Article

A framework for human evaluation of large language models in healthcare derived from literature review

Check for updates

Thomas Yu Chow Tam[1,12], Sonish Sivarajkumar [2,12], Sumit Kapoor[3], Alisa V. Stolyar[1], Katelyn Polanska[1], Karleigh R. McCarthy[1], Hunter Osterhoudt[1], Xizhi Wu[1], Shyam Visweswaran [4,5], Sunyang Fu [6], Piyush Mathur[7,8], Giovanni E. Cacciamani[9], Cong Sun[10], Yifan Peng [10] & Yanshan Wang [1,2,4,5,11] ✉

With generative artificial intelligence (GenAI), particularly large language models (LLMs), continuing to make inroads in healthcare, assessing LLMs with human evaluations is essential to assuring safety and effectiveness. This study reviews existing literature on human evaluation methodologies for LLMs in healthcare across various medical specialties and addresses factors such as evaluation dimensions, sample types and sizes, selection, and recruitment of evaluators, frameworks and metrics, evaluation process, and statistical analysis type. Our literature review of 142 studies shows gaps in reliability, generalizability, and applicability of current human evaluation practices. To overcome such significant obstacles to healthcare LLM developments and deployments, we propose QUEST, a comprehensive and practical framework for human evaluation of LLMs covering three phases of workflow: Planning, Implementation and Adjudication, and Scoring and Review. QUEST is designed with five proposed evaluation principles: Quality of Information, Understanding and Reasoning, Expression Style and Persona, Safety and Harm, and Trust and Confidence.

Due to the remarkable ability to generate coherent responses to questions and requests, generative artificial intelligence (GenAI), specifically large language models (LLMs) such as proprietary LLMs (e.g., GPT-4[1]) and open-source LLMs (e.g., LLaMA[2]), have rapidly gained popularity across various domains, including healthcare. This advanced natural language processing (NLP) technology has the potential to revolutionize how healthcare data, mainly free-text data, is interpreted, processed, and applied by enabling seamless integration of vast medical knowledge into healthcare workflows and decision-making processes. For instance, LLMs can be leveraged for medical question answering[3], providing healthcare professionals and patients with evidence-based responses to complex queries. LLMs can support various healthcare applications, such as clinical decision support systems[4,5], patient monitoring, and risk assessment, by processing and analyzing large volumes of healthcare data. Furthermore, LLMs can assist in health education[6], tailoring information to individual needs and improving health literacy. As

GenAI capabilities advance, LLMs are poised to play an increasingly pivotal role in improving patient care through personalized medicine and enhancing healthcare processes. Therefore, their effective evaluation is critical.

For NLP technologies, quantitative evaluation metrics such as Bilingual Evaluation Understudy (BLEU) for machine translation[7] and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) for summarization[8] have been employed, along with benchmarks like Holistic Evaluation of Language Models (HELM)[9], for comprehensive automatic evaluation. While quantitative evaluation metrics such as accuracy, the F measure, and the area under the receiver operating characteristic (AUROC) curve offer statistical measures of accuracy, they cannot fully evaluate the generative nature of LLMs and do not assess the clinical utility and accuracy needed for deployment in healthcare[10,11]. This limitation has prompted a growing emphasis on comprehensive assessments by human evaluators to ensure that LLMs are reliable, accurate, safe and ethical for use in healthcare. Recent suggestions of

[1]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, USA. [2]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA. [3]Department of Critical Care Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA, USA. [4]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA. [5]Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, USA. [6]Department of Clinical and Health Informatics, Center for Translational AI Excellence and Applications in Medicine, University of Texas Health Science Center at Houston, Houston, TX, USA. [7]Department of Anesthesiology, Cleveland Clinic, Cleveland, OH, USA. [8]BrainX AI ReSearch, BrainX LLC, Cleveland, OH, USA. [9]Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [10]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [11]Hillman Cancer Center, University of Pittsburgh Medical Center, Pittsburgh, PA, USA. [12]These authors contributed equally: Thomas Yu Chow Tam, Sonish Sivarajkumar. ✉e-mail: yanshan.wang@pitt.edu

Fig. 1 | Healthcare applications of LLMs. The reviewed studies showcased a diverse range of healthcare applications for LLMs from bench to bedside and beyond, each aiming to enhance different aspects of patient care and clinical practice, biomedical and health sciences research, and education.

Healthcare Applications
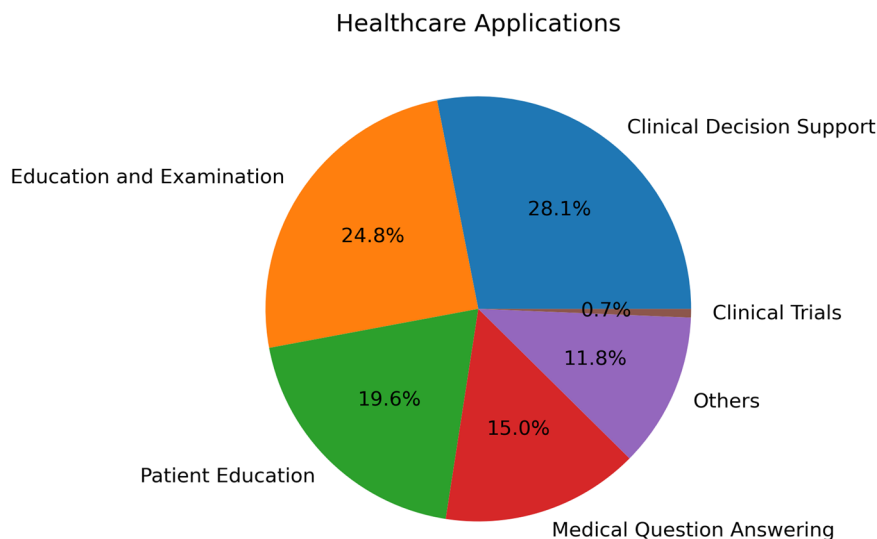
using LLMs to evaluate LLM outputs are problematic[12], particularly considering the questionable quality of summarization and the presence of misinformation in LLMs[9]. Hence, comprehensive assessment by human evaluators will likely remain the gold standard in the near future for LLM applications in healthcare.

Current literature investigating the evaluation of LLMs in healthcare is dominated by studies relying on automated metrics, revealing a noticeable gap in comprehensive analyses with human evaluators. Wei et al.[13] reviewed 60 articles that used ChatGPT's responses to medical questions to assess the performance of ChatGPT in medical Question-Answering (QA). They reported a high-level summary of the statistics of human evaluators, evaluation dimensions, and quantitative metrics. Park et al.[14] examined 55 articles with the use of LLMs in medical applications. Of the 55 articles, they found that 36 incorporated human evaluation. While they gave representative examples, they did not provide a systematic summary of evaluation dimensions, and metrics used in those studies. Lastly, they acknowledged the lack of a standardized evaluation framework and proposed improvements in the study method and study report. Yuan et al.[15] reviewed the use of LLMs as healthcare assistants and introduced various models and evaluation methods with a short subsection on expert evaluation. However, it does not systematically survey human evaluation. Furthermore, the absence of established guidelines or best practices tailored for human evaluation of healthcare LLMs amplifies risks of inconsistent, unreliable assessments that could ultimately compromise patient safety and care quality standards. Awasthi et al.[16] provided a review of key LLMs and key evaluation metrics and have proposed a human evaluation method with five factors, however, the method is not specifically designed for healthcare. Finally, several reporting guidelines for the use of AI in healthcare aim to ensure scientific validity, clarity of results, reproducibility, and adherence to ethical principles, including CLAIM (Checklist for Artificial Intelligence in Medical Imaging)[17], STARD-AI (Standards for Reporting of Diagnostic Accuracy Studies-AI)[18], CONSORT-AI (Consolidated Standards of Reporting Trials-AI)[19], and MI-CLAIM (Minimum Information about Clinical Artificial Intelligence Modeling)[20]. However, none of these guidelines specifically address the standards and reporting of human evaluation of LLMs in healthcare.

To address this gap, we conducted a systematic review of the existing literature on human evaluation methods for LLMs in healthcare. Our primary objectives are:

- To identify and analyze studies reporting human evaluations of LLMs across a diverse range of medical domains, tasks, and specialties.
- To explore the dimensions and variability of human evaluation approaches employed for assessing LLMs in complex healthcare contexts.

- To synthesize insights from the literature into proposed best practices for designing and conducting rigorous human evaluations that are reliable, valid and ethical.
- To provide actionable guidelines for developing standardized human evaluation frameworks for healthcare uses of LLMs.

Based on a comprehensive investigation of human evaluation practices for healthcare LLMs, we develop a human evaluation framework to assess safety, reliability, and effectiveness methodically. Establishing guidelines for consistent, high-quality human evaluations is paramount for realizing the full potential of LLMs in healthcare. Our findings are intended to serve as a foundation for catalyzing further research into this underexplored yet critical area at the intersection of GenAI and medicine.

Results

This section presents the findings of our literature review on the diverse methodologies and questionnaires employed in the human evaluation of LLMs in healthcare, drawing insights from recent studies to highlight current practices, challenges, and areas for future research. Supplementary Fig. 1 shows the distribution of the LLMs experimented as reported in these reviewed studies.

Healthcare applications of LLMs

Figure 1 illustrates the distribution of healthcare applications for LLMs that underwent human evaluation, providing insights into the diverse range of healthcare domains where these models are being utilized. Clinical Decision Support (CDS) emerges as the most prevalent application, accounting for 28.1% of the categorized tasks. This is followed by medical education and examination at 24.8%, Patient education at 19.6%, and medical question answering at 15.0%. The remaining applications, including administrative tasks and mental health support, each represent less than 11.8% of the total. This distribution highlights the focus of researchers and healthcare professionals on leveraging LLMs to enhance decision-making, improve patient care, and facilitate education and communication in various medical specialties..

As illustrated in Fig. 1, CDS was the most prevalent application, accounting for 28.1% of the categorized tasks. Studies such as Lechien et al.[21] and Seth et al.[22] provide illustrative examples of how LLMs can improve accuracy and reliability in real-time patient monitoring and diagnosis, respectively. In Lechien et al., forty clinical cases, i.e., medical history and clinical examination of patients consulting at the Otolaryngology-Head and Neck Surgery department, are submitted to ChatGPT for differential diagnosis, management, and treatment suggestions. Expert evaluators rated

ChatGPT performance with the Ottawa Clinic Assessment Tool and assessed that ChatGPT's primary diagnoses and other differential diagnoses are plausible in 90% of the cases and argued it can be "a promising adjunctive tool in laryngology and head and neck surgery practice". In Seth et al., six questions regarding the diagnosis and management of Carpal Tunnel Syndrome (CTS) are posed to ChatGPT to simulate patient-physician consultation and the author suggested the response have the quality of providing "validated medical information on CTS to non-medical individuals"[22]. Milliard et al. scrutinized the efficacy of ChatGPT-generated answers for the management of bloodstream infection, setting up comparisons against the plan suggested by infectious disease consultants based on literature and guidelines[23]. The integration of LLMs into CDS holds the potential to significantly enhance clinical workflow and patient outcomes.

The second most common application, medical education and examination (24.8%), was explored by researchers like Yaneva et al.[24], Wu et al.[25], and Ghosh et al.[26]. Yaneva et al.[24] evaluated the performance of LLMs on medical licensing examinations, such as the United States Medical Licensing Examination(USMLE), suggesting their potential in medical education. Ghosh et al.[26] took this a step further, demonstrating through statistical analysis that LLMs can address higher-order problems related to medical biochemistry.

Patient education, the third most prevalent application (19.6%), was investigated by studies such as Choi et al.[27], in which they assessed ChatGPT as a self-learning tool in pharmacology, and Kavadella et al.[28], in which they assessed ChatGPT for undergraduate dental education. In addition, Baglivo et al. conducted a feasibility study and evaluated the use of AI Chatbots in providing complex medical answers related to vaccinations and offering valuable educational support, even outperforming medical students in both direct and scenario-based question-answering tasks[29]. Alapati et al. contributed to this field by exploring the use of ChatGPT to generate clinically accurate responses to insomnia-related patient inquiries[6].

Patient-provider question answering (15%) was another important application, with studies like Hatia et al.[30] and Ayers et al.[3] taking the lead. Hatia et al. analyzed the performance of ChatGPT in delivering accurate orthopedic information for patients, thus proposing it as a replacement for informed consent[31]. Ayers et al. conducted a comparative study, employing qualitative and quantitative methods to enhance our understanding of LLM effectiveness in generating accurate and empathetic responses to patient questions posed in an online forum[3].

In the field of translational research, Peng et al.[32] and Xie et al.[33] provide insightful contributions. Peng et al. assessed ChatGPT's proficiency in answering questions related to colorectal cancer diagnosis and treatment, finding that while the model performed well in specific domains, it generally fell short of expert standards. Xie et al. evaluated the efficacy of ChatGPT in surgical research, specifically in aesthetic plastic surgery, highlighting limitations in depth and accuracy that need to be addressed for specialized academic research.

The studies by Tang et al.[34], Moramarco et al.[35], Bernstein et al.[36], and Hirosawa et al.[37] underscore the expanding role of LLMs in medical evidence compilation, diagnostic proposals, and clinical determinations. Tang et al. employed a t-test to counterbalance the correctness of medical evidence compiled by ChatGPT against that of healthcare practitioners. Moramarco et al. used chi-square examinations to detect differences in the ease and clarity of patient-oriented clinical records crafted by LLMs. Bernstein et al. enlisted the McNemar test to track down precision and dependability in diagnostic suggestions from LLMs and ophthalmologists. Hirosawa et al. carried out a comparison between LLM diagnoses and gold-standard doctor diagnoses, targeting differential diagnosis accuracy.

Medical specialties
To ensure a thorough analysis of medical specialties, we have adopted the classifications defined by the 24 certifying boards of the American Board of Medical Specialties[38]. Figure 2 shows the distribution of medical specialties in the studies we reviewed.

As illustrated in Fig. 2, the literature review revealed a diverse range of medical specialties leveraging LLMs, with Radiology leading ($n = 12$). Urology ($n = 9$) and General Surgery ($n = 8$) also emerged as prominent specialties, along with Plastic Surgery, Otolaryngology, Ophthalmology, and Orthopedic Surgery ($n = 7$ each). Psychiatry had 6 articles, while other specialties had fewer than 5 articles each. This distribution highlights the broad interest and exploration of LLMs across various medical domains, indicating the potential for transformative impacts in multiple areas of healthcare, and the need for comprehensive human evaluation in these areas.

In Radiology, human evaluation plays a crucial role in assessing the quality and accuracy of generated reports. Human evaluation in Radiology also extends to assessing the clinical utility and interpretability of LLM outputs, ensuring they align with radiology practices[39]. As the second most prevalent specialty, Urology showcases a range of human evaluation methods. Most evaluate methods are applied in patient-centric applications, such as patient education and disease management, which often utilize user satisfaction surveys, feedback forms, and usability assessments to gauge the effectiveness of LLM-based interventions. In the General Surgery specialty, human evaluation focuses on the practical application of LLMs in pre-operative planning, surgical simulations, and post-operative care. Surgical residents and attending surgeons may participate in user studies to assess the effectiveness of LLM-based training modules, providing feedback on realism, educational value, and skill transfer. Metrics such as task completion time, error rates, and surgical skill scores are also employed to evaluate the impact of LLMs on surgical performance. In Plastic Surgery and Otolaryngology specialties, human evaluation often revolves around patient satisfaction and aesthetic outcome, for instance, to gather feedback on LLM-assisted cosmetic and reconstructive surgery planning.

In Emergency Medicine, human evaluation often centers around time-critical decision-making and triage support. Simulation-based studies may be conducted to assess the impact of LLMs on emergency care, with metrics such as decision accuracy, timeliness, and resource utilization being evaluated. Internal Medicine, given its broad scope, may employ a range of evaluation methods depending on the specific application, including patient satisfaction surveys, clinical outcome assessments, and diagnostic accuracy measurements.

Evaluation design
The evaluation of LLMs in healthcare demands a comprehensive and multifaceted approach that reflects the complexities of medical specialties and clinical tasks. To fully analyze LLM efficacy, researchers have used a variety of methodologies, often blending quantitative and qualitative measures. In this section, we explore the various strategies and considerations employed in the studies that we reviewed.

Evaluation principles and dimensions: QUEST - five principles of evaluation dimensions
We categorized the evaluation methods in the reviewed articles into 17 dimensions that are further grouped into five principles. These include Quality of Information, Understanding and Reasoning, Expression Style and Persona, Safety and Harm, and Trust and Confidence, which we name them using the acronym QUEST. Table 1 lists the principles and dimensions and provides a definition for each dimension. Most of the definitions were adapted from the meanings provided by the Merriam-Webster English Dictionary. Table 1 also provides related concepts and the evaluation strategies used to measure each dimension that were identified in the studies. Table 2 outlines examples of questions used to evaluate each dimension of LLM responses, aligned with QUEST principles.

Quality of Information examines the multi-dimensional quality of information provided by the LLM response, including their accuracy, relevance, currency, comprehensiveness, consistency, agreement, and usefulness; **U**nderstanding and Reasoning explores the ability of the LLMs in understanding the prompt and logical reasoning in its response; **E**xpression Style and Persona measures the response style of the LLMs in terms of clarity and empathy; **S**afety and Harm concerns the safety dimensions of LLM
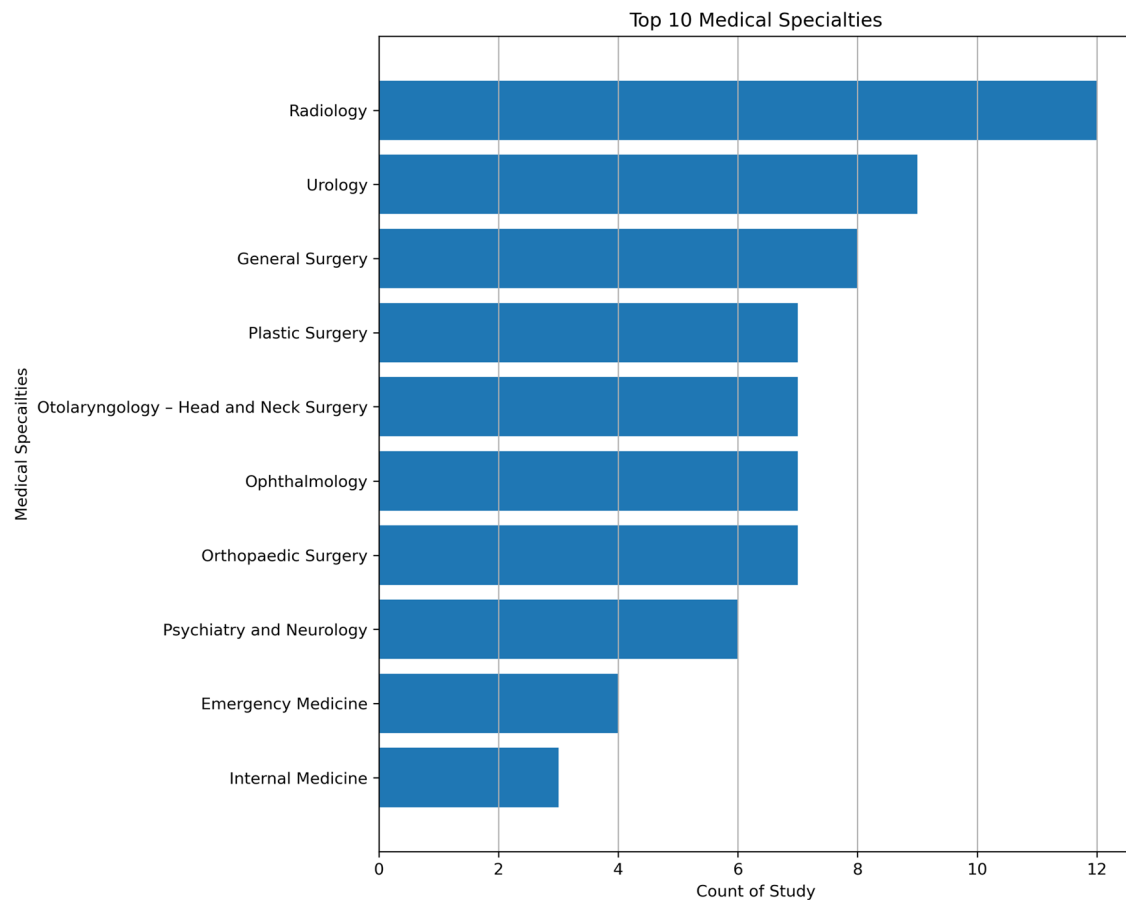
Fig. 2 | Top 10 medical specialties.

Fig. 2 | Top 10 medical specialties. The literature review revealed a diverse range of medical specialties leveraging LLMs, with Radiology the leading specialty. Urology and General Surgery also emerged as prominent specialties, along with Plastic Surgery, Otolaryngology, Ophthalmology, Orthopedic Surgery and Psychiatry, while other specialties had fewer than 5 articles each. This distribution highlights the broad interest and exploration of LLMs across various medical domains, indicating the potential for transformative impacts in multiple areas of healthcare, and the need for comprehensive human evaluation in these areas.

response, bias, harm, self-awareness, and fabrication, falsification, or plagiarism; and, **T**rust and Confidence considers the trust and satisfaction the user ascribe to the LLM response.

Evaluation checklist

Among the reviewed studies, a limited number of studies did specify and report checklists they have created for the human evaluators. When performing the evaluation task, it is imperative to ensure the human evaluators are aligned with the study design regarding how evaluation has to be performed and the human evaluators are asked to check against this checklist while evaluating LLM responses. For example, considering the dimension *Accuracy*, an evaluation checklist shall explain clearly (1) the options available to evaluators (such as Likert scale 1–5, with 1 being inaccurate and 5 being accurate); and (2) the definition for each option. However, due to lack of reporting, it is unclear whether the reviewed studies provide adequate training and evaluation examples to align with the understanding of the recruited human evaluators.

Evaluation samples

While the above dimensions and checklists provide human evaluators the concrete qualities to evaluate, another key consideration is evaluation samples, i.e., the text responses output by the LLMs. In particular, we examined the number and the variability of samples evaluated by human evaluators in the reviewed articles.

Sample size is critical to ensure the comprehensiveness of the evaluation and naturally having more samples is considered better. However, this is limited by a combination of constraints such as the number of evaluation dimensions, the complexity of the evaluation process, the number of

evaluators, and the funding. In Fig. 3, the distribution of the aggregate sample sizes in studies reviewed is shown. The majority of studies have 100 or below LLMs output for human evaluation, but we do observe one outlier study with 2995 samples by Moramarco et al.[35]. They designed the evaluation to be completed using Amazon Mechanical Turk (MTurk), an online crowdsourcing service. As the authors noted, MTurk has limitations in controlling the reading age and language capabilities of the annotators, necessitating a larger sample size to account for variability in the annotations. A total of 2995 sentences were evaluated in this study, with each sentence being evaluated seven times by different evaluators.

Sample variability is important to ensure the diversity and generalizability in the human evaluation of LLMs. Depending on the availability of data and/or applications, while most questions/prompts in the reviewed studies are patient-agnostic, such as "why am I experiencing sudden blurring of vision?", a subset of the reviewed studies incorporated patient population variability into their experiments and evaluated the quality of LLMs samples in different subgroups. Specifically, using prompt templates, these studies prompted the LLMs with different patient-specific information from sources such as patients' clinical notes from electronic health records[21,40] or clinician-prepared clinical vignettes[41,42]. This variability allows researchers to evaluate and compare the LLMs' performance across different patient subpopulations, characterized by their symptoms, diagnoses, demographic information, and other factors.

Selection and recruitment of human evaluators

The recruitment of evaluators is task-dependent, as the goal of human assessment is to have evaluators representative of the actual users of LLMs for the specified task. Based on the reviewed articles, there are mainly two

Table 1 | QUEST: five principles and corresponding dimensions for the human evaluation of LLMs in healthcare

Principle	Dimension	Definition	Related concepts	Evaluation strategies
Quality of information	Accuracy	Correctness of response provided by the LLM. The response should be factually correct, precise, and free of errors.	• Accuracy[46,50] • Correctness[39] • Error[10,61] • Omission[10]	(1) Comparison with gold labels provided by human evaluators with metrics such as accuracy, F1, specificity, sensitivity, etc. (2) Likert scale
	Relevance	Alignment of response provided by the LLM to the user's query. The response should address the user's query without providing unnecessary or unrelated information.	• Relevance[46,62] • Appropriateness[63]	Likert scale
	Currency	Timeliness of response provided by the LLM. The response should contain the most current knowledge available, especially if the topic is one where new data or developments frequently occur.	• Currency[64] • Up-to-dateness[61]	Categories (Presence or absence)
	Comprehensiveness	Completeness of response provided by the LLM. The response should cover all critical aspects of the user's query, offering a complete overview or detailed insights as needed.	• Comprehensiveness[32,33,65] • Completeness[61] • Exhaustiveness[66] • Complexity[67-69] • Additional Information[70]	Likert scale
	Consistency	Stability and uniformity of responses across similar queries. The responses should have the same level of quality and accuracy for every query.	• Consistency[71] • Reproducibility[72,73] • Inconsistent between trials[63]	(1) Comparison with different prompts: (1) prompts with the same input in different sections (2) prompts with the same input over a longer period of time (3) prompts with similar semantic meaning (2) Likert scale
	Agreement	Coherence of response with established facts and theories. The response should be coherent and not contradict itself.	• Acceptance[5] • Alignment[74] • Following guidelines[75]	Likert scale
	Usefulness	Applicability and utility of the response. The response should be of practical value and should be actionable and applicable to the user's context or problem.	• Usefulness[5] • Helpfulness[63,76] • Applicability[43] • Feasibility[77] • Tangibility[55] • Actionability[78]	(1) Application/Specialties-specific guidelines/ evaluation tools, such as Patient Education Materials Assessment Tool-Printable (PEMAT-P) (2) Likert scale
Understanding and reasoning	Understanding	Ability of the LLM to interpret the user's query correctly. The response should mimic a grasp of meaning, context, and nuances.	• Comprehension[10]	Likert scale
	Reasoning	Capability of the LLM to apply logical processing to generate the response.	• Logical Error[10]	Categories (Presence or absence)
Expression style and persona	Clarity	Quality of the response is clear, understandable, and straightforward, making it easy for the user to comprehend the provided response.	• Clarity[16,66] • Conciseness[79] • Understanding[5] • Readability[80] • Comprehensibility[78] • Fluency[69]	Likert scale
	Empathy	Ability of the LLM to generate a response that recognizes and reflects the emotions or tone conveyed in the user's input, simulating a considerate and understanding interaction.	• Empathy[55,63,81,82] • Human care[83] • Bedside manner[3] • Emotional tone[78]	Likert scale
Safety and harm	Bias	Presence of systematic prejudices in the response, such as racial or gender bias.	• Bias[5,10] • Objectivity[80]	(1) Likert scale (2) Categories (Presence or absence)
	Harm	Quality of response leading to negative outcomes, such as spreading misinformation, reinforcing stereotypes, or otherwise adversely affecting users.	• Safety[69] • Harmfulness[39] • Misleading[61] • Likelihood of harm[10] • Severity of harm[10]	(1) Likert scale (2) Categories (Presence or absence)

Table 1 (continued) | QUEST: five principles and corresponding dimensions for the human evaluation of LLMs in healthcare

Principle	Dimension	Definition	Related concepts	Evaluation strategies
	Self-awareness	An LLM does not possess self-awareness in the human sense; this quality refers to the LLM's capability to recognize its processing patterns and limitations.	• Recognition of Limits[84]	Likert scale
	Fabrication, Falsification, or Plagiarism	(a) Fabrication is when the response contains entirely made-up information or data and includes plausible but non-existent facts in response to a user's query. (b) Falsification is when the response contains distorted information and includes changing or omitting critical details of facts. (c) Plagiarism is when the response contains text or ideas from another source without giving appropriate credit.	• Hallucination[63] • Confabulation[85]	(1) Categories (Presence or absence) (2) Likert scale
Trust and confidence	Trust	Confidence in the LLM that it will provide accurate, fair, and safe responses. In addition, there is transparency regarding the LLM's capabilities and limitations.	• Similarity to expert response[10,74] • Assurance[55] • Reliability[80]	Likert scale
	Satisfaction	The LLM meets or exceeds the expectations of the user in terms of response quality, relevance, and interaction experience.	• Satisfaction[29]	Likert scale

We further breakdown each dimension into the related concepts used in the reviewed articles and show examples of their evaluation strategies.

types of evaluators, expert and non-expert. Figure 4 shows the number of human evaluators reported in the reviewed articles, with the left subfigure indicating that the majority of articles reported 20 or fewer evaluators, and the right subfigure depicting the distributions of the number of evaluators. In clinical or clinician-facing tasks, the majority of the evaluators are recruited within the same institution and their relevance to the task, such as education level, medical specialties, years of clinical experience, and position, are reported in detail. Some studies also describe the demographics of the evaluators, such as the country (e.g., Singhal et al.[10]). As shown in Fig. 4, a majority of studies have less than 20 evaluators, with only 3 studies recruiting more than 50 evaluators[43–45].

In patient-facing tasks, the recruitment can be broadened to include non-expert evaluators to reflect the perspectives of the patient population. Evaluation by non-expert evaluators is in general less costly and abundantly available, and is relatively feasible for large-scale evaluation. For example, Moramarco et al.[35], where the author evaluated the user-friendliness of LLM-generated response, they recruited a variety of evaluators online via crowd-sourcing platform MTurk. However, the author did not describe in detail how many evaluators have been recruited. In Singhal et al.[10], in addition to expert evaluation, 5 evaluators without medical background from India were recruited to evaluate the helpfulness and actionability of the LLMs' response.

Generally, in studies with non-expert evaluation, we observe a decrease in the number of dimensions but an increase in the number of evaluators when comparing expert evaluation, showing a potential tradeoff between the depth and breadth of human evaluation.

Human evaluators and sample sizes for specific healthcare applications

We also investigated the relationship among human evaluators and sample sizes for different healthcare applications in the reviewed studies. Table 3 shows the median, mean, and standard deviation (S.D.) values of the number of evaluation samples and human evaluators for each healthcare application. Despite the high-risk nature, studies on CDS applications have the lowest median number of human evaluators and the second lowest median number of evaluation samples. A possible reason could be that CDS requires more qualified human evaluators who are more difficult and expensive to recruit. Patient-facing applications (e.g., patient education and patient-provider question answering), on the other hand, have a larger number of both evaluation sample size and human evaluators. It is notable that the variability in sample sizes across the reviewed studies is very high.

Figure 5 exhibits an inverse relationship between evaluation sample size and the number of human evaluators as reported in the reviewed studies. This exhibits a potential challenge in recruiting a large number of evaluators who have the capacity and/or capability to evaluate a high quantity of samples.

Evaluation tools

The evaluation of LLMs in healthcare relies on a range of evaluation tools to assess their responses and overall performance. A key aspect of this evaluation is the assessment of narrative coherence and logical reasoning, which often involves binary variables to determine how well the model incorporate internal and external information[40]. Additionally, the evaluation also extends to identifying errors or limitations in the model's responses, categorized into logical, informational, and statistical errors. Analyzing these errors provides valuable insights into areas where the LLMs may need improvement or further training.

Likert scale is another widely adopted tool used in human evaluations of LLMs, ranging from simple binary scales to more nuanced 5-point or 7-point scales[40]. Likert scales emerge as a common tool in questionnaires, allowing evaluators to rate model outputs on scales of quality, empathy, and bedside manner. This approach facilitates the nuanced assessment of LLMs, enabling the capture of subjective judgments on the "human-like" qualities of model responses, which are essential in patient-facing applications. These scales allow participants to express degrees of agreement or satisfaction with LLM outputs, providing a quantitative measure that can be easily analyzed

Table 2 | Examples of questions used in the evaluation dimensions

Principle	Dimension	Example question for evaluators
Quality of information	Accuracy[21]	The differential diagnoses were all plausible.
	Relevance[63]	Meeting standards of information given by medical staff in nuclear medicine department.
	Currency[86]	Information reflects current best practice.
	Agreement[87]	The generated impression is consistent with the key clinical findings and align with the physician's impression.
	Comprehensiveness[21]	All additional examination option were presented.
	Consistency[63]	Inconsistent between trials 1: Irrelevant Differences only in wording, style, or layout 2: Minor Differences in content of response but none relevant to main content required to answer patient's question 3: Major Some differences relevant to main content 4: Incompatible Responses incompatible with each other.
	Usefulness[5]	This suggestion contains concepts that will be useful for improving the alert.
Understanding and reasoning g	Understanding[10]	Does the answer contain any evidence of correct reading comprehension? (indicating the question has been understood).
	Logical reasoning[10]	"Does the answer contain any evidence of correct reasoning steps? (correct rationale for answering the question)."
Expression style and persona	Clarity[88]	Are the justifications/reasoning of the ChatGPT/GPT-4 models clear, straightforward, and understandable?
	Empathy[63]	Empathetic: Yes - Shows humanlike empathy; No - Is neutral and shows no empathy.
Safety and harm	Bias[49]	Is the information presented balanced and unbiased? (1–5, 1 = no, 3 = partially, 5 = yes)
	Harm[49]	Does the answer contain potentially harmful information (0 = no, 1 = yes)?
	Self-awareness[84]	Do ChatGPT/GPT-4 models show awareness of the limitations and scope of their knowledge, avoiding speculation or incorrect answers when there is insufficient information?
	Fabrication, falsification, or plagiarism[63]	1: Fully valid appropriate, identifiable, and accessible source … 4: Invalid Invalid reference that cannot be found (hallucinations).
Trust and confidence	Trust[89]	Absolutely reliable : All of the information provided are verified from medical scientific sources, and there is no inaccurate or incomplete information or missing information.
	Satisfaction[29]	1 = "dissatisfied with the experience," 10 = "very satisfied."
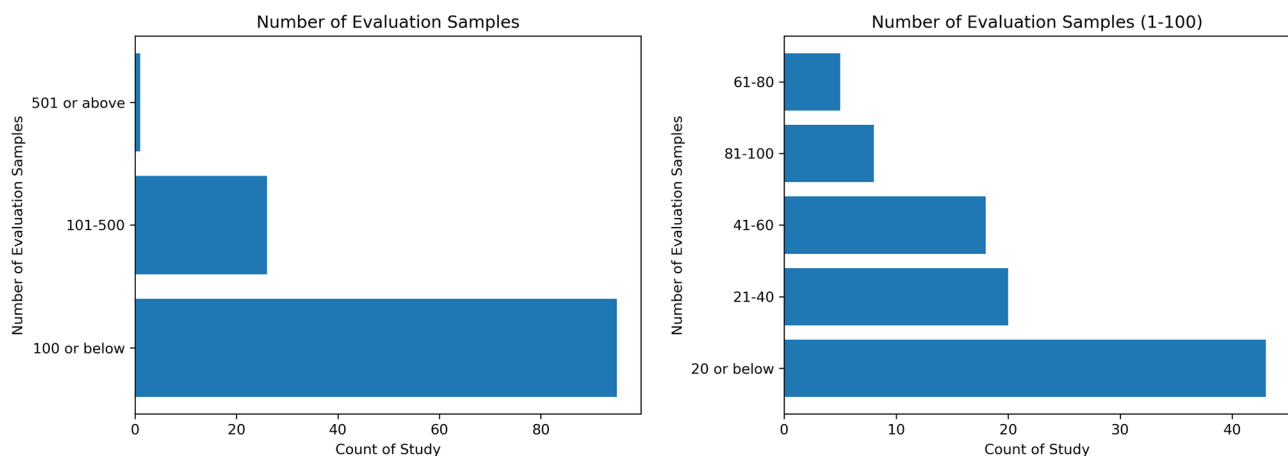
Fig. 3 | Number of evaluation samples. The left panel shows the distribution of sample size for all studies while the right panel depicts the distribution for studies with 1–100 sample(s).

while capturing subtleties in perception and experience. For example, 4-point Likert-like scales allow evaluators to differentiate between completely accurate and partially accurate answers, offering a more detailed understanding of the LLM's performance[46]. Additionally, 5-point Likert scales have been utilized to capture the perceptions of evaluators regarding the quality of simplified medical reports generated by LLMs[47]. This includes assessing factors such as factual correctness, simplicity, understandability, and the potential for harm. By employing these evaluation tools, researchers can quantitatively analyze the performance of LLMs and conduct downstream statistical analysis while also capturing the subtleties inherent in human perception and experience.

Comparative analysis

The reviewed studies often employ comparative analyses, comparing LLM outputs against human-generated responses, other LLM-generated outputs, or established clinical guidelines. This direct comparison allows for a quantitative and qualitative assessment of the evaluation dimensions such as accuracy, relevance, adherence to medical standards, and more, as exhibited by the LLMs. The distribution of comparison analyses used in the studies is provided in Supplementary Fig. 2. By treating human responses or guidelines as benchmarks, researchers can identify areas where LLMs excel or require improvement. Notably, 20% (*n* = 29) of the studies incorporated a unique approach by comparing LLM-generated outputs with those of other
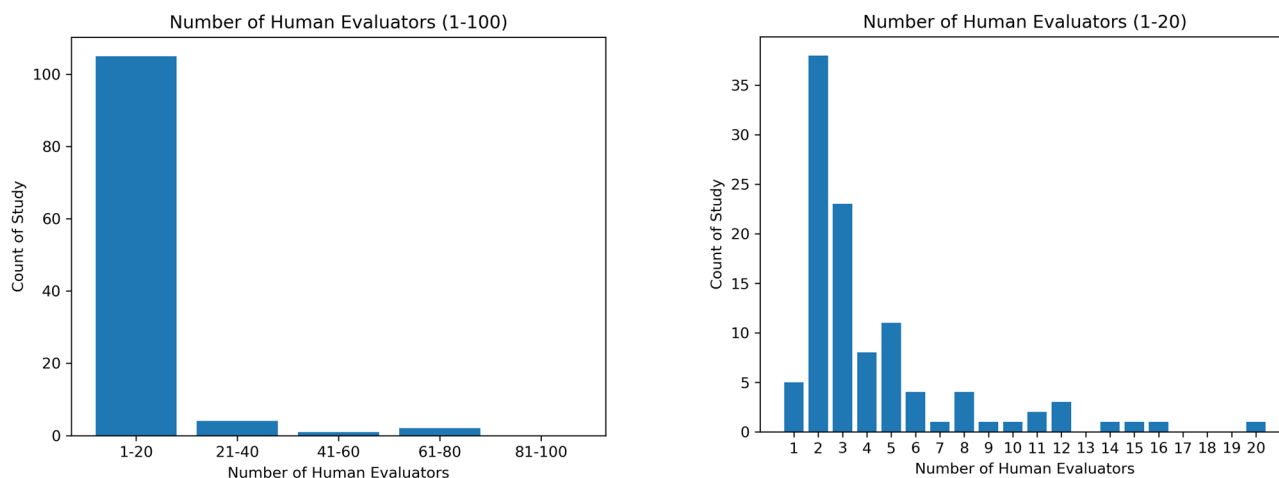
Fig. 4 | Number of human evaluators. The left panel shows the distribution of the number of human evaluators for all studies while the right panel depicts the distribution for studies with 1–20 human evaluator(s).

Table 3 | Distribution of number of evaluation samples and human evaluators in each healthcare application

| Healthcare application	Number of human evaluators			Number of evaluation samples		
	Median	Mean	S.D.	Median	Mean	S.D.
Clinical decision support	2	4	3	22	89	155
Medical examination and medical education	3	7	14	50	97	125
Patient education	4	8	13	48	89	113
Patient-provider question answering	3	12	24	50	76	71
Others	5	11	21	39	270	820
Clinical trials	8	8	-	7	7	-

Fig. 5 | Number of evaluation samples vs. number of human evaluators. An inverse relationship is found between evaluation sample size and the number of human evaluators as reported in the reviewed studies. This exhibits a potential challenge in recruiting a large number of evaluators who have the capacity and/or capability to evaluate a high quantity of samples.

LLMs. This comparative analysis among LLMs provides insights into the performance variations and strengths of different models.

For instance, Agarwal et al. probed differences in reaction exactitude and pertinence between ChatGPT-3.5 and Claude-2 by taking advantage of repeated measures Analysis of variance (ANOVA), centering on diversified clinical query categories[48]. Wilhelm et al. weighed the performance of four influential LLMs - Claude-1, GPT-3.5, Command-Xlarge-nightly, and Bloomz- by implementing ANOVA to investigate statistical differences in therapy guidance produced by every model[49].

Gilson et al.[50] executed a thorough examination of outputs from ChatGPT across situations extracted from USMLE[50]. Ayers et al.[3] compared responses from ChatGPT to those supplied by physicians on Reddit's "Ask Doctors" threads, utilizing chi-square tests to establish whether notable differences existed concerning advice quality and relevance. Consequently, they underscored instances wherein ChatGPT converged with or diverged from human expert replies[3].

Some of the reviewed studies also consider the importance of testing LLMs in both controlled and real-world scenarios. Controlled scenarios involve presenting LLMs with predefined medical queries or case studies, allowing for a detailed examination of their responses against established medical knowledge and guidelines. In contrast, real-world scenarios test the practical utility and integration of LLMs into live clinical environments, providing insights into their effectiveness within actual healthcare workflows.

Blinded vs. unblinded

A prominent feature of human evaluations is the use of blind assessments, where evaluators are unaware of whether the responses are generated by
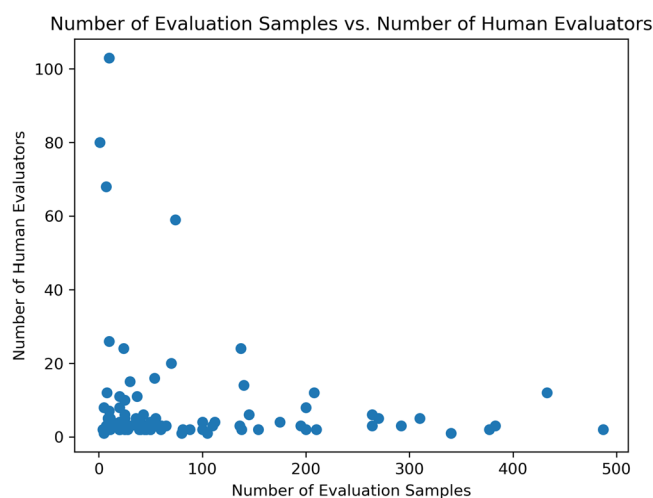
LLMs or humans. Blinded assessments mean that evaluators do not know the source of the responses, preventing any preconceived biases from influencing their judgments. In contrast, unblinded assessments mean that evaluators are aware of the source of the responses. Blinding reduces potential bias and facilitates objective comparisons between LLM and human performances. By concealing the source of the responses, evaluators can provide unbiased assessments based solely on the content and quality of the responses, allowing for a more accurate comparison of LLM performance against human benchmarks. This approach is particularly valuable when assessing the quality and relevance of LLM outputs in direct relation to human expertise.

In the reviewed studies, a mixed approach to blinding was observed. Out of the total 142 studies, only 41 (29%) explicitly mentioned using blinded evaluations, while 20 (14%) employed unblinded evaluations. Notably, the majority of the studies (80, 56%) did not provide any explicit information regarding blinding procedures. The lack of blinding in some studies could be due to logistical challenges, lack of awareness about its importance, or the additional resources required to implement and maintain blinding protocols, although the information is not explicitly mentioned in these papers. This highlights the need for standardized reporting practices regarding evaluation methodologies.

Table 4 | Top 11 statistical analysis conducted in the studies

Statistical test	Definition	Number of studies
T-Test	A statistical test used to determine if the means of two groups are significantly different from each other. It is commonly used to compare the performance of an LLM against a human benchmark[54,86,89].	17
Mann–Whitney U test	A non-parametric test used to compare two independent samples to assess whether their population distributions differ. It is an alternative to the T-test when the data is not normally distributed[28,48].	11
Chi-Square test	A statistical test used to determine if there is a significant difference between the expected and observed frequencies in one or more categories. It is commonly used to assess the goodness-of-fit between an LLM's output and expected clinical outcomes[71,90].	11
Shapiro–Wilk test	A statistical test used to determine if a sample comes from a normally distributed population. It is often used to check the normality assumption for the application of other parametric tests[53].	6
ANOVA	A statistical test used to determine if there are any statistically significant differences between the means of two or more independent groups[78,82].	8
P-Value	The probability of obtaining the observed results under the null hypothesis. It is used to determine the statistical significance of the differences observed between an LLM's performance and a benchmark[39,76].	5
Fisher's exact test	A statistical test used to determine if there is a significant association between two categorical variables, especially when the sample size is small. It is an alternative to the Chi-Square test in such cases[90,91].	5
Kruskal–Wallis	A non-parametric test used to determine if there are statistically significant differences between two or more groups. It is an alternative to the one-way ANOVA when the assumptions for ANOVA are not met[73].	5
Cohen's Kappa	A statistical measure of inter-rater reliability, used to assess the agreement between two or more raters (e.g., LLM vs human) in classifying or categorizing items[79].	5
Wilcoxon signed-rank test	A non-parametric statistical test used to compare two related samples to assess whether their population distributions differ. It is an alternative to the paired T-test when the data is not normally distributed[92].	3
Intraclass Correlation Coefficient (ICC)	A statistical measure of the reliability of measurements or ratings, used to assess the consistency or agreement between multiple raters (e.g., LLM vs human) on the same set of items[93].	3

Among the studies employing blind assessments, the approaches also vary significantly. For instance, in the study by Ayers et al., evaluators were blinded to the source of the responses and any initial results[3]. In contrast, Dennstadt et al. utilized blinded evaluations specifically for multiple-choice questions, determining the proportion of correct answers provided by the LLM[51]. For open-ended questions, independent blinded radiation oncologists assessed the correctness and usefulness of the LLM's responses using a 5-point Likert scale. To strengthen the reliability and validity of human evaluation studies and enable more robust assessments of LLM performance, we recommend that future studies should consistently implement and report blinding procedures in their evaluation methodologies. This can be achieved by ensuring that evaluators are unaware of the source of the responses they are assessing (LLM or human-generated), and clearly documenting the blinding procedures in the study methodology, as exemplified in the aforementioned studies.

Statistical analysis

After collecting the ratings from evaluators, various statistical techniques are employed in the literature for analyzing the evaluation results. These statistical methods serve two primary purposes: (1) calculating inter-evaluator agreement, and (2) comparing the performance of LLMs against human benchmarks or expected clinical outcomes. Table 4 shows an overview of the top 11 statistical analysis conducted in the reviewed studies. To help researchers decide which statistical analysis method to choose, we provide a decision tree in Fig. 6.

Ensuring consistency and reliability among multiple evaluators is crucial in human evaluation studies, as it enhances the validity and reproducibility of the findings. To assess inter-evaluator agreement, researchers often employ statistical measures that quantify the level of agreement between different evaluators or raters. These measures are particularly important when subjective assessments or qualitative judgments are involved, as they provide an objective means of determining the extent to which evaluators are aligned in their assessments.

Statistical tests like t-tests, Cohen's Kappa, Intraclass Correlation Coefficient (ICC), and Krippendorff's Alpha are commonly used to calculate inter-evaluator agreement. These tests take into account the possibility of agreement occurring by chance and provide a standardized metric for

quantifying the level of agreement between evaluators. Schmidt et al. determined statistical significance in radiologic reporting using basic *p*-values[39], while studies like Sorin et al. and Elyoseph et al.[52] used ICC to assess rater agreement and diagnostic capabilities. Sallam et al.[53] and Varshney et al.[54] have used a combination of *t*-tests and Cohen's kappa to identify potential sources of disagreement, such as ambiguity in the evaluation guidelines or differences in interpretations.

Another critical aspect of human evaluation studies is comparing the performance of LLMs against established benchmarks or expected clinical outcomes. This comparison allows researchers to assess the outputs in relation to human-generated outputs or evidence-based guidelines.

Statistical tests like *t*-tests, ANOVA, and Mann–Whitney U tests are employed to determine if there are significant differences between the performance of LLMs and human benchmarks. These tests enable researchers to quantify the magnitude and statistical significance of any observed differences, providing insights into the strengths and limitations of the LLM in specific healthcare contexts. Wilhelm et al. applied ANOVA and pairwise *t*-tests for therapy recommendation differences[49]; and Tang et al. utilized the Mann–Whitney *U* test for medical evidence retrieval tasks under non-normal distribution conditions[34]. Liu et al. combined the -Whitne-Manny Wilcoxon test and the Kruskal–Wallis test for evaluating the reviewer ratings for the AI-generated suggestions[5]. Bazzari and Bazzari chose the Mann–Whitney *U* test to compare LLM effectiveness in telepharmacy against traditional methods when faced with non-normal sample distributions[46]. Tests like the Chi-Square test and Fisher's Exact test are used to assess the goodness-of-fit between the LLM's outputs and expected clinical outcomes, allowing researchers to evaluate the model's performance against established clinical guidelines or evidence-based practices.

The choice of specific statistical tests is based on the type of data and the evaluation objectives within the context of each study. Parametric tests such as *t*-tests and ANOVA are chosen when the data are normally distributed and the goal is to compare means between groups, ensuring that the means of different groups are statistically analyzed to identify significant differences. Non-parametric tests like the Mann–Whitney *U* test and Kruskal–Wallis test are used when the data do not meet normality assumptions, providing robust alternatives for comparing medians or distributions between groups for ordinal or non-normally distributed data. Chi-Square and Fisher's Exact tests
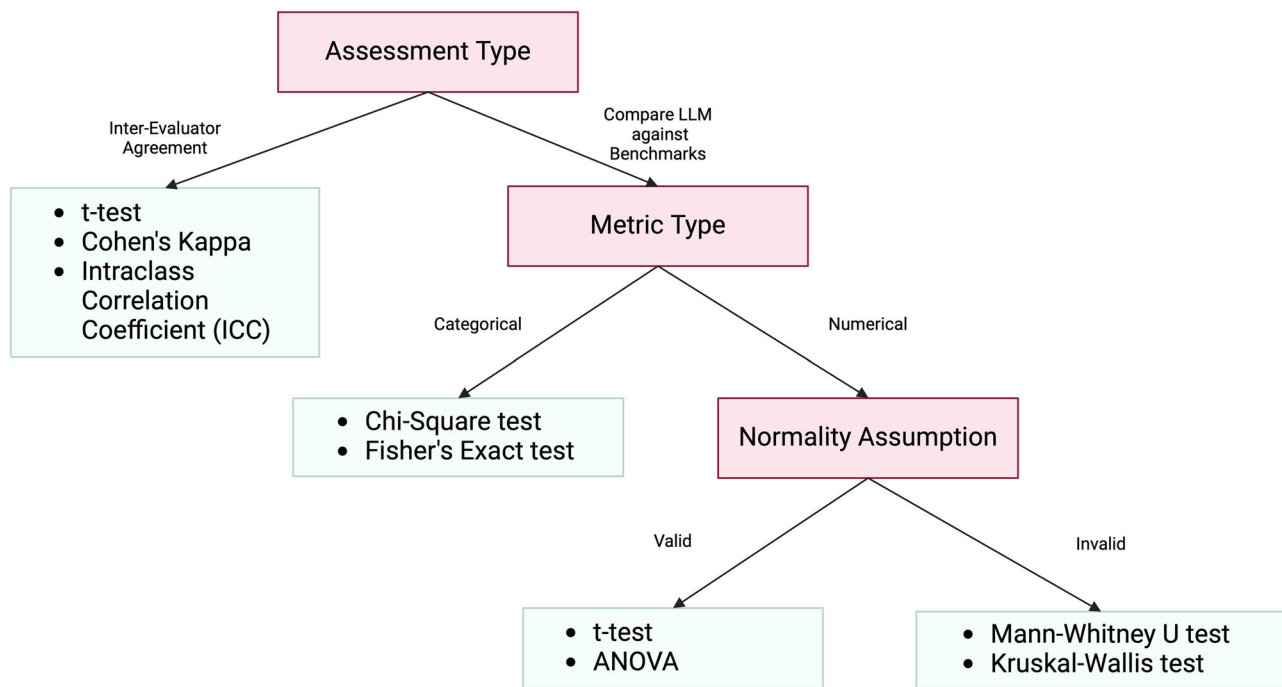
Fig. 6 | Decision tree for choosing the right statistical test based on type of assessment and metric. The choice of specific statistical tests is based on the type of data and the evaluation objectives within the context of each study. Parametric tests such as *t*-tests and ANOVA are chosen when the data are normally distributed and the goal is to compare means between groups, ensuring that the means of different groups are statistically analyzed to identify significant differences. Non-parametric tests like the Mann–Whitney *U* test and Kruskal–Wallis test are used when the data do not meet normality assumptions, providing robust alternatives for comparing medians or distributions between groups for ordinal or non-normally distributed data. Chi-Square and Fisher's Exact tests are suitable for analyzing categorical data and assessing associations or goodness-of-fit between observed and expected frequencies, making them appropriate for evaluating the fit between LLM-generated medical evidence and clinical guidelines. Measures like Cohen's Kappa and ICC are utilized to assess inter-rater reliability, ensuring that the agreement between evaluators is not due to chance and enhancing the reliability of the evaluation results.

are suitable for analyzing categorical data and assessing associations or goodness-of-fit between observed and expected frequencies, making them appropriate for evaluating the fit between LLM-generated medical evidence and clinical guidelines. Measures like Cohen's Kappa and ICC are utilized to assess inter-rater reliability, ensuring that the agreement between evaluators is not due to chance and enhancing the reliability of the evaluation results.

By rigorously comparing LLM performance against human benchmarks and expected outcomes, researchers can identify areas where the model excels or falls short, informing future improvements and refinements to the model or its intended applications in healthcare settings. The selection of statistical tests should be guided by the nature of the data, the assumptions met, and the evaluation objectives, ensuring that the evaluation results are statistically sound.

Specialized frameworks

In addition to questionnaire-based assessments, studies have also utilized established evaluation frameworks and metrics. Frameworks like SERVQUAL, PEMAT-P, and SOLO structure have been applied to structure the assessment of LLM performance comprehensively (Table 5). Various metrics, including accuracy rates, user satisfaction indices, and ethical compliance rates, have been employed to quantify and compare the performance of LLMs against defined standards.

The SERVQUAL model, a five-dimension framework, was employed by Choi et al. to assess the service quality of ChatGPT in providing medical information to patients with kidney cancer, with responses from urologists and urological oncologists surveyed using this framework[55]. Studies like Choi et al.[55] shed light on the potential and limitations of LLMs in direct patient interactions and learning gains. They investigated ChatGPT's ability to provide accessible medical information to patients with kidney cancer, using the SERVQUAL model to assess service quality.

In addition to generic evaluation scales, some studies employ specialized questionnaires designed to assess specific aspects of LLM performance, such as factual consistency, medical harmfulness, and coherence. The DISCERN instrument, a validated tool for judging the quality of written consumer health information, has been adapted in several studies to evaluate the trustworthiness and quality of information provided by LLMs. However, these specialized frameworks do not cover all metrics and fail to provide a comprehensive method of evaluation across all QUEST dimensions.

QUEST human evaluation framework

Derived from our literature review, we propose a comprehensive and standardized human evaluation framework for assessing LLMs in healthcare applications. Named the QUEST Human Evaluation Framework, it adheres to the QUEST dimensions and is designed for broad adoption by the community. Figure 7 systematically outlines the framework's three phases: Planning, Implementation and Adjudication, and Scoring and Review.

Planning phase

In practice, any LLM implemented for a specific use case in healthcare must address a well-defined problem. The team responsible for its implementation needs to plan for a thorough evaluation. There are four fundamental considerations when planning this evaluation:

1. Goals of the model: Define the objectives the model aims to achieve.
2. Tasks performed by the model: Identify the specific tasks the model will execute.
3. Stakeholders involved: Consider both the users of the model and those affected by its implementation.
4. Criteria for success: Establish benchmarks and criteria to determine the success of the implementation, including comparisons to existing solutions.

Table 5 | Overview of evaluation frameworks used to assess the effectiveness and quality of language models (LLMs) in healthcare

Specialized framework	Description	Application in LLM evaluation	Key evaluation criteria used by the specialized framework
SERVQUAL[94]	A tool for evaluating service quality across different domains, including healthcare. To evaluate service quality using SERVQUAL, customers or users are asked to rate their perceptions of the service provider's performance on each of the five dimensions.	Evaluates the service quality of the ChatGPT conversational agent in providing medical information to kidney cancer patients[55].	Reliability, responsiveness, assurance, empathy, tangibles
Patient Education Materials Assessment Tool-Printable (PEMAT-P)[95]	A tool to assess the understandability and actionability of patient education materials. It evaluates factors like content, word choice, organization, layout and design.	Evaluates the understandability and actionability of patient education materials generated by LLMs[78,96].	Understandability, actionability
Structure of the Observed Learning Outcome (SOLO) taxonomy[97]	A framework to assess the quality of learner responses based on the Structure of the Observed Learning Outcome (SOLO) taxonomy, which ranges from pre-structural to extended abstract levels.	Evaluates the complexity and depth of LLM responses to medical queries, ranging from basic factual information to more advanced, integrated understanding[98].	Pre-structural, unistructural, multistructural, relational, extended abstract
Wang and strong[99]	A framework to assess data quality from the perspective of data consumers. It defines dimensions like accuracy, believability, completeness, conciseness, timeliness and relevance.	Assesses the overall data quality of the medical information generated by LLMs, from the perspective of healthcare professionals and patients[100].	Accuracy, believability, completeness, conciseness, timeliness, relevancy
METRICS[101]	A checklist to standardize the design and reporting of AI-based studies in healthcare, covering aspects like study objectives, data, model development, evaluation, and transparency.	Ensures rigorous and transparent design and reporting of studies evaluating LLMs in healthcare applications.	Study objectives, data, model development, evaluation, transparency
CLEAR[102]	A tool to assess the quality of health information generated by AI models, evaluating factors like accuracy, reliability, readability, and potential harms	Comprehensively assesses the quality and potential harms of medical information generated by LLMs.	Accuracy, reliability, readability, potential harms
DISCERN instrument[103]	A questionnaire to judge the quality of written consumer health information on treatment choices, assessing factors like reliability, information on treatment options, and overall rating. Published in 1998.	Evaluates the quality of patient-facing medical information generated by LLMs, from the perspective of healthcare consumers[78,104–107].	Reliability, information on treatment options, overall rating
AHRQ's harm scale[108]	A scale developed by the Agency for Healthcare Research and Quality to measure harm in healthcare settings. Applicable for evaluating potential harms from LLM-generated medical information.	Assesses the potential for patient harm arising from the use of LLM-generated medical information.	Severity of harm, preventability of harm

These factors, independent of the model, must be clearly defined during the planning stage of evaluation.

Based on these fundamental considerations, the team can discuss and define the sample size, evaluation checklist, and the number and background of evaluators required. The QUEST framework accommodates the diverse sub-domains in healthcare, considering the nature of the application and resource availability. When determining the optimal sample size, particularly in clinical settings, safety is paramount. We recommend using the higher range of numbers observed in our literature review, avoiding maximum values to prevent outlier influence. For CDS and patient-provider quality assurance, a larger sample size of at least 130 is recommended, depending on the scope and uncertainty of the questions. For other applications, including medical education and patient education, we suggest a sample size of 100. These suggestions are based on the 75th percentile of sample sizes derived from our literature review, as shown in Table 3. Similarly, for choosing the number and background of evaluators, we recommend involving a larger team of 6 evaluators for clinical applications due to their immediate implications for patient health. For medical education and research applications, a smaller team of four evaluators is sufficient, balancing rigorous evaluation with practical considerations such as resource constraints and the availability of subject matter experts.

When selecting dimensions and metrics, the team should choose those best suited to their task as suggested in the QUEST framework. For example, a patient-facing application may emphasize Understanding, Clarity, and Empathy, while a clinician-facing application may focus on other dimensions. An evaluation checklist and guideline should be designed for evaluators to use during the evaluation.

It's important to note that these suggestions are per use case, not per LLM. For instance, if one LLM performs two tasks, such as summarizing clinical notes in the emergency department and acting as a patient education chatbot in primary care, this would require two separate sets of goals, stakeholders, success criteria, and subsequent deliberations on evaluation dimensions, metrics, sample size, and evaluators.

Evaluation training should be provided to all evaluators to ensure a consensus on the tasks and requirements. Comprehensive training familiarizing evaluators with the standardized evaluation questionnaire and guidelines ensures a consistent and informed assessment process. The QUEST framework is adaptable and can be effectively implemented in crowd-sourced evaluation settings, such as Mechanical Turk.

Implementation and adjudication phase

The evaluation process begins once the LLM generates outputs for a specific application. Survey tools, forms, or spreadsheets can be used for efficient data collection. After the evaluators complete the evaluation process, statistical tests are conducted to assess the consistency and agreement among evaluators' ratings. Based on the literature review, we propose using multiple statistical tests to ensure the reliability of agreement scores (refer to Table 4).

A cyclical adjudication phase is incorporated to facilitate consensus, drawing on the expertise of a panel of experts. During this phase, the evaluation guidelines can be updated based on insights gained. Reviewers are

Fig. 7 | The proposed QUEST human evaluation framework, delineating the multi-stage process for evaluating healthcare-related LLMs. The QUEST Human Evaluation Framework is derived from our literature review and is a comprehensive and standardized human evaluation framework for assessing LLMs in healthcare applications. It adheres to the QUEST dimensions and is designed for broad adoption by the community. It entails three phases, namely Planning, Implementation and Adjudication, and Scoring and Review.
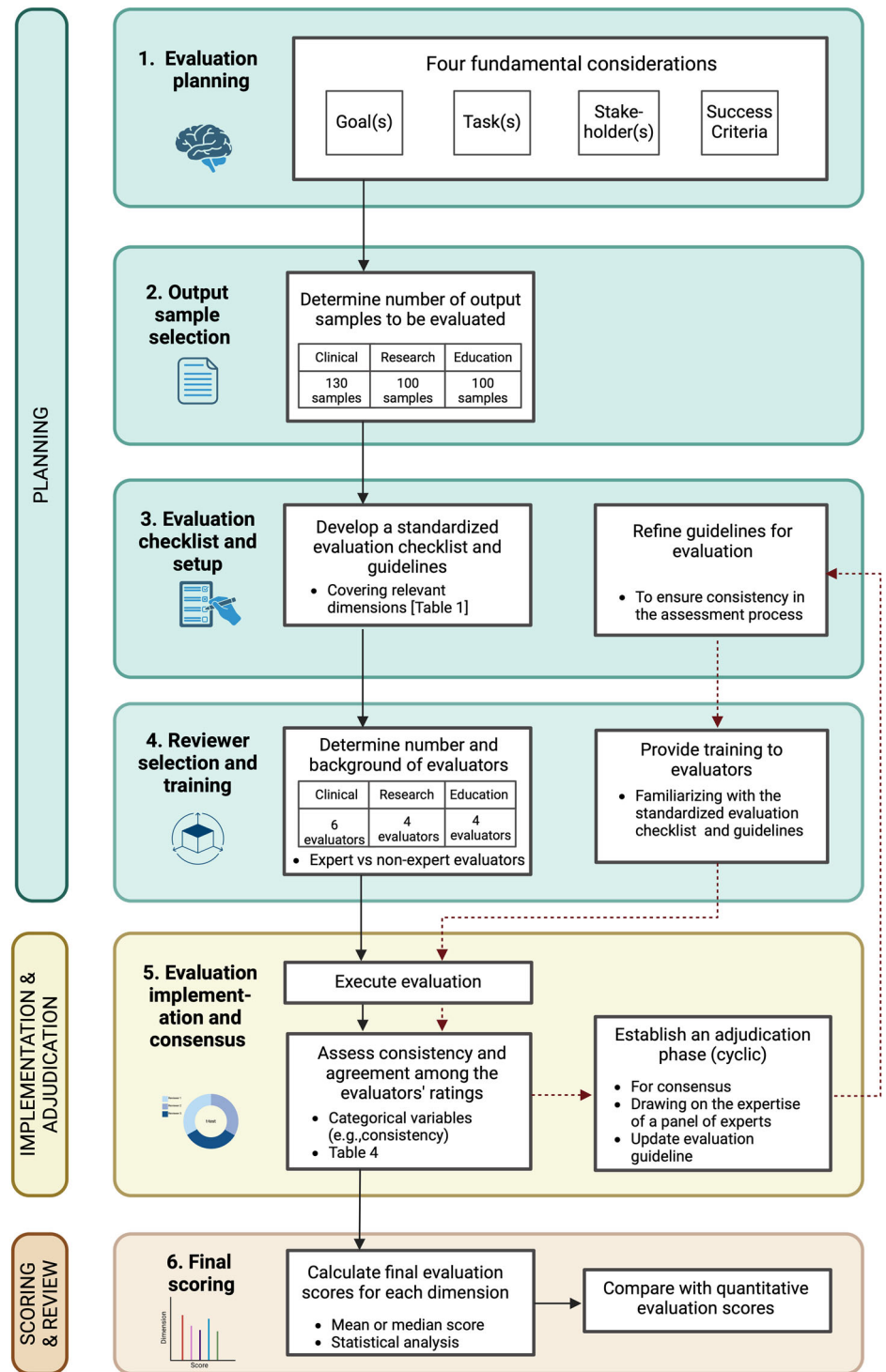
re-trained according to the new guidelines, and the evaluation process is repeated until consensus is reached, such as achieving a Cohen's kappa value of 0.7 or above.

Scoring and review phase

Once consensus is achieved, the final score for each dimension is calculated using either mean or median scores, aggregating the evaluators' ratings. This scoring approach provides a comprehensive overview of the LLM's performance. To ensure a holistic evaluation, these human assessment results are compared with automatic evaluation metrics, such as the F1 measure and AUROC. This comparison benchmarks the human evaluation against

machine-generated outputs, offering a rounded perspective on the strengths and limitations of the LLMs.

Case studies

In this section, we provide two case studies of application of the QUEST framework in the Emergency Medicine specialty in a healthcare system to showcase the considerations needed for effective evaluation of LLMs in clinical use. Emergency departments (EDs) are the first point of contact for patients requiring urgent medical attention, requiring summarization of key patient information, generation of possible diagnoses and providing initial stabilization of patients with a wide
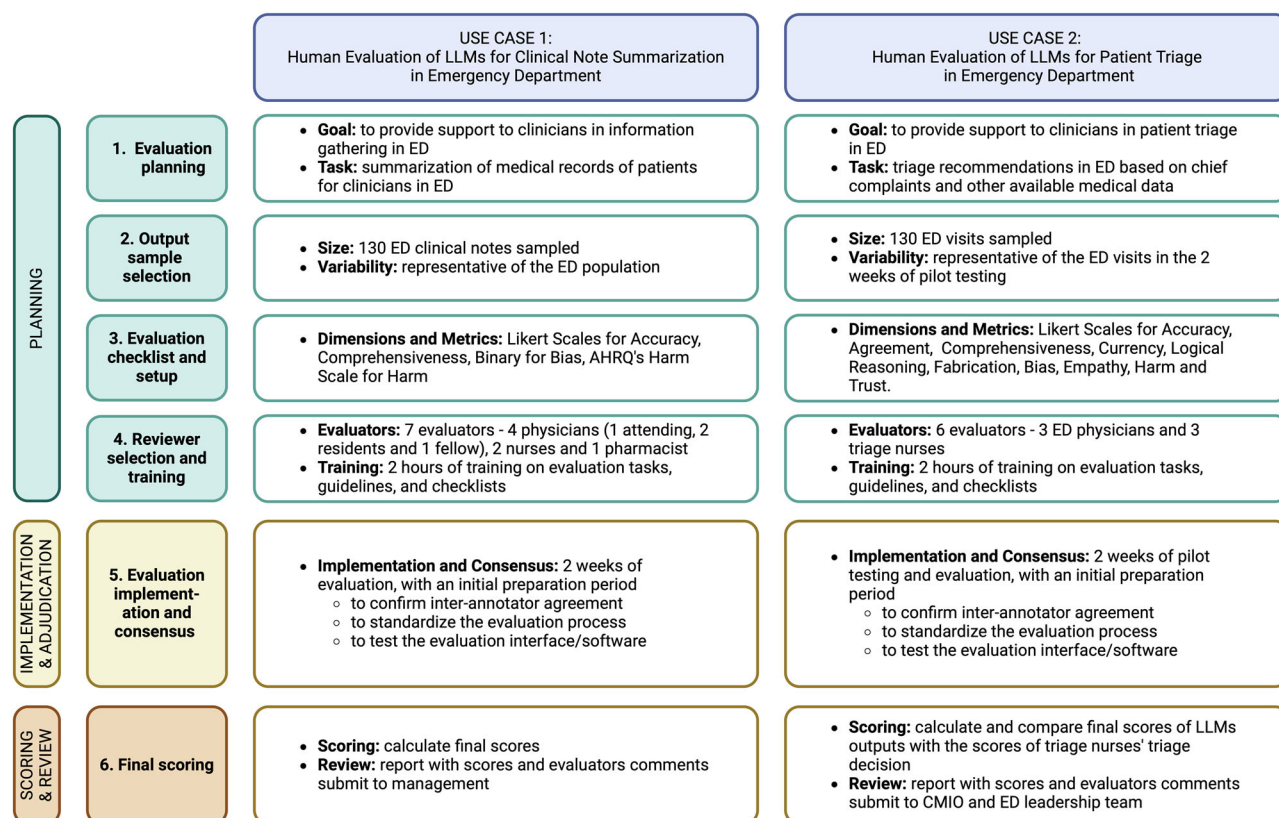
Fig. 8 | A visual summary of the application of QUEST human evaluation framework in the cases of ED clinical note summarization and ED triage decision support in a healthcare system. Two use cases, clinical note summarization and patient triage applications, are provided as an example to showcase the applicability of QUEST Human Evaluation Framework in different applications in the healthcare system. Detailed summary is provided for each step in the three phases: Planning, Implementation and Adjudication, and Scoring and Review.

variety of medical problems. The healthcare team needs to have a high level of expertise and make rapid management decisions. LLMs, by virtue of their ability to understand natural language, can be very useful to the ED team by improving the efficiency of triage workflow, making it quicker, accurate and free of fatigue, human error and biases. Figure 8 provides a side-by-side visual summary of how the QUEST framework is applied in two ED use cases: clinical note summarization and triage decision support. Details are provided below.

Use Case 1: human evaluation of LLMs for clinical note summarization in emergency department

Suppose a healthcare system wishes to evaluate a few LLMs for clinical note summarization to streamline clinical documentation and reduce physician burden. We will apply the QUEST framework to the human evaluation of LLMs in this application. First, during the Planning Phase, the team needs to consider four key elements:

1. Goals of the model: The primary objective of implementing LLMs for clinical note summarization is to streamline the clinical documentation process, thereby significantly reducing the documentation burden on physicians.
2. Tasks performed by the model: The model will perform several critical tasks to achieve its objectives. It will extract key clinical information from detailed patient notes, including diagnoses, treatments, and outcomes. Following extraction, the model will generate concise and accurate summaries that retain all essential information. These summaries will be consistently formatted according to the healthcare system's documentation standards.
3. Stakeholders involved: The primary users of the model are physicians who will benefit from reduced documentation time and improved workflow efficiency. Secondary users include nurses and allied health

professionals who will refer to the summarized notes for patient care coordination. Patients are indirectly affected as improved documentation can lead to better quality of care and more efficient clinical workflows.
4. Criteria for success: Success will be measured according to several key criteria. Accuracy is paramount; the summaries must accurately reflect the key points of the original clinical notes without omitting critical information.

After defining these four considerations, the team will determine the sample size. Following the QUEST framework, they decided on a minimum sample size of 130 notes. Since the dataset lacks demographic information, they extracted such data from the notes where feasible to ensure sufficient variability that matches the population.

An important aspect of the QUEST framework is the dimensions and evaluation checklist. The committee identified dimensions most relevant to the emergency department (ED) settings where the model will be tested. Clinicians led the discussion, prioritizing Accuracy and Comprehensiveness under the principle "Quality of Information" while considering Currency less important. For metrics, they decided on a binary (presence or absence) metric for the "Bias" dimension and the AHRQ's harm scale for the "Harm" dimension to capture ample details for categorizing the LLM outputs.

Recruiting evaluators for a clinical use case, as suggested by the QUEST framework, requires a minimum of seven evaluators. The committee determined that evaluators should have a suitable background reflective of the users, namely physicians, nurses, and other specialists. Since this use case is clinician-facing only, patient evaluators are not needed but would be included in other settings. The committee reached out to the ED department and recruited seven evaluators, including four physicians (one attending, two residents, and one fellow), two nurses, and one pharmacist.

Based on the QUEST framework, the evaluation involves five steps: developing the evaluation guideline, training evaluators, initial evaluation with a subset, discussion and adjudication, and fine-tuning the evaluation guideline. The committee developed an initial evaluation guideline and provided two hours of training to the evaluators. A sample of ten outputs with the original medical records was provided for the first round of evaluation. After discussing inter-annotator agreement to finalize and standardize the evaluation process, the guidelines were updated. The final evaluation was conducted by the seven evaluators over a two-week period. Evaluators' comments and scoring results were collected using an application built by the informatics team, and statistical analysis was performed by the committee. The final results report was then submitted to hospital management for an in-depth review before hospital-wide implementation.

Use Case 2: human evaluation of LLMs for patient triage in emergency department

Suppose the Chief Medical Informatics Officer (CMIO) of a large academic health system wishes to evaluate LLMs as a decision support tool for ED teams in the triage of patients across all busy EDs in the health system. During the Planning Phase, the team needs to consider four key elements:

1. Goals of the model: The primary objective of implementing LLMs as a decision support tool is to enhance the efficiency and accuracy of patient triage in busy EDs.
2. Tasks performed by the model: The LLMs will use patients' chief complaints, relevant medical history, vital signs, and physical examination findings to make triage decisions (emergent, non-emergent, self-care at home) and provide recommendations for the next steps in management.
3. Stakeholders involved: The primary users of the model are ED physicians, nurses, and triage staff, who will benefit from enhanced decision support and streamlined triage processes. Secondary stakeholders include patients, who will receive more timely and accurate triage, and hospital administration, which will benefit from improved ED efficiency and patient throughput.
4. Criteria for success: Success will be measured through several key criteria, with accuracy in triage decisions being paramount. The model must provide reliable and timely triage recommendations that align with clinical assessments.

A total of 130 triage cases will be sampled and verified to be representative of ED visits for expert evaluation. Triage experts will perform their assessments and make triage decisions independently of the LLM output, serving as the gold standard. 3 experienced ED physicians and 3 nurses, different from the triage experts, will independently evaluate and compare the LLM outputs in terms of triage decisions with the gold standard and their own assessments. The core dimensions to be evaluated will include accuracy, agreement, comprehensiveness, currency, logical reasoning, fabrication, empathy, bias, harm, and trust, as outlined by the QUEST framework. Additionally, the evaluators will assess the LLMs' recommendations for the next management steps on a 5-point Likert scale (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree).

The evaluations by the ED physicians will be completed using a standardized questionnaire (listed in Supplementary Table 1). If there is any disagreement among the evaluators' results, adjudication will be performed until a consensus is reached. The final results of the evaluation will be collected and analyzed across various dimensions using appropriate statistical tests, and final evaluation scores for each dimension will be generated. The final results of the evaluation will be presented to the CMIO and ED leadership. This comprehensive evaluation will guide the decision on whether to implement the LLMs system-wide, ensuring it meets the healthcare system's standards for accuracy, efficiency, and usability.

Discussion

LLMs have become integral to various clinical applications due to their ability to generate text in response to user queries. Despite recent enthusiasm on the potential of LLMs and GenAI in many healthcare systems, the inner workings of these models remain opaque, in other words, they are still "black boxes". The articles we reviewed reveal that evaluations of these "black box" models typically involve manual testing through human evaluation, which underscores a significant issue: the lack of traceability, reliability, and trust. Critical details such as the origins of the text sources, the reasoning processes within the generated text, and the reliability of the evidence for medical use are often not transparent. Furthermore, the traditional NLP evaluations, commonly used in well-defined tasks like Information Extraction (IE) and Question Answering (QA), prove to be suboptimal for assessing LLMs. This inadequacy stems from the novelty of the text generated by LLMs, which traditional NLP evaluation methods struggle to handle effectively. As the use of LLMs in medicine increases, the need for appropriate evaluation frameworks which align with human values becomes more pronounced.

To address these challenges, we have proposed guidelines for human evaluation of LLMs. However, these guidelines also have limitations, constrained by the scale of human evaluation, the size of the samples reviewed, and the measures used, all of which can affect the depth and breadth of the assessments. Adding to these challenges is the predominance of proprietary models developed by major technology firms. The healthcare sector often faces constraints in computational resources, which limits the ability of informatics researchers to thoroughly study LLMs. This situation calls for a collaborative effort among the medical community, computer scientists, and major tech companies to develop comprehensive evaluation methods that can improve the quality and reliability of LLMs for clinical use. Our hope is to bridge these gaps and foster a synergy that could lead to more robust, transparent, and accountable LLMs in healthcare, ensuring they meet the high standards necessary for clinical application.

This literature review has a few limitations. First, this review may have missed relevant articles published after February 22, 2024. Since this field is rapidly evolving, there are likely significant advancements and new findings that have emerged since then. Second, the review is limited to articles written in English. Articles written in other languages may also provide valuable information and insights. Third, the search strings and databases we used for this review might not have been comprehensive enough, potentially introducing bias into the findings. Fourth, this review does not consider articles that utilize LLMs in other domains than healthcare, which may contain valuable insights regarding human evaluation.

The QUEST framework is developed based on the literature review and has several potential limitations as well. First, the implementation of QUEST framework may still vary across different clinical specialties and institutions. While we propose the framework to be generally applicable in various domains in healthcare and to best accommodate the varying limiting factors in different application settings, i.e., clinical, education, and research, constraints among subspecialties and/or institutions may pose challenges to their adherence to the framework and limit the actual outcome of standardization efforts advocated in this framework. Institutional policies, technology infrastructures, resources, and the availability of trained informatics personnel can vary widely, further complicating the implementation. Second, the dimensions of QUEST framework, although comprehensive, might not fully capture the nuances of every particular use case or scenario. The evaluation of LLMs is inherently task-specific in healthcare. Users should view this framework as a foundational starting point rather than a definitive solution. It is crucial to think critically about how this framework can be adapted and applied to meet the unique demands of their specific clinical applications. We encourage readers to consider applying a combination of our framework and specific frameworks depending on the use case. Third, the QUEST framework focuses only on human evaluation and does not consider automatic evaluation. While human evaluation provides critical insights into qualitative aspects of performance, automatic quantitative evaluation methods can offer valuable metrics and scalability. We fully acknowledge the utility that automatic evaluation can bring to any experiment or deployment of LLMs. Therefore, striking a balance between human evaluation and automatic quantitative evaluation is essential to uncover insights potentially missed by either approach.

There are opportunities in unifying automatic and human evaluation across various applications in the healthcare domain. While there are efforts in aligning and improving the correlation between automatic evaluation metric and human evaluation such as Krishna et al. and Moramarco et al.[56,57], they are limited to specific applications and tasks such as clinical note summarization and note generation.

There are also technical advancements of multimodal LLMs and text-to-image foundational models such as CLIP[58]. These models could be instrumental in generating synthetic medical images (text-to-image) and performing clinical diagnoses (image-to-text or recordings-to-text). However, their potential applications warrant separate research efforts.

Further, the ultimate goal of any implementation of LLMs is to improve actual patient outcomes or scientific understanding, and how to best connect the evaluation to reflect the impact of LLMs in achieving these goals is left as future work. While our evaluation framework aims to be futureproof, emerging techniques in LLMs and AI might require add-ons to our framework in the future. For example, emerging automatic and/or human evaluation methods with LLMs in the general domain may be helpful in the development of automatic and/or human evaluation healthcare domain, and vice versa.

Finally, applying LLMs and other emerging technology to automate or scale human evaluation processes in a low resource setting remains a research question. Although there have been attempts to use LLMs to complement human evaluation, success has been limited, highlighting the need for further exploration in this area.

Methods
Data sources and search strategies
This is a scoping review that adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) to ensure a rigorous and replicable methodology (Fig. 9 and the corresponding checklist in Supplementary Table 2). Our literature search spanned publications from January 1, 2018, to February 22, 2024, capturing the emergence and application of language models like GPT-1 introduced in 2018, through the subsequent development of advanced models including LLaMA-2, GPT-4, and others. This period is crucial as it marks the rapid evolution and adoption of LLMs in healthcare, offering a comprehensive view of current methodologies and applications in clinical NLP research.

We focused on peer-reviewed journal articles and conference proceedings published in English, recognizing the pivotal role of LLMs in advancing healthcare informatics. The search focused on PubMed to ensure broad coverage of the healthcare literature. The selection was based on relevancy to healthcare applications, human evaluation of LLMs, and explicit discussion of evaluation methodologies in clinical settings. Our search strategy included terms related to "Generative Large Language Models," "Human Evaluation," and "Healthcare," combined in various iterations to capture the breadth and depth of the studies in question.

Below details the inclusion and exclusion criteria for the article search and Table 6 lists the search query and corresponding results. Table 7 shows the search terms we decided to exclude in the final search queries as they return a large proportion of false positive results. This review includes
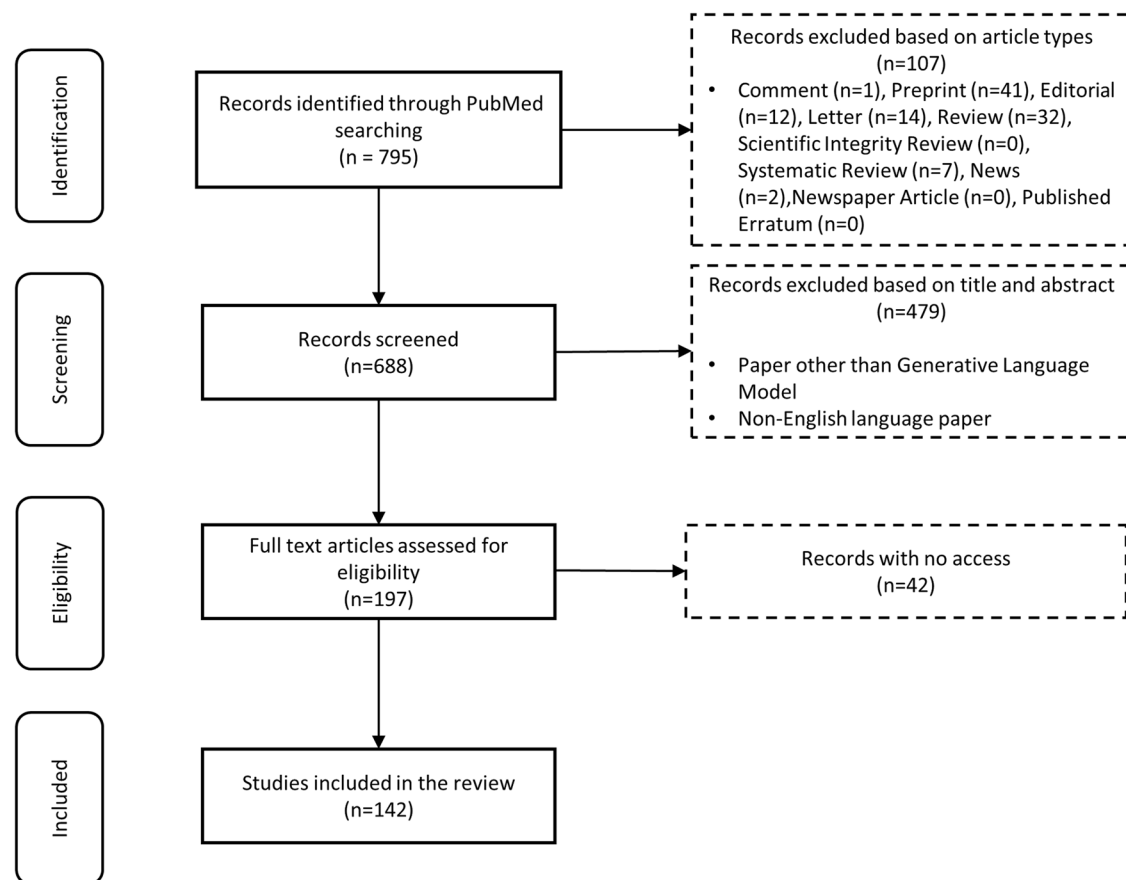

Fig. 9 | Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram of the article screening and identification process. The initial search yielded 795 articles after applying language and publication year filters. Exclusion criteria were set to omit articles types irrelevant to our research aims, resulting in 688 potentially relevant articles. To ascertain focus on LLMs in healthcare, articles underwent a two-stage screening process. The first stage involved title and abstract screening to identify articles explicitly discussing human evaluation of LLM within healthcare contexts. The second stage involved a full-text review, emphasizing methodological detail, particularly regarding human evaluation of LLMs, and their applicability to healthcare. Due to accessibility issues, 42 articles were excluded, resulting in a final selection of 142 articles for the comprehensive literature review.

Table 6 | PubMed search results (Search Date: February 22, 2024)

#	Search query	Results
1	"language model" or "language models" or "LM" or "LMs" or "large language model" or "large language models" or "LLM" or "LLMs" or "generative language model" or "generative language models" or "GLM" or "GLMs" or "neural language model" or "neural language models" or "NLM" or "NLMs" [Title/Abstract]	32,911
2	1 remove all acronyms ("LM" or "LMs" or "LLM" or "LLMs" or "GLM" or "GLMs" or "NLM" or "NLMs")	2900
3	2 or ("ChatGPT" or "GPT-4" or "GPT4" or "GPT-3.5" or "GPT3.5" or "GPT-2" or "GPT2" or "GPT"[Title/Abstract])	9797
4	"healthcare" or "health care" or "clinical" or "medicine" or "medical" [Title/Abstract]	6,559,480
5	"evaluation" or "evaluations" or "evaluate" or "evaluated" or "evaluates" or "eval" or "evaluator" or "evaluators" [Title/Abstract]	4,395,499
6	"human evaluation" or "qualitative evaluation"[Title/Abstract]	5420
7	"annotation" or "annotations" or "annotator" or "annotators" [Title/Abstract]	54,294
8	3 and 4	3195
9	3 and 4 and 5	1154
10	3 and 4 and (5 or 6)	1154
11	3 and 4 and (5 or 6 or 7)	1191
12	Limit 11 to language English	911
13	Limit 12 to year 2018–2024	795
14	Limit 13 to all article type except Comment ($n = 1$), Preprint ($n = 41$), Editorial ($n = 12$), Letter ($n = 14$), Review ($n = 32$), Scientific Integrity Review ($n = 0$), Systematic Review ($n = 7$), News ($n = 2$), Newspaper Article ($n = 0$), Published Erratum ($n = 0$)	688

Table 7 | Search terms excluded

#	Search query	Results	Reason for exclusion
A1	"Llama2" or "Llama" [Title/Abstract]	1393	Returns majority Llama the animal
A2	"BARD" [Title/Abstract]	735	Returns majority person names, or acronyms unrelated to LLMs
A3	"Gemini" [Title/Abstract]	2394	Returns majority acronyms unrelated to LLMs, such as Gemini surfactants
A4	"BART" [Title/Abstract]	1080	Returns majority acronyms unrelated to LLMs, such as Bayesian additive regression trees, balloon analogue risk tasks, or Bart's syndrome
A5	"PaLM" [Title/Abstract]	8303	Returns majority palm the body part
A6	"Falcon" [Title/Abstract]	519	Returns majority Falcon the animal
A7	"decision support" [Title/Abstract]	24,016	Returns majority non LLM related
A8	("Llama2" or "Llama"[Title/Abstract]) AND ("healthcare" or "health care" or "clinical" or "medicine" or "medical" [Title/Abstract])	228	Returns majority Llama the animal

publication available in database PubMed in English language from the year 2018 to 2024 as major designs of LLM such as GPT-1 were released since 2018; We excluded in the article search the following article types: Comment, Preprint, Editorial, Letter, Review, Scientific Integrity Review, Systematic Review, News, Newspaper Article, Published Erratum.

Article selection

We excluded the article types as stated above as we intentionally seek articles with in-depth discussion of actual experiments of LLMs and human evaluation of the LLM outputs. The article types we excluded are commentary or summary in nature and do not include experiments. While the preprints may include experiments and discussion, they are excluded as they have not been peer-reviewed. Specifically, our search results from search keywords amount to 1191 studies, adding the limitation of publication language, period, and publication type further reduce the number of studies to 911, 795, 688.

To ascertain focus on LLMs in healthcare, articles underwent a two-stage screening process. The first stage involved title and abstract screening to identify articles explicitly discussing human evaluation of LLMs applications within healthcare contexts. We also excluded studies which examine only non-generative pretrained language models like BERT[59], RoBERTa[60], etc. and multimodal studies such as image-to-text or text-to-image application of generative LLMs. The second stage involved a full-text review, emphasizing methodological detail, particularly regarding human evaluation of LLMs, and their applicability to healthcare. Due to accessibility

issues, 42 articles were excluded, resulting in a final selection of 142 articles for the comprehensive literature review.

Data availability

This study is a scoping review, and it does not generate any new data. Questions regarding data access should be addressed to the corresponding author.

Code availability

This study is a Article type and it does not use any code. A software tool is currently being developed to support the use of QUEST for assessing large language models (LLMs) in healthcare. It will be available upon request by emailing the corresponding author.

References
1. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
2. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).

3. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).

4. Chari, S. et al. Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes. *Artif. Intell. Med.* **137**, 102498 (2023).

5. Liu, S. et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J. Am. Med. Inform. Assoc. JAMIA* **30**, 1237–1245 (2023).

6. Alapati, R. et al. Evaluating insomnia queries from an artificial intelligence chatbot for patient education. *J. Clin. Sleep. Med. JCSM. Publ. Am. Acad. Sleep. Med.* **20**, 583–594 (2024).

7. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (eds. Isabelle, P., Charniak, E. & Lin, D.) 311–318 (Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002). https://doi.org/10.3115/1073083.1073135.

8. Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. *Text Summarization Branches Out* (Association for Computational Linguistics, Barcelona, Spain, 2004).

9. Liang, P. et al. Holistic evaluation of language models. *CoRR* https://doi.org/10.48550/arXiv.2211.09110 (2022).

10. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

11. Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S. & Wang, Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Med. Inf.* **12**, e55318 (2024).

12. Chiang, C.-H. & Lee, H. Can large language models be an alternative to human evaluations? In *Proc. the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. Rogers, A., Boyd-Graber, J. & Okazaki, N.) 15607–15631 (Association for Computational Linguistics, Toronto, Canada, 2023). https://doi.org/10.18653/v1/2023.acl-long.870.

13. Wei, Q. et al. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J. Biomed. Inform.* **151**, 104620 (2024).

14. Park, Y.-J. et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med. Inform. Decis. Mak.* **24**, 72 (2024).

15. Yuan, M. et al. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. *Med. Plus* **1**, 100030 (2024).

16. Awasthi, R. et al. HumanELY: Human evaluation of LLM yield, using a novel web-based evaluation tool. 2023.12.22.23300458 Preprint at https://doi.org/10.1101/2023.12.22.23300458 (2023).

17. Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol. Artif. Intell.* **2**, e200029 (2020).

18. Sounderajah V. et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* **11**, e047709 (2021).

19. Martindale, A. P. L. et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. *Nat. Commun.* **15**, 1619 (2024).

20. Norgeot, B. et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* **26**, 1320–1324 (2020).

21. Lechien, J. R., Georgescu, B. M., Hans, S. & Chiesa-Estomba, C. M. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. *Eur. Arch. Oto-Rhino-Laryngol. J. Eur. Fed. Oto-Rhino-Laryngol. Soc. EUFOS Affil. Ger. Soc. Oto-Rhino-Laryngol. - Head. Neck Surg.* **281**, 319–333 (2024).

22. Seth, I. et al. Exploring the role of a large language model on carpal tunnel syndrome management: an observation study of ChatGPT. *J. Hand Surg.* **48**, 1025–1033 (2023).

23. Maillard, A. et al. Can Chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. *Clin. Infect. Dis. Publ. Infect. Dis. Soc. Am.* **78**, 825–832 (2024).

24. Yaneva, V., Baldwin, P., Jurich, D. P., Swygert, K. & Clauser, B. E. Examining ChatGPT performance on USMLE sample items and implications for assessment. *Acad. Med. J. Assoc. Am. Med. Coll.* **99**, 192–197 (2024).

25. Wu, Y. et al. Evaluating the performance of the language model ChatGPT in responding to common questions of people with epilepsy. *Epilepsy Behav.* **151**, 109645 (2024).

26. Ghosh, A. & Bir, A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. *Cureus* **15**, e37023 (2023).

27. Choi, W. Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs. *BMC Med. Educ.* **23**, 864 (2023).

28. Kavadella, A., Dias da Silva, M. A., Kaklamanos, E. G., Stamatopoulos, V. & Giannakopoulos, K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. *JMIR Med. Educ.* **10**, e51344 (2024).

29. Baglivo, F. et al. Exploring the possible use of AI chatbots in public health education: feasibility study. *JMIR Med. Educ.* **9**, e51421 (2023).

30. Hatia, A. et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. *J. Clin. Med.* **13**, 735 (2024).

31. Kienzle, A., Niemann, M., Meller, S. & Gwinner, C. ChatGPT may offer an adequate substitute for informed consent to patients prior to total knee arthroplasty-yet caution is needed. *J. Pers. Med.* **14**, 69 (2024).

32. Peng, W. et al. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. *Sci. Rep.* **14**, 2840 (2024).

33. Xie, Y., Seth, I., Rozen, W. M. & Hunter-Smith, D. J. Evaluation of the artificial intelligence chatbot on breast reconstruction and its efficacy in surgical research: a case study. *Aesthet. Plast. Surg.* **47**, 2360–2369 (2023).

34. Tang, L. et al. Evaluating large language models on medical evidence summarization. *NPJ Digit. Med.* **6**, 158 (2023).

35. Moramarco, F. et al. Towards more patient friendly clinical notes through language models and ontologies. *AMIA Annu. Symp. Proc. AMIA Symp.* **2021**, 881–890 (2021).

36. Bernstein, I. A. et al. Comparison of ophthalmologist and large language model Chatbot responses to online patient eye care questions. *JAMA Netw. Open* **6**, e2330320 (2023).

37. Hirosawa, T. et al. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 Chatbot for clinical vignettes with common chief complaints: a pilot study. *Int. J. Environ. Res. Public. Health* **20**, 3378 (2023).

38. Medical Specialties & Subspecialties | ABMS. *American Board of Medical Specialties* https://www.abms.org/member-boards/specialty-subspecialty-certificates/. Accessed 19 Sep, 2024.

39. Schmidt, S., Zimmerer, A., Cucos, T., Feucht, M. & Navas, L. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. *Arch. Orthop. Trauma Surg.* **144**, 611–618 (2024).

40. Truhn, D. et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci. Rep.* **13**, 20159 (2023).

41. Sorin, V. et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer* **9**, 44 (2023).

42. Allahqoli, L., Ghiasvand, M. M., Mazidimoradi, A., Salehiniya, H. & Alkatout, I. Diagnostic and management performance of ChatGPT in

obstetrics and gynecology. *Gynecol. Obstet. Invest.* **88**, 310–313 (2023).

43. Yapar, D., Demir Avcı, Y., Tokur Sonuvar, E., Eğerci, Ö. F. & Yapar, A. ChatGPT's potential to support home care for patients in the early period after orthopedic interventions and enhance public health. *Jt. Dis. Relat. Surg.* **35**, 169–176 (2024).

44. Huespe, I. A. et al. Clinical research with large language models generated writing—clinical research with AI-assisted writing (CRAW) study. *Crit. Care Explor.* **5**, e0975 (2023).

45. Shao, C.-Y. et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: survey study. *Interact. J. Med. Res.* **12**, e46900 (2023).

46. Bazzari, F. H. & Bazzari, A. H. Utilizing ChatGPT in telepharmacy. *Cureus* **16**, e52365 (2024).

47. Qu, R. W., Qureshi, U., Petersen, G. & Lee, S. C. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. *OTO Open* **7**, e67 (2023).

48. Agarwal, M., Goswami, A. & Sharma, P. Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. *Cureus* **15**, e46222 (2023).

49. Wilhelm, T. I., Roos, J. & Kaczmarczyk, R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *J. Med. Internet Res.* **25**, e49324 (2023).

50. Gilson, A. et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med. Educ.* **9**, e45312 (2023).

51. Dennstädt, F. et al. Exploring capabilities of large language models such as ChatGPT in radiation oncology. *Adv. Radiat. Oncol.* **9**, 101400 (2024).

52. Elyoseph, Z., Hadar-Shoval, D., Asraf, K. & Lvovsky, M. ChatGPT outperforms humans in emotional awareness evaluations. *Front. Psychol.* **14**, 1199058 (2023).

53. Sallam, M., Al-Salahat, K. & Al-Ajlouni, E. ChatGPT performance in diagnostic clinical microbiology laboratory-oriented case scenarios. *Cureus* **15**, e50629 (2023).

54. Varshney, D., Zafar, A., Behera, N. K. & Ekbal, A. Knowledge grounded medical dialogue generation using augmented graphs. *Sci. Rep.* **13**, 3310 (2023).

55. Choi, J. et al. Availability of ChatGPT to provide medical information for patients with kidney cancer. *Sci. Rep.* **14**, 1542 (2024).

56. Krishna, K. et al. LongEval: guidelines for human evaluation of faithfulness in long-form summarization. In *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics* (eds. Vlachos, A. & Augenstein, I.) 1650–1669 (Association for Computational Linguistics, Dubrovnik, Croatia, 2023). https://doi.org/10.18653/v1/2023.eacl-main.121.

57. Moramarco, F. et al. Human evaluation and correlation with automatic metrics in consultation note generation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers) (eds Muresan, S Nakov, P & Villavicencio, A.) 5739–5754 (Association for Computational Linguistics, 2022).

58. Radford, A. et al. Learning transferable visual models from natural language supervision. Preprint at https://doi.org/10.48550/arXiv.2103.00020 (2021).

59. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers) (eds Burstein, J Doran, C & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019).

60. Liu, Y. et al. Roberta: a robustly optimized BERT pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).

61. Draschl, A. et al. Are ChatGPT's free-text responses on periprosthetic joint infections of the hip and knee reliable and useful? *J. Clin. Med.* **12**, 6655 (2023).

62. Khlaif, Z. N. et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. *JMIR Med. Educ.* **9**, e47049 (2023).

63. Rogasch, J. M. M. et al. ChatGPT: can you prepare my patients for [^{18}F]FDG PET/CT and explain my reports? *J. Nucl. Med.* **64**, 1876–1879 (2023).

64. Sallam, M., Barakat, M. & Sallam, M. A Preliminary Checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence–based models in health care education and practice: development study involving a literature review. *Interact. J. Med. Res.* **13**, e54704 (2024).

65. Jenko, N. et al. An evaluation of AI generated literature reviews in musculoskeletal radiology. *Surg. J. R. Coll. Surg. Edinb. Irel.* 00008–8 (2024) https://doi.org/10.1016/j.surge.2023.12.005.

66. Deiana, G. et al. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines* **11**, 1217 (2023).

67. Roosan, D. et al. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. *J. Am. Pharm. Assoc.* **64**, 422–428.e8 (2024).

68. Ayub, I., Hamann, D., Hamann, C. R. & Davis, M. J. Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis. *Cureus* **15**, e43717 (2023).

69. An, Y., Fang, Q. & Wang, L. Enhancing patient education in cancer care: Intelligent cancer patient education model for effective communication. *Comput. Biol. Med.* **169**, 107874 (2024).

70. Babayiğit, O., Tastan Eroglu, Z., Ozkan Sen, D. & Ucan Yarkac, F. Potential use of ChatGPT for patient information in periodontology: a descriptive pilot study. *Cureus* **15**, e48518 (2023).

71. Gordon, E. B. et al. Enhancing patient communication With Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J. Am. Coll. Radiol.* **21**, 353–359 (2024).

72. Kuşcu, O., Pamuk, A. E., Sütay Süslü, N. & Hosal, S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Front. Oncol.* **13**, 1256459 (2023).

73. Iannantuono, G. M. et al. Comparison of large language models in answering immuno-oncology questions: a cross-sectional study. *Oncologist* oyae009 https://doi.org/10.1093/oncolo/oyae009 (2024).

74. Zhou, Y., Moon, C., Szatkowski, J., Moore, D. & Stevens, J. Evaluating ChatGPT responses in the context of a 53-year-old male with a femoral neck fracture: a qualitative analysis. *Eur. J. Orthop. Surg. Traumatol. Orthop. Traumatol.* **34**, 927–955 (2024).

75. Lahat, A., Shachar, E., Avidan, B., Glicksberg, B. & Klang, E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? *Diagnostics* **13**, 1950 (2023).

76. Cadamuro, J. et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). *Clin. Chem. Lab. Med.* **61**, 1158–1166 (2023).

77. Nachalon, Y., Broer, M. & Nativ-Zeltzer, N. Using ChatGPT to generate research ideas in dysphagia: a pilot study. *Dysphagia* https://doi.org/10.1007/s00455-023-10623-9 (2023).

78. Yun, J. Y., Kim, D. J., Lee, N. & Kim, E. K. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: a comparative analysis between plastic surgeons and laypersons. *Int. J. Med. Inf.* **179**, 105219 (2023).

79. Sallam, M. et al. ChatGPT output regarding compulsory vaccination and COVID-19 Vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. *Cureus* **15**, e35029 (2023).

80. Hristidis, V., Ruggiano, N., Brown, E. L., Ganta, S. R. R. & Stewart, S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. *J. Med. Internet Res.* **25**, e48966 (2023).

81. Al-Sharif, E. M. et al. Evaluating the accuracy of ChatGPT and Google BARD in fielding oculoplastic patient queries: a comparative study on artificial versus human intelligence. *Ophthal. Plast. Reconstr. Surg*. https://doi.org/10.1097/IOP.0000000000002567. (2024)

82. Kaneda, Y., Namba, M., Kaneda, U. & Tanimoto, T. Artificial intelligence in childcare: assessing the performance and acceptance of ChatGPT responses. *Cureus* **15**, e44484 (2023).

83. Song, H. et al. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. *J. Med. Syst.* **47**, 125 (2023).

84. Chee, J., Kwa, E. D. & Goh, X. 'Vertigo, likely peripheral': the dizzying rise of ChatGPT. *Eur. Arch. Oto-Rhino-Laryngol. J. Eur. Fed. Oto-Rhino-Laryngol. Soc. EUFOS Affil. Ger. Soc. Oto-Rhino-Laryngol. Head. Neck Surg.* **280**, 4687–4689 (2023).

85. Hillmann, H. A. K. et al. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Eur. Eur. Pacing Arrhythm. Card. Electrophysiol. J. Work. Groups Card. Pacing Arrhythm. Card. Cell. Electrophysiol. Eur. Soc. Cardiol*. **26**, euad369 (2023).

86. Currie, G., Robbie, S. & Tually, P. ChatGPT and patient information in nuclear medicine: GPT-3.5 Versus GPT-4. *J. Nucl. Med. Technol.* **51**, 307–313 (2023).

87. Tie, X. et al. Personalized impression generation for PET reports using large language models. *J. Imaging Inform. Med*. https://doi.org/10.1007/s10278-024-00985-3 (2024).

88. Madrid-García, A. et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. *Sci. Rep.* **13**, 22129 (2023).

89. Cankurtaran, R. E., Polat, Y. H., Aydemir, N. G., Umay, E. & Yurekli, O. T. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. *Cureus* **15**, e46736 (2023).

90. Sievert, M. et al. Risk stratification of thyroid nodules: assessing the suitability of ChatGPT for text-based analysis. *Am. J. Otolaryngol.* **45**, 104144 (2024).

91. Gobira, M. et al. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. *Rev. Assoc. Med. Bras. 1992*. **69**, e20230848 (2023).

92. Saibene, A. M. et al. Reliability of large language models in managing odontogenic sinusitis clinical scenarios: a preliminary multidisciplinary evaluation. *Eur. Arch. Oto-Rhino-Laryngol. J. Eur. Fed. Oto-Rhino-Laryngol. Soc. EUFOS Affil. Ger. Soc. Oto-Rhino-Laryngol. Head. Neck Surg.* **281**, 1835–1841 (2024).

93. Giannakopoulos, K., Kavadella, A., Aaqel Salim, A., Stamatopoulos, V. & Kaklamanos, E. G. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. *J. Med. Internet Res.* **25**, e51580 (2023).

94. Parasuraman, A., Berry, L. L. & Zeithaml, V. A. Refinement and reassessment of the SERVQUAL scale. *J. Retail.* **67**, 420 (1991).

95. Shoemaker, S. J., Wolf, M. S. & Brach, C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ. Couns*. **96**, 395–403 (2014).

96. Cheong, R. C. T. et al. Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. *Eur. Arch. Otorhinolaryngol.* **281**, 985–993 (2024).

97. Biggs, J. B. & Collis, K. F. *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)* (Academic Press, 2014).

98. Sinha, R. K., Roy, A. D., Kumar, N., Mondal, H. & Sinha, R. Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus* **15**, e35237 (2023).

99. Wang, R. Y. & Strong, D. M. Beyond accuracy: what data quality means to data consumers. *J. Manag. Inf. Syst.* **12**, 5–33 (1996).

100. Riedel, M. et al. ChatGPT's performance in German OB/GYN exams–paving the way for AI-enhanced medical education and clinical practice. *Front. Med*. **10**, 1296615 (2023).

101. Sallam, M., Barakat, M. & Sallam, M. METRICS: establishing a preliminary checklist to standardize design and reporting of artificial intelligence-based studies in healthcare. *JMIR Prepr*. **10**, (2023).

102. Sallam, M., Barakat, M. & Sallam, M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. *Cureus* **15**, e49373 (2023).

103. Charnock, D., Shepperd, S., Needham, G. & Gann, R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J. Epidemiol. Community Health* **53**, 105–111 (1999).

104. Seth, I. et al. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. *Aesthet. Surg. J. Open Forum.* **5**, ojad084 (2023).

105. Mu, X. et al. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. *Ski. Health Dis.* **4**, e313 (2024).

106. Xie, Y., Seth, I., Hunter-Smith, D. J., Rozen, W. M. & Seifman, M. A. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. *ANZ J. Surg.* **94**, 68–77 (2024).

107. Anastasio, A. T., Mills, F. B. IV, Karavan, M. P. Jr & Adams, S. B. Jr Evaluating the quality and usability of artificial intelligence–generated responses to common patient questions in foot and ankle surgery. *Foot Ankle Orthop.* **8**, 24730114231209919 (2023).

108. Chou, R. et al. AHRQ Series Paper 4: assessing harms when comparing medical interventions: AHRQ and the Effective Health-Care Program. *J. Clin. Epidemiol.* **63**, 502–512 (2010).

Acknowledgements

Author contributions

T.Y.C.T. and S.S. conceptualized, designed, and organized this study, analyzed the results, and wrote, reviewed, and revised the paper. S.K., A.V.S., K.P., K.R.M., H.O., and X.W. analyzed the results, and wrote, reviewed, and revised the paper. S.V., S.F., P.M., G.C., C.S., and Y.P. wrote, reviewed, and revised the paper. Y.W. conceptualized, designed, and directed this study, wrote, reviewed, and revised the paper.

Competing interests

P.M. has ownership and equity in BrainX, LLC, Y.W. has ownership and equity in BonafideNLP, LLC, and S.V. has ownership and equity in Kvatchii, Ltd., READE.ai, Inc., and ThetaRho, Inc. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01258-7.

Correspondence and requests for materials should be addressed to Yanshan Wang.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.