucial for guiding diagnostic workflows and clinical decision-making (3).

AI-powered chatbots have recently emerged as first-line triage tools with the potential to streamline patient flow and reduce unnecessary emergency visits (4). However, their effectiveness depends on the quality of patients' unstructured free-text input, and any miscommunication or ambiguity can produce suboptimal outputs (5). Research on human-AI interaction suggests that individuals adapt their language and behavior based on perceptions of the recipient's nature (human vs. AI; 6-7). Such adaptations could compromise clinical utility: if patients alter their symptom descriptions when communicating with an AI chatbot, the medical suitability of those narratives may also change. Understanding these dynamics becomes crucial as digital-first strategies expand, and some patients even turn to medically unsupervised Large Language Models (LLMs) for self-diagnosis (8).

Prior work has primarily focused on the "output side" of human-AI interaction in the medical domain. For instance, identical medical advice was found to be judged as less reliable, less empathic and less worth following when labelled as AI- compared to human-generated (9). However, studies examining the "input side" of this interaction remain scarce. Notably, no study has systematically examined how the perceived recipient (human doctor v. AI chatbot) affects the generation of symptom reports, especially regarding their suitability for an initial medical assessment. It is essential to understand these differences to optimize digital triage approaches and ensure equitable clinical outcomes when patients interact with AI-supported platforms or LLMs.

We conducted a preregistered online experiment with a large UK-based sample ($N = 500$), stratified by demographics to approximate the composition of the UK adult population. Participants were asked to provide symptom reports for two common diseases (unusual headache and flu; randomized order), in the belief that their report would be received either by an AI chatbot or by a human physician (i.e., $N_{\text{cases}} = 1,000$). Participants were randomly allocated to one of two recipient types (AI or physician) and provided symptom reports for both diseases under the same condition. We then tested whether the suitability of these symptom reports for medical assessment is influenced by the believed recipient (AI chatbot vs. human

physician). A report was deemed as more suitable (on a scale from 1 to 5) the better the urgency of a required treatment could be inferred (see scoring scheme in the supporting information). Suitability assessments were conducted using LLM-based ratings, which were validated by physicians from the relevant specialties; both LLMs and physicians were blinded to the recipient type of each report.

We found that participants' symptom reports were 7.9% less suitable for an initial medical assessment when they believed they were interacting with an AI chatbot instead of a human physician (based on ratings from GPT-5.2; see Fig. 1), $b = 0.22$, $p < .001$ (Cohen's $d = 0.34$, 95% CI [0.16, 0.52]).

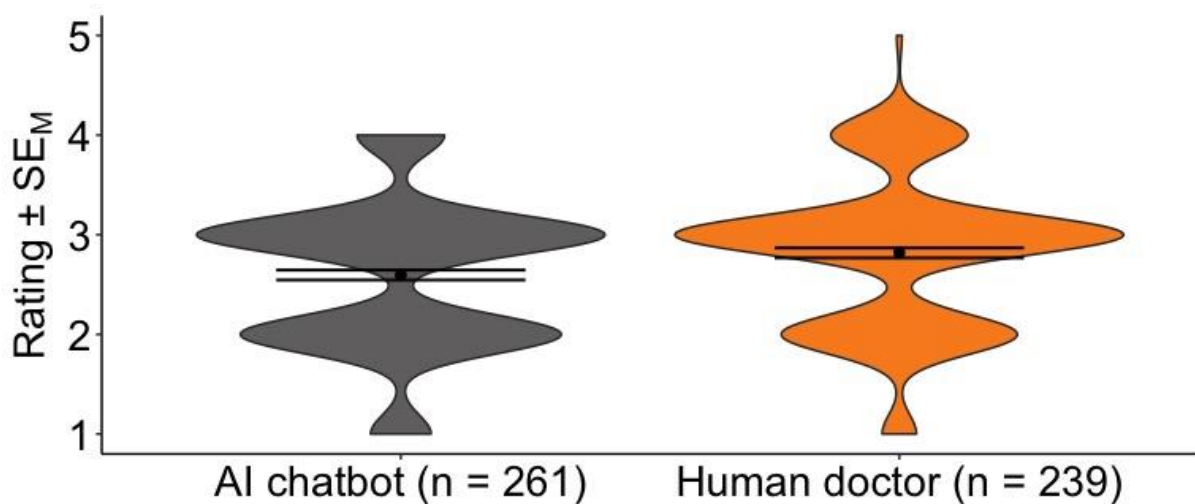


Figure 1. Average suitability ratings of symptom reports for an initial medical assessment by recipient type (AI chatbot: $M = 2.60$, $SE_M = 0.05$ vs. human doctor: $M = 2.82$, $SE_M = 0.05$; rated by GPT-5.2). We calculated mixed-effect regressions to test for differences between both conditions. Error bars show standard errors of individual means.

To validate these ratings, we randomly selected 200 symptom reports from each condition (i.e., 400 reports in total) and obtained additional assessments from four licensed physicians (two neurologists for the “headache” reports and two internists / pulmonologists for the “flu” reports; intraclass correlations (ICC) of human raters: $ICC(1,2) = .75$, 95% CI [0.70, 0.80]). Each report was independently evaluated by the two experts in the respective domain, and we calculated the average of both ratings (interrater reliability of GPT-5.2 and average human-coded suitability ratings: $ICC(1,1) = .70$, 95% CI [0.65, 0.75]). We also repeated our analyses with suitability ratings generated by other LLMs (DeepSeek R1, GPT-o3 mini, GPT-o4 mini, GPT-4o mini, GPT-4.1 mini) and compared these ratings with the assessments by the human experts. For each LLM, suitability ratings were significantly reduced for the AI

compared to the human physician group, $bs \geq 0.15$, $ps \leq .008$. Importantly, this effect was also evident for the reduced subset of human-coded responses, $b = 0.26$, $p = .021$. GPT-5.2 achieved the highest interrater reliability of all tested LLMs (see Table S3 in the supporting information for details). Table 1 shows average suitability ratings for both recipient conditions and different raters.

Table 1

Average suitability ratings (standard deviations in parentheses) for each combination of recipient and rater type.

		Rater type						
		Human expert	GPT-5.2	GPT-o3 mini	GPT-o4 mini	GPT-4o mini	GPT-4.1 mini	DeepSeek R1
Recipient type	Human physician	2.80 (1.00)	2.82 (0.77)	3.55 (0.76)	2.85 (0.80)	3.68 (0.97)	3.21 (0.89)	3.51 (0.79)
	AI chatbot	2.55 (0.95)	2.60 (0.76)	3.32 (0.77)	2.64 (0.83)	3.32 (1.04)	2.92 (0.92)	3.36 (0.83)

Exploratory analyses further revealed that symptom reports in the group with a human doctor as believed recipient were significantly more detailed (i.e., contained significantly more characters) than those in the AI chatbot group (human doctor: $M = 255.60$, $SD = 111.56$; AI chatbot: $M = 228.75$, $SD = 84.81$), $b = 26.85$, $p = .003$, and the level of detail positively affected suitability ratings (GPT-5.2), $b < 0.01$, $p < .001$. Crucially, there was a significant indirect effect of recipient type on suitability ratings via detail level, $b = -0.13$, 95% bootstrap CI [-0.21, -0.05]. This outcome suggests that suitability differences between perceived recipients may be driven by differences in the detail level of symptom reports.

Our results indicate that biases against medical AI encompass the input stage of human-AI interaction. While the present effect sizes are moderate, such biases may nevertheless be of high practical relevance. Given the rapid proliferation of AI-powered health applications (8), particularly for initial medical assessments (4), even modest effect sizes may scale substantially across large user populations. Thereby, such biases may constrain the effective implementation and beneficial impact of AI systems in healthcare delivery.

Although speculative, our exploratory analyses suggest that the present effect may be driven by less detailed requests when participants believe they are interacting with AI. In this case, important information for a valid medical assessment may be incomplete or missing. As

not directly measured, we can only speculate about reasons for this outcome. Potential explanations could be the so-called uniqueness neglect (the belief that AI cannot take the unique characteristics of the user into account; 10) as well as a perceived lack of transparency regarding AI's decision-making processes (11). In line with such reasoning, framing a medical AI as more (vs. less) capable of providing customized (vs. generic) responses, was found to increase the willingness to use such a tool (10). Similarly, providing additional explanations regarding AI's decision-making processes, increased acceptance of medical AI (11). When interacting with AI, individuals may abbreviate or withhold comprehensive symptom information. This could be due to general skepticism about the algorithm's diagnostic capabilities, privacy concerns, or because they regard this information as unnecessary for an AI. Recent evidence indicates that public skepticism toward medical AI extends beyond the technology itself to the physicians who use AI tools (12). This suggests that patients may provide less accurate reports of their symptoms when they suspect their physician is relying on AI assistance.

Overall, our study highlights further obstacles for AI adoption in healthcare. In line with recent findings (5), technical issues are not the main challenge for real-world deployment of medical AI; rather, it is the way humans interact with such tools. Future research should further elucidate human-AI interaction within this context to facilitate optimal AI integration. For instance, the presentation of the AI system (e.g., degree of human-likeness; 13; (believed) capability of taking the user's unique characteristics into account; 10) or explanatory features regarding the AI's decision-making processes (11) could moderate this effect. It may also be helpful to consistently remind users how to provide an AI tool with sufficient and suitable medical information. Furthermore, this bias may also depend on user-related characteristics like the severity and stigma of the medical condition (14).

As a limitation for the external validity of our findings, it should be noted that we used hypothetical scenarios with selected medical conditions for all participants. This design choice allowed us to conduct a highly controlled investigation of this previously unexplored research question. While earlier research indicates a strong correlation between such stated preferences and actual behavior in medical contexts (15), real-world patient reports may differ in several aspects from simulated symptom reporting (e.g., regarding emotional salience and informational completeness). To enhance the real-world implications of our findings, we explicitly selected common and relatable conditions, which one can easily remember or imagine; and for which medical consultation of LLMs is likely. Furthermore, we invited the participants of our experiment for a follow-up study to analyze how their previous experience with the presented conditions affects our findings. Importantly, our results remained robust

when only including participants who indicated that they had experienced the respective symptoms at the time when participating in the study (see supporting information for details). Future research can now build on our findings and transfer them to real-world care settings.

Methods

We assessed the medical suitability of each symptom report using OpenAI's GPT-5.2 model (model ID: gpt-5.2-2025-12-11) via the OpenAI API. The model parameters were set to temperature = 0, top_p = 1, and max_output_tokens = 16 to improve reproducibility and standardize response length. Reports were evaluated independently using separate API calls without retaining conversation history or shared context across inputs. For each scenario, we applied a fixed, standardized prompt template (system-level instruction) specifying the rating rubric and required output format (see supporting information for details).

This study was preregistered (<https://osf.io/cym9u>). Underlying data and analysis code are available in the OSF repository (<https://osf.io/8gkp9>). The study was conducted in accordance with the Declaration of Helsinki. The study protocol was approved by the ethics committee of the Institute of Psychology at the University of Wuerzburg (GZEK 2024-75) and all participants provided informed consent. Details on the sample, experimental procedure, and data analyses can be found in the supporting information.

Acknowledgements and funding sources

This study was supported by the faculty of humanities of the University of Wuerzburg (Grant number: OH_4_2025_3-4). We wish to thank Georg C. Auernig, MD, for sharing his expertise and providing case vignette ratings.

References

1. Topol, E.J. (2019). *Nature Medicine*, 25, 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
2. Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., ... & Herzog, S. M. (2025). *Proceedings of the National Academy of Sciences*, 122(24), e2426153122. <https://doi.org/10.1073/pnas.2426153122>
3. Popovich I., Szecket N., Nahill A. (2019). *Emergency Medicine Journal*, 36, 589-594. <https://doi.org/10.1136/emered-2019-208409>

4. Kopka, M., von Kalckreuth, N. & Feufel, M.A. (2025). *npj Digital Medicine*, 8, 178.
<https://doi.org/10.1038/s41746-025-01566-6>
5. Bean, A. M., Payne, R., Parsons, G., Kirk, H. R., Ciro, J., Mosquera, R., ... & Mahdi, A. (2025). *arXiv*. <https://doi.org/10.48550/arXiv.2504.18919>
6. Goergen, J., de Bellis, E., & Klesse, A. K. (2025). *Proceedings of the National Academy of Sciences*, 122(25), e2425439122.
<https://doi.org/10.1073/pnas.2425439122>
7. Mou, Y., & Xu, K. (2017). *Computers in Human Behavior*, 72, 432-440.
<https://doi.org/10.1016/j.chb.2017.02.067>
8. Haupt C.E., & Marks, M. (2023). *JAMA*, 329(16), 1349–1350.
<https://doi.org/10.1001/jama.2023.5321>
9. Reis, M., Reis, F. & Kunde, W. (2024). *Nature Medicine*, 30, 3098–3100.
<https://doi.org/10.1038/s41591-024-03180-7>
10. Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). *Journal of Consumer Research*, 46(4), 629-650. <https://doi.org/10.1093/jcr/ucz013>
11. Cadario, R., Longoni, C. & Morewedge, C.K. (2021). *Nature Human Behavior*, 5, 1636–1642. <https://doi.org/10.1038/s41562-021-01146-0>
12. Reis, M., Reis, F., Kunde, W. (2025). *JAMA Network Open*, 8(7), e2521643.
<https://doi.org/10.1001/jamanetworkopen.2025.21643>
13. Sestino, A., D'Angelo, A. (2023). In: Cesario, A., D'Oria, M., Auffray, C., Scambia, G. (eds) *Personalized medicine meets artificial intelligence*. Springer, Cham.
https://doi.org/10.1007/978-3-031-32614-1_17
14. Mayer, C. J., Mahal, J., Geisel, D., Geiger, E. J., Staatz, E., Zappel, M., ... & Ditzen, B. (2024). *Computers in Human Behavior*, 161, 108419.
<https://doi.org/10.1016/j.chb.2024.108419>
15. de Bekker-Grob, E. W., Donkers, B., Bliemer, M. C., Veldwijk, J., & Swait, J. D. (2020). *Social Science & Medicine*, 246, 112736.
<https://doi.org/10.1016/j.socscimed.2019.112736>