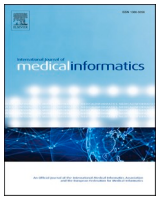
International Journal of Medical Informatics 209 (2026) 106250

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Review article

Large language models versus healthcare professionals in providing medical information to patient questions: A systematic review

Maud M.G. Jacobs [a,b,*], Jacobien H.F. Oosterhoff [c,d], Rintje Agricola [b,e], Walter van der Weegen [b,f]

[a] Department of Orthopedics, Radboudumc, Nijmegen, the Netherlands
[b] Sports & Orthopedics Research Centre, St. Anna Hospital, Geldrop, the Netherlands
[c] Faculty of Technology Policy and Management, Delft University of Technology, Delft, the Netherlands
[d] Department of Orthopaedic Surgery, University Medical Centre Groningen, Groningen, the Netherlands
[e] Department of Orthopedics and Sports Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands
[f] Department of Anesthesiology, Pain and Palliative Medicine, Radboudumc, Nijmegen, the Netherlands

ARTICLE INFO

Keywords:
Patient questions
Large Language Models (LLMs)
Natural Language Processing (NLP)
Healthcare
Artificial intelligence

ABSTRACT

Objective: The rapid expansion of digital healthcare has heightened the volume of patient communication, thereby increasing the workload for healthcare professionals. Large Language Models (LLMs) hold promises for offering automated responses to patient questions relayed through eHealth platforms, yet concerns persist regarding their effectiveness, accuracy, and limitations in healthcare settings. This study aims to evaluate the current evidence on the performance and perceived suitability of LLMs in healthcare, focusing on their role in supporting clinical decision-making and patient communication.

Materials and methods: A systematic search in PubMed and Embase up to June 11, 2025 identified 330 studies, of which 20 met the inclusion criteria for comparing the accuracy and adequacy of medical information provided by LLMs versus healthcare professionals and guidelines. The search strategy combined terms related to LLMs, healthcare professionals, and patient questions. The ROBINS-I tool assessed the risk of bias.

Results: A total of nineteen studies focused on medical specialties and one on the primary care setting. Twelve studies favored the responses generated by LLMs, six reported mixed results, and two favored the healthcare professionals' response. Bias components generally scored moderate to low, indicating a low risk of bias.

Discussion and conclusions: The review summarizes current evidence on the accuracy and adequacy of medical information provided by LLMs in response to patient questions, compared to healthcare professionals and clinical guidelines. While LLMs show potential as supportive tools in healthcare, their integration should be approached cautiously due to inconsistent performance and possible risks. Further research is essential before widespread adoption.

1. Background and significance

Digital transformation in healthcare is a major topic [1]. Manual documentation contributes significantly to the administrative burden, creating a need for automation [2]. To address workforce shortages while maintaining quality of care, organizations are encouraged to reflect on this process [3]. Digital innovations offer promising opportunities to improve both efficiency and care quality [4,5].

Digital healthcare improves accessibility [6] by enabling remote patient monitoring and large-scale data collection for quality improvement. However, these benefits come with unintended consequences. The rapid growth of digital services has nearly doubled patient message volumes in recent years [7]. Many patients turn to chat interfaces for medical advice due to long phone wait times [8], which increases workload stress [9] and contributes to burnout among healthcare professionals [10]. To manage this, routine questions are often delegated to support staff, while complex cases require physician input. Limited time and staffing can result in less experienced personnel answering messages, risking inaccurate responses and patient safety [11]. Some healthcare systems now charge for messages to reduce

* Corresponding author at: Radboudumc, Department of Orthopedics, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands.
E-mail address: maud.jacobs@radboudumc.nl (M.M.G. Jacobs).

https://doi.org/10.1016/j.ijmedinf.2025.106250
Received 23 September 2025; Received in revised form 24 December 2025; Accepted 28 December 2025
Available online 31 December 2025

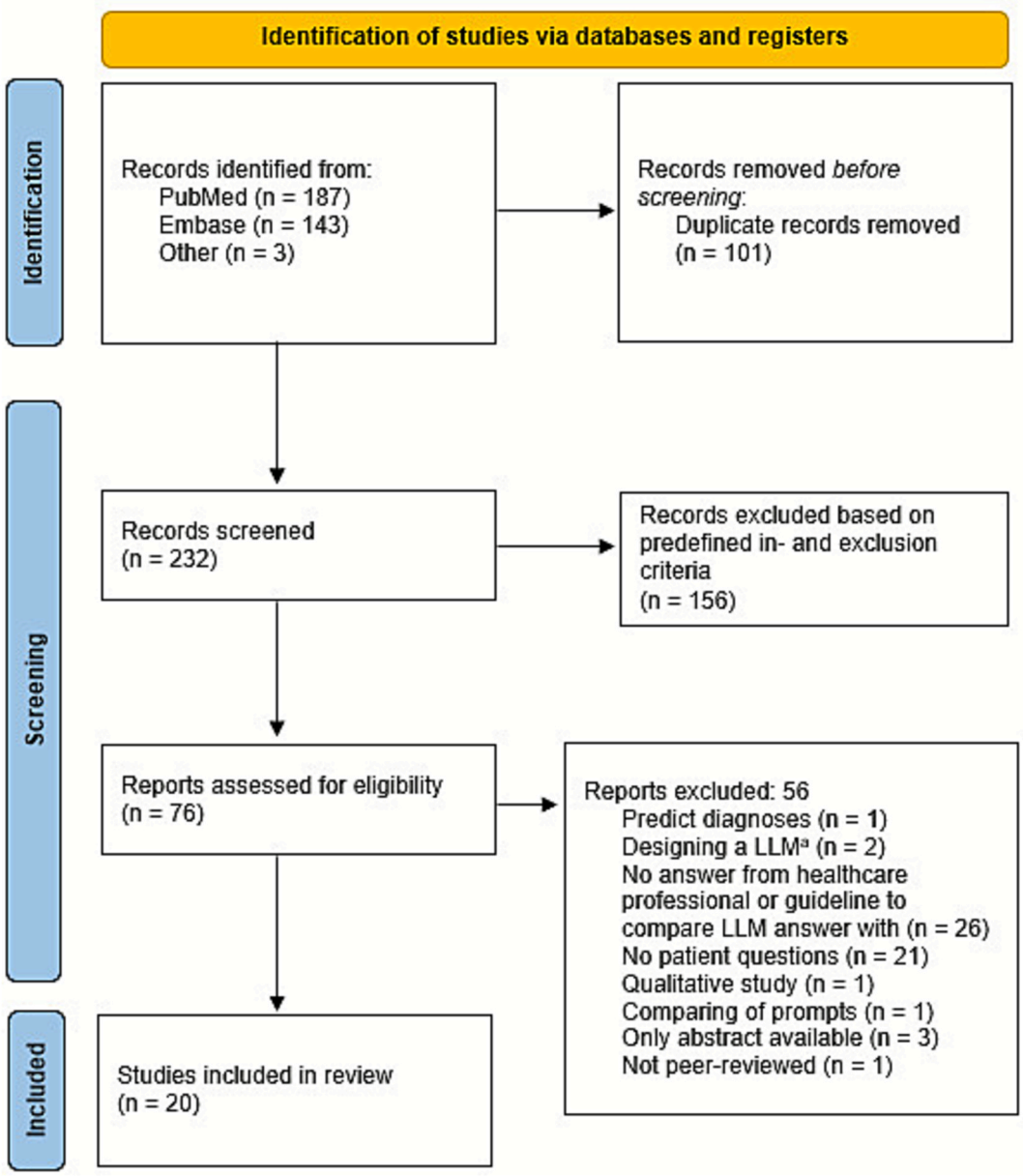


Fig. 1. PRISMA Flow Diagram. [a]LLM = Large Language Model.

volume, but this approach may leave questions unanswered [12].

Given these challenges, technological solutions such as artificial Intelligence (AI) have gained attention. AI, an integral part of computer science, is capable of analyzing complex medical data [13,14] and may help alleviate the growing burden of patient messaging [15]. Within AI, Large Language Models (LLMs) are built on transformer neural network architectures and advanced deep learning principles [16,17], enabling them to approximate human reasoning and generate coherent responses [18,19]. LLMs offer opportunities to automate repetitive tasks currently handled by healthcare personnel [20–23]. A promising application is the generation of draft responses to patient questions for clinician review [24].

Recent studies evaluating models such as Chat Generative Pre-Trained Transformer (ChatGPT) have shown promising results in terms of response preference and relevance [25–27]. However, their effectiveness in healthcare is still not fully investigated. Concerns remain regarding accuracy and safety of LLM-generated responses to patient questions [28–31]. Safety refers to the risk of harm caused by incorrect, biased, or inappropriate responses. Scientific evaluation is needed before these models can be reliably deployed in clinical practice.

2. Objective

Previous reviews have primarily focused on the use of LLMs for clinical research question answering [32] or have provided broad mappings of applications in patient care, particularly in areas such as medical text summarization, translation, and clinical documentation [23,33]. Other reviews emphasized that LLM performance is frequently evaluated through exam-style question-answering tasks in medicine, with results influenced by the prompting strategy used [34], and have also explored their potential integration into electronic health records (EHR's) [35].

However, only a few reviews have examined real-world patient interactions [36]. Existing literature often lack a comprehensive synthesis of evidence regarding LLM performance in actual patient communication scenarios. This review addresses this gap by comparing findings from studies involving real-world patient interactions, thereby providing insights into the suitability of LLMs for supporting clinical

decision-making and patient communication. Specifically, this study aims to evaluate the current evidence on LLM performance and perceived suitability in healthcare, focusing on their role in facilitating clinical decision-making and enhancing patient communication.

3. Materials and methods

The systematic review adhered to Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [37]. The protocol for this systematic review was registered in PROSPERO (Identification number: CRD42024520917).

3.1. Search strategy

A systematic search was conducted in PubMed and Embase in collaboration with a clinical librarian. The electronic databases PubMed and Embase were searched for publications up to 11 June 2025. Four domains of medical subject headings (MeSH) terms and keywords were combined with "AND" and within the domains combined with "OR". The domains included keywords related to large language models, question answering, patient and health personnel (Appendix A). In this review AI-related terminology were restricted to these keywords because ChatGPT was the first widely adopted LLM and is frequently referenced in healthcare literature [38]. The terms LLM and chatbot serve as umbrella categories, ensuring inclusion of comparable models while excluding unrelated AI systems. Records from both PubMed and Embase searches were imported into the web application Rayyan for collaborative screening [39].

3.2. Eligibility criteria

The search is limited to 2018 onwards, as this period marks the introduction of transformer-based architectures and the emerge of large language models (LLMs) [40]. Earlier systems were rule-based or smaller NLP models, which differ fundamentally from LLMs [24,41].

Inclusion criteria encompassed published peer-reviewed observational studies addressing whether LLMs match healthcare professionals in terms of accuracy and adequacy of medical information provided in response to patient questions. Patient questions had to originate from a platform where patients asked questions directly to healthcare professionals or were derived from medical websites that included frequently asked questions or were created by doctors based on what they commonly hear in practice. These studies evaluated LLMs in clinical contexts by comparing their responses to those provided by healthcare professionals or based on clinical guidelines. Assessors (independent) judged the performance of LLMs, allowing for an analysis of their accuracy, suitability, and acceptance in healthcare practice.

Exclusion criteria were as follows; 1) studies published in a language other than English or Dutch; 2) publications addressing LLMs responses outside the healthcare setting; 3) publications originating from non-academic sources (e.g., newspapers, internet websites and magazines); 4) publications that did not offer full-text availability and 5) articles in wherein an LLM was utilized for educational purposes, such as assessing the model's ability to provide accurate responses to a medical examination.

3.3. Study selection

After removal of duplicates and non-English or non-Dutch articles, all titles and abstracts underwent independent screening using Rayyan by two authors to determine eligibility. Subsequently, the full-text manuscripts of selected records were independently screened by these two authors. In case of uncertainty, consensus was reached through discussion. A flow diagram illustrating the study selection process is presented in Fig. 1. Tables summarizing the characteristics of the studies were included in this review. In addition to summarizing study characteristics in tables, the review provides a descriptive synthesis highlighting key similarities and differences, including study design, populations, interventions, and outcomes.

3.4. Data synthesis

LLMs delivering medical information to patient questions constituted the intervention group. Their responses were compared to reference responses obtained from the included studies; these were either answers previously formulated by healthcare professionals to the same questions or recommendations extracted from clinical guidelines, and this comparison was already performed within the included studies. Accuracy (how closely LLM responses aligned with those of healthcare professionals or clinical guidelines in terms of quality and similarity) and adequacy (correctness and the response's fulfillment of the question's requirements, including comprehensiveness, relevance, clarity, readability, thoroughness, and empathy) of medical information served as primary outcomes. Each study was analyzed separately to thoroughly evaluate the performance of the LLM versus healthcare professionals. Overall conclusions, based on significant differences, were compared between the studies.

3.5. Risk of bias assessment

To assess the risk of bias, the standardized tool for Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) was employed [42]. Ratings were assigned by one author, with green indicating low bias, orange signifying intermediate bias, and red indicating high bias. The ROBINS-I tool examines bias in pre-intervention (confounding, selection bias), at intervention (classification bias), and post-intervention (bias due deviations from intended interventions, missing data bias, outcome measurement bias, reporting bias) (see supplementary table 1, Appendix B). When a study exhibited high bias, it was included but explicitly noted in the discussion that the study's reliability was questionable.

4. Results

4.1. PRISMA flow diagram

A total of 333 studies were identified through three sources: PubMed (n = 187), Embase (n = 143), and other sources such as snowballing and reference lists (n = 3). After removing duplicates (n = 101), 232 unique records remained. No records were excluded based on language. Of these, 156 were excluded during title and abstract screening based on predefined in- and exclusion criteria, leaving 76 for full-text review. After full-text screening, 56 were excluded, resulting in 20 studies included in the systematic review. See Fig. 1.

4.2. Study characteristics

The majority of studies included in the review (n = 20) were published in 2024 (n = 12), with others from 2023 (n = 1) and 2025 (n = 7). The number of patient questions varied widely, ranging from 251 [43] to eight [44]. Several sources were used to derive these questions: three used patient questions from eHealth platforms [31,45,46], six studies used questions composed by physicians [29,44,47–50], ten used FAQs from patient websites [43,51–59], and one study used a combination of physician-composed and FAQ-based questions [60]. All studies used ChatGPT, with various versions, and many included other LLMs such as Bard, Claude, Gemini, and Perplexity. Four studies [29,50,54,59] used prompts to instruct the LLM to respond as a physician. Most studies were conducted in hospital settings, with only one in primary care [44]. See supplementary table 2 (Appendix B).

Supplementary table 3 (Appendix B) provides an overview of the methods used to evaluate the agreement between LLM-generated

responses and those of healthcare professionals. Most studies assessed outcomes such as accuracy, clarity, empathy, reproducibility, and readability. Nearly all studies (n = 17) used a Likert-type scale to rate these outcomes, with the exception of two studies [48,49] that used binary or categorical scoring based on guideline consistency or reproducibility. Assessments were conducted by a diverse group of (independent) raters, including healthcare professionals (e.g., physicians, specialists, dietitians) (n = 93); non-physician raters (n = 3); authors or researchers (n = 10); patients (n = 31); PhD students and professors (n = 3). The identities of the non-physicians, authors of the article, early-career physician and professor were not specified in the studies. For the team of fellowship-trained neurosurgeons, the exact number of assessors involved was not reported. The number of assessors per study ranged from one [29,49] to 42 [45]. Four studies [29,47,56,58] used validated readability instruments such as the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL).

4.3. Performance of LLM and healthcare professionals in answering patient questions

4.3.1. Accuracy and correctness

ChatGPT often scores high on accuracy, sometimes even outperforming healthcare professionals (HCPs) [45,49,54,60]. In some cases, HCPs score higher or the difference is not statistically significant [31,43,56]. Accuracy rates vary from 47 % to 95 %, depending on the model version (e.g., ChatGPT 3.5 vs. ChatGPT 4.0), language, and context. ChatGPT rarely gives completely incorrect answers but may provide incomplete or partially correct information [51,53]. Supplementary table 4 (Appendix B) summarizes key study findings.

4.3.2. Readability and comprehensibility

FRE and FKGL scores show that ChatGPT's responses are less readable than those of HCP's [56]. However, in some studies, ChatGPT is rated as more comprehensible, especially by laypeople [31,45].

4.3.3. Empathy

HCPs are rated as more empathetic [31], but in some cases, ChatGPT scores higher [46,56]. Empathy scores vary by language (English vs. Italian) [50].

4.3.4. Thoroughness, completeness, and actionability

ChatGPT often performs well in terms of thoroughness, completeness and actionability [31,44,56,58,59]. In some cases, it is considered overly detailed or overly cautious, leading to overestimation of conditions [52].

4.3.5. Reproducibility and consistency

ChatGPT shows high consistency in providing similar answers to repeated questions [49,51,53]. Reproducibility rates are often above 90 %, which is crucial for reliability in medical contexts.

4.3.6. Statistical significance and methodology

Studies compared their (Likert)-scores using correlation coefficients or p-values to indicate significant differences. Most studies use T-tests, ANOVA, or Kruskal-Wallis to assess differences.

4.3.7. Comparison between LLMs

Each study used ChatGPT, and in some cases, ChatGPT was also compared to other LLMs. ChatGPT is frequently compared to other models like Claude, Gemini, Perplexity, and Copilot. ChatGPT generally scores highest, especially in accuracy, empathy, and relevance [46,47,52,55,57].

4.3.8. Evaluator differences

There are notable differences in ratings between physicians, non-physicians and patients. Patients tend to rate ChatGPT more positively, especially in empathy and satisfaction [45]. Non-physician raters rate healthcare professionals more positively [31].

4.3.9. Best topics for LLMs

In thyroid-related topics [45], ChatGPT outperformed specialists in accuracy, comprehensiveness, compassion, and satisfaction. For rare respiratory diseases [46], ChatGPT 4.0 scored highest on correctness, comprehensibility, relevance, and empathy. In total knee replacement [60], it achieved 88 % accuracy and perfect relevance.

4.4. Risk of bias assessment

The studies were assessed using the ROBINS-I tool [42]. Supplementary table 5 (Appendix B) presents the evaluation across different domains, with intermediate denoting missing or conflicting information.

All studies scored low on confounding and bias due to missing data, as questions posed to an LLM typically do not experience dropout or interact with other aspects. Additionally, bias in the classification of interventions was well-described, as there were only two groups: the LLM and the healthcare professional, with any subgroups outlined beforehand. The same applied to bias due to deviations from intended interventions, as no unexpected deviations occurred in these investigations. However, one study [44] posed fewer than ten questions and one other study [47] only posed fifteen questions, therefore scoring intermediate on bias. Studies with intermediate bias in outcome measurement lacked new chats, blinded and/or independent assessors, more than one assessor, or validated methods. Several studies scored intermediate for reporting bias due to unclear results from missing or inconsistent values. None of the included studies had a high risk of bias.

5. Discussion

This systematic review is, to our knowledge, the first to provide an overview of current comparisons between LLMs and healthcare professionals in answering patient questions. Twelve out of twenty included studies identified ChatGPT as a viable alternative, demonstrating high accuracy, similarity to human responses, scientific correctness, and comprehensibility [44–46,48–51,53,54,57,59,60]. Six studies reported mixed results [29,43,47,55,56,61], and two studies deemed LLMs unsuitable due to language complexity, poor quality, reduced empathy, and response accuracy [31,52].

5.1. Interpretation of key findings

5.1.1. Variation in study characteristics and clinical contexts

Two studies [44,47], of which one viewed LLMs as a viable alternative and one reported mixed results, used a limited number of questions (<20). With fewer questions, the reduced variation in topic and complexity makes the results more vulnerable to chance, which can affect the reliability of both positive and negative outcomes. However, studies that included approximately 150 questions and similarly concluded that LLMs can serve as a viable alternative, suggesting that positive findings are not solely attributable to small sample sizes [48,59].

LLMs were seen as a viable alternative to answer patient question related to nutrition, ophthalmology, colon and rectal cancer, total knee replacement, hypertension, thyroid-related topics, glaucoma and oral cancer, rare respiratory diseases, ovarian cancer, neck masses, prolonged disorders of consciousness, but not in dermatology and laboratory results. However, it is difficult to argue that dermatology and laboratory results represent inherently more challenging topics for LLMs compared to other domains. In fact, one study demonstrated that LLMs can outperform humans on certain patient questions in dermatology [62], and similar findings were reported in a study evaluating LLM-generated responses to patient questions about laboratory test results [63].

One study using raw patient questions from an eHealth platform reported high scores for accuracy and empathy in favor of ChatGPT, with its responses also rated as more readable [45]. Another study using similar patient-generated questions also found ChatGPT to perform well in terms of empathy [46]. This demonstrates strong potential for using LLMs to answer real patient questions, which was the specific focus of this study.

5.1.2. Influence of measurement methods on reported performance

Measurement methods, predominantly Likert-scale or similar, had negligible influence, as all studies focused on comparing LLM responses to those of healthcare professionals or clinical guidelines. The same applies to the choice of LLM used; all studies included ChatGPT, with various versions of ChatGPT used in those studies that considered LLMs as a viable alternative as well as those viewing LLMs as unsuitable alternatives. Therefore, this is unlikely to have significantly influenced the outcomes. The choice for ChatGPT likely reflects the fact that ChatGPT is widely accessible [23], and some research indicates that ChatGPT performs better than other LLMs in certain evaluation settings [63]. Newer LLMs like ChatGPT 4.0 and Claude 2.1 consistently outperform older models such as ChatGPT 3.5 in accuracy, clarity, and empathy. They often match or exceed the performance of healthcare professionals, especially in specialized topics [46–48,57].

5.1.3. Role of assessors in evaluating LLM responses

Including laypersons, such as non-physicians and patients, as assessors contributed to a more diverse and potentially more representative evaluation of LLMs. In one case, lay assessors judged the LLM as unsuitable, despite being unaware they were comparing it to a healthcare professional, suggesting their judgement was unbiased [31]. In contrast, another study found that patients considered the LLM a suitable alternative [45]. These observations provide limited insight, as they do not establish a consistent pattern regarding patient attitudes toward LLMs in healthcare. However, other research has shown that patients can view chatbots (or LLM-based solutions) positively, provided that the system delivers accurate, contextually appropriate responses and is easy to use [64].

Interestingly, four studies used a prompt to guide ChatGPT; while one of them reported mixed results [29], the other three demonstrated that LLMs were capable of providing high-quality and appropriate responses [50,54,59]. However, these findings highlight that the quality of the prompt is critical for eliciting the best possible answers from the LLM [34 63]. Notably, the studies with the largest number of assessors ($n = 37$ and $n = 42$) concluded that ChatGPT responses were accurate and adequate, reinforcing its potential as a reliable alternative when evaluated by a broader group [44,45].

5.1.4. Impact of study bias on interpretation of findings

Among the studies with the lowest risk of bias [29,43,45,46,50–52,58,60], two concluded that it was unclear whether the LLM outperformed humans, one found humans to be superior, and six studies indicated that the LLM performed better. This pattern suggest that studies with a lower risk of bias more frequently conclude that LLMs represent a suitable alternative to human performance.

5.2. Strengths and limitations

The study benefits from its well-defined inclusion and exclusion criteria that emphasize studies comparing LLM responses with those of healthcare professionals or clinical guidelines on patient questions conducted by independent assessors. By assessing both accuracy and adequacy across various aspects, the study effectively evaluates the quality of LLM responses. To our knowledge, no previous systematic review on studies comparing LLM responses to healthcare professional or guideline responses for patient questions has been published. Variability in reporting and measurement aspects makes drawing conclusive inferences challenging. Despite using the ROBINS-I tool, bias in included studies remains possible due to incomplete descriptions, impacting result validation.

While the synthesis of findings provides useful insights, these results should be interpreted with caution. Differences in LLM performance may be overstated due to the small number of included studies and the limited question sets used. Such variations could reflect sampling variability rather than true differences in model capability. Therefore, these findings should be considered exploratory. This review provides an early snapshot of a rapidly evolving field, where both the evidence base and the technology are advancing quickly, making some conclusions time-sensitive. Acknowledging this dynamic context enhances transparency and underscores the need for future research using larger and more diverse datasets to confirm these results and maintain relevance.

Another limitation of this review is the narrow search strategy focused on the terms LLM, ChatGPT, and chatbot, which may have excluded studies using broader keywords. Future reviews should consider a wider range of terms to capture related work and provide a more comprehensive overview.

5.3. Future research

This systematic review suggests that LLMs may be suitable alternatives to healthcare professionals in answering patient questions, but consistent with prior studies [29,31,43,47,52,61] highlighting deficiencies in complexity, poor quality, reduced empathy, and response accuracy. The lack of empathy, intuition, and experience compared to healthcare professionals [30] may be improved in the future, enabling LLMs to serve better as an alternative or aid for healthcare professionals in answering patient questions. One excluded study, focused on responses to patient complaints rather than medical questions, still showed that ChatGPT 4.0 outperformed human responses in empathy, completeness, and satisfaction [65]. This study provides significant evidence supporting the effectiveness of LLMs in the resolution of patient complaints.

LLMs can in future automatically generate a response after which the healthcare professional can check it and send it to the patient, this way time can be saved. Besides, the prompt can be improved so that the LLM is encouraged to respond as a doctor and information on, for example, wound care, movements after surgery can be added to the LLM so that it will always give the right advice which is in line with the human doctor.

Despite promising results, future research should focus on newer LLMs and their performance across medical domains. Key gaps include the limited use of real patient-generated questions, lack of cost-effectiveness data, and overreliance on simulated or FAQ-based inputs. To advance the field, studies should incorporate real-world data, diverse assessors, and questions from e-health communication platforms. Additionally, future research should explore why LLMs perform better in certain clinical domains and less effectively in others. It will also be important to investigate the underlying factors driving these performance differences, such as topic complexity and data availability.

5.4. Practical implications

Despite the potential of LLMs in answering patient questions, several challenges remain. LLMs may help alleviate the growing burden of patient messaging and offer opportunities to automate repetitive tasks currently handled by healthcare personnel [15,20–23]. A promising application is the generation of draft responses to patient questions for clinician review . If successfully implemented, these capabilities could reduce administrative workload and improve response times, potentially enhancing patient satisfaction. However, data privacy must be safeguarded to protect patient information. Technical integration requires IT expertise, which may not be readily available. The costs of implementing LLMs are still unclear, and their sustainability, in terms of energy consumption and long-term employability, must still be

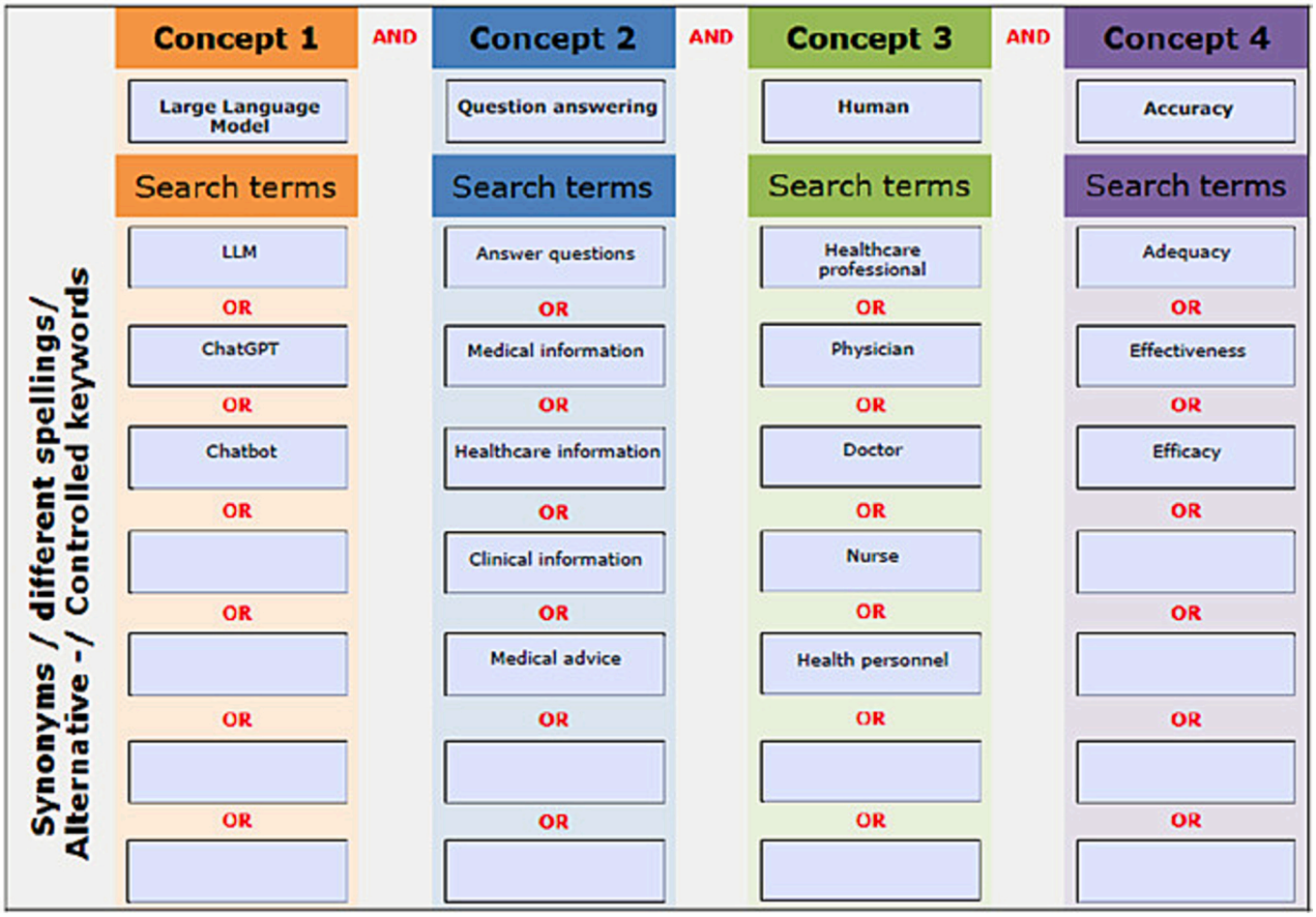


Fig. 2. Search Planning Form.

investigated. In addition, successful implementation will require clinical protocols, training for healthcare professionals, and especially in early stages, a human-in-the-loop to ensure safe and effective use of LLMs in patient communication.

5.5. Prospective applications of LLMs in healthcare

Prospective applications include using LLMs as clinical decision-support tools, patient education aids, and triage assistants. These roles are justified by their ability to process large volumes of medical text and generate contextually relevant outputs. For example, summarizing guidelines can help clinicians quickly access key recommendations, while drafting patient-friendly explanations supports clear communication and health literacy. Assisting with routine questions may reduce administrative burden and free up time for direct patient care. In all cases, professional oversight remains essential to ensure safety and accuracy.

6. Conclusion

The review summarizes current evidence on the accuracy and adequacy of medical information provided by LLMs in response to patient questions, compared to healthcare professionals and clinical guidelines. While LLMs show potential as supportive tools in healthcare, their integration should be approached cautiously due to inconsistent performance and possible risks. Further research is essential before widespread adoption.

7. CRediT authorship contribution statement

M.J and W.v.d.W. extracted the data from the original studies, conducted the initial analyses and designed the data collection instruments. All authors contributed (from first draft to final version) to the writing of the manuscript. M.J. and W.v.d.W. carried out the statistical analyses and contributed to the writing of the methods and results section. J.O. and R.A. provided guidance and expertise in the overall conceptualization of the review and critically reviewed the manuscript. All authors approved the final manuscript as submitted and agreed for accountability for all aspects of the work.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Each author certifies that they have no commercial associations (e.g., consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted article.

Acknowledgements

The authors would like to express their sincere gratitude to Mrs. I. ter Hoeven for her assistance in creating the search string. ChatGPT and Copilot have been used to translate or clarify sentence structure.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

APPENDIX A: SEARCH PLANNING FORM

Fig. 2 contains the search planning form.

The finalized PubMed search string was as follows: (“Large Language Model” OR “LLM” OR “ChatGPT” OR “Chatbot”) AND (“question answering” OR “answer questions” OR “medical information” OR “healthcare information” OR “clinical information” OR “medical advice”) AND (“human” OR “healthcare professional” OR “physician” OR “doctor” OR “nurse” OR ‘’health personnel’’[MeSH]) AND (“accuracy” OR “adequacy” OR “effectiveness” OR “efficacy”).

The finalized Embase search string was as follows: ((large language model or LLM or ChatGPT or chatbot).ti,ab. AND (question answering or answer questions or medical information or healthcare information or clinical information or medical advice).ti,ab. AND (human or healthcare professional or physician or doctor or nurse or health personnel).ti,ab. AND (accuracy or adequacy or effectiveness or efficacy).ti,ab.).

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2025.106250.

References

[1] K. Ghosh, M. Dohan, H. Veldandi, M. Garfield, Digital Transformation in Healthcare: Insights on Value Creation, J. Comput. Inf. Syst. (2022) 1–11.
[2] K. Rego, T. Petzold, Relating digitalization and quality management in health care organizations: a systematic review, Health Care Manage. Rev. 50 (2) (2025).
[3] K. Sippli, S. Deckert, J. Schmitt, M. Scheibe, Healthcare effects and evidence robustness of reimbursable digital health applications in Germany: a systematic review, npj Digital Med. 8(1):495 (2025).
[4] O. Steidle, K. Rego, T. Petzold, Digitale Gesundheitsversorgung, Anforderungen an Eine Erfolgreiche Transformation. Gesundheitswesen. 86 (08/09) (2024) 549–552.
[5] F.A. Bernardi, D. Alves, N. Crepaldi, D.B. Yamada, V.C. Lima, R. Rijo, Data Quality in Health Research: Integrative Literature Review, J. Med. Internet Res. 25 (2023) e41446.
[6] D. Erku, R. Khatri, A. Endalamaw, E. Wolka, F. Nigatu, A. Zewdie, et al., Digital Health Interventions to Improve Access to and Quality of Primary Health Care Services: a Scoping Review, Int. J. Environ. Res. Public Health 20 (19) (2023).
[7] A.J. Holmgren, N.L. Downing, M. Tang, C. Sharp, C. Longhurst, R.S. Huckman, Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use, J. Am. Med. Inform. Assoc. 29 (3) (2022) 453–460.
[8] J.A. Vaucel, N. Enaud, C. Paradis, C. Bragança, A. Courtois, M. Lan, et al., Poison control centres and alternative forms of communication: comparison of response rates between text message and telephone follow-up, Clin. Toxicol. (Phila.) 60 (8) (2022) 947–953.
[9] D.J. Reid, F.J. Reid, Text or talk? Social anxiety, loneliness, and divergent preferences for cell phone use, Cyberpsychol. Behav. 10 (3) (2007) 424–435.
[10] M. Tai-Seale, E.C. Dillon, Y. Yang, R. Nordgren, R.L. Steinberg, T. Nauenberg, et al., Physicians' Well-being Linked to In-Basket Messages Generated by Algorithms In Electronic Health Records, Health Aff (millwood). 38 (7) (2019) 1073–1078.
[11] C.A. Sinsky, T.D. Shanafelt, J.A. Ripp, The Electronic Health Record Inbox: Recommendations for Relief, J. Gen. Intern. Med. 37 (15) (2022) 4002–4003.
[12] A.J. Holmgren, M.E. Byron, C.K. Grouse, J. Adler-Milstein, Association between billing Patient Portal Messages as e-Visits and Patient Messaging volume, JAMA 329 (4) (2023) 339–342.
[13] A.N. Ramesh, C. Kambhampati, J.R. Monson, P.J. Drew, Artificial intelligence in medicine, Ann. R. Coll. Surg. Engl. 86 (5) (2004) 334–338.
[14] J. Bajwa, U. Munir, A. Nori, B. Williams, Artificial intelligence in healthcare: transforming the practice of medicine, Future Healthc J. 8 (2) (2021) e188–e194.
[15] P. Hamet, J. Tremblay, Artificial intelligence in medicine, Metabolism 69s:S36–s40 (2017).
[16] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Contreras, S. Siegel, A. Bumin, et al., Transformers and large language models in healthcare: a review, Artif. Intell. Med. 154 (2024) 102900.
[17] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, et al., Medical foundation large language models for comprehensive text analysis and beyond, npj Digital Med. 8 (1):141 (2025).
[18] K.I. Alohali, L.A. Almusaeeb, A.A. Almubarak, A.I. Alohali, R.A. Muaygil, Reasoning-based LLMs surpass average human performance on medical social skills, Sci. Rep. 15 (1) (2025) 36453.
[19] X. Luo, A. Rechardt, G. Sun, K.K. Nejad, F. Yáñez, B. Yilmaz, et al., Large language models surpass human experts in predicting neuroscience results, Nat. Hum. Behav. 9 (2) (2025) 305–315.
[20] M. Hussein, M. Pavlova, M. Ghalwash, W. Groot, The impact of hospital accreditation on the quality of healthcare: a systematic literature review, BMC Health Serv. Res. 21 (1) (2021) 1057.
[21] Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare [Internet]. 2025; 13(6):[603 p.].
[22] J. Seo, D. Choi, T. Kim, W.C. Cha, M. Kim, H. Yoo, et al., Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study, J. Med. Internet Res. 26 (2024) e58329.
[23] Anderle L, Abidi A, Ünlü S, Steidle O, Janßen T, Petzold T. LLM usage for clinical decision support and documentation support in hospitals – a systematic review. 2025;Preprint.
[24] J. Clusmann, F.R. Kolbinger, H.S. Muti, Z.I. Carrero, J.N. Eckardt, N.G. Laleh, et al., The future landscape of large language models in medicine, Commun Med (lond). 3 (1) (2023) 141.
[25] J.W. Ayers, A. Poliak, M. Dredze, E.C. Leas, Z. Zhu, J.B. Kelley, et al., Comparing physician and Artificial Intelligence Chatbot responses to Patient questions posted to a Public Social Media Forum, JAMA Intern. Med. 183 (6) (2023) 589–596.
[26] B.R. Sosa, M. Cung, V.J. Suhardi, K. Morse, A. Thomson, H.S. Yang, et al., Capacity for large language model chatbots to aid in orthopedic management, research, and patient queries, J. Orthop. Res. 42 (6) (2024) 1276–1282.
[27] Y. Li, J. Li, M. Li, E. Yu, D. Rhee, M. Amith, et al., VaxBot-HPV: a GPT-based chatbot for answering HPV vaccine-related questions, JAMIA Open. 8(1):ooaf005 (2025).
[28] T. Abdullahi, R. Singh, C. Eickhoff, Learning to Make Rare and complex Diagnoses with Generative AI Assistance: Qualitative Study of Popular Large Language Models, JMIR Med Educ. 10 (2024) e51391.
[29] A. Cocci, M. Pezzoli, M. Lo Re, G.I. Russo, M.G. Asmundo, M. Fode, et al., Quality of information and appropriateness of ChatGPT outputs for urology patients, Prostate Cancer Prostatic Dis. 27 (1) (2024) 103–108.
[30] I. Altamimi, A. Altamimi, A.S. Alhumimidi, A. Altamimi, M.H. Temsah, Artificial Intelligence (AI) Chatbots in Medicine: a Supplement, not a Substitute, Cureus 15 (6) (2023) e40922.
[31] K. Reynolds, D. Nadelman, J. Durgin, S. Ansah-Addo, D. Cole, R. Fayne, et al., Comparing the quality of ChatGPT- and physician-generated responses to patients' dermatology questions in the electronic medical record, Clin. Exp. Dermatol. 49 (7) (2024) 715–718.
[32] L. Wang, J. Li, B. Zhuang, S. Huang, M. Fang, C. Wang, et al., Accuracy of Large Language Models when Answering Clinical Research questions: Systematic Review and Network Meta-Analysis, J. Med. Internet Res. 27 (2025) e64486.
[33] F. Busch, L. Hoffmann, C. Rueger, E.H. van Dijk, R. Kader, E. Ortiz-Prado, et al., Current applications and challenges in large language models for patient care: a systematic review, Commun Med (lond). 5 (1) (2025) 26.
[34] F. Liu, H. Zhou, B. Gu, X. Zou, J. Huang, J. Wu, et al., Application of large language models in medicine, Nat. Rev. Bioeng. 3 (6) (2025) 445–464.
[35] X. Du, Z. Zhou, Y. Wang, Y.-W. Chuang, Y. Li, R. Yang, et al., Performance and improvement strategies for adapting generative large language models for electronic health record applications: a systematic review, Int. J. Med. Inf. 205 (2026) 106091.
[36] S. Bedi, Y. Liu, L. Orr-Ewing, D. Dash, S. Koyejo, A. Callahan, et al., Testing and Evaluation of Health Care applications of Large Language Models: a Systematic Review, JAMA 333 (4) (2025) 319–328.
[37] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, PLoS Med. 6 (7) (2009) e1000097.
[38] J. Li, A. Dada, B. Puladi, J. Kleesiek, J. Egger, ChatGPT in healthcare: a taxonomy and systematic review, Comput. Methods Programs Biomed. 245 (2024) 108013.
[39] Ouzzani M HH, Fedorowicz Z, et al. A web and mobile app for systematic reviews [Internet]. 2016 [cited 2024, 27 February]. Available from: https://www.rayyan.ai/.
[40] Radford A, Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. 2018.
[41] D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: state of the art, current trends and challenges, Multimed. Tools Appl. 82 (3) (2023) 3713–3744.
[42] J.A. Sterne, M.A. Hernán, B.C. Reeves, J. Savović, N.D. Berkman, M. Viswanathan, et al., ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions, BMJ 355 (2016) i4919.
[43] E. Jo, S. Song, J.H. Kim, S. Lim, J.H. Kim, J.J. Cha, et al., Assessing GPT-4's Performance in Delivering Medical Advice: Comparative Analysis with Human experts, JMIR Med Educ. 10 (2024) e51282.
[44] D. Kirk, E. van Eijnatten, G. Camps, Comparison of answers between ChatGPT and Human Dieticians to Common Nutrition questions, J Nutr Metab. 2023 (2023) 5548684.
[45] S. Guo, R. Li, G. Li, W. Chen, J. Huang, L. He, et al., Comparing ChatGPT's and Surgeon's responses to Thyroid-related questions from patients, J. Clin. Endocrinol. Metab. 110 (3) (2025) e841–e850.
[46] M.T. Weber, R. Noll, A. Marchl, C. Facchinello, A. Grünewaldt, C. Hügel, et al., MedBot vs RealDoc: efficacy of large language modeling in physician-patient communication for rare diseases, J. Am. Med. Inform. Assoc. 32 (5) (2025) 775–783.

[47] J.A. Carlson, R.Z. Cheng, A. Lange, N. Nagalakshmi, J. Rabets, T. Shah, et al., Accuracy and Readability of Artificial Intelligence Chatbot responses to Vasectomy-Related questions: Public Beware, Cureus 16 (8) (2024) e67996.
[48] S. Zhou, X. Luo, C. Chen, H. Jiang, C. Yang, G. Ran, et al., The performance of large language model-powered chatbots compared to oncology physicians on colorectal cancer queries, Int. J. Surg. 110 (10) (2024) 6509–6517.
[49] A. Almagazzachi, A. Mustafa, A. Eighaei Sedeh, A.E. Vazquez Gonzalez, A. Polianovskaia, M. Abood, et al., Generative Artificial Intelligence in Patient Education: ChatGPT takes on Hypertension questions, Cureus 16 (2) (2024) e53441.
[50] S. Bagnato, C. Boccagni, J. Bonavita, Assessing the Accuracy of ChatGPT in Answering questions about Prolonged Disorders of Consciousness, Brain Sci. 15 (4) (2025) 392.
[51] A.A. Alqudah, A.J. Aleshawi, M. Baker, Z. Alnajjar, I. Ayasrah, Y. Ta'ani, et al., Evaluating accuracy and reproducibility of ChatGPT responses to patient-based questions in Ophthalmology: an observational study, Medicine (Baltimore) 103 (32) (2024) e39120.
[52] A. Meyer, A. Soleman, J. Riese, T. Streichert, Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum, Clin. Chem. Lab. Med. 62 (12) (2024) 2425–2434.
[53] S.G. Kerci, B. Sahan, An Analysis of ChatGPT4 to Respond to Glaucoma-Related questions, J. Glaucoma 33 (7) (2024) 486–489.
[54] Y.-T. Xiong, H.-N. Liu, Y.-M. Zeng, Z.-Z. Zhan, W. Liu, Y.-C. Wang, et al., Exploring the capabilities of GenAI for oral cancer consultations in remote consultations, BMC Oral Health 25 (1) (2025) 269.
[55] M. Baturu, M. Solakhan, T.G. Kazaz, O. Bayrak, Frequently asked questions on erectile dysfunction: evaluating artificial intelligence answers with expert mentorship, Int J Impot Res 37 (4) (2025) 310–314.
[56] J.M. Roy, E. Atallah, K. Piper, S. Majmundar, N. Mouchtouris, D.M. Self, et al., Comparison of quality, empathy and readability of physician responses versus chatbot responses to common cerebrovascular neurosurgical questions on a social media platform, Clin. Neurol. Neurosurg. 255 (2025) 108986.
[57] S. Hack, S. Alsleibi, N. Saleh, E.E. Alon, N. Rabinovics, E. Remer, Are chatbots a reliable source for patient frequently asked questions on neck masses? Eur. Arch. Otorhinolaryngol. 282 (8) (2025) 4273–4282.
[58] J. Jesus-Ribeiro, E. Roza, B. Oliveiros, J.B. Melo, M. Carreño, Comparative assessment of artificial intelligence chatbots' performance in responding to healthcare professionals' and caregivers' questions about Dravet syndrome, Epilepsia Open. (2025).
[59] H.H. Chou, Y.H. Chen, C.T. Lin, H.T. Chang, A.C. Wu, J.L. Tsai, et al., AI-driven patient support: evaluating the effectiveness of ChatGPT-4 in addressing queries about ovarian cancer compared with healthcare professionals in gynecologic oncology, Support Care Cancer 33 (4) (2025) 337.
[60] S. Zhang, Z.Q.G. Liau, K.L.M. Tan, W.L. Chua, Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement, Knee Surg Relat Res. 36 (1) (2024) 15.
[61] R.J. Davis, O. Ayo-Ajibola, M.E. Lin, M.S. Swanson, T.N. Chambers, D.I. Kwon, et al., Evaluation of Oropharyngeal Cancer Information from Revolutionary Artificial Intelligence Chatbot, Laryngoscope 134 (5) (2024) 2252–2257.
[62] A.B. Murthy, V. Palaniappan, S. Radhakrishnan, S. Rajaa, K. Karthikeyan, A Comparative Analysis of the Performance of Large Language Models and Human Respondents in Dermatology, Indian Dermatol. Online J. 16 (2) (2025) 241–247.
[63] Z. He, B. Bhasuran, Q. Jin, S. Tian, K. Hanna, C. Shavor, 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. 26 (2024) e56655.
[64] A.A. Moore, J.R. Ellis, N. Dellavalle, M. Akerson, M. Andazola, E.G. Campbell, et al., Patient-facing chatbots: Enhancing healthcare accessibility while navigating digital literacy challenges and isolation risks-a mixed-methods study, Digit Health. 11 (2025) 20552076251337321.
[65] L.P.X. Yong, J.Y.M. Tung, Z.Y. Lee, W.S. Kuan, M.T. Chua, Performance of Large Language Models in Patient Complaint Resolution: Web-based Cross-Sectional Survey, J. Med. Internet Res. 26 (2024) e56413.