medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs)

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

0. Key Points

- **Question:** How are healthcare applications of large language models (LLMs) currently evaluated?
- **Findings:** Studies rarely used real patient care data for LLM evaluation. Administrative tasks such as generating provider billing codes and writing prescriptions were understudied. Natural Language Processing (NLP)/Natural Language Understanding (NLU) tasks like summarization, conversational dialogue, and translation were infrequently explored. Accuracy was the predominant dimension of evaluation, while fairness, bias and toxicity assessments were neglected. Evaluations in specialized fields, such as nuclear medicine and medical genetics were rare.
- **Meaning:** Current LLM assessments in healthcare remain shallow and fragmented. To draw concrete insights on their performance, evaluations need to use real patient care data across a broad range of healthcare and NLP/NLU tasks and medical specialties with standardized dimensions of evaluation.

1. Abstract

Importance: Large Language Models (LLMs) can assist in a wide range of healthcare-related activities. Current approaches to evaluating LLMs make it difficult to identify the most impactful LLM application areas.

Objective: To summarize the current evaluation of LLMs in healthcare in terms of 5 components: evaluation data type, healthcare task, Natural Language Processing (NLP)/Natural Language Understanding (NLU) task, dimension of evaluation, and medical specialty.

Data Sources: A systematic search of PubMed and Web of Science was performed for studies published between 01-01-2022 and 02-19-2024.

Study Selection: Studies evaluating one or more LLMs in healthcare.

Data Extraction and Synthesis: Three independent reviewers categorized 519 studies in terms of data used in the evaluation, the healthcare tasks (the what) and the NLP/NLU tasks (the how) examined, the dimension(s) of evaluation, and the medical specialty studied.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Results:
Only 5% of reviewed studies utilized real patient care data for LLM evaluation. The most popular healthcare tasks were assessing medical knowledge (e.g. answering medical licensing exam questions, 44.5%), followed by making diagnoses (19.5%), and educating patients (17.7%). Administrative tasks such as assigning provider billing codes (0.2%), writing prescriptions (0.2%), generating clinical referrals (0.6%) and clinical notetaking (0.8%) were less studied. For NLP/NLU tasks, the vast majority of studies examined question answering (84.2%). Other tasks such as summarization (8.9%), conversational dialogue (3.3%), and translation (3.1%) were infrequent. Almost all studies (95.4%) used accuracy as the primary dimension of evaluation; fairness, bias and toxicity (15.8%), robustness (14.8%), deployment considerations (4.6%), and calibration and uncertainty (1.2%) were infrequently measured. Finally, in terms of medical specialty area, most studies were in internal medicine (42%), 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 focused on accuracy of question answering for medical exams, without consideration of real patient care data. Dimensions like fairness, bias and toxicity, robustness, and deployment considerations received limited attention. To draw meaningful conclusions and improve LLM adoption, future studies need to establish a standardized set of LLM applications and evaluation dimensions, perform evaluations using data from routine care, and broaden testing to include administrative tasks as well as multiple medical specialties.

Keywords: Large Language Models, Generative Artificial Intelligence, Healthcare, Dimensions of Evaluation, Evaluation Metrics.

2. Introduction

The adoption of Artificial Intelligence (AI) in healthcare is rising, catalyzed by the emergence of Large Language Models (LLMs) like OpenAI's ChatGPT [1 2 3 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]. Prominent applications such as facilitating clinical note-taking have already been implemented by several health systems in the U.S., and there is excitement in the medical community for improving healthcare efficiency, quality, and patient outcomes [7 8]. A recent report estimates that LLMs could unlock a substantial portion of the $1 trillion in untapped healthcare efficiency improvements, including an estimated savings ranging from 5 to 10 percent of US healthcare spending or approximately $200 billion to $360 billion annually based on 2019 figures [9 10].

New and revolutionary technologies are often met with excitement about their many potential uses, leading to widespread and often unfocussed experimentation across different healthcare applications. Thus, as expected, the performance of LLMs in real-world healthcare settings too, remains inconsistently conducted and evaluated [11 12]. For instance, Cadamuro et al. assessed ChatGPT-

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

4's diagnostic ability by evaluating relevance, correctness, helpfulness, and safety, finding responses to be generally superficial and sometimes inaccurate, lacking in helpfulness and safety [13]. In contrast, Pagano et al. also assessed diagnostic ability, but focused solely on correctness, concluding that ChatGPT-4 exhibited a high level of accuracy comparable to clinician responses [14].Thus, we hypothesize that the current evaluation landscape lacks the uniformity, thoroughness, and robustness necessary to effectively guide the deployment of LLMs in a real-world setting.

Our research aimed to assess the evaluation landscape of LLMs' healthcare applications with the goal of accelerating their time to impact and successful integration. Through a systematic review of 519 studies, we comprehensively categorized how LLMs have been evaluated in healthcare along 5 axes: evaluation data type used, healthcare task, NLP/NLU task, dimension of evaluation, and medical specialty. To enable the categorization of the diverse range of applications and their evaluation setups, we used two categorization frameworks: the first describes healthcare applications of LLMs in terms of their constituent healthcare and NLP/NLU tasks, and the second describes dimensions of evaluation and associated metrics. These frameworks were then applied systematically to characterize the current state of evaluations to quantify the variability in LLM application evaluations and identify areas for further exploration.

Among the studies reviewed, there was no existing categorization framework that was consistently used. This lack of standardization required the creation of categorization framework, where we define each category as well as show illustrative examples of each. While such categorization can be a limitation, given that it provides a way to consistently discuss the testing and evaluations of LLM applications, it may have use beyond this review.

Our results show that evaluations of LLM applications in healthcare have been unevenly distributed both in terms of dimensions of evaluation used and in terms of medical specialty and application.

3. Methods

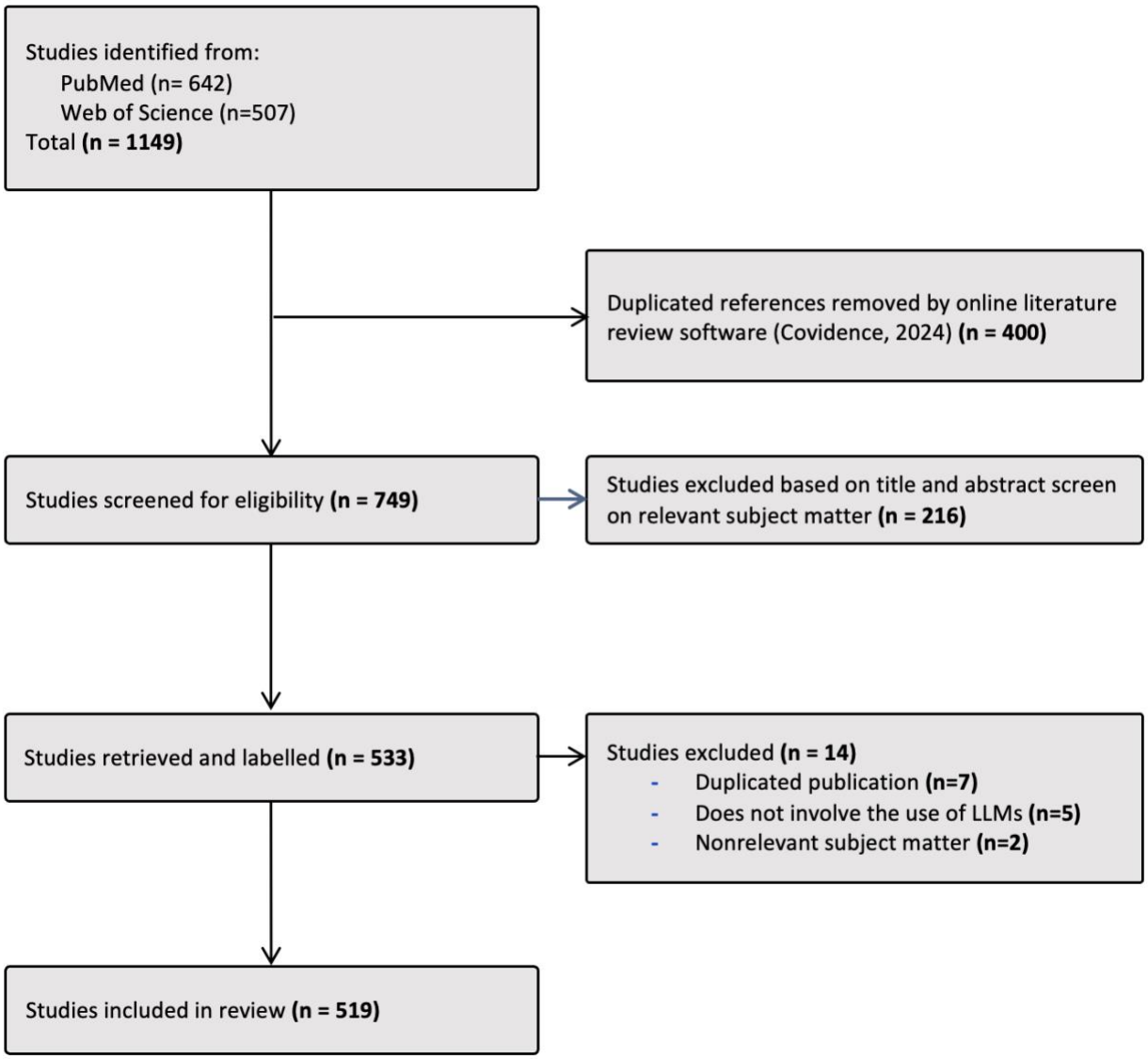
3.1 Design

A systematic review was conducted following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines as shown in **Figure 1** [15]**.**

Figure 1: PRISMA Flow Diagram

This diagram shows the process of screening and selecting the categorized 519 studies.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



3.2 Information sources

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 as detailed in **Supplement 1.** Our search focused on titles and abstracts to identify studies on evaluation of LLMs' healthcare applications. This two-year period aimed to capture publications evaluating LLM healthcare applications since the public launch of ChatGPT in November 2022. Given our hypothesis that the current landscape lacks the necessary elements needed to truly assess LLM performance in healthcare, we included a broad spectrum of studies. Citations were imported into EndNote 21 (Clarivate) for analysis.

3.3 Categorization framework

Each study was categorized by evaluation data type, healthcare task, NLP/NLU task, dimension of evaluation, and medical specialty. Healthcare task categories were developed using publicly

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

available healthcare task and competency lists and were refined by consulting board-certified MDs [16 17] as outlined in **Table 1.** NLP/NLU categories and dimension of evaluations were developed using the Holistic Evaluation of Language Models (HELM) and Hugging Face frameworks [18 19] as shown in **Tables 2** and **3.** Medical specialties were adapted from Accreditation Council for Graduate Medical Education (ACGME) residency programs. [20]

Table 1: Healthcare task definitions and examples

This table lists the range of healthcare tasks that the 519 studies were categorized into, with definition and example for each task category.

Healthcare Tasks	Definition	Example
Enhancing medical knowledge	The process of enhancing the skills, knowledge, and capabilities of healthcare professionals to meet the evolving needs of healthcare delivery.	Measuring the performance of GPT on Neurosurgery Written Board examinations (Ali et al)[21]
Making diagnoses	The process of identifying the nature or cause of a disease or condition through the examination of symptoms, medical history, and diagnostic tests.	Comparing the performance of GPT and Physicians for diagnostic accuracy (Fraser et al)[22]
Educating patients	Providing patients with information and resources to help them understand their health conditions, treatment options etc. for more informed decision-making around their care.	Using GPT for Patient Information in periodontology (Babayiğit et al)[23]
Making treatment recommendations	The process of providing treatment recommendations for patients to manage or cure their health conditions.	Using GPT for therapy recommendations in mental health (Patient Information in Periodontology (Wilhelm et al)[24]
Communicating with patients	The exchange of information between healthcare providers and patients. This could be done via patient messaging platforms, or via Chatbots integrated into the provider workflow.	Using GPT to communicate with palliative care patients (Srivastava et al)[25]
Care coordination and planning	The process of organizing and integrating healthcare services to ensure that patients receive the right care at the right time, involving communication and collaboration	Measuring the reliability and quality of nursing care planning generated (Dağcı et al)[26]
Triaging patients	Clinical triage is the process of prioritizing patients based on the severity of their condition and the urgency of their need for care.	Measuring the accuracy of patient triage in parasitology examination (Huh)[27]
Carrying out a literature review	A literature review is a critical summary and evaluation of existing research or literature on a specific topic.	Examining the validity of ChatGPT in identifying relevant Nephrology literature (Suppadungsuk et al)[28]
Synthesizing data for research	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.	Synthesizing radiologic data for effective clinical decision-making (Rao et al)[29]
Generating clinical referrals	A referral is an order that a medical provider places to send their patient to a specialized physician or department for further evaluation, diagnosis, or treatment.	Assistance in optimizing Emergency Department radiology referrals and imaging selection (Barash et al)[30]
Generating medical reports	An image-captioning task of producing a professional report according to input image data.	Assessing the feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients (Chung et al)[31]
Managing clinical knowledge	The process of ensuring clinical knowledge bases is correct, consistent, complete, and current.	Using GPT models for phenotype concept recognition (Groza et al)[32]

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Providing asynchronous care	A proactive way to ensure that everyone assigned to a clinic is up to date on basic preventive care - like cancer screenings or immunizations - and that they receive extra help if they have lab numbers that are high.	Asynchronously answering patient questions pertaining to erectile dysfunction (Razdan et al)[33]
Clinical note-taking	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	Using GPT models for taking notes during primary care visits (Kassab et al)[34]
Enhancing surgical operations	The process of supporting healthcare professionals, such as surgical technologists, nurses, and other staff, during surgical procedures	Using GPT to pinpoint innovations for future advancements in general surgery (Lim et al)[35]
Conducting medical research	Medical research generation, including writing papers, refers to the process of conducting original research in medicine or healthcare and documenting the findings in academic papers.	Using GPT models for sentiment analysis of COVID-19 survey data (Lossio-Ventura et al)[36]
Biomedical data mining	The process of searching and extracting data regarding a patient's health	Using GPT models to mine and generate biomedical text (Chen et al)[37]
Generating provider billing codes	Medical billing is the process of submitting and following up on claims with health insurance companies to receive payment for healthcare services provided to patients.	Using GPT models to predict diagnosis-related group (DRG) codes for hospitalized patients (Wang et al)[38]
Writing prescriptions	The process by which a healthcare provider, typically a physician or other qualified medical professional, orders medications or treatments for a patient	Prescription of kidney stone prevention treatment (Alumtrakul et al)[39]

Table 2: Definition of NLP/NLU tasks

This table lists the range of NLP/NLU tasks that the 519 studies were categorized into, with definition and examples for each task category.

NLP/NLU Task	Definition	Examples
Summarization	For a clinical document ***D*** of length ***L***, generate a concise summary such that length of the summary ***l << L***.	"Summarize the impression section of a radiology report"
Question Answering	For a clinical question ***Q***, with or without reference to a context ***T***, generate a response ***R***.	- "What are the symptoms of Type 2 diabetes?" - “What is the recommended dosage of ibuprofen for a 40 year-old male with mild fever?”
Information Extraction	For a clinical document ***D***, extract structured information with semantic labels ***s_1, ..., s_n.***	"Extract the mentions of adverse drug events, disease exacerbations and surgical interventions from a patient's history."
Text Classification	For a clinical document ***D*** of length ***L***, assign a label or class ***P***	"Categorize clinical notes into classes such as 'diagnosis', 'treatment' or 'prognosis'
Translation	For a clinical document ***D*** in language ***M***, generate another document ***D'*** in language ***M'*** where ***D == D'***	"Translate a patient's old lab test results from Spanish to English"
Conversational Dialogue	For a history of chat messages ***m_1,.., m_n*** generate the next response ***m_{n+1}***	"Using a patient's history of chat messages, help them reschedule their appointment"

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3: Dimensions of evaluation for LLM response

This table lists the range of dimensions of evaluations that the 519 studies were categorized into, with definitions, metrics and reviewer-generated example responses where each dimension is evaluated for a simple input question, "What are the symptoms of Type 2 diabetes?"

Dimension of Evaluation	Definition	Metric Examples	Illustrative response demonstrating each dimension of evaluation
Accuracy	Measures how close the LLM output is to the true or expected answer	Human evaluated correctness, ROUGE, MEDCON	**Correct Response** - "Common symptoms of Type 2 Diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision."
Calibration and Uncertainty	Measures how uncertain or underconfident an LLM is about its output for a specific task	Human evaluated uncertainty, calibration error, Platt scaled calibration slope	**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 healthcare provider for a more accurate and certain response."
Robustness	Measures the LLM's resilience against adversarial attacks and perturbations like typos.	Human evaluated robustness, exact match on LLM input with intentional typos, F1 on LLM input with intentional use of word synonyms	**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."
Factuality	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	Human evaluated factual consistency, citation recall, citation precision	**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 - https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-1-diabetes"
Comprehensiveness	Measures how well an LLM's output coherently and concisely addresses	Human evaluated comprehensiveness, fluency, UniEval relevance	**Comprehensive Response**: "Symptoms of Type 2 Diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, blurred vision, slow

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

	all aspects of the task and reference provided		wound healing, and tingling or numbness in the hands or feet. "
Fairness, bias and toxicity	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	Human evaluated toxicity, counterfactual fairness, performance disparities across race	**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."
Deployment considerations	Measures the technical and parametric details of an LLM to generate a desired output	Cost, latency, inference runtime	**Response with runtime**: "The model provides information about Type 2 Diabetes symptoms in less than 0.5 second, ensuring quick access to essential health information."

3.4 Eligibility criteria and screening

Screening was conducted by SB, YL, and LOE using the Covidence software (Covidence, 2024) as outlined in **Figure 1.** Included studies used LLMs for healthcare tasks and evaluated their performance. Excluded articles were those focused on multimodal tasks or basic biological science research with LLMs.

3.5 Data extraction and labeling

We adopted a paired review approach, wherein each study was categorized into evaluation data type, healthcare tasks, NLP/NLU tasks, dimension(s) of evaluation, and medical specialty by at least one human reviewer (SB, YL, or LOE) and GPT-4, based on the title and abstract. Note that GPT-4 was used as a force multiplier while the final categories were assigned by the human reviewers. In instances of disagreements regarding category assignments, the methods sections of the studies were retrieved, and final categories were determined through reviewer consensus. The prompts given to GPT-4 can be found in **Supplement 2.**

Each study received one or more healthcare tasks, NLP/NLU task, and dimension of evaluation labels as appropriate, hence the percentages sum above 100% in **Table 4**. In addition, each study could be assigned more than one medical specialty based on the evaluation conducted.

4. Results

749 relevant studies were screened for eligibility. After applying the inclusion and exclusion criteria described in **3.4**, 519 studies were included in the analysis using the frameworks developed by the authors.

4.1 Categorization framework for healthcare tasks, NLP/NLU tasks and dimensions of evaluation

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We deconstructed each healthcare application of an LLM into its constituent healthcare task (**Table 1**), i.e. the clinical and non-clinical task it is used for (the "what"), and the NLP/NLU task (**Table 2**), i.e. the language processing task being performed (the "how"). Examples of a healthcare task are diagnosing a patient's disease, recommending a treatment for osteoarthritis. Examples of the language-processing job to be accomplished – which is not necessarily specific to the medical domain are summarizing the impression section of a radiology report, answering questions about the symptoms of type 2 diabetes etc.

An example of how healthcare tasks and NLP/NLU combine for a healthcare application of LLM is how Gan et al. evaluated LLM performance for mass-casualty triaging [40]. The healthcare task (the "what") is triaging patients while the NLP/NLU tasks (the "how") are information extraction (extracting detailed patient information from the triage questionnaire scenarios, including age, symptoms, and vital signs), text classification (classifying the triage questionnaire scenarios into different triage levels), and question answering (generating final decision responses to the triage questionnaire).

We initially compiled a list of healthcare tasks using publicly available resources [16] [17]. Subsequently, through consultation with three board-certified MDs, we refined the list through iterative discussions to establish the final categories for classification, as outlined in **Table 1.** To compile a list of common NLP/NLU tasks, we referred to sources such as the Holistic Evaluation of Language Models (HELM) study and the Hugging Face task framework to derive 6 categories: 1) Summarization, 2) Question answering, 3) Information extraction, 4) Text classification (such as clinical notes, research articles, and documents), 5) Translation, and 6) Conversational dialogue (**Table 2**) [18] [19].

We categorized the most common dimensions of evaluation used in the reviewed studies based on the list outlined in **Table 3**. These dimensions include: 1) Accuracy, 2) Calibration and uncertainty, 3) Robustness, 4) Factuality, 5) Comprehensiveness, 6) Fairness, bias, and toxicity, and 7) Deployment considerations. Fairness, bias, and toxicity were grouped together for ease of analysis, due to their infrequent occurrence in the reviewed studies, and relevance to ethical evaluation of LLMs. Additionally, we compiled common metrics for each dimension (**eTable 1**) to serve as a starting framework for researchers designing studies to assess LLM performance in healthcare applications.

4.2 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 (SME) generated questions.

4.3 Categorizing articles based on healthcare tasks and NLP/NLU tasks

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The studies we examined had a predominant focus on evaluating LLMs for their medical knowledge (**Table 4)**, primarily through assessments such as the USMLE. This trend assumes that because we assess medical professionals' readiness for entering clinical practice through board-style examinations, mirroring this type of evaluation for LLMs is adequate to certify their fitness-for-use. Making diagnoses, educating patients and making treatment recommendations were the other common healthcare tasks studied. While these tasks represent critical aspects of healthcare delivery, validating the utility of LLMs in supporting them requires assessment with real patient care data. The limited examination of administrative tasks like assigning provider billing codes, writing prescriptions, generating clinical referrals, and clinical notetaking suggests a gap in studying LLMs' use for high-value, immediately impactful administrative tasks. These tasks are often labor intensive, presenting a ripe opportunity for testing LLMs to enhance efficiency in these areas [41].

Among the NLP/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 a quarter of the studies focused on text classification and information extraction tasks. Tasks such as summarization, conversational dialogue, and translation remained underexplored. This gap is significant because condensing patient records into concise summaries, translating medical content into simpler languages or the patient's native language, and facilitating conversations through chatbots are often touted benefits of using LLMs and could substantially alleviate physician burden.

4.4 Categorizing articles based on the dimensions of evaluation

As seen in **Table 4**, accuracy and comprehensiveness were overwhelmingly the top two most examined dimensions, whereas factuality, fairness, bias, and toxicity, robustness, deployment considerations, and calibration and uncertainty were infrequently assessed. This suggests a potential gap in assessing the broader capabilities and suitability of LLMs for real-world deployment. While accuracy and comprehensiveness are crucial for ensuring the reliability and effectiveness of LLMs in healthcare tasks, dimensions like fairness, bias, and toxicity are equally vital for addressing ethical concerns and ensuring equitable outcomes. Similarly, robustness and deployment considerations are essential for assessing the sustainability of integrating LLMs into healthcare systems. The limited assessment of calibration and uncertainty raises questions about the extent to which researchers are addressing the need for LLMs to provide uncertainty quantifications, particularly in healthcare scenarios.

Table 4: Frequency of publications examining each dimension of evaluation across healthcare and NLP/NLU task categories

The first column lists healthcare tasks followed by NLP/NLU tasks (separated by a double line); the first row lists the dimensions of evaluation used in each study examined. The percentages in the last row are the percentage of studies in which a specific dimension was evaluated and the percentages on the last column indicate the percentage of studies in which a specific healthcare task or NLP/NLU task was evaluated.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

	Accuracy	Comprehensiveness	Factuality	Robustness	Fairness, Bias and Toxicity evaluation	Deployment Metrics	Calibration and Uncertainty	TOTAL NUMBER OF PAPERS	
Healthcare Task								Number	%
Enhancing medical knowledge	222	91	44	33	16	10	3	231	**44.5%**
Making Diagnoses	100	38	11	11	14	4	0	101	**19.5%**
Educating patients	88	68	32	21	18	3	2	92	**17.7%**
Making treatment recommendations	47	22	9	8	3	1	0	48	**9.2%**
Communicating with patients	35	29	8	15	22	1	0	39	**7.5%**
Care coordination and planning	36	24	4	5	7	1	0	39	**7.5%**
Triaging patients	24	7	5	2	8	8	0	24	**4.6%**
Carrying out a literature review	18	7	3	2	2	2	0	18	**3.5%**
Synthesizing data for research	16	7	2	3	2	2	0	17	**3.3%**
Generating medical reports	8	8	2	0	3	0	0	9	**1.7%**
Conducting medical research	8	7	3	3	3	1	0	9	**1.7%**
Providing asynchronous care	8	5	3	3	1	1	0	8	**1.5%**
Managing clinical knowledge	5	5	1	1	0	0	0	6	**1.2%**
Clinical note-taking	6	2	1	1	0	0	1	4	**0.8%**
Generating clinical referrals	3	0	0	0	0	1	0	3	**0.6%**
Enhancing surgical operations	3	3	1	1	0	0	0	3	**0.6%**
Biomedical data mining	2	0	0	2	1	2	0	2	**0.4%**
Generating provider billing codes	1	0	0	0	0	0	0	1	**0.2%**
Writing prescriptions	1	0	0	0	0	0	0	1	**0.2%**
NLP/NLU Task									
Question Answering	398	194	71	61	54	14	5	437	**84.2%**
Text Classification	29	10	6	5	10	2	0	145	**27.9%**
Information Extraction	29	12	8	5	4	6	0	128	**24.7%**
Summarization	29	21	7	3	8	0	1	46	**8.9%**
Conversational Dialogue	6	6	1	1	5	1	0	17	**3.3%**
Translation	5	1	2	2	1	1	0	16	**3.1%**
TOTAL NUMBER OF PAPERS	495	244	95	77	82	24	6		
%	**95.4%**	**47.0%**	**18.3%**	**14.8%**	**15.8%**	**4.6%**	**1.2%**		

4.5 Distribution of studies by medical specialty

We categorized studies according to the Accreditation Council for Graduate Medical Education (ACGME) residency programs, augmented to include additional categories to capture studies investigating applications in dental specialties, treatment of genetic disorders and generic healthcare applications [20]. Notably, over a fifth of the studies were categorized as generic, indicating a significant focus on healthcare applications that are relevant to many specialties, rather than a specific specialty. Among the specialties, internal medicine, surgery, and ophthalmology were the top specialties. 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 2**. The distribution of studies across specialties underscores the potential for LLMs to contribute to a wide range of medical specialties, but also signals opportunities for further exploration within less represented areas such as nuclear medicine, physical medicine, and medical genetics.

5. Discussion

Our systematic review of 519 studies summarizes existing evaluations of LLMs across medical specialties, the underlying healthcare task, NLP/NLU task, and dimension of evaluation. Doing such categorization captures the heterogeneity of current LLM applications in healthcare while providing a concrete way to discuss their testing and evaluation in a consistent manner. While our categorization we created may be viewed as a limitation, given the precise definitions and illustrative examples of each category, it may have utility beyond this review.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Overall, we identified six limitations in the current efforts and suggest how to address them in future. Our findings call for an urgent need to develop consensus-driven guidance for evaluating LLMs in medicine, in a manner similar to the creation of the blueprint for trustworthy AI by The Coalition for Health AI for traditional AI models[42].

The need for evaluations based on real patient care data

One striking finding is that only 5% of the studies used real patient care data for evaluation, with most studies using a mix of medical exam questions, patient vignettes and subject matter expert generated questions [43 14 44]. Shah et al noted that testing LLMs with hypothetical medical questions is like assessing a car's performance with multiple-choice questions before certifying it for road use [11]. Real patient care data encompasses the complexities of clinical practice, providing a more thorough evaluation of LLM performance that will closely mirror real-world performance [14 45 46 6].

Real-world LLM evaluations provide valuable insights that may be overlooked in simulations or synthetic environments. For instance, while LLMs have been touted for potentially saving time and enhancing clinician experience, Garcia et al. found that the mean utilization rate for drafting patient messaging responses in an EHR system was only 20%, resulting in a reduction in burnout score but no time savings [47].

Given the importance of using real patient care data, systems need to be created to ensure their use in evaluating LLMs' healthcare applications. The Office of the National Coordinator for Health Information Technology (ONC) recently passed HT-1, the first federal regulation to set specific reporting requirements for developers of AI tools through their 'model report cards'[48]. ONC and other regulators should look to embed a mandate for the use of patient care data in the evaluation process of LLM tools into its requirements.

The need to standardize the task formulations and dimensions of evaluation

There is a lack of consensus on which dimensions of evaluation to examine for a given healthcare task or NLP/NLU task. For instance, for a medical education task, Ali et al. tested the performance of GPT-4 on a written board examination focusing on output accuracy as the sole dimension [21]. Another study tested the performance of ChatGPT on the USMLE, focusing on output accuracy, factuality and comprehensiveness as primary dimensions of evaluation [49].

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

Prioritize immediately impactful, administrative applications

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Current research predominantly focuses on medical knowledge tasks, such as answering medical exam questions (44.5%), or complex healthcare tasks, as well as making diagnoses (19.5%) and making treatment recommendations (9.2%). However, there are many administrative tasks in healthcare that are often labor-intensive, requiring manual input and contributing to physician burnout [41]. Particularly, areas such as assigning provider billing codes (1 study), writing prescriptions (1 study), generating clinical referrals (3 studies), and clinical note-taking (4 studies); all of which remain under-researched and could greatly benefit from a systematic evaluation of using LLMs for those tasks [38 39 50 51].

The need to bridge gaps in LLM utilization across clinical specialties

The substantial representation of generic healthcare applications, accounting for over a fifth of the studies, underscores the potential of LLMs in addressing needs applicable to many specialties, such as summarizing medical reports. In contrast, the scarcity of research in particular specialties like nuclear medicine (3 studies), physical medicine (2 studies), and medical genetics (1 study) suggests an untapped potential for using LLMs in these complex medical domains that often present intricate diagnostic challenges and demand personalized treatment approaches[52 53 54 55]. 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.

The need for a realistic accounting of financial impact

Generative AI is projected to create $200 billion to $360 billion in healthcare cost savings through productivity improvements [10]. However, the implementation of these tools could pose a significant financial burden to health systems. In a recent review by Sahni and Carrus, defining the cost and benefit of deploying AI was highlighted as one of the greatest challenges [56]. It is key for health systems to capture this, to accurately estimate and budget for increased implementation and computing costs [57].

Within this review, only one study conducted a financial impact or cost-effectiveness analysis. Rau et al. investigated the use of ChatGPT to develop personalized imaging, demonstrating "an average decision time of 5 minutes and a cost of €0.19 for all cases, compared to 50 minutes and €29.99 for radiologists" [58]. However, this analysis was a parallel implementation of the LLM solution compared with the traditional radiologist approach, thus not providing a realistic assessment of the *added* value of LLM integration into existing clinical workflows and its corresponding financial impact.

While the dearth of real-world testing is understandable given the infancy of LLM applications in healthcare, it is imperative to establish realistic assessments of these tools before reallocating resources from other healthcare initiatives. Notably, such assessments should estimate the total cost of implementation, which includes not only the cost to run the model but also expenses associated with monitoring, maintenance, and any necessary infrastructure adjustments.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The need to better define and quantify bias

Recent studies have highlighted a concerning trend of LLMs perpetuating race-based medicine in their responses [59]. This phenomenon can be attributed to the tendency of LLMs to reproduce information from their training data, which may contain human biases [60].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, gender, or age impact bias in the model's output [61] [62] [63]. 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 towards mitigating harmful biases perpetuated by LLMs [64].

The need to 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[65]. The FDA 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 healthcare [66] [67].

In the ‘Conclusion’ sections of many studies, only a select few researched why the deployment of the LLM did not produce satisfactory results (e.g. ineffective prompt engineering) [68]. A deeper examination of failure modes and why the exercise was deemed unsuccessful or inaccurate (e.g. the reference data was factually incorrect or outdated), is necessary to further improve the use of LLMs in healthcare settings.

6. Conclusion

The evaluation of LLMs lacks standardized task definitions and dimensions of evaluation. This systematic review underscores the need for evaluating LLMs using real patient care data, particularly on administrative healthcare tasks like generating provider billing codes, writing prescriptions, and clinical note-taking. It highlights the need to expand testing criteria beyond accuracy to include fairness, bias, toxicity, robustness, and deployment considerations across different medical specialties. Establishing shared task definitions and rigorous testing and evaluation standards are crucial for the safe integration of LLMs in healthcare. Realistic financial accounting and robust reporting of failures are essential to accurately assess their value and safety in clinical settings. Broadly, there is an urgent need to develop a nationwide consensus and guidance for evaluating LLMs in healthcare, so that we may realize the tremendous promise these groundbreaking technologies have to offer.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Author contributions: SB, YL, LOE, NHS conceived of the study, defined the main outcomes and measures. SB, LOE, YL searched the literature to identify the publications to review and categorized the publications. SB, LOE, YL and NHS drafted the manuscript. SB designed the GPT-4 based screening strategy with input from DD. SB developed the NLP/NLU task and dimensions of evaluation framework. LOE developed the healthcare task framework. DD guided the creation and categorization of healthcare tasks. AC and AS guided the creation of NLP/NLU task categorization. JAF guided the creation of the dimensions of evaluation categorization, SK helped select HELM dimensions to reuse. MK refined the medical specialty categorization, and MW critiqued the review methodology and figure organization. LL and HH assessed the usefulness of the frameworks for other analyses. NRS guided LOE and YL on all aspects of performing systematic reviews. AM reviewed and edited the manuscript for framing the discussion. KS and TT assessed the relevance of the results for developing consensus LLM testing and evaluation guidance for CHAI. MAP critiqued the deployment concerns in health systems and reviewed the categories. All authors reviewed, edited and approved of the final manuscript.

Acknowledgements: We thank Nicholas Chedid for extensive guidance in the development of the healthcare task categorization.

Supplement 1. Search terms for PubMed as of 02/19/2024

(("Large Language Model" [Title/Abstract] OR "ChatGPT" [Title/Abstract] OR "Generative AI" [Title/Abstract]) AND ("Health" [Title/Abstract] OR "Medical" [Title/Abstract] OR "Clinical" [Title/Abstract] OR "Medicine" [Title/Abstract]) AND ("Test" [Title/Abstract] OR "Evaluate" [Title/Abstract] OR "Performance" [Title/Abstract] OR "Assess" [Title/Abstract]))

Search terms for Web of science as of 02/19/2024

(TS=("Large Language Model" OR "ChatGPT" OR "Generative AI")
AND
TS=("Health" OR "Medical" OR "Clinical" OR "Medicine")
AND
TS=("Test" OR "Evaluate" OR "Performance" OR "Assess"))

Supplement 2. Prompts used to extract and assign categories for human review

Prompt 1
"You are assisting in a systematic review of large language models in healthcare. Summarize the {entity_type} mentioned in this research abstract in 25 words"

Prompt 2

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

"Using the generated summaries, identify and categorize the following text based on {entity_type}:\n\n{text}\n\nCategories:"
Where entity_type can be *NLP task*, *medical specialty* or *metric* and categories is the list of possible values for each entity_type to make categorization into, for the NLP task, metric and medical specialty.

eTable 1 - Examples of metrics for each dimension of evaluation

The first row represents the names of the dimensions of evaluation in our designed framework. Under each dimension there are metrics. The bold italicized cells represent metric subclasses for each dimension and regular font cells under each subclass represent the metrics.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

	Robustness metrics	Factuality metrics	Comprehensiveness metrics	Fairness, bias and toxicity metrics	Deployment considerations metrics
	• ***Human evaluated robustness***	• ***Human evaluated factuality***	• ***Readability***	• **Privacy**	• **Hardware Requirements**
	• ***Match (based on perturbation type: Typos)***	• ***Entity Precision***	- Coherence	• **Fairness**	- Number of parameters
	- Exact match	- Citation precision	- Conciseness	- Conterfactual Fairness	- Latency
	- Quasi-exact match	- Concept precision	- Relevance	- Performance Disparities	- Throughput (tokens/s)
caling)	• ***Match (based on perturbation type: Synonyms)***	• ***Entity Recall***	- Fluency	• **Bias and Stereotyping**	- Memory
Scaling)	- Exact match	- Citation recall	- Consistency	- Racial	• **Efficiency**
	- Quasi-exact match	- Concept recall	• ***Completeness***	- Gender	- Inference runtime
	• ***F1 (based on perturbation type: Typos)***	• ***Data statistic evaluations***	• ***UniEval***	- Age	- Observed inference runtime (s)
	• ***F1 (based on perturbation type: Synonyms)***	- Compression	- UniEval Coherence	- Sycophancy	- Denoised inference runtime (s)
	• ***Reciprocal Rank (based on perturbation type: Typos)***	- Coverage	- UniEval Relevance	• **Toxicity**	- Idealized inference runtime (s)
	• ***Reciprocal Rank (based on perturbation type: Synonyms)***	- Density	- UniEval Consistency	• **Legal and Regulatory compliance**	- Estimated training emissions (kg CO2)
	• ***NDCG (based on perturbation type: Typos)***		- UniEval Fluency	• **Potential for Harm**	- Estimated training energy cost (MWh)
	• ***NDCG (based on perturbation type: Synonyms)***		- UniEval Overall		- Time spent on designing prompts
			• ***DDx completion score***		• **Cost**
			• ***Self-BLEU***		
			• ***Lexical matches***		
			• ***Abbreviations (Abbr)***		

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

eTable 2 - Frequency of publications by medical specialty

This table shows the different medical specialties of the 519 studies, along with three additional categories: Generic, Dentistry, and Medical Genetics

Specialty	Number of papers	%
Generic	133	25.6%
Internal Medicine	85	16.4%
Surgery	59	11.4%
Opthalmology	36	6.9%
Radiology	34	6.6%
Otolaryngology	26	5.0%
Psychiatry	19	3.7%
Dentistry	17	3.3%
Orthopedics	17	3.3%
Emergency Medicine	15	2.9%
Urology	15	2.9%
Neurology	14	2.7%
Pathology	14	2.7%
Family Medicine	10	1.9%
Dermatology	9	1.7%
Obstetrics and Gynecology	9	1.7%
Anesthesia	7	1.3%
Pediatrics	6	1.2%
Radiation Oncology	5	1.0%
Nuclear Medicine	3	0.6%
Physical Medicine	2	0.4%
Medical Genetics	1	0.2%

1 Stafie CS, Sufaru IG, Ghiciuc CM et al. Exploring the Intersection of Artificial Intelligence and Clinical Healthcare: A Multidisciplinary Review. *Diagnostics.* 2023;13(12):1995. doi:https://doi.org/10.3390/diagnostics13121995

2 Kohane IS. Injecting Artificial Intelligence into Medicine. *NEJM AI.* 2024;1(1). doi:https://doi.org/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:https://doi.org/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:https://doi.org/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 arXiv:2402.17177. 2024 Feb 27

6 Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus.* 2023 May 21;15(5):e39305. doi: 10.7759/cureus.39305. PMID: 37378099; PMCID: PMC10292051

7 Landi H. Abridge clinches $150M to build out generative AI for medical documentation. Fierce Healthcare. Published February 23rd 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:https://doi.org/10.1038/s41591-023-02700-1

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

9 Bhasker S, Bruce D, Lamb J et al. Tackling healthcare's biggest burdens with generative AI. McKinsey. www.mckinsey.com. Published July 10, 2023. https://www.mckinsey.com/industries/healthcare/our-insights/tackling-healthcares-biggest-burdens-with-generative-ai

10 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.

11 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

12 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. Published 2023 Jul 29. doi:10.1038/s41746-023-00879-8

13 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. Published 2023 Apr 24. doi:10.1515/cclm-2023-0355

14 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. Published 2023 Nov 28. doi:10.1186/s10195-023-00740-4

15 Page MJ, McKenzie JE, Bossuyt PM et al. The PRISMA 2020 statement: an Updated Guideline for Reporting Systematic Reviews. *British Medical Journal.* 2021;372(71). doi:https://doi.org/10.1136/bmj.n71

16 USMLE Physician Tasks/Competencies. 2020. https://www.usmle.org/sites/default/files/2021-08/USMLE_Physician_Tasks_Competencies.pdf

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

18 Liang P, Bommasani R, Lee T et al. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research.* Published online February 1, 2023. Accessed February 2024. https://openreview.net/forum?id=iO4LZibEqW

19 Tasks - Hugging Face. huggingface.co. https://huggingface.co/tasks

20 Residency & Fellowship Programs. Graduate Medical Education. https://med.stanford.edu/gme/programs.html

21 Ali R, Tang OY, Connolly ID, et al. *Performance of CHATGPT and GPT-4 on Neurosurgery Written Board Examinations*. Published online March 29, 2023. doi:10.1101/2023.03.25.23287743

22 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 and uHealth*. 2023;11. doi:10.2196/49995

23 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*. Published online November 8, 2023. doi:10.7759/cureus.48518

24 Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: Comparative study. *Journal of Medical Internet Research*. 2023;25. doi:10.2196/49324

25 Srivastava R, Srivastava S. Can Artificial Intelligence Aid Communication? considering the possibilities of GPT-3 in palliative care. *Indian Journal of Palliative Care*. 2023;29:418-425. doi:10.25259/ijpc_155_2023

26 Dağcı M, Çam F, Dost A. Reliability and quality of the nursing care planning texts generated by CHATGPT. *Nurse Educator*. Published online November 22, 2023. doi:10.1097/nne.0000000000001566

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

27 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. *Journal of Educational Evaluation for Health Professions*. 2023;20:1. doi:10.3352/jeehp.2023.20.1

28 Suppadungsuk S, Thongprayoon C, Krisanapan P, et al. Examining the validity of chatgpt in identifying relevant nephrology literature: Findings and implications. *Journal of Clinical Medicine*. 2023;12(17):5550. doi:10.3390/jcm12175550

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

30 Barash Y, Klang E, Konen E, Sorin V. CHATGPT-4 assistance in Optimizing Emergency Department radiology referrals and Imaging Selection. *Journal of the American College of Radiology*. 2023;20(10):998-1003. doi:10.1016/j.jacr.2023.06.009

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

32 Groza T, Caufield H, Gration D, et al. An evaluation of GPT models for phenotype concept recognition. *BMC Medical Informatics and Decision Making*. 2024;24(1). doi:10.1186/s12911-024-02439-w

33 Razdan S, Siegal AR, Brewer Y, Sljivich M, Valenzuela RJ. Assessing chatgpt's ability to answer questions pertaining to erectile dysfunction: Can our patients trust it? *International Journal of Impotence Research*. Published online November 20, 2023. doi:10.1038/s41443-023-00797-z

34 Kassab J, Hadi El Hajjar A, Wardrop RM, Brateanu A. Accuracy of online artificial intelligence models in Primary Care Settings. *American Journal of Preventive Medicine*. Published online February 2024. doi:10.1016/j.amepre.2024.02.006

35 Lim B, Seth I, Dooreemeah D, Lee CH. Delving into new frontiers: Assessing chatgpt's proficiency in revealing uncharted dimensions of general surgery and pinpointing innovations for future advancements. *Langenbeck's Archives of Surgery*. 2023;408(1). doi:10.1007/s00423-023-03173-z

36 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 Mental Health*. 2024;11. doi:10.2196/50150

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

38 Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-Llama : Tuning llama model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine*. 2024;7(1). doi:10.1038/s41746-023-00989-3

39 Aiumtrakul N, Thongprayoon C, Arayangkool C, et al. Personalized medicine in urolithiasis: AI chatbot-assisted dietary management of oxalate for Kidney Stone Prevention. *Journal of Personalized Medicine*. 2024;14(1):107. doi:10.3390/jpm14010107

40 Gan RK, Ogbodo JC, Wee YZ, Gan AZ, González PA. Performance of Google bard and ChatGPT in mass casualty incidents triage. Am J Emerg Med. 2024;75:72-78. doi:10.1016/j.ajem.2023.10.034

41 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. doi:https://doi.org/10.34172/ijhpm.2021.129

42 Coalition for Health AI. Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare. Published April 4th 2023. https://coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

43 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 Nov 27;11:e49886. doi: 10.2196/49886.

44 Surapaneni KM. Assessing the Performance of ChatGPT in Medical Biochemistry Using Clinical Case Vignettes: Observational Study. JMIR Med Educ. 2023 Nov 7;9:e47191. doi: 10.2196/47191.

45 Choi HS, Song JY, Shin KH, Chang JH et al. 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:https://doi.org/10.3857/roj.2023.00633

46 Fleming SL, Lozano A, Haberkorn WJ et al. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. Dec 2023. arXiv:2308.14089; https://doi.org/10.48550/arXiv.2308.14089.

47 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:https://doi.org/10.1001/jamanetworkopen.2024.3201

48 Office of the National Coordinator for Health Information Technology. Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing. Federal Register. January 9, 2024;89(6):[page numbers]. Available from: Federal Register.

49 Gilson A, Safranek CW, Huang T et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment [published correction appears in JMIR Med Educ. 2024 Feb 27;10:e57594]. JMIR Med Educ. 2023;9:e45312. Published Feb 8 2023. doi:https://doi.org/10.2196/45312

50 Heston TF. Safety of Large Language Models in Addressing Depression. Cureus. 2023;15(12):e50729. doi:https://doi.org/10.7759/cureus.50729

51 Pushpanathan K, Lim ZW, Er Yew SM et al. Language Model Chatbots' Accuracy, Comprehensiveness, and Self-Awareness in Answering Ocular Symptom Queries. iScience. 2023;26(11):108163. doi:https://doi.org/10.1016/j.isci.2023.108163

52 Currie G, Barry K. ChatGPT in Nuclear Medicine Education. July 2023. J Nucl Med Technol. 2023 Sep;51(3):247-254. doi:https://doi.org/10.2967/jnmt.123.265844

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

54 Walton N, Gracefo S, Sutherland N et al. Evaluating ChatGPT as an Agent for Providing Genetic Education. bioRxiv (Cold Spring Harbor Laboratory). Published online October 29, 2023. doi:https://doi.org/10.1101/2023.10.25.564074

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

56 Sahni NR, Carrus B. Artificial Intelligence in U.S. Health Care Delivery. July 2023. The New England Journal of Medicine. 2023;389(4):348-358. doi:https://doi.org/10.1056/nejmra2204673

57 Jindal JA, Lungren MP, Shah NH. Ensuring useful adoption of generative artificial intelligence in healthcare. J Am Med Inform Assoc. Published online March 7, 2024. doi:https://doi.org/10.1093/jamia/ocae043

58 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. July 2023;308(1). doi:https://doi.org/10.1148/radiol.230970

59 Omiye JA, Lester JC, Spichak S et al. Large language models propagate race-based medicine. NPJ Digit Med. October 2023; 6(1):195. doi:https://doi.org/10.1038/s41746-023-00939-z

medRxiv preprint doi: https://doi.org/10.1101/2024.04.15.24305869; this version posted April 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

60 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:https://doi.org/10.1073/pnas.2313790120

61 Guleria A, Krishan K, Sharma V et al. ChatGPT: ethical concerns and challenges in academics and research. September 2023. J Infect Dev Ctries. 2023;17:1292–1299. doi:https://doi.org/10.3855/jidc.18738

62 Hanna JJ, Wakene AD, Lehmann CU et al. Assessing Racial and Ethnic Bias in Text Generation for Healthcare-Related Tasks by ChatGPT. medRxiv (Cold Spring Harbor Laboratory). Published online August 28, 2023. doi:https://doi.org/10.1101/2023.08.28.23294730

63 Levkovich I, Elyoseph Z. Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette Study. JMIR Mental Health. September 2023;10(1):e51232. doi:https://doi.org/10.2196/51232

64 Heming C, Abdalla M, Ahluwalia M et al. Benchmarking Bias: Expanding Clinical AI Model Card to Incorporate Bias Reporting of Social and Non-Social Factors. Accessed March 2024. https://arxiv.org/pdf/2311.12560.pdf

65 Thomas, D. Revolutionizing Failure Modes and Effects Analysis with ChatGPT: Unleashing the Power of AI Language Models. J Fail. Anal. and Preven. May 2023;23(3):911–913. https://doi.org/10.1007/s11668-023-01659-y

66 Research C for DE and. FDA Adverse Event Reporting System (FAERS) Public Dashboard. FDA. Published online October 29, 2020. https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard

67 MAUDE - Manufacturer and User Facility Device Experience. Fda.gov. Published 2012. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm

68 Galido PV, Butala S, Chakerian M et al. A Case Study Demonstrating Applications of ChatGPT in the Clinical Management of Treatment-Resistant Schizophrenia . Cureus. Published online April 26, 2023; 15(4): e38166. doi:https://doi.org/10.7759/cureus.38166