BMC Medical Informatics and Decision Making

SYSTEMATIC REVIEW

Open Access

A systematic review of large language model (LLM) evaluations in clinical medicine

Sina Shool[1], Sara Adimi[2], Reza Saboori Amleshi[1], Ehsan Bitaraf[1], Reza Golpira[2] and Mahmood Tara[1,2*]

Abstract

Background Large Language Models (LLMs), advanced AI tools based on transformer architectures, demonstrate significant potential in clinical medicine by enhancing decision support, diagnostics, and medical education. However, their integration into clinical workflows requires rigorous evaluation to ensure reliability, safety, and ethical alignment.

Objective This systematic review examines the evaluation parameters and methodologies applied to LLMs in clinical medicine, highlighting their capabilities, limitations, and application trends.

Methods A comprehensive review of the literature was conducted across PubMed, Scopus, Web of Science, IEEE Xplore, and arXiv databases, encompassing both peer-reviewed and preprint studies. Studies were screened against predefined inclusion and exclusion criteria to identify original research evaluating LLM performance in medical contexts.

Results The results reveal a growing interest in leveraging LLM tools in clinical settings, with 761 studies meeting the inclusion criteria. While general-domain LLMs, particularly ChatGPT and GPT-4, dominated evaluations (93.55%), medical-domain LLMs accounted for only 6.45%. Accuracy emerged as the most commonly assessed parameter (21.78%). Despite these advancements, the evidence base highlights certain limitations and biases across the included studies, emphasizing the need for careful interpretation and robust evaluation frameworks.

Conclusions The exponential growth in LLM research underscores their transformative potential in healthcare. However, addressing challenges such as ethical risks, evaluation variability, and underrepresentation of critical specialties will be essential. Future efforts should prioritize standardized frameworks to ensure safe, effective, and equitable LLM integration in clinical practice.

Keywords Systematic review, Large language models, LLM evaluation, Clinical medicine, Artificial intelligence in medicine, Deep learning in healthcare, Natural language processing

*Correspondence:
Mahmood Tara
smtara@gmail.com
[1]Center for Technology and Innovation in Cardiovascular Informatics, Rajaie Cardiovascular Medical and Research Center, Iran University of Medical Sciences, Tehran, Iran
[2]Rajaie Cardiovascular Medical and Research Center, Iran University of Medical Sciences, Tehran 1995614331, Iran

Background

Background Large language models (LLMs) are advanced AI systems based on transformer architectures, designed to process and generate human language by modeling the probabilistic relationships between tokens in a sequence. Unlike traditional AI models, LLMs are pre-trained on massive datasets, enabling them to learn complex linguistic patterns and adapt to diverse tasks through fine-tuning or prompting. This differentiates LLMs from broader categories like generative AI and neural networks, which may encompass non-linguistic or less context-sensitive models [1].

LLMs can be categorized into three primary types:

- **Encoder-only models (e.g., BERT, DeBERTa)**: Specializing in understanding text for tasks such as classification and sentiment analysis.
- **Decoder-only models (e.g., GPT-series, PaLM)**: Excelling in text generation and language modeling.
- **Encoder-decoder models (e.g., T5, ChatGLM)**: Designed for tasks requiring both understanding and generation, such as summarization and translation.

In healthcare, LLMs have shown potential in various applications. For instance, ChatGPT has demonstrated utility in medical education by generating differential diagnoses and answering exam-style questions, achieving performance comparable to human experts in USMLE tests. Similarly, models like MedPaLM-2 and MedPrompt have been fine-tuned for specific medical tasks, ranging from electronic health record (EHR) analysis to generating patient discharge summaries. Despite these advances, challenges such as mitigating biases, ensuring data security, and addressing ethical concerns remain critical for their broader adoption [1].

The advent of large language models (LLMs) like ChatGPT in healthcare marks a significant shift, potentially transforming medical practices across patient data management, clinical research, and direct care. As digital technologies progress, research explores LLMs' practical applications and efficacy within clinical environments. Notable studies, including those by Cascella et al., assess ChatGPT's implementation viability, revealing its broad utility from enhancing patient communications to aiding clinical decision-making [2].

LLMs promise substantial advancements by swiftly processing extensive medical literature and data, potentially revolutionizing decision support systems, personalizing interactions, and supporting complex tasks like surgical planning as Tustumi et al. discuss [3]. Such innovations aim not only for increased efficiency but also for improved diagnostic accuracy and patient management. Yet, deploying these sophisticated tools invites critical discussions on their reliability, security, and ethical use, especially given the sensitive nature of healthcare. As highlighted in *Nature Medicine*, these technologies present both significant opportunities and challenges in the medical field [4]. Furthermore, Lahat and Klang argue that LLMs can help meet rising demands for specialized medical services and enhance telehealth, crucial for addressing global health disparities [5].

The rising importance of LLMs necessitates improved evaluation frameworks and interdisciplinary efforts to enhance their clinical integration and ensure safety and effectiveness. This systematic review aims to examine the evaluations of LLMs within medical and clinical fields.

Methods

A comprehensive literature search was conducted on January 15, 2025, using databases such as PubMed, Scopus, Web of Science, arXiv, and IEEE Xplore. The search employed keywords and MeSH terms related to "evaluation," "large language models," "artificial intelligence chatbot," and "medical and clinical practice," as detailed in Appendix Table (Table S1).

Inclusion criteria

The review included original research articles assessing LLMs within medical contexts, requiring that both abstracts and full texts were accessible. No limitations were imposed regarding publication date or language.

Exclusion criteria

Non-original articles, including reviews, letters, editorials, and conference papers, were excluded, along with articles lacking abstracts, those not specifying evaluation parameters, or those focusing on non-LLM models. Multimodal Large Language Models (MLLMs), Large Vision Language Models (e.g., ChatGPT 4v, LVLM, llava), Vision-Language Processing (VLP) models, Vision models, Small Language Models, and general Language Models (only Large Language Models would be included) were also excluded.

Study selection

The initial search identified multiple records, which were deduplicated and screened for relevance. Articles failing to meet inclusion criteria were systematically excluded per PRISMA guidelines [5]. The study selection process adhered to the PRISMA guidelines, and a PRISMA flow diagram was used to illustrate the selection process.

Data extraction

The remaining articles underwent detailed data extraction, removing entries without accessible abstracts or full texts, missing DOIs, duplicates, and non-original research. The process involved answering 11 key questions, as outlined in Table (Table 1), ensuring a thorough

Table 1 Key questions for data extraction

Q1	Based on the article provided, which medical field does this article pertain to?
Q2	Is the language of the article a non-English language? (yes = 1, No = 0)
Q3	Is an LLM or GPT mentioned in the article used for educational purposes in medical/clinical field? (yes = 1, No = 0)
Q4	Is an LLM or GPT mentioned in the article used for examination and evaluating purposes in medical/clinical field? (yes = 1, No = 0)
Q5	Is the evaluation of the LLM or GPT conducted by humans or compared with humans? (yes = 1, No = 0)
Q6	What is the name of the LLM(s) or GPT(s) version evaluated in the article?
Q7	What is the targeted group of interest for the LLM or GPT mentioned in the article (e.g., doctors, nurses, students, patients)?
Q8	How are the responses of the LLM evaluated?
Q9	What is the gold standard against which the LLM's responses are compared?
Q10	What tools, scales, or set of questions are used in the evaluation, and how many questions are there?
Q11	What parameters are assessed to measure the LLM's responses?

and unbiased review of evaluation of LLM performance in healthcare contexts.

Titles and abstracts were independently screened by two reviewers to assess relevance against the inclusion and exclusion criteria. Full-text articles of potentially eligible studies were retrieved and independently evaluated by the same reviewers. Any disagreements regarding study eligibility were resolved through discussion. If consensus could not be reached, a third reviewer was consulted to adjudicate and reach a final decision.

The percentages represent the proportion of studies within each group that evaluated a specific parameter. This approach ensures a clear understanding of how widely a parameter was assessed in relation to its group context.

Human evaluation methods varied across studies, including expert raters, peer evaluations, and crowdsourcing. However, few studies reported using standardized rubrics or guidelines, which may affect reliability and consistency. This variability highlights the need for more standardized evaluation frameworks to ensure uniformity in future assessments. While this review focuses on identifying evaluation parameters, future studies could systematically categorize and analyze evaluation methods.

Results

Study selection and data extraction

A comprehensive search across PubMed, Scopus, Web of Science, arXiv, and IEEE Xplore yielded 25,156 studies, from which 2754 duplicates and 328 additional records were removed (Fig. 1). This resulted in 22,074 records being screened by title and abstract, leading to the exclusion of 20,198 for not meeting inclusion criteria. Following this, data extraction was performed on 1876 articles that passed the initial screening. Of these, 586 articles were excluded due to reasons such as inaccessible abstracts or full texts, lack of DOI, duplication, and non-original research types.

Following a detailed full-text review, an additional 529 articles were excluded. Ultimately, this rigorous and meticulous effort culminated in 761 articles from which data was fully extracted, as documented in Appendix Table (Table S2). [This appendix table represents a cornerstone of the study, containing the most comprehensive data compilation from the included articles. Due to its considerable length and detail—spanning over 100 pages—it could not be incorporated into the main manuscript but is made available in its entirety to ensure transparency and to highlight the exhaustive work underlying this research. Readers are strongly encouraged to consult Appendix Table S2 to fully appreciate the depth and scope of the extracted data.]

The evaluation of publications from 2019 to 2025 shows a notable exponential increase in research output, particularly evident from 2021 onwards. In 2019, only 1 article was published, increasing to 3 in 2020, 6 in 2021, 7 in 2022, and dramatically rising to 160 in 2023. This trend continued into 2024, with 557 articles published, followed by 27 articles in early 2025, highlighting a marked growth in research activity over this period.

Summary of LLMs evaluated

The studies evaluated a total of 1,534 instances of LLMs. Among these, the majority were general-domain LLMs, accounting for 1,435 records (93.55%). In contrast, medical-domain LLMs were assessed in 99 records, making up 6.45% of the total.

Among the 1,435 general-domain LLMs, the majority were decoder-only models, accounting for 1,340 records (93.4%). Encoder-decoder models were evaluated in 21 records (1.5%), while encoder-only models were assessed in 74 records (5.2%).

In the medical-domain LLMs (99 records), decoder-only models dominated with 79 records (79.8%). Encoder-decoder models were mentioned in 4 records (4.0%), and encoder-only models in 14 records (14.1%). For 2 records (2.0%), the architecture type was not explicitly detailed.

Among the 1,340 decoder-only general-domain LLMs, ChatGPT was the most frequently evaluated, with 242 records (18.1%), followed by ChatGPT-4 (175 records, 13.1%), GPT-4 (165 records, 12.3%), and ChatGPT-3.5 (139 records, 10.4%). Google PaLM 2/Bard/Gemini was assessed in 118 records (8.8%), while GPT-3.5 appeared

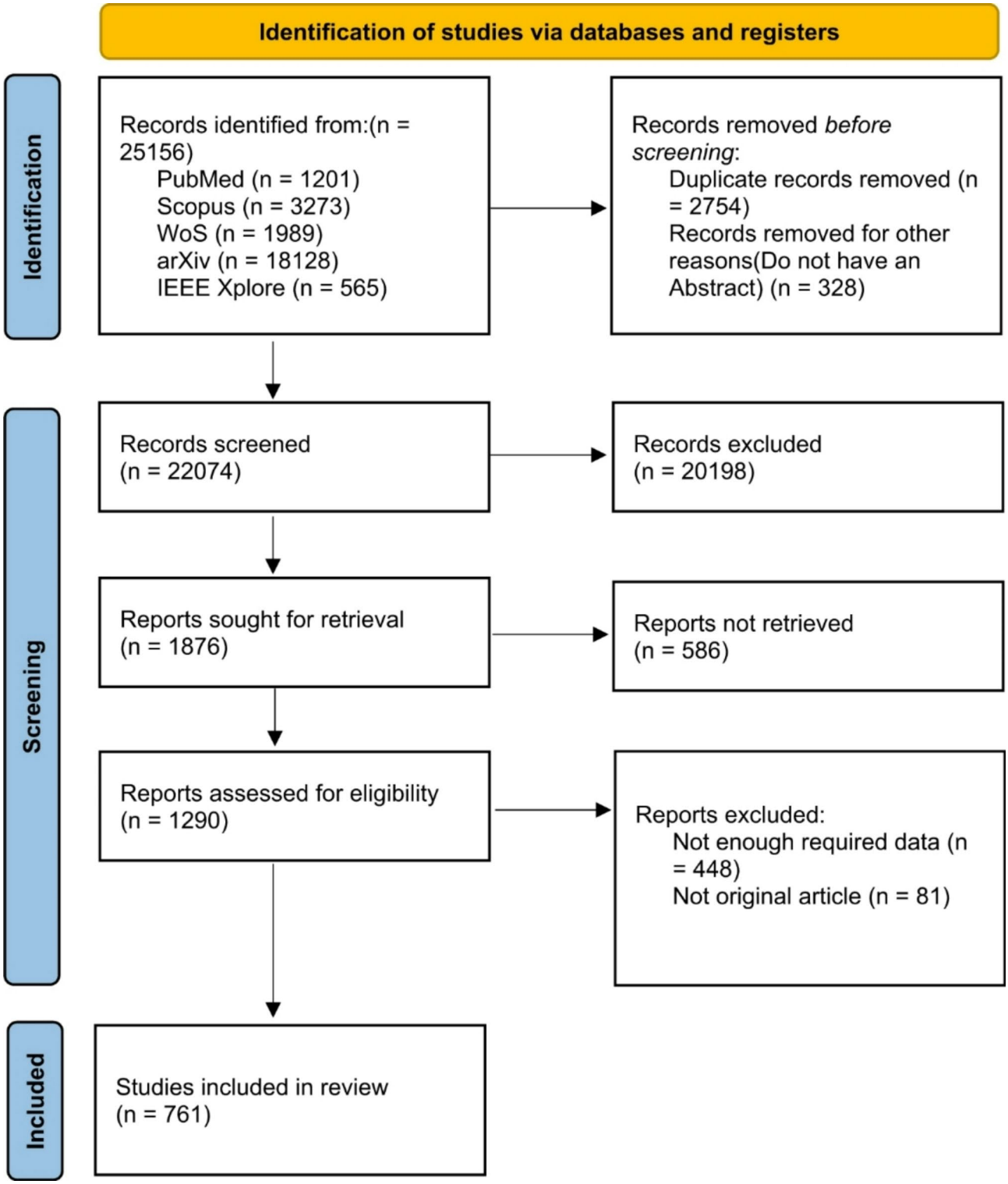

Fig. 1 PRISMA flow diagram for systematic reviews which included searches of databases

in 58 records (4.3%). Other models included Meta Llama 2 (46 records, 3.4%), Microsoft Copilot/Bing (44 records, 3.3%), Meta Llama 3 (41 records, 3.1%), and Anthropic Claude (39 records, 2.9%).

Smaller groups of models included Mistral (27 records, 2.0%), Qwen (2.5-72b) (20 records, 1.5%), GPT-3.5 Turbo (19 records, 1.4%), GPT-4o (14 records, 1.0%), ChatGPT-4o and Mixtral (each with 11 records, 0.8%), and Llama

(10 records, 0.7%). Models such as Baichuan (9 records, 0.7%), Perplexity AI (8 records, 0.6%), GPT models, Vicuna, PMC-LLaMA, and Gemma (each with 7 records, 0.5%) followed.

A variety of models were evaluated in fewer than six records, such as GPT-4 Turbo, InternLM, ChatGPT (customized), GPT-4o mini, and GPT-2, all with five records (0.4%). Models like GPT-3, ChatGPT3.5-turbo, ERNIE Bot, and ChatGPT-3 were each assessed in four records (0.3%). Numerous other models, including Yi-C, OpenAI o1-mini, ChatGPT+, and WizardLM, were evaluated in three records (0.2%).

The remaining models, including OpenAI o1-preview, Falcon, InstructGPT, and others, were assessed in two or fewer records (0.1%), with many—such as Baize-Healthcare, Alpaca, and DanteLLM_instruct_7b-v0.2-boosted—evaluated only once.

Among the 21 general-domain encoder-decoder LLMs, ChatGLM was the most frequently evaluated, appearing in 9 records (42.9%). Both Flan-T5 and GLM-4 were each evaluated in 4 records (19.0%), while BART was assessed in 3 records (14.3%). FLAN-UL2 was evaluated in 1 record (4.8%).

Among the 74 encoder-only general-domain LLMs, BERT was the most frequently evaluated, appearing in 34 records (45.9%), followed by RoBERTa in 9 records (12.2%) and BioBERT in 7 records (9.5%). SciBERT and ALBERT were each assessed in 3 records (4.1%). AfroX-LMR and M-BERT were evaluated in 2 records each (2.7%).

Several models were evaluated only once (1.4%), including AfriBERTa-large, AfroLM-active-l, Camem-BERT-with-Dates, KoBERT, ELECTRA, DNA-BERT, CH-BERT, AlphaBERT, SentenceBERT, DistilBERT, DeBERTa, ColBERT, and CliRoberta (domain-adaptive pre-trained LLM).

Among the 79 decoder-only medical-domain LLMs, the most frequently evaluated were Meditron and HuatuoGPT, each appearing in 10 records (12.7%). BioMistral followed with 6 records (7.6%), while BioGPT was evaluated in 5 records (6.3%). PULSE, MedAlpaca, and Asclepius were each assessed in 4 records (5.1%).

Other models included MMed-Llama, which was evaluated in 3 records (3.8%), and several models, including DocOA, ChatMed, BianQue, BenTsao, and BioMedLM, each assessed in 2 records (2.5%).

The remaining models, such as SenseNova, CollectiveSFT-7B, Clinical Camel (70B), GutGPT, Doctor PuJiang (Dr. PJ), ChatDoctor, AntGLM-Med-10, MedLlama2, MedGPT-7B, MedicalGPT, EyeGPT (fine-tuned version of Llama2), Drug-GPT, DermGPT, Aeyeconsult (based on GPT-4), MedLM Medium, MedPaLM, Med42 (based on Llama-2), HyperCLOVA X, Hermes7b_ITA (Nous-Hermes-llama-2-7b), EthioLLM-large, EthioLLM, ACS-GPT, and DrBode models, were each evaluated in 1 record (1.3%).

Among the 4 encoder-decoder medical-domain LLMs, all were evaluated in a single record (25.0% each). These included MOPH (a Chinese-specific ophthalmic LLM), BiomedNLP, CLINGEN (a knowledge-infused LLM model), and Clinical-T5-Large.

Among the 14 encoder-only medical-domain LLMs, GatorTron and BioClinicalBERT were the most frequently evaluated, each appearing in 3 records (21.4%). The remaining models, including MoLFormer-XL (Protein-specific LLMs), CancerBERT, MentalBERT, ClinicalBERT, BioMed-RoBERTa, and BioALBERT, were each evaluated in 1 record (7.1%).

Among the medical-domain LLMs with architecture not explicitly detailed, two models were evaluated, each appearing in 1 record (50.0%). These included LICT (Large language model-based Identifier for Cell Types) and ClinicLLM (an LLM trained on [HOSPITAL]'s clinical notes).

Major specialties evaluated

In total, the studies analyzed 781 records, providing a comprehensive overview of the distribution of medical specialties in this research. Surgery was the most frequently evaluated specialty, accounting for 220 records (28.2%). Within surgery, ophthalmology was the most common subspecialty, with 55 records (25.0%), followed by orthopedics with 44 records (20.0%), and urology and otolaryngology each with 31 records (14.1%). Plastic surgery accounted for 20 records (9.1%), while general surgery was represented in 12 records (5.5%). Less common subspecialties included obstetrics and gynecology with 6 records (2.7%), neurosurgery and bariatric surgery with 5 records each (2.3%), hand surgery with 3 records (1.4%), vascular, laparoscopic, and spine surgery each with 2 records (0.9%), and trauma and thoracic surgery each with 1 record (0.5%).

Internal medicine was the second most frequently evaluated specialty, with 119 records (15.2%). Within internal medicine, oncology was the predominant subspecialty, accounting for 56 records (47.1%), followed by endocrinology with 22 records (18.5%), gastroenterology and hepatology with 18 records (15.1%), rheumatology with 8 records (6.7%), nephrology with 6 records (5.0%), hematology with 5 records (4.2%), pulmonology with 3 records (2.5%), and general internal medicine with 1 record (0.8%).

Medical informatics was the third most commonly evaluated specialty, with 112 records (14.3%), followed by radiology with 64 records (8.2%) and general medicine with 53 records (6.8%). Medical education was assessed in 52 records (6.7%), while neurology was evaluated in 40 records (5.1%), and psychiatry in 30 records (3.8%).

Emergency medicine was represented in 21 records (2.7%), cardiology in 15 records (1.9%), dermatology in 11 records (1.4%), and pediatrics in 10 records (1.3%). Pathology accounted for 7 records (0.9%), radiation oncology for 5 records (0.6%), and infectious diseases and anesthesiology each for 4 records (0.5%). Nuclear medicine and geriatrics each accounted for 3 records (0.4%), and family medicine and sports medicine each for 2 records (0.3%). Finally, chronic diseases, patient education, physical medicine and rehabilitation (physiatry), and sleep medicine were each evaluated in 1 record (0.1%).

Target audience for LLMs evaluation

The Target audience for LLMs evaluation includes a total of 976 instances, distributed across various targeted groups of interest. Doctors constitute the largest group, with 306 instances (31.4%), followed by patients, accounting for 260 instances (26.6%). Medical and healthcare professionals and researchers make up significant portions, with 100 instances (10.2%) and 98 instances (10.0%), respectively.

Students and residents together represent 103 instances, with students contributing 72 instances (7.4%) and residents 31 instances (3.2%). Smaller groups include healthcare providers at 19 instances (1.9%), caregivers with 10 instances (1.0%), nurses and general people each with 7 instances (0.7%), and families or parents of patients contributing 13 instances (1.3%). Educators (6 instances, 0.6%) and learners (3 instances, 0.3%) are the least represented categories.

Lastly, 44 instances (4.5%) fall under the "others or not mentioned" category, representing data that does not pertain to the primary groups of interest.

Grouping and evaluation criteria for LLM studies

The studies were categorized into various groups based on specific criteria. as follows:

- **Group A-e**: Studies where the language assessed was exclusively English, as determined by the answer to Q2.
- **Group A-ne**: Studies where the languages assessed included non-English languages or languages other than English, as determined by the answer to Q2.
- **Group B-h**: Studies where evaluations were conducted directly by humans or compared with human evaluations (e.g., experts or others), as determined by the positive answer to Q5.
- **Group B-nh**: Studies where evaluations were not conducted directly by humans or were not compared with human evaluations, as determined by the negative answer to Q5.
- **Group C**: Studies where LLMs were explicitly used for educational purposes in the medical or clinical field, as determined by the positive answer to Q3.
- **Group D**: Studies where LLMs were specifically used for examination and evaluation purposes in the medical or clinical field, as determined by the positive answer to Q4.

This categorization highlights the diverse applications and evaluation contexts of LLMs in medical research, demonstrating the various ways these models are integrated and assessed within the field.

Evaluation parameters

A comprehensive analysis of evaluation parameters across the 761 studies, as summarized in Table S2 (column Q11), revealed 2,239 instances of parameter usage. After filtering for parameters that appeared in more than 1% of the total instances, 16 parameters were identified as the most frequently evaluated. These parameters, recorded verbatim from the studies, reflect the diverse approaches used to assess large language models (LLMs) in various contexts, particularly in the medical and clinical fields.

Figure 2 presents the percentage distribution of these parameters, both across the total dataset and within specific study groups, highlighting variations in focus depending on the application or evaluation criteria.

1. **Accuracy** was the most commonly assessed parameter, appearing in 419 instances, representing 21.78% of evaluations in Group A-e, 22.99% in Group A-ne, 21.64% in Group B-h, 21.84% in Group B-nh, 20.38% in Group C, and 24.31% in Group D.
2. **Consistency** was evaluated in 33 instances (2.19% in Group A-e, 2.20% in Group A-ne, 1.46% in Group B-h, 2.15% in Group B-nh, 2.23% in Group C, and 3.45% in Group D).
3. **Performance** was recorded in 34 instances, with notable percentages in Group A-ne (6.95%), Group B-h (4.68%), and Group D (5.17%), compared to 1.95% in Group A-e, 1.99% in Group B-nh, and 2.65% in Group C.
4. **Reliability** was assessed in 46 instances, with relatively higher percentages in Group B-nh (2.70%) and Group C (2.97%) compared to other groups.
5. **Clarity** was evaluated in 35 instances but remained less prominent in most groups (< 1.0% in Group B-h and Group D), with slightly higher percentages in Group B-nh (2.10%) and Group C (2.44%).
6. **Quality** was reported in 43 instances, with its highest percentage in Group C (3.82%) and modest levels in Group A-e (2.88%) and Group B-nh (2.76%).

		Total	% in Group A-e	% in Group A-ne	% in Group B-h	% in Group B-nh	% in Group C	% in Group D
1	**Accuracy**	419	21.78	22.99	21.64	21.84	20.38	24.31
2	**Consistency**	33	2.19	2.20	1.46	2.15	2.23	3.45
3	**Performance**	34	1.95	6.95	4.68	1.99	2.65	5.17
4	**Reliability**	46	2.53	1.07	1.46	2.70	2.97	1.55
5	**Clarity**	35	1.75	1.60	<1.0	2.10	2.44	1.38
6	**Quality**	43	2.88	<1.0	2.63	2.76	3.82	2.07
7	**Readability**	95	4.29	4.81	4.68	4.41	6.48	<1.0
8	**Reasoning**	13	<1.0	2.14	1.17	1.05	1.17	2.76
9	**Comprehensiveness**	47	2.24	3.20	1.46	2.59	3.29	<1.0
10	**Completeness**	49	2.34	2.14	<1.0	2.81	2.34	<1.0
11	**Correctness**	34	1.80	<1.0	<1.0	1.82	2.76	3.10
12	**Safety**	21	1.07	1.60	<1.0	1.32	1.80	<1.0
13	**Appropriateness**	24	1.41	<1.0	<1.0	1.65	1.27	<1.0
14	**Relevancy**	43	2.24	<1.0	<1.0	2.43	1.80	<1.0
15	**Sensitivity**	31	1.66	<1.0	2.05	1.49	<1.0	<1.0
16	**Specificity**	30	1.41	<1.0	1.46	1.32	<1.0	<1.0
	Each Group % of Total		**91.60**	**8.40**	**15.30**	**81.00**	**42.10**	**25.90**

Fig. 2 Distribution of evaluation parameters in total and across groups

7. **Readability** was a major focus, evaluated in 95 instances, showing significant emphasis in Group C (6.48%) and consistent usage across Groups A-e (4.29%), A-ne (4.81%), and B-h (4.68%).
8. **Reasoning** appeared in 13 instances, with notable percentages in Group A-ne (2.14%) and Group D (2.76%), while being evaluated at < 1.0% in other groups.
9. **Comprehensiveness**, assessed in 47 instances, was most emphasized in Group C (3.29%) and Group A-ne (3.20%), with smaller percentages in other groups.
10. **Completeness**, appearing in 49 instances, was consistently evaluated across most groups, with the highest percentage in Group B-nh (2.81%).
11. **Correctness** was recorded in 34 instances, with its highest emphasis in Group D (3.10%) and Group C (2.76%).

12. **Safety** was assessed in 21 instances, with small percentages across all groups, peaking at 1.80% in Group C.
13. **Appropriateness** appeared in 24 instances, with modest evaluation levels across groups, peaking at 1.65% in Group B-nh.
14. **Relevancy**, reported in 43 instances, was evaluated most prominently in Group B-nh (2.43%).
15. **Sensitivity**, assessed in 31 instances, showed a higher focus in Group B-h (2.05%) compared to other groups.
16. **Specificity** was recorded in 30 instances, with modest levels across all groups, peaking at 1.46% in Group B-h.

Discussion

Evaluation types

The assessment of large language models (LLMs) in healthcare requires advanced evaluation methodologies that prioritize context-specific metrics, safety, and accuracy, surpassing traditional benchmarks. These methodologies must also address critical concerns such as data privacy, ethical implications, and risks posed by inaccuracies or biases. Additionally, the unique demands of healthcare require LLMs to interpret and generate specialized medical content with high reliability and contextual relevance [6, 7].

This study highlights a dramatic increase in research interest in LLMs in healthcare, with publications surging from a single study in 2019 to 557 in 2024. This exponential growth underscores the expanding capabilities and clinical potential of LLMs, particularly in diagnostics, decision support, medical education, and patient communication. However, the lack of standardized evaluation tools, variability in study designs, and ethical concerns such as data privacy and hallucination risks represent key barriers to effective evaluation of LLMs in clinical settings. Addressing these issues requires interdisciplinary efforts and the development of robust frameworks tailored to clinical contexts [8–10].

Barriers

Clinical evaluations of LLMs necessitate interdisciplinary collaboration to meet the intricate demands of medical practice, requiring rigorous validation and optimization for diverse clinical applications. The growing use in healthcare underscores the urgent need for standardized evaluation frameworks to assess their performance and safety effectively [11–13]. While LLMs offer significant advancements, their rapid development raises ethical concerns, including the potential erosion of human expertise, reduced interpersonal interactions, and risks of misuse. For instance, AI-generated medical advice could diminish the role of human empathy in patient care. Ensuring responsible development and deployment through regulatory oversight is critical to mitigate these risks and balance innovation with societal well-being.

Frameworks

While no single evaluation framework has been universally adopted, several studies propose initial guidelines, emphasizing metrics such as transparency, explainability, and clinical relevance. These frameworks could serve as a foundation for future systematic evaluations. Our analysis of 761 studies provides a comprehensive overview of the evaluation parameters and applications of LLMs in healthcare. The studies focused predominantly on general-domain LLMs (93.55%), with decoder-only architectures like ChatGPT and GPT-4 being the most frequently evaluated models. Medical-domain LLMs, accounting for 6.45% of studies, demonstrated early but promising specialization, with models such as Meditron and HuatuoGPT being the most assessed. However, the limited evaluation of encoder-decoder and encoder-only models, both in general and medical domains, reveals a gap in exploring alternative architectures. Its extensive use in clinical settings underscores its versatility and superior performance in diagnostics and generating differential diagnoses, reflecting its enhanced linguistic and contextual processing capabilities [14–16].

Applications and trends

The analysis underscores the evaluation of a wide array of LLMs, with around one thirds of studies focusing on GPT models like GPT-4, and specialized or customized variants, reflecting tailored explorations for specific clinical tasks. Other models, including Google Bard, Microsoft Bing, and BERT variants, along with Claude, Llama, and PaLM2, are also reviewed, pointing to a vibrant AI research landscape in healthcare. Yet, the limited assessment of these models highlights the necessity for standardized evaluation frameworks to enable effective comparisons [17]..

The evaluation of LLMs in healthcare reveals significant variations in research focus and application, underscoring the critical need to align research priorities with clinical demands. Surgery emerged as the most frequently evaluated specialty, representing 28.2% of all studies, reflecting its prominent role in healthcare. However, critical specialties such as cardiology (1.9%) and emergency medicine (2.7%) remain significantly underrepresented despite their global importance. These findings highlight the necessity for future research to target high-burden and underserved areas to maximize the potential impact of LLMs in clinical practice.

Subspecialty analysis

The analysis of subspecialties within surgery emphasizes the dominance of ophthalmology (25.0%), orthopedics (20.0%), and urology and otolaryngology (14.1% each). Despite their importance, general surgery (5.5%) and other subspecialties, including neurosurgery and vascular surgery, were evaluated far less frequently, highlighting potential gaps in research coverage. Similarly, internal medicine—a key specialty—was the second most evaluated area (15.2%), with oncology (47.1%) leading among its subspecialties. However, other critical areas, such as nephrology (5.0%) and pulmonology (2.5%), were minimally represented, signaling the need for broader evaluations within this domain. Future research should focus on aligning LLM evaluations with the specific clinical needs of diverse medical specialties to ensure their effective and responsible integration into healthcare practice [4, 6].

Parameter evaluations

A total of 2,239 parameter evaluations were identified, with accuracy emerging as the most frequently assessed metric (419 instances, 21.78%). This reflects the critical importance of producing precise and reliable outputs in clinical settings. Other frequently evaluated parameters, such as readability (95 instances, 4.29%) and reliability (46 instances, 2.53%), emphasize the need for outputs that are both clear and dependable. Less commonly assessed parameters, including safety, bias, and appropriateness, highlight areas requiring more focused research to address potential risks and ethical challenges in clinical applications.

Our grouping framework, based on language, application purposes, and evaluation methods, revealed distinct patterns in LLM usage and assessment:

Group A-e and Group A-ne studies collectively emphasized accuracy as a key evaluation parameter, with usage rates of 21.78% and 22.99%, respectively. This reflects the critical need for precise and dependable outputs, regardless of whether the language focus was exclusively English or included non-English languages. Group A-ne studies, which included non-English languages, showed a higher focus on performance (6.95%) and comprehensiveness (3.20%), reflecting the challenges of evaluating multilingual capabilities.

Group B-h, involving direct human evaluations, emphasized accuracy (21.64%) and correctness (1.80%), highlighting the role of expert validation in ensuring the clinical utility of LLM outputs.

Group B-nh, which relied on automated or indirect evaluations, focused on metrics like completeness (2.81%) and quality (2.76%), reflecting the need for reliable outputs in contexts without human oversight.

Group C, addressing educational applications, placed significant emphasis on readability (6.48%) and comprehensiveness (3.29%), crucial for effective knowledge dissemination.

Group D, targeting examination and evaluation purposes, highlighted accuracy (24.31%) and correctness (3.10%) as key metrics, underscoring the importance of dependable outputs in high-stakes contexts.

These group-specific analyses provide valuable insights into how LLMs are assessed across diverse research contexts, reflecting the tailored objectives and priorities of each group. Notably, the limited focus on ethical parameters like safety and bias across all groups highlights a critical gap that must be addressed to ensure equitable and responsible LLM integration.

Limitations

The evaluation of LLMs in healthcare highlights their varied applications and categorization by language and methods. Nonetheless, several issues persist, including an excessive focus on accuracy, which does not adequately capture the complexity of model performance in clinical settings. Essential factors like safety, fairness, and bias are often neglected, and many studies rely on closed-ended tasks, failing to mirror the complexity of clinical decision-making which requires comprehensive, open-ended reasoning [8, 18].

Additionally, the categorization of studies by language and purpose reveals varied applications of LLMs and underscores a lack of standardized evaluation practices. This fragmentation impedes a unified understanding of LLMs' capabilities across medical fields. Moreover, the application of LLMs in clinical settings faces challenges, as the lack of domain-specific training data can cause inaccuracies, especially in precise fields like radiology or genetics [6, 19].

A notable limitation of this study is the restriction of the search strategy to titles of published studies. This approach, while providing a focused scope, may have excluded relevant studies identifiable through abstracts. Future systematic reviews in this domain should consider expanding the search strategy to include both titles and abstracts to ensure a more comprehensive capture of eligible studies.

The lack of standardized definitions for some evaluation parameters and the variability in human evaluation practices are recognized as limitations of this review. While this work identifies trends in parameter usage and grouping criteria, future research should explore standardizing these definitions and frameworks within specific medical specialties or model types. Some findings, such as the distribution of clinical specialties and target audiences, provide contextual insights but are not directly aligned with the primary focus on evaluation methods. Future reviews could streamline the analysis to align more closely with evaluation frameworks.

Future directions and perspectives

To fully realize the potential of large language models (LLMs) in healthcare, future efforts should prioritize several key areas. First, enhancing interpretability is critical to developing transparent models that clinicians can trust for reliable decision support. Establishing robust validation frameworks tailored to the complexities of clinical settings is equally essential to ensure the accuracy and applicability of LLM outputs. Ethical considerations, such as safeguarding data privacy, mitigating bias, and addressing the societal impacts of automation, must also be a primary focus. Additionally, the evolving roles of healthcare professionals require exploration, as these technologies may shift their responsibilities from decision-makers to supervisors of AI-generated insights. To mitigate risks associated with misuse or overreliance on LLMs, the development of comprehensive governance frameworks is imperative, ensuring their deployment aligns with ethical and safety standards. Finally, addressing barriers to adoption, including resource constraints and resistance to change, will require interdisciplinary collaboration and targeted education efforts to foster acceptance and successful integration into healthcare practices.

Conclusions

This systematic review underscores the expanding role of LLMs in clinical medicine, highlighting their potential to revolutionize medical diagnostics, education, and patient care. While their applications are diverse, critical challenges remain, including the need for standardized evaluation frameworks, attention to ethical considerations, and the underrepresentation of high-priority medical specialties. Addressing these challenges through interdisciplinary collaboration and robust governance will be essential for the responsible deployment of LLMs. Future research should focus on enhancing model interpretability, tailoring evaluations to clinical complexities, and addressing disparities in specialty-specific applications. By aligning technological advancements with clinical needs, LLMs can drive significant improvements in healthcare outcomes.

Abbreviations

LLM Large Language Model

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02954-4.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

SS, EB, and MT conceptualized the study, developed the methodology, and conducted the initial literature review. SS and RSA were responsible for data extraction, analysis, and synthesis of the findings. EB, SA, and RG contributed to the interpretation of the results and provided critical revisions to the manuscript. MT supervised the project, provided expert guidance on the clinical applications of LLMs, and contributed to the final review of the manuscript. All authors read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 30 September 2024 / Accepted: 26 February 2025
Published online: 07 March 2025

References

1. Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J et al. A survey of large language models in medicine: progress, application, and challenge. ArXiv Preprint. 2023;arXiv:231105112.
2. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33.
3. Tustumi F, Andreollo NA, Aguilar-Nascimento, JEd. Future of the language models in healthcare: the role of chatGPT. ABCD arquivos brasileiros de cirurgia digestiva (são paulo). 2023;36:e1727.
4. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res. 2023;25:e49324.
5. Lahat A, Klang E. Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services? J Telemed Telecare. 2024;30(9).
6. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models in medical applications: a survey. ArXiv Preprint. 2024;arXiv:240507468.
7. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus. 2023;15(5):e39305.
8. Nazi ZA, Peng W. Large language models in healthcare and medical domain: A review. ArXiv Preprint. 2023;arXiv:240106775.
9. Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. Front Med. 2024;11:1380148.
10. Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J et al. A survey of large language models in medicine: principles, applications, and challenges. arXiv preprint. 2023;arXiv:231105112.
11. Busch F, Hoffmann L, Rueger C, van Dijk EHC, Kader R, Ortiz-Prado E et al. Systematic review of large language models for patient care: current applications and challenges. medRxiv. 2024:2024.03.04.24303733.
12. Park Y-J, Pillai A, Deng J, Guo E, Gupta M, Paget M, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inf Decis Mak. 2024;24(1):72.
13. Perlis RH, Fihn SD. Evaluating the application of large language models in clinical research contexts. JAMA Netw Open. 2023;6(10):e2335924–e.

14. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. J Med Internet Res. 2024;26:e56110.
15. Mackey BP, Garabet R, Maule L, Tadesse A, Cross J, Weingarten M. Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. Discover Artif Intell. 2024;4(1):38.
16. Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. BMC Digit Health. 2024;2(1):4.
17. Alessandri-Bonetti MGR, Naegeli M, Liu HY, Egro FM. Assessing the soft tissue infection expertise of ChatGPT and Bard compared to IDSA recommendations. Ann Biomed Eng. 2023.
18. Liu F, Zhou H, Hua Y, Rohanian O, Clifton L, Clifton DA. Large language models in healthcare: A comprehensive benchmark. MedRxiv. 2024:2024.04.24.24306315.
19. García-Méndez S, de Arriba-Pérez F. Large language models and healthcare alliance: potential and challenges of two representative use cases. Ann Biomed Eng. 2024;52(8):1928–31.

Publisher's note