Study Protocol for a Two-Phase Randomized Evaluation of Large Language Models in Adherence to Evidence-Based Health Communication Guidelines for Breast and Prostate Cancer Screening: The Role of User Prompt Specificity and Minimal Interventions (BOOST-AI)

Protocol version: 1.0

Last updated: 21.10.2024

Dr. Felix G. Rebitschek[1, 2], rebitschek@uni-potsdam.de, 1st author (corresponding)

M.Sc. Christoph Wilhelm[1, 3], christoph.wilhelm@uni-potsdam.de, senior author

----

[1]Harding Center for Risk Literacy, Faculty of Health Sciences Brandenburg, University of Potsdam, Virchowstr. 2, 14482 Potsdam, Germany

[2]Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[3]Institute of Health and Nursing Sciences, Martin Luther University Halle-Wittenberg, Magdeburger Str. 8, 06112 Halle (Saale), Germany

Trial Registration: This clinical trial is registered in the German Clinical Trials Register (DRKS).

Registration Number: DRKS00035228

Registration Date: October 15, 2024

Permalink to Trial Registration: https://drks.de/search/de/trial/DRKS00035228

Introduction

In Western healthcare systems, there is a growing emphasis on ensuring that health information is communicated in a way that allows individuals to make well-informed decisions about their medical treatments. Central to this is the clear presentation of patient-relevant benefits and risks of medical procedures.[1] To achieve this, evidence-based guidelines, such as the Guideline for the Development of Evidence-based Patient Information[2] and the Working Group for Good Practice in Health Information (GPHI)[3], provide a structured framework for health communication. These guidelines stress the importance of using precise numerical data to present risks and benefits, contextualizing information within a relevant reference class, and offering balanced explanations of possible outcomes. Despite the availability of these comprehensive frameworks, their practical application remains inconsistent. As the internet increasingly becomes the primary source of health information for people in Europe and the U.S.,[4,5] many online resources fall short of adhering to these evidence-based standards.[6] As a result, much of the health information available online lacks the necessary accuracy and rigor, and in some cases even spreads misinformation, particularly in areas such as cancer.[7,8] This leaves individuals vulnerable to receiving misleading information, which can lead to poorly informed health decisions and a compromised ability to evaluate risks and benefits effectively. The rapid emergence of artificial intelligence (AI) tools, particularly large language models (LLMs) such as OpenAI's ChatGPT, Google's Gemini, and Mistral AI's Le Chat, has opened new possibilities for digital health communication and laypeople are increasingly turning to these platforms to seek answers to health-related questions.[9] However, a critical concern remains: Can these AI-driven systems consistently provide information that adheres to established guidelines for communicating patient-relevant benefits and harms, especially when prompted by laypeople?

While previous studies have compared LLM responses to those of physicians, exploring aspects such as empathy and accessibility,[10,11] others have focused on LLM accuracy in specific areas such as lab test interpretation,[12] cancer screening recommendations,[13] and random cancer-related queries,[14] just to name a few, revealing both the strengths and limitations of these models. However, in all of these studies, the interactions were based on artificial prompts generated by researchers, rather than authentic inquiries from laypeople. This controlled setup allowed for a systematic evaluation but did not capture the variability and complexity of real-world questions relevant to patients in a chat situation.

In contrast, our research addresses a critical gap in the evaluation of LLMs by investigating how well these AI tools adhere to evidence-based guidelines when communicating about breast and prostate cancer screenings. The primary research question is: Can LLMs provide accurate, guideline-based information on the risks, benefits, and outcomes of cancer screenings, particularly when prompted by laypeople? This question is crucial, given the previous mentioned increasing reliance on AI-driven platforms for health information

Ultimately, previous research has focused on the general capabilities of LLMs, but our study aims to determine whether these models could support informed decision-making in high-stakes medical contexts, focusing on improving both user interaction and model reliability—by providing clear, accurate, and evidence-based health communication. LLMs like ChatGPT, Google Gemini, and Mistral AI will be tested to determine if they can be reliable sources of education, particularly in scenarios where individuals seek information about cancer screening. The ultimate goal is to assess whether these AI systems can be effectively integrated into healthcare systems to complement traditional health communication tools, enhancing patient understanding and supporting informed health decisions without replacing healthcare professionals.

Finally, we need to evaluate the effectiveness of basic evidence-based strategies of information search. Consequently, we study different prompting strategies on the quality of LLM responses. Specifically, we will compare standard prompting (control group) with enhanced prompting (intervention group who receives a brief premise for an evidence-based search) to assess how the specificity of user input affects the adherence of LLM-generated responses to evidence-based health communication guidelines.

Objective

The primary objective is to evaluate how well LLMs, including OpenAI's ChatGPT, Google Gemini, and Mistral AI, adhere to evidence-based guidelines when responding to breast and prostate cancer screening prompts. Specifically, we aim to:

1. Assess the evidence-based quality of LLM responses by evaluating the extent to which the health risk communication provided by LLMs aligns with established evidence-based guidelines. This includes assessing whether LLM-generated responses clearly communicate risks, benefits, and outcomes of breast and prostate cancer screening.
   RQ1: Is the health risk communication provided by LLMs evidence-based? The hypothesis is that LLMs frequently fail to meet standard criteria for evidence-based health risk communication, with more than 50% of responses deviating from these guidelines across multiple presentation criteria.
2. Analyze the effect of prompt specificity on LLM response quality by determining whether more specific, well-informed prompts lead to a higher quality of evidence-based responses from LLMs. This objective focuses on how the inclusion of key decision-making information in prompts affects the clarity and accuracy of LLM outputs.
   RQ2: Does more informed prompting, particularly in terms of health decision-making preparation, result in better evidence-based health risk communication from LLMs? We hypothesize that there is a moderate to strong positive correlation between the specificity of the prompts and the quality of evidence-based health risk communication provided by LLMs.
3. Compare the evidence-based quality of responses generated by laypeople with varying levels of informed prompting by comparing the quality of LLM responses based on prompts generated by laypeople who provide varying levels of detail and information. We will assess whether layperson-generated prompts result in more evidence-based responses than low-informed prompts, but less so than moderately informed prompts.
   RQ3: Are LLM responses generated by laypeople more evidence-based than those from low-informed prompting, but less so than moderately informed prompting? We hypothesize that Layperson-generated prompts produce more evidence-based responses compared to low-informed prompts, but less so than moderately informed prompts that include about 50% of key health decision-making information.
4. Evaluate the impact of a minimal boosting intervention on the quality of LLM responses by assessing whether a minimal boosting intervention—where users are encouraged to provide more specific prompts by considering the consequences of their health decisions—can improve the evidence-based quality of LLM responses.
   RQ4: Does a minimal boosting intervention increase evidence-based responses? We hypothesize that reminding users to consider the consequences of their choices will lead to an increase in evidence-based responses, improving the overall adherence of LLM outputs to guideline-based communication standards.
5. Test the digital native hypothesis: RQ5: Without boosting intervention, digital natives (defined as participants under 30 years of age) do not prompt a higher quality of LLM responses than people from the age of 30 years and more.

Trial Design

The study is divided into two phases. Phase 1 uses a content analysis design to evaluate LLM responses to standardized prompts. Phase 2 is a randomized between-subjects experiment with a 1:1 allocation ratio, comparing standard prompting (control group) and enhanced prompting (intervention group). Each group will interact with LLMs under similar conditions, but with differing levels of instruction specificity.

Given that this study is entirely web-based, there are no onsite requirements for integrating the AI intervention into the trial setting. All interactions with the LLMs will occur remotely, facilitated through the SoSci Survey platform, hosted by the University of Potsdam. The integration of the AI systems (Open AI ChatGPT, Google Gemini, and Mistral AI Le Chat) is managed through an API, ensuring seamless communication between the survey platform and the LLMs.

For offsite requirements, participants will need access to a digital device (computer, tablet) with a stable internet connection to interact with the LLMs and complete the survey.

There is no need for any physical infrastructure or on-premises installations, as the entire intervention process, from user prompts to LLM responses, will be executed in the cloud through secure, encrypted connections.

Methods

This protocol was developed in adherence to the Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension.[15]

Phase 1: Systematic Evaluation of LLM Responses to Predefined Prompts

In the first phase of the study, we will systematically assess the quality of outputs generated by OpenAI's ChatGPT, Google Gemini, and Mistral AI. The LLMs will respond to a set of predefined, standardized English-language prompts related to BC and PC screening. These prompts will encompass a range of typical patient inquiries, such as the risks and benefits of cancer screenings, the interpretation of screening results, and the overall recommendation for undergoing screening. Each LLM will respond to multiple iterations of the same prompts, allowing for an assessment of both consistency and variability in the responses.

In total, six (decision preparation elements) times three distinct (in specificity varied) prompts will be repeated 20 times across three LLMs and two screening topics, resulting in 2,160 total responses to be scored. An additional scoring across six prompts (full decision preparation) will be based on nine evaluation criteria of the compliance with context information standards (e.g. declaration of conflict of interest). Each LLM response will be evaluated against predefined quality metrics, by independent human raters using the validated MappInfo tool for the evidence-based quality assessment of digital health information[16] and a new checklist derived from the Guideline for the Development of Evidence-based Patient Information[2]. This process will help identify patterns of strengths and weaknesses in how LLMs communicate crucial health information.

Phase 2: User-Generated Prompts and Evaluation with a Minimal Boosting Intervention

The second phase of the study shifts from predefined prompts to user-generated prompts to explore how laypeople's input affects the quality of LLM-generated health information. In this phase, participants will be asked to generate their own prompts regarding breast and prostate cancer screening. Participants will be randomized into two groups:

Control group: Participants will generate prompts with no specific guidance, reflecting a typical layperson’s inquiry based on their own knowledge or concerns.

Intervention group: Participants will receive the minimal intervention (boosting) - a brief premise for an evidence-based search: “Please consider the OARS rule: You need to know your options, the advantages and risks of each, and how steady they are to happen.”, which implies simple instructions encouraging them to consider the possible consequences of their health decisions and to provide more specific, detailed prompts.

We will follow a similar approach to Phase 1, using predefined quality metrics to assess the responses. Furthermore, this phase will analyze whether the boosting intervention improves the overall quality of LLM responses by increasing the specificity and relevance of the user-generated prompts. Specifically, the evaluation will focus on two key aspects: (1) the proportion of evidence-based responses, measuring how often LLMs provide information aligned with established guidelines; (2) the impact of the boosting intervention, determining whether simple guidance leads to significant improvements in LLM outputs compared to unguided prompts.

Analysis

Phase 1:

Descriptive statistics will be used to summarize the quality scores of the LLM responses across the different levels of prompt specificity. The mean scores with standard deviations will be calculated for each LLM and prompt category. Additionally, the percentage of responses meeting the predefined evidence-based criteria will be computed. We test the assumption of prompt specificity and explore potential differences between LLMs with the help of ANOVA across repetitions. Two independent researchers will code the data. Interrater reliability will be assessed using Cohen’s kappa coefficient to evaluate agreement between the coders. Any discrepancies will be resolved through consensus discussions to ensure consistent and accurate evaluation.

Phase 2:

Descriptive statistics will be calculated for the demographic variables gender, age, and education, as well as for LLM usage frequency, prior experience, and preference for shared decision-making.

The primary analysis will focus on comparing the quality of LLM-generated responses across groups using inference statistical methods for assessing differences between participants assigned to the prompting conditions (independent variable). To evaluate the responses (the dependent variable) the responses will be coded using two predefined scales to assess its adherence to evidence-based (EB) health communication standards and pooled additionally (so, three dependent variables for independent analyses). The coding will be based on the MappInfo tool for the evidence-based quality assessment of digital health information[16] and the Guideline for the Development of Evidence-based Patient Information[2]. The coding will be done by two independent raters, and inter-rater reliability will be assessed using Cohen's kappa. Any discrepancies between raters will be resolved through discussion and consensus. The code book will be provided in the appendix of the final study.

Secondary analyses will include a statistical test based on age split (<30 vs 30+ years) on the difference in the quality of LLM responses prompted by participants of the control condition and there will be an assessment of participant preferences for informed decision-making processes and their experiences with LLMs, using descriptive statistics and mean comparisons. A subgroup analysis will explore potential differences in responses based on demographic characteristics.

All statistical tests will be conducted at a significance level of $p < .050$ (and adjusted downwards according to Bonferroni procedure for multiple testing), and effect sizes will be calculated to determine the meaning of findings. Results will be presented in detail with supporting tables and figures.

Inclusion Criteria

Phase 1, LLMs:

The inclusion criteria for selecting the LLMs in the study will be the following:

- State-of-the-Art Performance: Only LLMs that represent current, state-of-the-art models in natural language processing, such as OpenAI's ChatGPT, Google Gemini, and Mistral AI, will be considered.
- Accessibility for Public Use: The LLMs must be accessible to the general public, ensuring that their capabilities reflect real-world use cases and that the findings can be generalized to typical interactions by laypeople.
- Multidomain Knowledge: The LLMs must demonstrate the ability to handle a wide range of topics, including healthcare and cancer-related information, ensuring their relevance for answering complex, domain-specific queries.

Phase 2, Participants:

- Age: Participants must be 18 years or older to ensure they have the capacity to make informed decisions and engage meaningfully with health-related prompts.
- Language Proficiency: Participants must have proficiency in English, as the study involves interacting with LLMs in English and understanding health-related information presented in this language.
- Access to Digital Devices: Participants must have access to and be able to use digital devices (e.g., computers, tablets) with internet access, as the study involves generating and submitting prompts to LLMs online.
- Geographic Location: Participants should reside in regions where access to healthcare is comparable to international standards, such as the U.K., to ensure that the health information provided by LLMs is relevant and applicable to their context.

Exclusion Criteria

Phase 1, LLMs:

The exclusion criteria for selecting the LLMs in the study will be the following:

- Limited Access or Restricted Use: LLMs that are not publicly accessible or require proprietary access for specialized use will be excluded, as they do not represent general-use cases for laypeople.
- Domain-Specific Models: LLMs that are specifically trained or fine-tuned for niche domains (e.g., exclusively healthcare-specific models) will be excluded, as they do not reflect the broader, general-purpose models used by the public.
- Non-English Language Proficiency: LLMs that primarily operate in languages other than English or demonstrate limited proficiency in understanding and generating English-language responses will be excluded.

Phase 2, Participants:

The exclusion criteria for selecting the Participants in the study will be the following:

- Health Professionals: Individuals with professional expertise in healthcare, particularly in cancer screening or health communication, will be excluded to avoid bias and ensure that participants reflect the general lay population.
- Previous Experience with LLM Studies: Participants who have taken part in similar studies involving LLMs will be excluded to prevent familiarity with the technology from influencing the results.

Procedure and material

Phase 1:

Researchers will prompt LLMs via API and collect the responses.

Phase 2:

Participants will receive a brief introduction outlining the study's purpose, which involves interacting with one of three LLMs [OpenAI ChatGPT (gpt-3.5-turbo)[17], Google Gemini (1.5-Flash)[18], or Mistral AI Le Chat (mistral-large-2402)][19], all preset as a "helpful assistant" to obtain health information of ether BC or PC screening. Following informed consent, participants will be asked to provide basic demographic details such as gender, age, and education level through a questionnaire interface built with SoSci Survey. They will then be given the choice to receive information about either BC or PC screening. Participants will engage with the chatbot by entering their queries (prompts) through the LLM's API, and the generated responses will be collected alongside the prompts for systematic analysis. At the end of the session, participants will be asked about their frequency of LLM usage, their prior experience with such models, and their perspectives on how an informed decision-making process should ideally be conducted for them, by choosing one from four options.

A pre-test with n=20 participants will be conducted to verify the effectiveness of the randomization procedures, ensuring balanced group assignment, the technical functionality of the survey platform and its integration with the LLM APIs. Additionally, the average completion time will be measured and potential issues related to participant burden will be identified. Feedback on user experience will be collected to address any usability concerns. Based on these findings, necessary adjustments will be made at the protocol for the main trial.

Tests and Outcomes

Primary Outcome:

Expert scoring of how evidence-based LLM communication is (mappInfo tool in addition to a novel score based on the guideline EB health information)

Secondary Outcome:

Self-reported use and experience with LLM; preference in shared decision-making

Sample Size

Phase 1:

There are no human participants involved.

Phase 2:

To analyze a moderate ANOVA main effect (partial eta squared = .06) comparing two between-subjects conditions (with and without intervention), we are aiming for a minimum sample size of n= 237 participants. To ensure a representative sample reflecting the simplified census data of Great Britain in terms of sex, age, and ethnicity, we will recruit n= 300 participants from a diverse pool via Prolific[20].

Recruitment

Participants will be recruited via the online platform Prolific[20].

Assignment of Interventions

Participants will be assigned to either the control group or the intervention group via block randomization. In the intervention group, participants will receive additional instructions designed to improve the quality of the prompts they generate for the LLMs.

Allocation

To achieve a balanced distribution of participants across the study groups, we will employ a block randomization (1:1 allocation ratio) with the BLOCKRAND() function in SoSci Survey. The assignment is concealed by this function.

Blinding

Participants and researchers will be blinded to both the assigned LLM and the type of prompting instructions to prevent any bias in interaction and response evaluation. Coding will happen independently by two researchers.

Data Management

Data will be collected via the SoSci Survey platform, and all datasets will be anonymized and stored securely on password-protected servers to ensure participant confidentiality. Contact details will not be collected, as Prolific—responsible for participant recruitment—operates with an anonymized data collection model. Participant demographic data, such as gender and age, will be collected separately within the study, despite these being part of the quota set by Prolific.

Data handling is fully anonymized. Prolific manages recruitment, while SoSci Survey, hosted by the University of Potsdam, handles the questionnaires. Communication with the LLMs is facilitated through an API within SoSci Survey, and LLM providers do not have access to the study data. Payment processing is managed entirely by prolific.co, and no financial data is collected or stored on our end. Additionally, no names, addresses, birth dates, or IP addresses will be recorded. We will not record the recruitment IDs generated by prolific.co, and due to the large number of simultaneous participants and the variability in time between recruitment and survey completion, it will not be feasible to link questionnaire responses with recruitment IDs.

The anonymized research data will initially be stored on a password-protected account under the control of the project lead on the SoSci Survey server (hosted by the University of Potsdam). For further analysis, the data will be transferred to password-protected computers under the supervision of project team members at the Harding Center for Risk Literacy. Prolific will not have access to the research data at any point.

Statistical Methods

Data will be analyzed using descriptive statistics and inferential statistical tests, including t-tests and ANOVA for group comparisons, and linear regression models for more detailed analyses of prompt specificity effects. All statistical tests will be conducted at a significance level of $p<0.05$, and effect sizes will be calculated to determine the practical significance of findings.

Harms

No physical or psychological risks are anticipated for participants. Anticipated survey completion time is six minutes. Participation is entirely voluntary, and participants can withdraw at any time.

Errors of the used LLMs will be identified through content analysis conducted by independent human raters using predefined quality metrics, and categorized based on their nature and impact on the reliability of the information provided. The errors will then be analyzed to determine patterns of LLM performance issues. If errors are not analyzed in specific instances, it will be due to limitations in the scope of the study, such as focusing primarily on guideline adherence rather than linguistic or technical issues outside the study's objectives.

Monitoring

The study will be monitored by the research team at the Harding Center for Risk Literacy. Regular audits will be conducted to ensure compliance with the study protocol and data management procedures.

Ethics and Dissemination

Ethics approval was obtained from the University of Potsdam's Ethical Committee (Approval No. 52/2024) on August 29, 2024. The complete study protocol is publicly available on the website of the Harding Center (https://www.hardingcenter.de/de/forschung/projekt-eb-llm). Findings from the study will be disseminated through academic publications, conference presentations, and open-access databases to contribute to the field of digital health communication. A de-identified, aggregate-level data will be made available upon reasonable request to qualified researchers who agree to comply with ethical guidelines for data sharing and usage.

Protocol Amendments

All amendments to the protocol will be submitted to the ethical committee for review and approval. Any major changes will be communicated to participants and will be listed, along with justifications, in a new version of the study protocol and the final study publication.

Consent

Participants will provide informed consent electronically prior to participating in the study. The consent form includes detailed information about the study, participant rights, and data protection measures.

Confidentiality

All participant data will be anonymized, and no personally identifiable information will be collected or stored. Data will be stored securely, and access will be restricted to the research team.

Funding and Conflicts of Interest

This study will not receive any external funding. There are no financial conflicts of interest to declare. Furthermore, we confirm that neither the research team nor any individual involved in this study will receive any monetary or personal benefit from the engagement of prolific for participant recruitment. The commissioning of funds for various studies using Prolific is managed through the relevant department for procurement of the University of Potsdam.

Role of Study Sponsor and Funders

As this study is not funded by any external sponsor, there is no external influence on the study design, data collection, management, analysis, interpretation of data, or the writing of the final report. The decision to submit the report for publication rests solely with the research team of the Harding Center.

The research will be conducted independently, without involvement from any third-party sponsor or funder.

Composition, Roles, and Responsibilities

The study will be coordinated by the research team at the Harding Center, with oversight by the project lead Dr. Felix G. Rebitschek. Given the nature of this trial and the lack of external funding, there is no steering committee or endpoint adjudication committee involved. Data management will be handled by the research team, ensuring compliance with data privacy regulations and ethical standards. There is no need for a separate data monitoring committee, as the study presents minimal risk and involves anonymized, non-invasive interactions with AI-based language models.

Contribution

Christoph Wilhelm (CW) and Dr. Felix G. Rebitschek (FGR) were responsible for the conceptualization and development of the methodology. CW wrote this protocol. CW is the guarantor of the mauscript. FGR contributed equaly, provided key resources, and oversaw the project administration. He reviewed, edited, and supervised this protocol. Both autors significantly contributied to the writing, reviewing, and editing of this manuscript.

1. Klein WMP, Chou W-YS, Vanderpool RC. Health Information in 2023 (and Beyond): Confronting Emergent Realities With Health Communication Science. *JAMA* 2023; **330**(12): 1131-2.

2. Luhnen J, Albrecht M, Hanssen K, Hildebrandt J, Steckelberg A. [Guideline for the Development of Evidence-based Patient Information: insights into the methods and implementation of evidence-based health information]. *Z Evid Fortbild Qual Gesundhwes* 2015; **109**(2): 159-65.

3. Arbeitsgruppe G. [Good practice guidelines for health information]. *Z Evid Fortbild Qual Gesundhwes* 2016; **110-111**: 85-92.

4. Andreassen HK, Bujnowska-Fedak MM, Chronaki CE, et al. European citizens' use of E-health services: A study of seven countries. *BMC Public Health* 2007; **7**(1): 53.

5. Calixte R, Rivera A, Oridota O, Beauchamp W, Camacho-Rivera M. Social and Demographic Patterns of Health-Related Internet Use Among Adults in the United States: A Secondary Data Analysis of the Health Information National Trends Survey. *Int J Environ Res Public Health* 2020; **17**(18).

6. Riera R, de Oliveira Cruz Latorraca C, Padovez RCM, et al. Strategies for communicating scientific evidence on healthcare to managers and the population: a scoping review. *Health Research Policy and Systems* 2023; **21**(1): 71.

7. Loeb S, Langford AT, Bragg MA, Sherman R, Chan JM. Cancer misinformation on social media. *CA Cancer J Clin* 2024; **74**(5): 453-64.

8. Johnson SB, Parsons M, Dorff T, et al. Cancer Misinformation and Harmful Information on Facebook and Other Social Media: A Brief Report. *JNCI: Journal of the National Cancer Institute* 2022; **114**(7): 1036-9.

9. Park Y-J, Pillai A, Deng J, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making* 2024; **24**(1): 72.

10. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 2023; **183**(6): 589-96.

11. Haver HL, Gupta AK, Ambinder EB, et al. Evaluating the Use of ChatGPT to Accurately Simplify Patient-centered Information about Breast Cancer Prevention and Screening. *Radiol Imaging Cancer* 2024; **6**(2): e230086.

12. He Z, Bhasuran B, Jin Q, et al. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. *J Med Internet Res* 2024; **26**: e56655.

13. Huo B, McKechnie T, Ortenzi M, et al. Dr. GPT will see you now: the ability of large language model-linked chatbots to provide colorectal cancer screening recommendations. *Health and Technology* 2024; **14**(3): 463-9.

14. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncol* 2023; **9**(10): 1437-40.

15. Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *The Lancet Digital Health* 2020; **2**(10): e549-e60.

16. Kasper J, Luhnen J, Hinneburg J, et al. MAPPinfo - mapping quality of health information: Validation study of an assessment instrument. *PLoS One* 2023; **18**(10): e0290027.

17. OpenAI. GPT-4 Overview2023. https://openai.com/index/gpt-4/ (accessed 23.09.2024).

18. Google. Google Gemini Overview2023. https://gemini.google.com/ (accessed 24.09.2024).

19. AI M. Mistral AI Overview2023. https://mistral.ai/ (accessed 24.09.2024).

20. Prolific. Prolific: Online research participant platform2024. https://www.prolific.com/ (accessed 25.09.2024).

436