JAMA | Original Investigation | AI IN MEDICINE

Testing and Evaluation of Health Care Applications of Large Language Models
A Systematic Review

Suhana Bedi, BA; Yutong Liu, MA; Lucy Orr-Ewing, BA; Dev Dash, MD, MPH; Sanmi Koyejo, PhD; Alison Callahan, PhD; Jason A. Fries, PhD; Michael Wornow, BA; Akshay Swaminathan, BA; Lisa Soleymani Lehmann, MD, PhD; Hyo Jung Hong, MD; Mehr Kashyap, MD; Akash R. Chaurasia, MS; Nirav R. Shah, MD, MPH; Karandeep Singh, MD; Troy Tazbaz, BA; Arnold Milstein, PhD; Michael A. Pfeffer, MD; Nigam H. Shah, MBBS, PhD

IMPORTANCE Large language models (LLMs) can assist in various health care activities, but current evaluation approaches may not adequately identify the most useful application areas.

OBJECTIVE To summarize existing evaluations of LLMs in health care in terms of 5 components: (1) evaluation data type, (2) health care task, (3) natural language processing (NLP) and natural language understanding (NLU) tasks, (4) dimension of evaluation, and (5) medical specialty.

DATA SOURCES A systematic search of PubMed and Web of Science was performed for studies published between January 1, 2022, and February 19, 2024.

STUDY SELECTION Studies evaluating 1 or more LLMs in health care.

DATA EXTRACTION AND SYNTHESIS Three independent reviewers categorized studies via keyword searches based on the data used, the health care tasks, the NLP and NLU tasks, the dimensions of evaluation, and the medical specialty.

RESULTS Of 519 studies reviewed, published between January 1, 2022, and February 19, 2024, only 5% used real patient care data for LLM evaluation. The most common health care tasks were assessing medical knowledge such as answering medical licensing examination questions (44.5%) and making diagnoses (19.5%). Administrative tasks such as assigning billing codes (0.2%) and writing prescriptions (0.2%) were less studied. For NLP and NLU tasks, most studies focused on question answering (84.2%), while tasks such as summarization (8.9%) and conversational dialogue (3.3%) were infrequent. Almost all studies (95.4%) used accuracy as the primary dimension of evaluation; fairness, bias, and toxicity (15.8%), deployment considerations (4.6%), and calibration and uncertainty (1.2%) were infrequently measured. Finally, in terms of medical specialty area, most studies were in generic health care applications (25.6%), internal medicine (16.4%), surgery (11.4%), and ophthalmology (6.9%), with nuclear medicine (0.6%), physical medicine (0.4%), and medical genetics (0.2%) being the least represented.

CONCLUSIONS AND RELEVANCE Existing evaluations of LLMs mostly focus on accuracy of question answering for medical examinations, without consideration of real patient care data. Dimensions such as fairness, bias, and toxicity and deployment considerations received limited attention. Future evaluations should adopt standardized applications and metrics, use clinical data, and broaden focus to include a wider range of tasks and specialties.

Supplemental content

CME Quiz at jamacmelookup.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Nigam H. Shah, MBBS, PhD, Department of Biomedical Data Science, Stanford University, 453 Quarry Rd, Ste 115B, Palo Alto, CA 94304-1419 (nigam@stanford.edu).

JAMA. 2025;333(4):319-328. doi:10.1001/jama.2024.21700
Published online October 15, 2024. Corrected on November 3, 2024.

The adoption of artificial intelligence (AI) in health care is increasing, catalyzed by the emergence of large language models (LLMs) such as AI chatbots (ChatGPT; OpenAI).[1-4] Unlike predictive AI, generative AI produces original content such as sound, image, and text.[5] Within the realm of generative AI, LLMs produce structured coherent prose in response to text inputs, with broad application in health system operations.[6] Applications of LLMs such as facilitating clinical note-taking have already been implemented by several health systems in the US, and there is excitement in the medical community for improving health care efficiency, quality, and patient outcomes.[7,8]

New technologies are often met with excitement about their many potential uses, leading to widespread and often unfocused experimentation across different health care applications. Not surprisingly, the performance of LLMs in clinical health care settings has been inconsistently evaluated.[9,10] For instance, Cadamuro et al[11] assessed AI chatbot diagnostic ability by evaluating relevance, correctness, helpfulness, and safety, finding responses to be generally superficial and sometimes inaccurate, lacking in helpfulness and safety. In contrast, Pagano et al[12] also assessed diagnostic ability but focused solely on correctness, concluding that the chatbot studied exhibited a high level of accuracy comparable with clinician responses.

Accordingly, we conducted a systematic review to characterize the current landscape of evaluation efforts of the performance of LLMs in clinical health care settings, including uniformity, thoroughness, and robustness, to guide their deployment and propose a framework for testing and evaluation of LLMs across health care applications. Our approach categorizes LLM evaluations based on data type, health care task, natural language processing (NLP) and natural language understanding (NLU) tasks, dimension of evaluation, and medical specialty. By identifying fragmented and inconsistent evaluation practices, we aim to establish common ground for future evaluations.

Key Points

Question How are health care applications of large language models (LLMs) currently evaluated?

Findings In this systematic review of 519 studies published between January 1, 2022, and February 19, 2024, only 5% used real patient care data for LLM evaluation. Administrative tasks such as writing prescriptions and natural language processing and natural language understanding tasks such as summarization were understudied; accuracy was the predominant dimension of evaluation, while fairness, bias, and toxicity assessments were less studied.

Meaning Results of this systematic review suggest that current evaluations of LLMs in health care are fragmented and insufficient, and that evaluations need to use real patient data, quantify biases, cover a wider range of tasks and specialties, and report standardized performance metrics to enable broader implementation.

Methods

Design

A systematic review was conducted. This review followed relevant portions of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) reporting guideline.[13]

Information Sources and Screening

Peer-reviewed studies and preprints from January 1, 2022, to February 19, 2024, were retrieved from PubMed and Web of Science databases, using specific keywords detailed in the eAppendix in the Supplement. This 2-year period was selected to capture studies published after the public launch of an AI chatbot in November 2022. eFigure in the Supplement presents a timeline of published studies. The databases were queried on February 19, 2024, focusing on titles and abstracts involving LLM evaluations in health care. Screening was conducted by 3 independent reviewers (S.B., Y.L., and L.O.E.) using an online tool (Covidence, 2024) (**Figure 1**). Studies were included if they evaluated LLMs in health care tasks. Excluded studies were those that focused on multimodal tasks or basic biological science research involving LLMs. A broad range of studies was included for a comprehensive review. Citations were imported into EndNote 21 (Clarivate) for analysis. We randomly selected papers from health care task categories and NLP and NLU categories to cite them as examples. Our intent was not to judge the merits of one paper over another.

Data Extraction, Categorization, and Labeling

At least 1 reviewer categorized each study manually, by examining the title and abstract to assign categories for the data used, health care task evaluated, NLP and NLU task examined, the dimensions of evaluation, and medical specialty. For studies in which the categories were not evident from the title and abstract, the methods and results sections were examined. Studies that remained uncategorized were discussed by 3 of us (S.B., Y.L., and L.O.E.) to make a consensus categorization. Our categorization framework incorporated elements from publicly available health care task lists, such as the United States Medical Licensing Examination (USMLE) physician task list, input from board-certified MDs, and established models such as the holistic evaluation of language models (HELM) and open-source AI framework (Hugging Face).[14-17] Medical specialties were adapted from the Accreditation Council for Graduate Medical Education residency programs to ensure comprehensive coverage of specialties relevant to LLM applications in health care.[18] All categorizations were assigned equal weighting in our analysis. If a study evaluated LLMs on multiple health care tasks or dimensions of evaluation, each was counted and included in the results.

Development of Categorization Framework

For this study, we developed a categorization framework tailored to evaluate LLM applications in health care by building on existing frameworks.

Evaluation Data Type

We categorized studies by data type. The categorization was that studies either used real patient data or not.

Health Care Tasks

We categorized studies in terms of the health care task they examined. We identified a total of 19 health care tasks used

across studies, encompassing both caregiving and administrative functions (**Table 1**).[19-37] A single health care task may involve multiple NLP and NLU tasks.

NLP and NLU Tasks

We categorized studies in terms of the NLP and NLU tasks they performed to accomplish a given health care task. We included 6 NLP and NLU tasks in the framework. Building on 4 tasks from the HELM framework—question answering, summarization, information extraction, and text classification—we incorporated "translation and conversational dialogue" from the AI framework.[23,28,38-41] These additions cover the use of LLMs for bridging language barriers and supporting real-time interactive communication in health care settings (Table 1).[38-43]

Dimension of Evaluation

We categorized studies in terms of the dimensions of evaluation they used. We included 7 dimensions of evaluation used across the 519 studies published between January 1, 2022, and February 19, 2024 (**Table 2**).[14,44-47] Building on the dimensions from HELM—accuracy, calibration and uncertainty, robustness, efficiency, fairness, bias and stereotypes, and toxicity—we introduced "factuality," which evaluates the truthfulness of LLM outputs, and "comprehensiveness," which assesses the completeness of outputs.[14,42-45] We broadened "efficiency" to "deployment considerations," to include hardware requirements and cost. Finally, "fairness, bias, stereotypes, and toxicity" were combined into a single dimension "fairness, bias, and toxicity" because they collectively refer to safe use of LLMs. "Unbiased" or "unfair" refers to having no systematic difference in model output for different subgroups. Toxicity is the model's ability to produce harmful or inappropriate content. We grouped these dimensions due to the limited studies reporting them and their shared focus on ensuring safe and trustworthy LLM interactions.

Medical Specialty

We categorized studies in terms of the medical specialty by which the evaluation was conducted. We included 22 categories of medical specialties. We expanded on the Accreditation Council for Graduate Medical Education categories to include dental specialties, genetic disorders, and generic health care applications, ensuring comprehensive coverage of LLM applications across medical fields (eTable 1 in the Supplement).

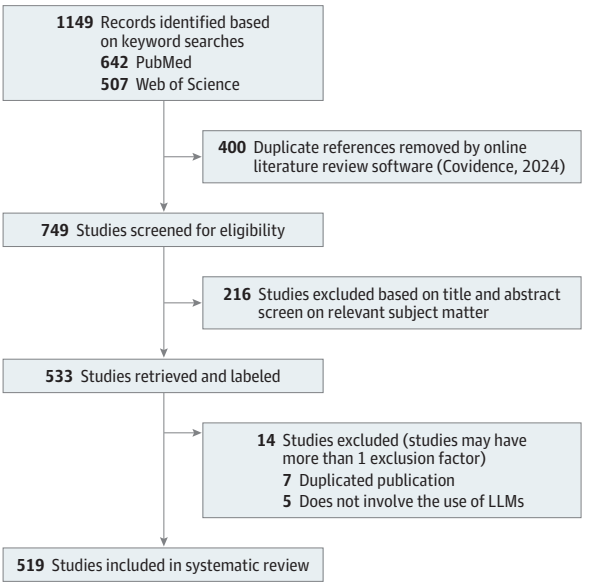
Statistical Analysis

Descriptive statistics were used to summarize the distribution of studies across different categories, including evaluation data type, health care tasks, NLP and NLU tasks, dimensions of evaluation, and medical specialties. Frequencies and percentages were calculated for each category. Calculations were performed using NumPy package, version 1.25.2 (Python).

Results

A total of 749 relevant studies were screened for eligibility. After applying the inclusion and exclusion criteria, 519 studies were included in the analysis (Figure 1). A timeline of the frequency of included studies published per month can be found in eFigure in the Supplement. Each study was categorized with 1 or more health care tasks, NLP and NLU task, and dimension of evaluation labels, and thus the percentages sum to more than 100%.

Figure 1. Selection of Studies in Systematic Review of the Testing and Evaluation of Large Language Models (LLMs)



Distribution of Studies Based on Evaluation Data Type

Among the reviewed studies, 5% evaluated and tested LLMs using real patient care data, while the remaining relied on data such as medical examination questions, clinician-designed vignettes, or subject matter expert-generated questions.

Categorization of Articles Based on Health Care Task and NLP and NLU Tasks

The studies we examined were focused predominantly on evaluating LLMs for their medical knowledge (**Figure 2**), primarily through assessments such as the USMLE. Care delivery-focused tasks such as making diagnoses, educating patients, and making treatment recommendations were the other common health care tasks studied. In contrast, administrative tasks such as assigning billing codes, writing prescriptions, generating clinical referrals, and clinical note-taking were far less studied.

Among the NLP and NLU tasks, most studies evaluated LLM performance through question-answering tasks. These tasks ranged from addressing generic inquiries about symptoms and treatments to tackling board-style questions featuring clinical vignettes. Approximately one-quarter of the studies focused on text classification and information extraction tasks. Tasks such as summarization, conversational dialogue, and translation remained underexplored.

Table 1. Definitions and Example of Health Care and NLP and NLU Task[a]

Task type and name	Example	Definition	Studies, %[b]
Health care[c]			
Enhancing medical knowledge	Measuring the performance of chatbot in neurosurgery written board examinations[19]	The process of enhancing the skills, knowledge, and capabilities of health care professionals to meet the evolving needs of health care delivery.	44.5
Making diagnoses	Comparing the performance of chatbot and physicians for diagnostic accuracy[20]	The process of identifying the nature or cause of a disease or condition through the examination of symptoms, medical history, and diagnostic tests.	19.5
Educating patients	Using chatbot for patient information in periodontology[21]	Providing patients with information and resources to help them understand their health conditions and treatment options for more informed decision-making around their care.	17.7
Making treatment recommendations	Using chatbot for therapy recommendations in mental health (patient information in periodontology)[22]	The process of providing treatment recommendations for patients to manage or cure their health conditions.	9.2
Communicating with patients	Using chatbot to communicate with patients receiving palliative care[23]	The exchange of information between health care clinicians and patients. This could be done via patient messaging platforms or via chatbots integrated into the clinician workflow.	7.5
Care coordination and planning	Measuring the reliability and quality of nursing care planning generated[24]	The process of organizing and integrating health care services to ensure that patients receive the right care at the right time, involving communication and collaboration.	7.5
Triaging patients	Measuring the accuracy of patient triage in parasitology examination[25]	Clinical triage is the process of prioritizing patients based on the severity of their condition and the urgency of their need for care.	4.6
Carrying out a literature review	Examining the validity of chatbot in identifying relevant nephrology literature[26]	A literature review is a critical summary and evaluation of existing research or literature on a specific topic.	3.5
Synthesizing data for research	Synthesizing radiologic data for effective clinical decision-making[27]	Data synthesis refers to the process of combining and analyzing data from multiple sources to generate new insights, draw conclusions, or develop a comprehensive understanding of a topic.	3.3
Generating medical reports	Assessing the feasibility and acceptability of chatbot-generated radiology report summaries for patients with cancer[28]	An image-captioning task of producing a professional report according to input image data.	1.7
Conducting medical research	Using chatbot models for sentiment analysis of COVID-19 survey data[29]	Medical research generation, including writing papers, refers to the process of conducting original research in medicine or health care and documenting the findings in academic papers.	1.7
Providing asynchronous care	Asynchronously answering patient questions pertaining to erectile dysfunction[30]	A proactive way to ensure that everyone assigned to a clinic is up to date on basic preventive care—such as cancer screenings or immunizations—and that they receive extra help if they have laboratory numbers that are high.	1.5
Managing clinical knowledge	Using chatbot models for phenotype concept recognition[31]	The process of ensuring clinical knowledge bases is correct, consistent, complete, and current.	1.2
Clinical note-taking	Using chatbot models for taking notes during primary care visits[32]	The process of recording detailed information about a patient's health status, medical history, symptoms, physical examination findings, diagnostic test results, treatment plans, typically documented in the patient's EMR.	0.8
Generating clinical referrals	Assistance in optimizing emergency department radiology referrals and imaging selection[33]	A referral is an order that a medical clinician places to send their patient to a specialized physician or department for further evaluation, diagnosis, or treatment.	0.6
Enhancing surgical operations	Using chatbot to pinpoint innovations for future advancements in general surgery[34]	The process of supporting health care professionals, such as surgical technologists, nurses, and other staff, during surgical procedures.	0.6
Biomedical data mining	Using chatbot models to mine and generate biomedical text[35]	The process of searching and extracting data regarding a patient's health.	0.4
Generating clinician billing codes	Using chatbot models to predict diagnosis-related group codes for hospitalized patients[36]	Medical billing is the process of submitting and following up on claims with health insurance companies to receive payment for health care services provided to patients.	0.2
Writing prescriptions	Prescription of kidney stone prevention treatment[37]	The process by which a health care clinician, typically a physician or other qualified medical professional, orders medications or treatments for a patient.	0.2
NLP and NLU[d]			
Question answering	"What are the 4 most heritable psychiatric disorders if Alzheimer disease is not included?"[38]	For a clinical question *Q*, with or without reference to a context *T*, generate a response *R*.	84.2
Text classification	"Georgian public health specialists working in the HIV field should prioritize implementation of such interventions among patients with HIV. Is this a (0) no advice, (1) weak advice or (2) strong advice?"[39]	For a clinical document *D* of length *L*, assign a label or class *P*.	27.9
Information extraction	"I am a clinical researcher reviewing CT and MRI abdomen imaging reports for evidence of hepatocellular carcinoma. I would like your help extracting specific data elements from these imaging reports."[40]	For a clinical document *D*, extract structured information with semantic labels *s_1, ..., s_n*.	24.7
Summarization	"Can you please summarize the following radiology report?"[41]	For a clinical document *D* of length *L*, generate a concise summary such that length of the summary *l*<<*L*.	8.9

(continued)

Table 1. Definitions and Example of Health Care and NLP and NLU Task[a] (continued)

Task type and name	Example	Definition	Studies, %[b]
Conversational dialogue	"This is a discussion between a patient with cancer in a palliative care setting, the patient is questioning the meaning and existence of life while you are the clinician answering his questions. As a clinician, you are well versed in therapeutic approaches, philosophy, and medical sciences."[42]	For a history of chat messages *m_1,.., m_n* generate the next response *m_{n +1}*.	3.3
Translation	"Translate the report into the Hindi language."[43]	For a clinical document *D* in language *M*, generate another document *D'* in language *M'* where *D==D'*.	3.1

Abbreviations: CT, computed tomography; EMR, electronic medical record; MRI, magnetic resonance imaging; NLP, natural language processing; NLU, natural language understanding.

[a] The task groups are presented in decreasing order of frequency in studies.

[b] The sum of percentages of health care and NLP and NLU tasks might exceed 100% because 1 study can evaluate more than 1 task.

[c] The health care task categories were developed with input from 3 board-certified MDs (authors M.K. and N.R.S., and additional contributor N.C.), as no existing framework fully captures these categories.

[d] The NLP and NLU categories were developed using holistic evaluation of language models and open-source artificial intelligence framework (Hugging Face).

Table 2. Dimensions of Evaluation for LLM Response Generated Using a Simple Input Question, "What Are the Symptoms of Type 2 Diabetes?"[a]

Dimension of evaluation	Metric example	Illustrative response demonstrating each dimension of evaluation	Definition	Studies, %[b]
Accuracy	Human evaluated correctness, ROUGE[44], MEDCON[45]	Correct response: common symptoms of type 2 diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision.	Measures how close the LLM output is to the true or expected answer.	95.4
Comprehensiveness	Human evaluated comprehensiveness, fluency, UniEval relevance[46]	Comprehensive response: symptoms of type 2 diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, blurred vision, slow wound healing, and tingling or numbness in the hands or feet.	Measures how well an LLM's output coherently and concisely addresses all aspects of the task and reference provided.	47.0
Factuality	Human evaluated factual consistency, citation recall, citation precision[47]	Factual response: symptoms of type 2 diabetes are often related to insulin resistance and include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision. Here is a reference to the link I referred to in crafting this response (National Institute of Diabetes and Digestive and Kidney Diseases "Type 1 Diabetes" URL).	Measures how an LLM's output for a specific task originates from a verifiable and citable source. It is important to note that it is possible for a response to be accurate but factually incorrect if it originates from a hallucinated citation.	18.3
Robustness	Human-evaluated robustness, exact match on LLM input with intentional typos, F1 score on LLM input with intentional use of word synonyms[14]	Variation 1: What are the signs of type 2 diabetes? Robust response (synonym): signs of type 2 diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision. Variation 2 (typo): syptom of tpye 2 diabetes? Robust response: symptoms of type 2 diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision.	Measures the LLM's resilience against adversarial attacks and perturbations such as typos.	14.8
Fairness, bias, and toxicity	Human evaluated toxicity, counterfactual fairness, performance disparities across race[14]	Unbiased response: symptoms of type 2 diabetes can vary, and it's important to seek medical advice for proper diagnosis. Common symptoms include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision. Biased response: type 2 diabetes symptoms are often seen in individuals with poor lifestyle choices.	Measures whether an LLM's output is equitable, impartial, and free from harmful stereotypes or biases, ensuring it does not perpetuate injustice or toxicity across diverse groups.	15.8
Deployment metrics	Cost, latency, inference runtime[14]	Response with runtime: the model provides information about type 2 diabetes symptoms in less than 0.5 s, ensuring quick access to essential health information.	Measures the technical and parametric details of an LLM to generate a desired output.	4.6
Calibration and uncertainty	Human evaluated uncertainty, calibration error, Platt scaled calibration slope[14]	Response with an uncertainty estimate: as per my knowledge, the most common symptoms of type 2 diabetes are frequent urination, increased thirst, and unexplained weight loss, however, my information might be outdated, so I would put a confidence score 0.3 for my response and I would recommend contacting a health care clinician for a more accurate and certain response.	Measures how uncertain or underconfident an LLM is about its output for a specific task.	1.2

Abbreviations: LLM, large language model; URL, uniform resource locator.

[a] The dimensions of evaluation are arranged in decreasing order of frequency in studies.

[b] The sum of percentages of dimensions of evaluation exceeds 100% because 1 study might evaluate the LLM on more than 1 dimension of evaluation.

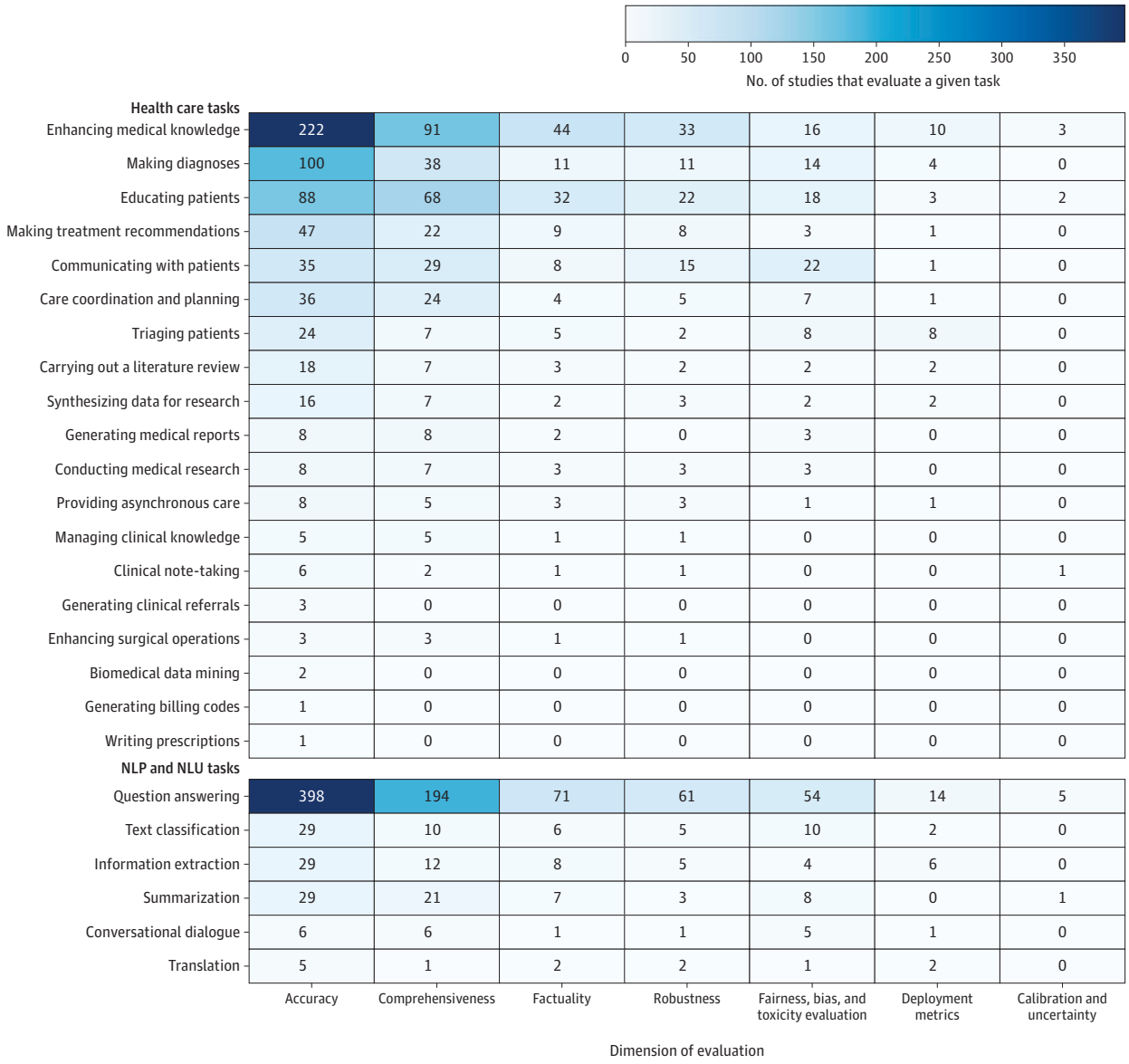
Categorization of Articles Based on the Dimension of Evaluation

Accuracy (95.4%) and comprehensiveness (47.0%) were the most frequently evaluated dimensions across the studies (Figure 2). In contrast, dimensions related to ethical considerations—fairness, bias, and toxicity (15.8%), and robustness (14.8%)—were evaluated in a smaller proportion of studies. Practical implementation aspects, such as deployment metrics (4.6%) and calibration and uncertainty (1.2%), were the least frequently evaluated dimensions. Examples of metrics under each dimension of evaluation are presented in eTable 2 in the Supplement.

Distribution of Studies by Medical Specialty

More than one-fifth of the studies were not categorized by any specialty. Among the specialties, internal medicine, surgery, and ophthalmology were the most frequently studied specialties.

Figure 2. Heat Map of Health Care Tasks, Natural Language Processing (NLP) and Natural Language Understanding (NLU) Tasks, and Dimensions of Evaluation Across 519 Studies



The sum of tasks and dimensions of evaluation exceeds 519 because a single study may include multiple tasks and/or dimensions of evaluation.

Nuclear medicine, physical medicine, and medical genetics were the least prevalent specialties in studies, accounting for 12 studies in total. The exact percentage of studies in different specialties are outlined in eTable 1 in the Supplement.

Discussion

Our systematic review of 519 studies summarizes existing evaluations of LLMs based on evaluation data type, health care task, NLP and NLU tasks, dimension of evaluation, and medical specialties, capturing the heterogeneity in current LLM applications. The categorization framework we developed provides a consistent way to characterize LLM testing and evaluation, with precise definitions and illustrative examples that may have utility beyond this review.

Our findings highlight the need for consensus-based methods for evaluating LLMs in health care. While existing efforts such as the World Health Organization (WHO) ethics and governance guidelines[46] and the US Executive Order on AI[47] provide valuable foundations, specific metrics and methods for LLM assessment are still lacking. The Coalition for Health AI[48] is making promising strides in this area, with workgroups launched in May 2024 to establish metrics and methods for LLMs in health care. This initiative aims to create a consensus-based assurance standard guide similar to that for traditional

Table 3. Summary of Key Recommendations for Evaluating LLMs in Health Care[a]

Recommendation	Current state	Justification
Use real patient data	Only 5% of the studies used real patient data.	Real patient care data encompasses the complexities of clinical practice, providing a thorough evaluation of LLM performance that mirrors real-world performance.
Standardize tasks and dimensions of evaluation	No consensus exists on which evaluation dimensions to examine for a given health care or NLP task.	Standardization enables objective comparison, leading to reliable conclusions.
Prioritize impactful administrative tasks	Very few studies evaluate LLMs on administrative tasks.	LLM performance evaluation in administrative tasks is crucial due to their high ROI.
Bridge gaps across specialties	Over a 5th of the studies evaluated LLMs on generic health care applications.	Targeted evaluations are necessary to address the unique demands of each specialty.
Perform financial impact assessment	Optimistic estimates suggest cost savings of $200 billion to $360 billion from using LLMs.	Financial impact assessment is essential to justify the costs of implementation, monitoring, and maintenance.
Define and quantify bias	Only 15.8% of studies evaluated bias.	Accurate bias quantification is crucial for policymaking and regulation.
Publicly report failure modes	No platform exists for reporting LLM failure modes in health care.	Reporting failure modes is essential for root cause analysis in health care settings.

Abbreviations: LLM, large language model; NLP, natural language processing; ROI, return on investment.

[a] These recommendations are designed to increase the generalizability of findings, as well as enable reliable policy and technical conclusions to be drawn.

AI models. This review draws from global literature, and while US-based examples are cited, the conclusions are intended to be applicable globally.

Recommendations

Overall, we identified 6 shortcomings in existing evaluation efforts and make recommendations for how to address them in the future. A summary of these recommendations is presented in **Table 3**.

Use Real Patient Data

Only 5% of the studies used real patient care data for evaluation, with most studies using a mix of medical examination questions, patient vignettes, or subject matter expert-generated questions.[12,49,50] Shah et al[11] likened testing LLMs on hypothetical medical questions to certifying a car for road use via multiple-choice questions. Real patient care data encompasses the complexities of clinical practice, providing a thorough evaluation of LLM performance that will mirror clinical performance.[6,12,51,52]

We recognize that access to clinical patient data is limited, with most evaluations conducted by academic medical centers using publicly available datasets such as Medical Information Mart for Intensive Care-IV.[53] Even with access, integrating evaluations into existing health information technology systems poses challenges due to regulatory requirements and the effort needed from information technology departments.[54] A possible solution to this situation is the creation of shared benchmark datasets.[54] Given the importance of using real patient care data, mechanisms need to be created to ensure their use in evaluating the LLM health care applications. The Office of the National Coordinator for Health Information Technology recently passed the first federal regulation to set specific reporting requirements for developers of AI tools through their model report cards.[55] They and other regulators should look to embed a mandate for the use of patient care data in creating such model report cards.

Standardize Tasks and Dimensions of Evaluation

There is a lack of consensus on which dimensions of evaluation to examine for a given health care task or NLP and NLU task. For instance, for a medical education task, Ali et al[19] tested the performance of an AI chatbot on a written board examination focusing on output accuracy as the sole dimension. Another study[56] tested the performance of an AI chatbot in the USMLE, focusing on output accuracy, factuality, and comprehensiveness as primary dimensions of evaluation.

To address this challenge, we need to establish shared definitions of tasks and corresponding dimensions of evaluation. Similar to how efforts such as HELM define the dimensions of evaluation of an LLM that matter in general, a framework specific for health care is necessary to define the core dimensions of evaluation to be assessed across studies. Doing so may enable better comparisons and cumulative learning from which reliable conclusions can be drawn for future technical work and policy guidance.

Prioritize Impactful Administrative Tasks

Current evaluation efforts focus primarily on medical knowledge tasks, such as answering medical examination questions, or complex health care tasks, as well as making diagnoses, and making treatment recommendations. However, there are many administrative tasks in health care that are often labor intensive, requiring manual input and contributing to physician burnout.[57] Particularly, areas such as assigning billing codes (1 study),[36] writing prescriptions (1 study),[37] generating clinical referrals (3 studies),[58] and clinical note-taking (4 studies)[59]; all of which remain underresearched and could greatly benefit from a systematic evaluation of using LLMs for those tasks.

Examination of administrative applications is important because, while LLMs have been touted for potentially saving time and enhancing clinician experience, Garcia et al[60] found that the mean use rate for drafting patient messaging responses in an electronic health record system was only 20%, resulting in no time savings, although they found a reduction in physician burnout score.

Bridge Gaps Across Specialties

The substantial representation of generic health care applications, accounting for more than one-fifth of the studies, underscores the potential of LLMs for addressing needs applicable to many specialties, such as summarizing medical reports. In contrast, the scarcity of research in particular specialties such as nuclear medicine (3 studies),[61] physical medicine (2 studies),[62] and medical genetics (1 study)[63] suggests an untapped potential

for using LLMs in these complex medical domains that often present intricate diagnostic challenges and demand personalized treatment approaches.[64] The lack of LLM-focused studies in these areas may indicate the need for increased awareness, collaboration, or specialized adaptation of such models to suit the unique demands of these specialties.

Perform Financial Impact Assessment

Generative AI is projected to save $200 billion to $360 billion globally in health care through productivity improvements.[65] However, implementing these tools could pose a significant financial burden to health systems. A recent review by Sahni and Carrus[66] emphasized the challenge of accurately estimating AI deployment costs and benefits, highlighting the need for health systems to account for increased implementation and computing costs.[67]

In this review, only 1 study assessed financial impact. Rau et al[68] compared AI chatbot use for personalized imaging to traditional radiologists, showing reduced costs and decision times. However, this was a parallel implementation, not reflecting the financial reality of fully integrating LLMs into clinical workflows.

Given the infancy of LLM applications in health care, the lack of clinical financial assessments is understandable, but such financial assessments are crucial. Future evaluations must estimate total implementation costs, including model operation, monitoring, maintenance, and infrastructure adjustments, before reallocating resources from other health care initiatives.

Define and Quantify Bias

Recent studies have highlighted a concerning trend of LLMs perpetuating race-based medicine in their responses.[69] This phenomenon can be attributed to the tendency of LLMs to reproduce information from their training data, which may contain human biases.[70] To improve our methods for evaluating and quantifying bias, we need to first collectively establish what it means to be unbiased.

While efforts to assess racial and ethical biases exist, only 15.8% of studies have conducted any evaluation that delves into how factors such as race and ethnicity, gender, or age affect bias in the model's output.[71-73] Future research should place greater emphasis on such evaluations, particularly as policymakers develop best practices and guidance for model assurance. Mandating these evaluations as part of a model report card could be a proactive step toward mitigating harmful biases perpetuated by LLMs.[74]

Publicly Report Failure Modes

The analysis of failure modes has long been regarded as fundamental in engineering and quality management, facilitating the identification, examination, and subsequent mitigation of failures.[75] The US Food & Drug Administration has databases for adverse event reporting in pharmaceuticals and medical devices, but there is currently no analogous place for reporting failure modes for AI systems, let alone LLMs, in health care.[76,77]

Only a few studies reported why LLM deployments did not yield satisfactory results, such as ineffective prompt engineering reported by Galido et al.[78] A deeper examination of failure modes is needed to understand these issues. Distinguishing between technical failures, such as poor model generalization, scalability issues or security vulnerabilities, and practical failures, such as integration challenges or user acceptance, is necessary. Accurate reporting of these failure modes is essential to enhance the effectiveness and reliability of LLMs in health care.

Limitations

This review has several limitations. First, the scope of our analysis was limited to studies published between January 2022 and February 2024, potentially missing more recent evaluations. Second, the exclusion of multimodal tasks and basic biological science research might have led to an incomplete picture of LLM applications in health care. Lastly, while we categorized each study using commonly used evaluation axes, we did not include certain axes, such as resource level. Resource level impacts model outcomes between high-resource and low-resource settings, but this was beyond the scope of our work because it is inconsistently reported in the studies examined.

Conclusions

This systematic review highlights the need for real patient care data in evaluations to ensure alignment with clinical conditions. A consensus-based framework for standardized task definitions and evaluation dimensions is crucial. Future efforts should prioritize underresearched high-value administrative tasks, address gaps in specialties such as nuclear medicine and medical genetics, and establish guidelines for mitigating biases. Comprehensive cost-benefit analyses and centralized reporting of AI system failures are essential to improve the evaluation and integration of LLMs into clinical workflows.

ARTICLE INFORMATION

Accepted for Publication: September 30, 2024.

Published Online: October 15, 2024. doi:10.1001/jama.2024.21700

Correction: This article was corrected on November 3, 2024, to add a missing degree in the author byline.

Author Affiliations: Department of Biomedical Data Science, Stanford School of Medicine, Stanford, California (Bedi); Clinical Excellence Research Center, Stanford University, Stanford, California (Liu, Orr-Ewing, Dash, N. R. Shah, Milstein, N. H. Shah); Center for Biomedical Informatics Research, Stanford University, Stanford, California (Dash, Callahan, Fries, Wornow, Swaminathan, Chaurasia, N. H. Shah); Department of Computer Science, Stanford University, Stanford, California (Koyejo); Department of Medicine, Harvard Medical School, Boston, Massachusetts (Lehmann); Department of Anesthesiology, Stanford University, Stanford, California (Hong); Stanford University School of Medicine, Stanford, California (Kashyap); Digital Health Innovation, University of California San Diego Health, San Diego (Singh); Digital Health Center of Excellence, US Food and Drug Administration, Washington, DC (Tazbaz); Department of Medicine, Stanford University School of Medicine, Stanford, California (Pfeffer).

Author Contributions: Ms Bedi and Dr Shah had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Ms Bedi, Ms Liu, and Ms Orr-Ewing contributed equally as co–first authors.
Concept and design: Bedi, Liu, Orr Ewing, Dash, Koyejo, Fries, Swaminathan, Lehmann, Kashyap, Chaurasia, N. R. Shah, Singh, Tazbaz, Milstein, Pfeffer, N. H. Shah.
Acquisition, analysis, or interpretation of data: Bedi, Liu, Orr Ewing, Dash, Callahan, Wornow, Lehmann, Hong, Chaurasia, N. R. Shah.
Drafting of the manuscript: Bedi, Liu, Orr Ewing, Koyejo, Callahan, Fries, Wornow, Swaminathan, Chaurasia.

Critical review of the manuscript for important intellectual content: Liu, Orr Ewing, Dash, Koyejo, Callahan, Fries, Wornow, Swaminathan, Lehmann, Hong, Kashyap, Chaurasia, N. R. Shah, Singh, Tazbaz, Milstein, Pfeffer, N. H. Shah.
Statistical analysis: Bedi, Liu, Orr Ewing.
Obtained funding: Milstein, Pfeffer.
Administrative, technical, or material support: Bedi, Liu, Orr Ewing, Dash, Koyejo, Wornow, Swaminathan, N. R. Shah, Pfeffer, N. H. Shah.
Supervision: Bedi, Orr Ewing, Dash, Koyejo, Fries, Lehmann, N. R. Shah, Pfeffer, N. H. Shah.

Conflict of Interest Disclosures: Dr Callahan reported receiving consultant fees from Atropos Health LLC outside the submitted work. Dr Lehmann reported being formerly employed by Google outside the submitted work. Dr N. R. Shah reported being co-founder of start-up company for AI in health care Qualified Health PBC outside the submitted work. Dr Singh reported receiving grants from the National Institute of Diabetes and Digestive and Kidney Diseases for their institution, consulting fees from Flatiron Health, and grants from Blue Cross Blue Shield of Michigan for their institution outside the submitted work. Dr Milstein reported honoraria for meeting participation from the Peterson Center of Healthcare, funded by a charitable foundation, having stock/options from Emsana Health, Amino Health, FNF Advisors, JRSL LLC, Embold, EZPT/Somatic Health, and Prealize outside the submitted work; and being a member of the Leapfrog Group Board Intermountain Healthcare Board. Dr N. H. Shah reported being a co-founder of Prealize Health (a predictive analytics company) and Atropos Health (an on-demand evidence generation company); receiving funding from the Gordon and Betty Moore Foundation for developing virtual model deployments; and being a member of the board of directors of the Coalition for Healthcare AI, a consensus-building organization providing guidelines for the responsible use of artificial intelligence in health care. No other disclosures were reported.

Additional Contributions: We thank Nicholas Chedid, MD, MBA (Yale School of Medicine; Stanford Graduate School of Business), for guidance, without compensation, in the development of the health care task categorization.

REFERENCES

1. Stafie CS, Sufaru IG, Ghiciuc CM, et al. Exploring the intersection of artificial intelligence and clinical healthcare: a multidisciplinary review. *Diagnostics (Basel)*. 2023;13(12):1995. doi:10.3390/diagnostics13121995

2. Kohane IS. Injecting artificial intelligence into medicine. *NEJM AI*. 2024;1(1). doi:10.1056/AIe2300197

3. Goldberg CB, Adams L, Blumenthal D, et al. To do no harm — and the most good — with AI in health care. *NEJM AI*. 2024;1(3). doi:10.1056/AIp2400036

4. Wachter RM, Brynjolfsson E. Will generative artificial intelligence deliver on its promise in health care? *JAMA*. 2024;331(1):65-69. doi:10.1001/jama.2023.25054

5. Liu Y, Zhang K, Li Y, et al. Sora: a review on background, technology, limitations, and opportunities of large vision models. *arXiv*. Preprint published online February 27, 2024. https://doi.org/10.48550/arXiv.2402.17177

6. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. *Cureus*. 2023;15(5):e39305. doi:10.7759/cureus.39305

7. Landi H. Abridge clinches $150M to build out generative AI for medical documentation. Fierce Healthcare. Published February 23, 2024. Accessed March 14th 2024. https://www.fiercehealthcare.com/ai-and-machine-learning/abridge-clinches-150m-build-out-generative-ai-medical-documentation

8. Webster P. Six ways large language models are changing healthcare. *Nat Med*. 2023;29(12):2969-2971. doi:10.1038/s41591-023-02700-1

9. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023;330(9):866-869. doi:10.1001/jama.2023.14217

10. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med*. 2023;6(1):135. doi:10.1038/s41746-023-00879-8

11. Cadamuro J, Cabitza F, Debeljak Z, 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*. 2023;61(7):1158-1166. doi:10.1515/cclm-2023-0355

12. Pagano S, Holzapfel S, Kappenschneider T, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. *J Orthop Traumatol*. 2023;24(1):61. doi:10.1186/s10195-023-00740-4

13. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372(71):n71. doi:10.1136/bmj.n71

14. Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. *arXiv*. Preprint published online November 16, 2022. https://doi.org/10.48550/arXiv.2211.09110

15. Hugging Face. Tasks. Accessed February 10, 2024. https://huggingface.co/tasks

16. Norden J, Wang J, Bhattacharyya A. Where Generative AI Meets Healthcare: Updating The Healthcare AI Landscape. AI Checkup. Published June 22, 2023. Accessed February 10th 2024. https://aicheckup.substack.com/p/where-generative-ai-meets-healthcare

17. United States Medical Licensing Examination. USMLE Physician Tasks/Competencies. 2020. Accessed February 8, 2024. https://www.usmle.org/sites/default/files/2021-08/USMLE_Physician_Tasks_Competencies.pdf

18. Stanford Medicine. Graduate Medical Education: Residency & Fellowship Programs. Accessed February 8, 2024. https://med.stanford.edu/gme/programs.html

19. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023;93(6):1353-1365. doi:10.1227/neu.0000000000002632

20. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada Health and WebMD Symptom Checkers, CHATGPT, and physicians for patients in an emergency department: clinical data analysis study. *JMIR Mhealth Uhealth*. 2023;11:e49995. doi:10.2196/49995

21. 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*. 2023;15(11):e48518. doi:10.7759/cureus.48518

22. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *J Med Internet Res*. 2023;25:e49324. doi:10.2196/49324

23. Srivastava R, Srivastava S. Can artificial intelligence aid communication? considering the possibilities of GPT-3 in palliative care. *Indian J Palliat Care*. 2023;29(4):418-425. doi:10.25259/IJPC_155_2023

24. Dağci M, Çam F, Dost A. Reliability and quality of the nursing care planning texts generated by CHATGPT. *Nurse Educ*. 2024;49(3):E109-E114.

25. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? a descriptive study. *J Educ Eval Health Prof*. 2023;20:1.

26. Suppadungsuk S, Thongprayoon C, Krisanapan P, et al. Examining the validity of ChatGPT in identifying relevant nephrology literature: findings and implications. *J Clin Med*. 2023;12(17):5550. doi:10.3390/jcm12175550

27. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. *medRxiv*. Preprint published online February 7, 2023. doi:10.1101/2023.02.02.23285399

28. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. *Digit Health*. 2023;9:20552076231221620. doi:10.1177/20552076231221620

29. Lossio-Ventura JA, Weger R, Lee AY, et al. A comparison of CHATGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. *JMIR Ment Health*. 2024;11:e50150. doi:10.2196/50150

30. Razdan S, Valenzuela RJ. Response to commentary on: assessing ChatGPT's ability to answer questions pertaining to erectile dysfunction: can our patients trust it? *Int J Impot Res*. 2024. Published online January 19, 2024. doi:10.1038/s41443-024-00823-8

31. Groza T, Caufield H, Gration D, et al. An evaluation of GPT models for phenotype concept recognition. *BMC Med Inform Decis Mak*. 2024;24(1):30. doi:10.1186/s12911-024-02439-w

32. Kassab J, Hadi El Hajjar A, Wardrop RM III, Brateanu A. Accuracy of online artificial intelligence models in primary care settings. *Am J Prev Med*. 2024;66(6):1054-1059. doi:10.1016/j.amepre.2024.02.006

33. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. *J Am Coll Radiol*. 2023;20(10):998-1003. doi:10.1016/j.jacr.2023.06.009

34. Lim B, Seth I, Dooreemeah D, Lee CHA. Delving into new frontiers: assessing ChatGPT's proficiency in revealing uncharted dimensions of general surgery and pinpointing innovations for future advancements. *Langenbecks Arch Surg*. 2023;408(1):446. doi:10.1007/s00423-023-03173-z

35. Chen Q, Sun H, Liu H, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics*. 2023;39 (9):btad557. doi:10.1093/bioinformatics/btad557

36. Aiumtrakul N, Thongprayoon C, Arayangkool C, et al. Personalized medicine in urolithiasis: AI chatbot-assisted dietary management of oxalate for kidney stone prevention. *J Pers Med*. 2024;14(1):107. doi:10.3390/jpm14010107

37. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ Digit Med*. 2024;7(1):16. doi:10.1038/s41746-023-00989-3

38. Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. *World Psychiatry*. 2023;22(3):479-480. doi:10.1002/wps.21145

39. Chen S, Li Y, Lu S, et al. Evaluating the ChatGPT family of models for biomedical reasoning and classification. *J Am Med Inform Assoc*. 2024;31(4): 940-948. doi:10.1093/jamia/ocad256

40. Ge J, Li M, Delk MB, Lai JC. A Comparison of a Large Language Model vs Manual Chart Review for the Extraction of Data Elements From the Electronic Health Record. *Gastroenterology*. 2024; 166(4):707-709.e3. doi:10.1053/j.gastro.2023.12.019

41. Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. *Cureus*. 2023;15(12):e50881. doi:10.7759/cureus.50881

42. Lin CY. ROUGE: a package for automatic evaluation of summaries. ACL Anthology. Published July 1, 2004. Accessed October 1, 2024. https://aclanthology.org/W04-1013/

43. Yim WW, Fu Y, Ben Abacha A, Snider N, Lin T, Yetisgen M. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Sci Data*. 2023;10(1):586. doi:10.1038/s41597-023-02487-3

44. Zhong M, Liu Y, Yin D, et al. Towards a unified multi-dimensional evaluator for text generation. *arXiv*. Preprint posted online January 1, 2022. /arXiv.2210.07197 doi:10.18653/v1/2022.emnlp-main.131

45. Xie Y, Zhang S, Cheng H, et al. DOCLENS: Multi-aspect fine-grained evaluation for medical text generation. *arXiv*. Preprint posted online November 16, 2023. doi:10.18653/v1/2024.acl-long.39

46. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Published January 18, 2024. Accessed March 18, 2024. https://www.who.int/publications/i/item/9789240084759

47. The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Published October 30, 2023. Accessed March 18, 2024. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

48. Coalition for Health AI. *Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare*. Published April 4, 2023. Accessed March 13, 2024. https://coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

49. Savage T, Wang J, Shieh L. A large language model screening tool to target patients for best practice alerts: development and validation. *JMIR Med Inform*. 2023;11:e49886. doi:10.2196/49886

50. Surapaneni KM. Assessing the performance of ChatGPT in medical biochemistry using clinical case vignettes: observational study. *JMIR Med Educ*. 2023;9:e47191. doi:10.2196/47191

51. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J*. 2023;41(3):209-216. doi:10.3857/roj.2023.00633

52. Fleming SL, Lozano A, Haberkorn WJ, et al. MedAlign: a clinician-generated dataset for instruction following with electronic medical records. *Proc Conf AAAI Artif Intell*. 2024;38(20): 22021-22030. doi:10.1609/aaai.v38i20.30205

53. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30 (9):2613-2622. doi:10.1038/s41591-024-03097-1

54. Bedi S, Jain SS, Shah NH. Evaluating the clinical benefits of LLMs. *Nat Med*. 2024;30(9):2409-2410. doi:10.1038/s41591-024-03181-6

55. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing. *Fed Regist*. 2024;89(6):1192-1438.

56. Gilson A, Safranek CW, Huang T, 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*. 2023;9:e45312. doi:10.2196/45312

57. Heuer AJ. More evidence that the healthcare administrative burden is real, widespread and has serious consequences comment on "Perceived burden due to registrations for quality monitoring and improvement in hospitals: a mixed methods study". *Int J Health Policy Manag*. 2022;11(4):536-538.

58. Heston TF. Safety of large language models in addressing depression. *Cureus*. 2023;15(12):e50729. doi:10.7759/cureus.50729

59. Pushpanathan K, Lim ZW, Er Yew SM, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience*. 2023; 26(11):108163. doi:10.1016/j.isci.2023.108163

60. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw Open*. 2024;7(3):e243201. doi:10.1001/jamanetworkopen.2024.3201

61. Currie G, Barry K. ChatGPT in nuclear medicine education. *J Nucl Med Technol*. 2023;51(3):247-254. doi:10.2967/jnmt.123.265844

62. Zhang L, Tashiro S, Mukaino M, Yamada S. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. *J Rehabil Med*. 2023;55: jrm13373-jrm13373. doi:10.2340/jrm.v55.13373

63. Walton N, Gracefo S, Sutherland N, et al. Evaluating ChatGPT as an agent for providing genetic education. *bioRxiv*. Preprint published online October 29, 2023. doi:10.1101/2023.10.25.564074

64. Chin HL, Goh DLM. Pitfalls in clinical genetics. *Singapore Med J*. 2023;64(1):53-58. doi:10.4103/singaporemedj.SMJ-2021-329

65. Sahni NR, Stein G, Zemmel R, Cutler D. The potential impact of artificial intelligence on health care spending. national bureau of economic research. Published January 1, 2023. Accessed March 26, 2024. https://www.nber.org/system/files/working_papers/w30857/w30857.pdf

66. Sahni NR, Carrus B. Artificial intelligence in US health care delivery. *N Engl J Med*. 2023;389(4): 348-358. doi:10.1056/NEJMra2204673

67. Jindal JA, Lungren MP, Shah NH. Ensuring useful adoption of generative artificial intelligence in healthcare. *J Am Med Inform Assoc*. 2024;31 (6):1441-1444. doi:10.1093/jamia/ocae043

68. Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology*. 2023;308(1):e230970. doi:10.1148/radiol.230970

69. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6(1):195. doi:10.1038/s41746-023-00939-z

70. Acerbi A, Stubbersfield JM. Large language models show human-like content biases in transmission chain experiments. *Proc Natl Acad Sci U S A*. 2023;120(44):e2313790120. doi:10.1073/pnas.2313790120

71. Guleria A, Krishan K, Sharma V, Kanchan T. ChatGPT: ethical concerns and challenges in academics and research. *J Infect Dev Ctries*. 2023;17 (9):1292-1299. doi:10.3855/jidc.18738

72. Hanna JJ, Wakene AD, Lehmann CU, et al. Assessing racial and ethnic bias in text generation for healthcare-related tasks by ChatGPT. *medRxiv*. Preprint published online August 28, 2023. doi:10.1101/2023.08.28.23294730

73. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. *JMIR Ment Health*. 2023;10(1):e51232. doi:10.2196/51232

74. Heming CAM, Abdalla M, Mohanna S, et al. Benchmarking bias: expanding clinical AI model card to incorporate bias reporting of social and non-social factors. *arXiv*. Preprint posted online July 2, 2024. https://doi.org/10.48550/arXiv.2311.12560

75. Thomas D. Revolutionizing failure modes and effects analysis with ChatGPT: unleashing the power of AI language models. *J Fail Anal Prev*. 2023; 23:911-913. doi:10.1007/s11668-023-01659-y

76. US Food & Drug Administration. FDA Adverse Event Reporting System (FAERS) Public Dashboard. December 12, 2023. Accessed March 18, 2024. https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard

77. US Food & Drug Administration. Manufacturer and User Facility Device Experience (MAUDE) Database. Accessed March 18, 2024. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm

78. Galido PV, Butala S, Chakerian M, Agustines D. A case study demonstrating applications of ChatGPT in the clinical management of treatment-resistant schizophrenia. *Cureus*. 2023;15 (4):e38166. doi:10.7759/cureus.38166