

Large Language Model Architectures in Healthcare: A Scoping Review of Research Perspectives

Florian Leiser, Richard Guse, Ali Sunyaev

Submitted to: Journal of Medical Internet Research
on: December 19, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 26

 Figures..... 27

 Figure 1..... 28

 Figure 2..... 29

Multimedia Appendixes..... 30

 Multimedia Appendix 1..... 31

 Multimedia Appendix 2..... 31

 Multimedia Appendix 3..... 31

Large Language Model Architectures in Healthcare: A Scoping Review of Research Perspectives

Florian Leiser^{1*} MSc; Richard Guse^{2*} MSc; Ali Sunyaev³ PhD

¹Research Group Critical Information Infrastructures Institute of Applied Informatics and Formal Description Methods Karlsruhe Institute of Technology Karlsruhe DE

²Chair of Information Infrastructures School of Computation, Information and Technology Technical University of Munich, Campus Heilbronn Heilbronn DE

*these authors contributed equally

Corresponding Author:

Florian Leiser MSc

Abstract

Objective: Large language models (LLMs) can provide benefits for most medical applications. However, while most studies investigate how LLMs are used, they neglect the importance of selecting suitable LLM architectures.

Methods: We conduct a scoping review, to identify which LLMs are used in healthcare. Our search included manuscripts from PubMed, arXiv, and medRxiv. We used open and selective coding to assess the 114 identified manuscripts regarding 11 dimensions related to usage and technical facets.

Results: We identify four foci that emerged previously in manuscripts with a focus on investigating LLM performance. We find that GPT-based models are used for general purposes like exam preparation or communicating symptoms and treatments, while BERT-based models are used for knowledge discovery and model improvements.

Conclusions: Healthcare professionals should consider the intended purpose to select the best-suited LLM architecture. Our study suggests that GPT-based model can aid general-purpose tasks in healthcare like report generation or treatment recommendations. However, other LLM families like BERT seem to allow easier extensions of their models for domain-specific tasks.

(JMIR Preprints 19/12/2024:70315)

DOI: <https://doi.org/10.2196/preprints.70315>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, I will be able to make the full manuscript visible to all users.

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Large Language Model Architectures in Healthcare: A Scoping Review of Research Perspectives

Abstract

Background: Large language models (LLMs) can support healthcare professionals in their daily work, for example, when writing and filing reports or communicating diagnoses. With the rise of LLMs, current research investigates how LLMs could be applied in medical practice and their benefits for physicians in clinical workflows. However, most studies neglect the importance of selecting suitable LLM architectures.

Objective: In this literature review, we aim to provide insights on the different LLM model architecture families (i.e., BERT-based or GPT-based models) used in previous research. We report on the suitability and benefits of different LLM model architecture families for various research foci.

Methods: To this end, we conduct a scoping review to identify which LLMs are used in healthcare. Our search included manuscripts from PubMed, arXiv, and medRxiv. We used open and selective coding to assess the 114 identified manuscripts regarding 11 dimensions related to usage and technical facets and the research focus of the manuscripts.

Results: We identify four research foci that emerged previously in manuscripts, with LLM performance being the main focus. We find that GPT-based models are used for communicative purposes like exam preparation or patient interaction. In contrast, BERT-based models are used for medical tasks such as knowledge discovery and model improvements.

Conclusions: Our study suggests that GPT-based models are better suited for communicative purposes like report generation or patient interaction. BERT-based models seem to be better suited for innovative applications like classification or knowledge discovery. This could be due to the architectural differences where GPT processes language unidirectionally and BERT bidirectionally, allowing more in-depth understanding of the text. In addition, BERT-based models seem to allow more straightforward extensions of their models for domain-specific tasks which generally lead to better results. In summary, healthcare professionals should consider the benefits and differences of the LLM architecture families when selecting a suitable model for their intended purpose.

Trial Registration: Not applicable

Keywords: Large Language Models; Scoping Review; ChatGPT; Generative AI, Digital Health; Medical Informatics

Introduction

Background

The last years brought an unmatched rise of large language models (LLMs) like OpenAI's ChatGPT [1,2], Google's BERT [3], and Meta's LLaMA [4]. LLMs have dramatically extended the abilities of natural language processing (NLP) through generating text by repeatedly adding the most likely following words [5]. Most LLMs are based on large amounts of general-purpose text data on which the models were trained [6–8]. Thereby, LLMs provide human-like responses to user-given prompts and can analyze [9], explain [10], and generate text [11]. These abilities offer several opportunities to facilitate and improve workflows in various domains, including healthcare.

To that end, LLMs have been increasingly applied and tested, for example, to assist physicians in diagnosis and treatment decisions [12], researchers in identifying disease phenotypes [13], or students in medical exam preparation [14]. Despite these opportunities, the use of LLMs in healthcare is subject to biases [15], hallucinations (i.e., presenting incorrect information in a factual

correct form [14]), or a limited understanding of the complexity of current medical nomenclature [16]. Since healthcare-specific nomenclature is not fully covered in general-purpose LLMs, their applicability in healthcare is restricted. In addition, the model architectures of general-purpose LLMs differ. For example, BERT-based models use bidirectional encoding, meaning the input text is read in both directions, making them especially suitable for language understanding [3]. In contrast, GPT-based models are built on unidirectional decoders, allowing them to excel in text generation [1]. If practitioners and researchers in healthcare use general-purpose LLMs without being aware of their benefits and limits, the LLM's unsuitability might risk patient health.

With the ever-evolving LLM landscape, it is crucial for healthcare researchers and practitioners to have a thorough overview and keep up with the current trends of using LLMs in healthcare. Having insights into models' performances on specific tasks and the application of LLM architecture families to specific healthcare improvements could significantly ease the usage of LLMs in healthcare. However, prior literature reviews focus primarily on issues in applying LLMs in healthcare and lack a holistic overview of technical instantiations and research foci.

Prior Work

Due to the underrepresentation of medical nomenclature in the training of most general-purpose LLMs, models used in healthcare require unique adaptation. One prominent example is MedPaLM, which can pass questions in the style of the United States Medical Licensing Examination (USMLE) [17]. Medical LLMs have by now been developed and adapted for different use cases like documentation, communication, decision support, picture interpretation, or treatment plan creation [18].

Previous review studies investigated the opportunities and limitations of LLMs in healthcare [19–21]. These reviews have started to aggregate the knowledge scattered across various medical specialties and focus on opportunities and risks (e.g., hallucinations, legal and ethical concerns, or privacy issues) for using LLMs in biomedicine [21] or on the development and applications of LLMs in healthcare (e.g., information and knowledge discovery) [20]. These reviews have already provided a considerable description of the applications of LLMs and related them to the patient care journey [20,21]. For specific applications, one review also included an overview of selected LLM architectures in biomedical and health fields alongside their performance [21]. These reviews made a critical step in providing an overview of the emerging application areas and the performance of LLM architectures for specific biomedical datasets [20,21]. However, they focus on a selection of LLMs and do not include the breadth of available LLM architectures in healthcare research. Moreover, the application areas to which the LLM architectures are applied could help researchers and practitioners to choose the right LLM architecture for a specific use case. Since LLM architectures differ in how they process text [1,3], specific architectures can be more suitable for a particular medical task. Hence, a thorough investigation of current LLM architectures in healthcare is crucial as it can aid practitioners in selecting suitable LLM model architectures for their medical tasks.

Objectives

In this study, we investigate the use of LLM model architectures in current studies in healthcare. While our results regarding the target use are similar to other reviews (e.g., [21]), we specifically provide an analysis of model architecture families for four research foci (i.e., LLM Performance, LLM Societal Impact, LLM Comprehensibility, and LLM Innovation). For each research focus, we assess the used model architectures, the integration of LLMs in clinical practice, and the data types used in each manuscript. With that, we aim to provide insights into LLMs' maturity, practical implementation and innovation stages in healthcare. We conduct a scoping review [22] to determine the breadth and boundaries of the literature on model architectures and the use of LLMs in the medical domain [23].

Methods

In our study, we investigated PubMed as the primary database for identifying manuscripts on LLMs in healthcare. Since the topic is emerging and novel models are introduced rapidly, we also included the pre-print databases arXiv and medRxiv in our search. We searched for *TITLE-ABS-KEY(GPT OR LLM? OR Large Language Model? OR LLaMA OR Bard OR MedPaLM)* to focus on LLMs rather than the medical domain. We included specific model architectures in our search string to gain insights into the technical manifestations of widely recognized LLMs. As shown in Figure 1, our initial search in January 2024 yielded 1,842 hits, with arXiv having the most results (n=813). Before screening the manuscripts, we excluded 479 manuscripts published before 2018, which marked the introduction of GPT-1 [24] as well as duplicates (n=117) and non-English publications (n=37). For analyzing the manuscripts, we followed the PRISMA-ScR guidelines which are summarized in Appendix 3 [25]. During the screening of the title and abstract of each manuscript, we investigated the topic of the manuscripts and excluded papers with no focus on healthcare (n=837). We aligned the decisions on borderline cases by having two authors investigate the manuscript and reach a consensus. Excluded manuscripts were, for example, related to llamas instead of the LLaMA architecture.

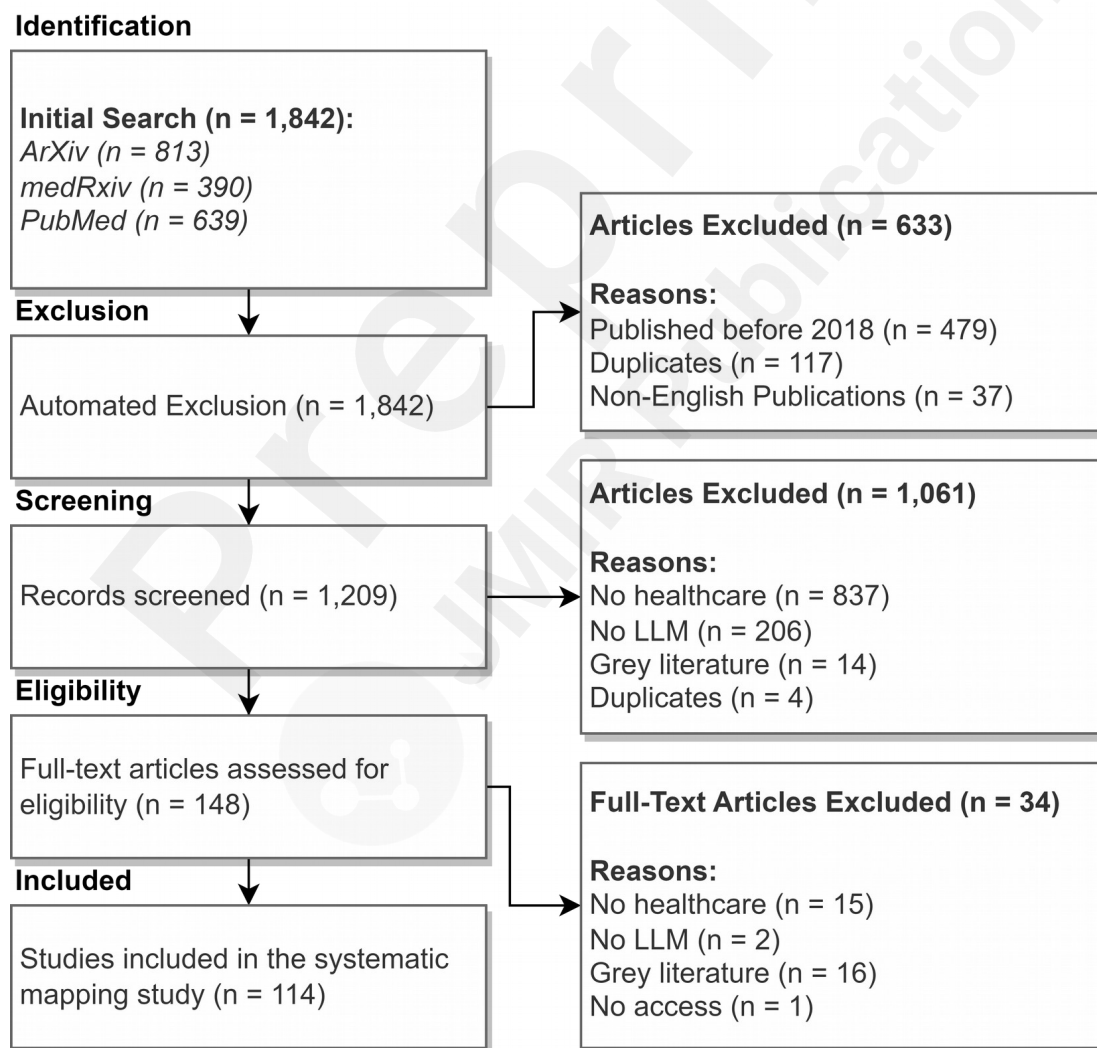


Figure 1. PRISMA flowchart for inclusion and exclusion of manuscripts.

In addition, we excluded papers with no focus on LLMs (n=206) or grey literature (n=14). We also found four additional duplicates. During full-text analysis of the remaining 148 manuscripts, we

further found 15 manuscripts with no focus on healthcare, 2 manuscripts without LLM focus, 16 grey literature manuscripts like blog posts, and excluded one publication with access restrictions. All included manuscripts can be seen in Appendix 1.

Our scoping review used a combination of predefined categories and open coding as shown in Table 1. We coded the identified manuscripts along nine dimensions split between three usage, five technical dimensions, and an overarching *research focus* dimension. To align the understanding of all dimensions and generate codes inductively, we double-coded 20 manuscripts during full-text analysis (11.5%). For these 20 manuscripts, we ensured that the author team understood the dimensions and their codes coherently. With the initial set of dimensions and codes, we proceeded with single-coding for all dimensions. The coding process was monitored through regular meetings where we discussed coding of all dimensions. If the coding author was uncertain about a dimension, other members of the author team re-investigated the manuscript and decided. We introduced new codes only if the entire author team agreed. In coding the identified manuscripts, we allowed multiple codes within each dimension to represent the variety of characteristics discussed within the manuscripts. Therefore, most coding dimensions excel 114, for example, since studies use multiple data modalities or LLM architectures.

An overview of all dimensions and corresponding tables and figures can be found in Appendix 2. For usage dimensions, we investigated *medical specialty* (e.g., radiology, pathology, or cardiology), *target audience* (e.g., physicians, patients, or students), and *target use* (e.g., diagnostic support or patient communication). These dimensions provided relevant information for the targeted application of LLMs in healthcare.

We further investigated the technical aspects of each manuscript. The coding dimension *model integration* differentiated between studies focusing on LLMs on a conceptual level, their usage, implementation, or evaluation. Relatedly, we captured the *model novelty* (i.e., developing a new model, applying an existing model, extending an existing model, or not using a model at all). This categorization helped us understand the extent of current and new developments of LLMs in healthcare. Furthermore, we extracted the specific *model architectures* like GPT-3.5 or BERT to highlight the variety of models in the manuscripts. We also noted the *data type* utilized in the studies such as clinical notes, images, or electronic health records. We documented the *evaluation metrics* used to assess the models' performance, such as accuracy, precision, recall, or F1-score.

Finally, we coded the targeted contributions of the manuscript as research themes. Similar to the other coding dimensions, we openly coded the research themes (e.g., correctness, patient well-being). Since several contributions discussed across the identified research themes were within similar contexts (i.e., fairness (n=1), bias (n=4), and societal impact (n=2)), we combined and aggregated coding themes into research foci. After revising the research themes and double-coding by the involved authors, we combined the themes into four research foci.

Table 1. Overview of coding dimensions for the scoping review.

	Coding Dimension	Description
Usage	Medical Specialty	Medical domain of the conducted study
	Target Audience	Stakeholders targeted by the LLM application
	Target Use	Use case of the LLM application
Technical	Model Integration	Focus of the LLM application
	Model Novelty	Maturity of model development
	Model Architecture	Applied LLM architecture
	Data Type	Used data type
	Evaluation Metrics	Used evaluation metrics
	Research Theme	Investigated improvement focus of the manuscript

Results

Usage Dimensions

Medical Specialty

In examining the medical specialty (see Table 2Error: Reference source not found), we find a frequent usage of LLMs in medical research (22.81%), health education (14.91%), and hospital medicine (14.91%). LLMs in medical research are used for predictions [26–29], labeling [27,30,31], or fine-tuning techniques [32,33] without focusing on a specific medical domain. Most studies using LLMs in health education (82.35%) examine the performance of LLMs in medical domain-agnostic exams like the USMLE [34], the Medical College Admission Test (MCAT) [35], or the Chinese National Medical Licensing Examinations [36]. Other studies investigating LLMs for medical exams conduct domain-specific exams like neurology board questions [37]. Studies that focus on hospital medicine investigate the usage of LLMs as a replacement for doctors' medical advice [33,38–40].

Target Use

In our sample, we identify nine target uses of applying LLMs in healthcare. Within our samples of 114 studies, we identify 161 instances of LLM target use cases in total. The most prominent target use was decision support in 27.19% of the studies, followed by information retrieval (21.05%) and communication (21.05%). Target use is related to our research foci, where the application of LLMs for communication and explanation targets at LLMs comprehensiveness. For example, the authors in [41] compare the perceived accuracy and comprehensiveness of GPT-3.5, GPT-4, and BARD on myopia-related questions.

Target Audience

We identify eight different stakeholders in our sample. Most studies provide insights for physicians (60.53%), researchers (28.07%), patients (20.18%), and students (14.91%). Concerning the research foci, we find that physicians are more targeted in studies on the comprehensibility of LLMs (81.25%), for example, to understand breast cancer survival [42]. Students are especially targeted for educative purposes to assess passing current medical licensing exams with LLM support (e.g., [34,43]).

Technical Dimensions

Model Integration

In our sample, two studies (1.75%) focus conceptually on LLM usage from an ethical or regulatory viewpoint [98, 101]. Most manuscripts focus on using LLMs in clinical practice (n=65, 57.02%). Twenty-seven studies (23.68%) implement LLM models by extending existing LLM architectures or developing new models. However, only two manuscripts report on a full implementation already evaluated in a real-world setting [29, 84]. Twenty papers (17.54%) evaluate different model architectures, for example, to investigate racial presumptions in Claude, BERT, GPT-3.5, and GPT-4 [44].

Table 2. Medical specialty.

Medical Specialty	n (in %)
-------------------	----------

Anesthesiology	3 (2.63)
Cardiology	6 (5.26)
Clinical Laboratory Sciences	6 (5.26)
Dermatology	3 (2.63)
Emergency Medicine	2 (1.75)
Endocrinology	3 (2.63)
Gastroenterology	4 (3.51)
Health Education	17 (14.91)
Hospital Medicine	17 (14.91)
Intensive Care Medicine	2 (1.75)
Internal Medicine	7 (6.14)
Medical Research	26 (22.81)
Nephrology	2 (1.75)
Neurology	6 (5.26)
Neurosurgery	3 (2.63)
Oncology	6 (5.26)
Ophthalmology	5 (4.39)
Orthopedic Surgery	1 (0.88)
Otolaryngology	2 (1.75)
Pathology	5 (4.39)
Pediatrics	2 (1.75)
Psychiatry	4 (3.51)
Pulmonology	2 (1.75)
Radiology	7 (6.14)

Model Novelty

Most papers apply LLMs (n=86, 75.44%) without focusing on technical changes, and only a fourth of the manuscripts extended existing models (n=29, 25.44%) by incorporating additional data, knowledge, or human assessment with six manuscripts doing both. We identify only three manuscripts (2.63%) developing novel retrained model architectures in our sample. The three models include chain-of-thought-reasoning into the inference step of LLMs [45], incorporate physician behavior to teach ChatGPT to reply in a doctor-specific manner [38], or retrain GPT-3 on biomedical data sets containing clinical texts by question-answer challenges [46].

Most model extensions are part of the LLM innovation research focus. While most studies use GPT-based architectures, studies in the innovation research focus on BERT-based and GPT-based models 23 times (35.94%) each. The higher relative percentage indicates that BERT-based models are easier to extend for novel models than GPT-based models.

Model Architecture

We observe that authors frequently apply multiple LLM architectures. This resulted in 57 different LLM architectures used across all studies. Thirty-four of these architectures are only named once, highlighting the variety of LLMs developed for healthcare. In general, GPT-based (n=139, 53.66%) and BERT-based models (n=66, 25.46%) provide the vast majority of applied LLM architectures. Most of the studies apply OpenAI's state-of-the-art models GPT-3 (n=12), GPT-3.5 (n=74), and GPT-4 (n=41), although BERT (n=17) and LLaMA (n=10) are frequently used as well. An overview of the different model architecture families and their use in different research foci are shown in Figure 2.

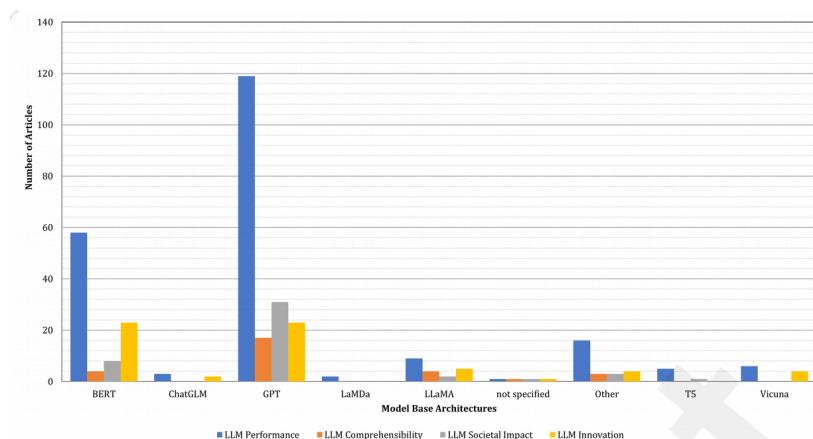


Figure 2. Number of manuscripts per LLM model family and research focus.

Data Types

Our sample accounts for multiple data modalities used as input or output dimensions of the LLMs. This results in 200 mentions of 11 data modalities within our sample as shown in Error: Reference source not found. The most significant data type manifests as symptom descriptions ($n=44$), treatment options ($n=31$), and diagnosis ($n=29$). Since these modalities comprise the steps during diagnosis, they are also frequently used in combination, for example, to assess the treatment advice generated by ChatGPT if patients present themselves with symptom descriptions [40]. These data are also heavily used in manuscripts investigating the understandability of LLMs. The understandability of LLM is especially important when the models communicate their reasoning with patients. Interaction with patients, in turn, is required when treatment options, diagnosis, or symptom descriptions are provided.

In contrast, only very few manuscripts focus on traditional medical data modalities like genomic data ($n=2$), imaging ($n=6$), or electronic health records ($n=13$). These data modalities are usually used by physicians and, therefore, are used in innovation-driven papers, like [32] who extend GPT-3.5 based on existing patient-doctor dialogues to generate a response for the diagnosis report given an ophthalmology image. The remaining papers revolve around traditionally text-related data like patient communication or questions ($n=8$), clinical notes ($n=12$), or medical reports ($n=21$). Other frequently used data modalities include exam questions ($n=28$), for example, the performance of GPT-3.5 in the USMLE [34], the MCAT [35], or on Korean general surgery board exams [47].

Table 3. Data types used in the retrieved literature.

Data Type	n (in %)
Adverse Effects	7 (6.14)
Clinical Notes	12 (10.53)
Diagnosis	29 (25.44)
EHR	13 (11.4)
Exam Questions	28 (24.56)
Genomic Data	2 (1.75)
Images	6 (5.26)
Patient Communication	7 (6.14)
Reports	21 (18.42)
Symptom Description	44 (38.6)
Treatment Options	31 (27.19)

Research Foci

In our analysis, we analyzed the targeted improvements of the manuscripts. We found twelve research themes which we aggregated into four research foci as shown in Table 4. Half of the 114 included manuscripts encompass multiple research foci (n=57).

Research Focus 1: LLM Performance

The research focus on LLM performance encompasses improvements in model performance, ensuring correctness (including hallucinations), and assessing prediction consistency. It underscores the importance of building models that improve performance by providing correct, hallucination-free, and consistent results. Overall, the studies show that domain-specific models improve performance and correctness.

Correctness

Ensuring correct predictions of LLMs constitutes the most prominent research theme. Correct predictions are especially crucial in healthcare as errors can lead to severe consequences for patients (e.g., inappropriate treatment). 89 manuscripts test the correctness of LLMs (e.g., [48–50]). However, the predictive performance of LLMs varies based on the source and structure of the input data [51].

The potential of general-purpose LLMs for medical applications is highlighted in a manuscript that explores whether ChatGPT could generate helpful suggestions for improving a clinical decision support system (CDSS). Clinicians rated the generated suggestions as highly relevant and clear, with low bias [52]. Other manuscripts evaluate the performance of general-purpose LLMs on medical and nursing exams, for example, on the Faculty of Public Health's diplomat exam [53]. However, for healthcare-specific tasks, like biological named entity recognition [54], general-purpose LLMs perform insufficiently.

Table 4. Overview of identified research foci and their themes.

Focus	Theme	Goal	n (in %)
Performance LLM	Correctness	Ensuring predictive performance	89 (78.07)
	Performance Improvement	Improving LLM performance	14 (12.28)
	Hallucinations	Analyzing and preventing hallucinations	9 (7.89)
	Consistency	Comparing LLM performance across contexts or over years	8 (7.02)
Societal Impact LLM	Ethics	Discuss ethical questions for LLM use in high-risk and sensitive contexts of healthcare	12 (10.53)
	Bias	Evaluate and mitigate biases like stereotypes influencing health-related examinations in LLM	6 (5.26)
	Confidentiality	Evaluate risks and attack vectors in LLM use for healthcare data and discuss potential solutions	4 (3.51)
	Regulation	Discuss issues related to accountability and transparency requirements that need to be enforced by law to safely use LLMs in healthcare	3 (2.63)

LLM Comprehensibility	Explainability	Enhance the plausibility and comprehensibility of health information provided by LLMs with additional explanations	12 (10.53)
	Interpretability	Retrace the inner mechanics of the reasoning processes of the LLM	5 (4.39)
LLM Innovation	Novel Applications	Develop novel LLM applications in various healthcare use cases and research	14 (12.28)
	Novel Training Techniques	Improve and develop novel training techniques for LLMs, for example, to improve the quality of LLM responses	10 (8.77)

Therefore, another common theme is the development of specific-purpose LLMs and testing their correctness. For example, HuatuoGPT extends ChatGPT with real-world data from doctors to outperform other LLMs in terms of human evaluations and benchmarks, yielding highly accurate results in medical contexts [38]. Another study tests the application of LLMs in clinical trial enrollment. For that, GPT-3.5 is used to extract medical entity tags from patient information and classify these to identify suitable clinical trials. The application achieves a high accuracy for this task [55].

Performance Improvement

Within the 14 studies in our sample, performance improvement is achieved by adapting LLMs to a specific domain or developing new, specialized LLMs with domain-specific data in LLM pretraining. GatorTronGPT is one example of a clinical LLM trained on clinical and general English text using a GPT-3 architecture and outperforms general-purpose LLMs in medical research [46]. Further, domain adaptation of LLMs improves the predictive performance, for example, to classify nuclear medicine reports [56]. One manuscript compares models ranging from 220 million to 175 billion parameters and finds that smaller, domain-specific models often match or exceed the performance of larger general-purpose models [48].

Hallucinations

In our sample, we found nine studies investigating hallucinations of LLMs. Hallucinations occur especially in summaries of textual information [57] or in 12.5% of GPT-based cancer treatment recommendations [58]. Another study finds that using LLMs for drug-drug interactions also produced many hallucinations [59].

Consistency

The consistency of LLMs evaluates performance across different contexts or over time. One manuscript evaluates ChatGPT's performance on the Japanese National Nursing Exam across multiple years [60]. In another case, researchers assess the consistency across GPT-3.5 and GPT-4 on questions related to heart failure [61]. Overall, the importance of reproducibility in complex decision-making tasks is highlighted [62].

Research Focus 2: LLM Societal Impact

The second research focus pertains to the societal impact of LLMs. The societal considerations surrounding LLMs in healthcare highlight complex challenges related to ethics, bias, confidentiality, and regulation. While LLMs have the potential to improve healthcare provision, their deployment

must be carefully managed to mitigate bias, ensure patient data confidentiality, and ensure ethical appropriateness. Moreover, the absence of clear regulations and accountability frameworks presents a critical barrier to their widespread and safe use.

Ethics

Twelve studies investigate ethical concerns like appropriate empathy of LLMs. For instance, one manuscript finds that LLM responses to patient questions express appropriate empathy [39]. On the contrary, another manuscript suggests that LLMs fail to resolve high-risk scenarios, such as suicide prevention [63]. In their manuscript, only a few models provide essential crisis resources like suicide hotlines, highlighting the need for model refinement to comply with ethical standards [63].

Bias

Six studies show that LLMs can replicate and amplify harmful stereotypes in medical content. For example, LLMs propagate race-based misconceptions when responding to historically involved race-based medicine [44]. Another study shows location-based biases in cardiovascular disease risk assessment by ChatGPT [64]. Bias extends beyond location, for example, to gender as ChatGPT changes its predictions more frequently when race or gender descriptors are added to medical texts [65]. While these findings highlight the importance of addressing bias in LLMs to prevent exacerbating healthcare inequalities, research also discusses potential solutions. For example, LLMs are shown to help identify bias within a corpus of text [30].

Confidentiality

Ensuring data confidentiality is still an issue for LLMs in healthcare [54]. Studies highlight how adversarial attacks could be used to generate misinformation or leak private patient data [66]. Research tries to tackle these issues with a privacy-aware data augmentation approach for LLM-based patient-trial matching [67] or using GPT-3.5 to identify confidential content within clinical notes [68].

Regulation

Three studies emphasize the need for better regulation to ensure safe usage of LLMs in healthcare [15,57,69]. This lack of regulation poses questions about who is held accountable for errors in diagnosis or treatment recommendations [15,57] and the transparency of LLM functionality [69].

Research Focus 3: LLM Comprehensibility

The third research focus deals with the comprehensibility of LLMs, particularly in terms of explainability and interpretability. On the one hand, explainability provides insights into LLM predictions and recommendations. Interpretability, on the other hand, offers explanations for the internal processes of LLMs. Both explainability and interpretability are crucial for healthcare, as healthcare professionals need to rely on clear, accurate information to make decisions.

Explainability

In our sample, twelve studies focus on improving the explainability of LLMs either by enhancing LLMs with additional information or using self-explanations. For example, LLMs incorporate additional information in the form of in-text citations and references [69] or include causal graphs to explore the effect of different genetic perturbations on the survival rates of breast cancer patients [42].

LLMs self-explaining their reasoning helps students understand complex medical conditions and generate personalized explanations of MCAT-related questions [35]. Similarly, one manuscript focuses on using LLMs to explain its reasoning for diagnosing certain mental health conditions [70].

Interpretability

Five studies focus on the inner reasoning processes of LLMs. For example, one manuscript assesses the reasoning of ChatGPT behind the responses for 119 public health exam questions [53]. In another study, model interpretability is enhanced by using the Alpaca architecture to map radiologist

findings to radiological images easing comprehensibility for physicians [12]. Another approach uses GPT-3 to generate interpretability by grouping patients based on the similarity of their embeddings for medication prescriptions [71].

Research Focus 4: LLM Innovation

The fourth research focuses on LLM innovation through novel applications and training techniques. This focus explores the adaptation and extension of LLMs to accelerate, improve, and automate healthcare tasks. At the same time, novel training techniques, such as multimodal learning and context-enriched approaches, are extending the capabilities of LLMs to more complex medical contexts.

Novel Applications

Fourteen studies present novel applications of LLMs in specific areas. One manuscript reports that LLMs could reduce the dimensionality of complex healthcare data like ICD-10 codes while maintaining the original embeddings [72]. Another application uses LLMs to create patient clusters based on pharmacy records for more personalized treatment recommendations [71]. In biomedical research, LLMs trained on additional medical literature identify and validate disease-relevant targets [26]. Studies also show that LLMs can be used for theme-driven analyses (e.g., of opioid use disorder) uncovering both frequent and clinically undocumented patterns [27].

In healthcare education, LLMs are used to create exam questions to accelerate exam generation, standardize assessments, and reduce bias in questions [43]. When preparing for exams, LLMs could respond to individual learning styles and needs in various medical domains [39,73].

Novel Training Techniques

Ten studies in our sample develop novel training techniques to adapt LLMs to healthcare. These approaches show that medical data can be directly used to pre-train LLMs (e.g., [32,48,56]). Adapting LLMs or prompts to medical domains improves their performance [32,45,74,75] and could simulate expert thought processes [45].

In a more general setting, the research explores the integration of external knowledge sources to improve the accuracy of LLaMA in diagnosing disease symptoms [33] or the reliability of ChatGPT responses [69]. Further, LLMs are used to clean and annotate medical data to fine-tune additional LLMs [74].

Other studies develop a novel multimodal model fine-tuned on images, diagnostic reports, and LLM-based dialogues to enhance condition diagnosis and response generation [32,75]. These multimodal training approaches provide LLMs with a broader range of inputs to increase diagnosing versatility.

Discussion

Principal Results

With the rise of LLMs in the past years, research utilizing LLMs in healthcare has significantly increased. We found that the manuscripts in our literature sample targeted a broad variety of stakeholders in the healthcare sector such as physicians (60.53%), researchers (28.07%), patients (20.18%), or medical students (14.91%).

We identified four primary research foci for LLMs in healthcare: LLM performance, LLM societal impact, LLM comprehensibility, and LLM innovation. With several manuscripts tackling two research foci simultaneously, most of our manuscripts focused on the performance of LLMs (85.09%).

A large number of studies focusing on LLM performance show the importance of accurate, reliable LLM results in paving the way for the application of LLMs in healthcare. This research focus

supports the idea that general-purpose LLMs are not yet ready to be applied to healthcare tasks and instead need domain adaptation. As the performance of general-purpose LLMs progresses, we found that the predictive performance seems to be better in BERT-based models when investigating a limited sample of 13 manuscripts that used BERT-based and GPT-based models. While the performance difference was often marginal, GPT-based models seem to have a higher recall but lower precision [13,27] or are used to augment training data for BERT-based models [74,76]. This could result from the model architectures. While both model architectures use transformer techniques, BERT-based models use a bidirectional encoder [3] and GPT-based models use a unidirectional decoder [1]. Therefore, BERT can consider words in both reading directions when understanding text, while GPT processes text unidirectionally (usually from left to right) and is optimized for generating texts. This shows an interesting conflict of LLM architectures and highlights the benefits of BERT-based models for medical tasks like knowledge discovery or report generation [74], while GPT-based models are better suited for language generation and communicative purposes (e.g., exam preparation [14] or dialogues with patients [82]).

In addition, the consistency of LLMs was rarely investigated in our sample. Systematic evaluation over time is missing to allow reliable decisions on the application of LLMs. Some studies have started to compare domain-adapted models against domain-agnostic models [54,77]. Our findings suggest that domain-adapted BERT-based models provide the best predictive performance. Further research should strengthen these findings by conducting computational benchmark evaluations between models. Relatedly, hallucinations were only evaluated for general-purpose LLMs in the retrieved manuscripts. Moreover, many hallucinations were present in the target use case of text summarization, although the textual information was directly given to the LLM. For other use cases (e.g., knowledge discovery), hallucinations have not been examined in our sample. Several mitigation strategies for hallucinations were developed, such as enabling users to detect hallucinations themselves [78] or reinforcing model inference by incorporating human feedback [78]. As a second research focus, the societal impact of LLMs was investigated (21.93%). Our results hint that LLMs in healthcare come along with race-based misconceptions that reproduce harmful stereotypes [44] and provide insufficient support in high-risk scenarios such as suicide prevention [63]. Similar to research on AI, societal and ethical impact should be considered when using LLMs in healthcare [79]. This indicates that further research is needed to determine how bias in healthcare can be prevented. This is in line with existing reviews [19–21] that also note various challenges of applying LLMs in healthcare (e.g., hallucination, fairness, bias, privacy, and legal concerns) underlying the relevance of this research focus. In our sample, societal impact is predominantly evaluated with GPT-based models instead of other model architectures like BERT. When accounting for the societal impact of LLMs, an investigation of multiple model architectures could provide new insights. Further manuscripts highlighted confidentiality and privacy issues when using and adapting LLM with sensitive healthcare data. Our results indicate that future research still needs to determine privacy-preserving strategies when using LLMs in healthcare. A possible avenue for future research could be to introduce privacy-preserving training approaches like federated learning that have been shown to work in healthcare settings [80]. The third research focus on model comprehensibility included a limited number of sixteen studies (14.04%). In this selection, a minimal tendency can be made toward extending state-of-the-art LLM models instead of simply applying them. In contrast, despite this limited number of manuscripts, the focus suggests that LLM comprehensibility is tailored toward physicians in 13 of 16 studies (81.25%). In addition, the frequent use of LLMs for decision support (56.25%) and classification tasks (25.00%) further emphasizes explanations targeted toward medical personnel. While AI research has already highlighted the need for explanations on model reasoning to improve model reliance [81], in healthcare not only the physician's perspective should be highlighted. Patients' perspectives should mainly be investigated when assessing model comprehensibility. To that end, our sample suggests that the comprehensibility studies instead comprised patient-specific data types like symptom

descriptions (50.00%), treatment options (43.75%), EHR (18.75%), or images (18.75%). As LLMs become increasingly integrated into healthcare, improving explainability and regulations will be crucial to ensure the safe and responsible use of LLMs and create an environment where LLMs can be used effectively in healthcare settings (e.g., [82]).

As a fourth research focus, we identified studies focusing on model innovation. In our literature review, we found that 75.44% of the studies applied existing models and 25.44% of the studies extended models (with six studies doing both). The studies in our sample rarely implemented new models (2.63%). The studies mostly used state-of-the-art general-purpose models such as GPT-3.5, GPT-4, or BERT. While this holds true for the entire data set, innovation (i.e., model innovation or novel applications) was comparatively more performed on BERT-based models (35.94%) than GPT-based models (35.94%). This includes healthcare-specific extensions of LLMs like BioBERT, ClinicalBERT, or MedBERT that enhance the performance of these general-purpose models for the healthcare domain by including external knowledge sources. This again builds on the above-mentioned differences in model architectures where BERT-based model architectures are better suited for knowledge discovery tasks and GPT-based model architectures rather than solving communicative purposes.

These insights help practitioners to select appropriate LLM model architectures. If the focus of the usage of the LLM is to ease communication, GPT-based models seem more appropriate. When LLMs are used for innovative purposes (like knowledge discovery), our sample suggests the superiority of BERT-based models. However, most general-purpose models still require domain adaptation and could benefit from the incorporation of domain knowledge [83,84].

Limitations

This study is not without limitations. First, all codings and assessments of the dimensions were subjectively assessed by the authors of this manuscript. While we tried to minimize the subjectiveness through double-coding and regular meetings to discuss and ensure a similar understanding of all dimensions and manifestations, the final assessment is still subject to subjective assessments. Second, our comparison of GPT-based and BERT-based models focuses on a limited sample of 13 manuscripts that report on the performance of both. Future benchmark evaluations of these manuscripts are needed to support these tendencies. Third, we conducted the assessment based on a literature search in January 2024. While we included early works from pre-publication databases like arXiv or medRxiv to reduce the impact of later publication, maintaining a current and up-to-date review in this fast-changing field is challenging. For example, the introduction of GPT Playground might have facilitated the implementation and adaptation of GPT-based models. Fourth, in our search term, we focused on LLMs and did not include further related concepts (e.g., transformer-based models). Although we included established LLMs in our search query, this might have led to the exclusion of relevant studies. Last, we assume that manuscripts retrieved from pre-publication databases have sufficient scientific rigor. While we excluded only two publications that we deemed of insufficient quality, manuscripts included in our sample might not hold up to highest scientific standards.

Conclusions

In this literature review, we identified four research foci for LLMs in healthcare and highlighted specifics to each focus. Most manuscripts in our review focused on the performance of LLMs. Our sample shows a tendency that GPT-based models provide higher recall and are more heavily involved in societal questions that require interaction and communication. In contrast, BERT-based models show higher precision and are more intensely used for innovation like classification or knowledge discovery. This is in line with the architectural design of the models where the unidirectional design of GPT is better suited for generative purposes. In contrast, the bidirectional

design of BERT is better suited for question-answering. We also find that more manuscripts use diagnosis-related data like treatment recommendations or symptom descriptions than traditional medical data like images or EHR. This, again, highlights the novel ways to communicate with LLMs. While these findings seem to hold through the manuscripts comparing different LLMs, future research should conduct thorough comparative benchmark evaluations.

With this review, we aid healthcare professionals in understanding the benefits of specific LLM technologies. GPT-based models are better suited for communicative purposes due to their enhanced ability to generate text. However, in the current state, general-purpose LLMs might be able to pass medical exams but are unable to explain or reason their predictions. When incorporating medical-specific knowledge, the performance of LLMs was further increased. Since we only investigate published literature, future research should extend the comparison of performance and bias of domain-adapted and general-purpose LLM models. Without these benchmarks, we should teach physicians to use LLMs as supporting tools while raising awareness to potential model insufficiencies.

Acknowledgments

We would like to acknowledge the assistance of our internal friendly reviewers who helped us revise this manuscript before submission. We used the generative AI tools ChatGPT by OpenAI and Grammarly to revise the writing of the manuscript.

Conflicts of Interest

None declared.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers

CDSS: Clinical decision support system

EHR: Electronic health records

GPT: Generative-pretraining transformer

LLM: Large language model

MCAT: Medical College Admission Test

NLP: Natural language processing

USMLE: United States Medical Licensing Exam

References

1. OpenAI. Introducing ChatGPT. 2023. Available from: <https://openai.com/blog/chatgpt> [accessed Aug 27, 2023]
2. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. 2023. Available from: <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
3. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019. Available from: <https://arxiv.org/abs/1810.04805>
4. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: Open and Efficient Foundation Language Models. 2023. Available from: <https://arxiv.org/abs/2302.13971>
5. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G,

- Günemann S, Hüllermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, Stadler M, Weller J, Kuhn J, Kasneci G. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 2023;103:102274. doi: <https://doi.org/10.1016/j.lindif.2023.102274>
6. Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson Ú, Oprea A, Raffel C. Extracting Training Data from Large Language Models. 30th USENIX Security Symposium (USENIX Security 21) USENIX Association; 2021. p. 2633–2650. Available from: <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>
 7. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, others. A survey of large language models. *arXiv preprint arXiv:230318223* 2023;
 8. Contreras Kallens P, Kristensen-McLachlan RD, Christiansen MH. Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science Wiley*; 2023 Feb;47(3). doi: 10.1111/cogs.13256
 9. Chew R, Bollenbacher J, Wenger M, Speer J, Kim A. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:230614924* 2023;
 10. He X, Bresson X, Laurent T, Perold A, LeCun Y, Hooi B. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. *arXiv preprint arXiv:230519523* 2023;
 11. Zhao Z, Song S, Duah B, Macbeth J, Carter S, Van MP, Bravo NS, Klenk M, Sick K, Filipowicz AL. More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. *Proceedings of the 15th Conference on Creativity and Cognition* 2023. p. 368–370.
 12. Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z, Ma C, Shu P, Chen C, Kim S, Dai H, Zhao L, Sun L, Zhu D, Liu J, Liu W, Shen D, Li X, Li Q, Liu T. Radiology-GPT: A Large Language Model for Radiology. *arXiv*; 2024. Available from: <http://arxiv.org/abs/2306.08666> [accessed Oct 12, 2024]
 13. Shyr C, Hu Y, Bastarache L, Cheng A, Hamid R, Harris P, Xu H. Identifying and Extracting Rare Diseases and Their Phenotypes with Large Language Models. *Journal of Healthcare Informatics Research* 2024 Jun;8(2):438–461. doi: 10.1007/s41666-023-00155-0
 14. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. *Medical Teacher* 2024 May 3;46(5):657–664. doi: 10.1080/0142159X.2023.2271159
 15. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine* 2023 Jul;6(1):120. doi: 10.1038/s41746-023-00873-0
 16. Jin Q, Yang Y, Chen Q, Lu Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. Wren J, editor. *Bioinformatics* 2024 Feb 1;40(2):btae075. doi: 10.1093/bioinformatics/btae075
 17. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H,

- Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. *Nature Springer Science and Business Media LLC*; 2023 Jul;620(7972):172–180. doi: 10.1038/s41586-023-06291-2
18. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: A taxonomy and systematic review. *Computer Methods and Programs in Biomedicine Elsevier BV*; 2024 Mar;245:108013. doi: 10.1016/j.cmpb.2024.108013
 19. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nature medicine Nature Publishing Group US New York*; 2023;29(8):1930–1940.
 20. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. *Health Care Science Wiley Online Library*; 2023;2(4):255–263.
 21. Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X, Yang Y, Chen Q, Kim W, Comeau DC, others. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics Oxford University Press*; 2024;25(1):bbad493.
 22. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 2005 Feb;8(1):19–32. doi: 10.1080/1364557032000119616
 23. Paré G, Trudel M-C, Jaana M, Kitsiou S. Synthesizing information systems knowledge: A typology of literature reviews. *Information & management Elsevier*; 2015;52(2):183–199.
 24. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018;
 25. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Chang C, McGowan J, Stewart L, Hartling L, Aldcroft A, Wilson MG, Garritty C, Lewin S, Godfrey CM, Macdonald MT, Langlois EV, Soares-Weiser K, Moriarty J, Clifford T, Tunçalp Ö, Straus SE. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* 2018 Oct 2;169(7):467–473. doi: 10.7326/M18-0850
 26. Zagirova D, Pushkov S, Leung GHD, Liu BHM, Urban A, Sidorenko D, Kalashnikov A, Kozlova E, Naumov V, Pun FW, Ozerov IV, Aliper A, Zhavoronkov A. Biomedical generative pre-trained based transformer language model for age-related disease target discovery. *Aging* 2023 Sep;15(18):9293–9309. doi: 10.18632/aging.205055
 27. Romano W, Sharif O, Basak M, Gatto J, Preum SM. Theme-Driven Keyphrase Extraction to Analyze Social Media Discourse. *Proceedings of the International AAAI Conference on Web and Social Media* 2024. p. 1315–1327. doi: 10.1609/icwsm.v18i1.31391
 28. Moradi M, Blagec K, Haberl F, Samwald M. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:210902555* 2021;

29. Hu M, Alkhairy S, Lee I, Pillich R, Bachelder R, Ideker T, Pratt D. Evaluation of large language models for discovery of gene set function. 2023. doi: 10.21203/rs.3.rs-3270331/v1
30. Susnjak T. Applying BERT and ChatGPT for Sentiment Analysis of Lyme Disease in Scientific Literature. In: Gilbert L, editor. *Borrelia burgdorferi* New York, NY: Springer US; 2024. p. 173–183. doi: 10.1007/978-1-0716-3561-2_14 ISBN:978-1-07-163560-5
31. Alrdahi H, Han L, Šuvalov H, Nenadic G. Medmine: Examining pre-trained language models on medication mining. arXiv preprint arXiv:230803629 2023;
32. Gao W, Deng Z, Niu Z, Rong F, Chen C, Gong Z, Zhang W, Xiao D, Li F, Cao Z, Ma Z, Wei W, Ma L. OphGLM: Training an Ophthalmology Large Language-and-Vision Assistant based on Instructions and Dialogue. 2023. Available from: <https://arxiv.org/abs/2306.12174>
33. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus* 2023 Jun; doi: 10.7759/cureus.40895
34. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. Dagan A, editor. *PLOS Digital Health Public Library of Science (PLoS)*; 2023 Feb;2(2):e0000198. doi: 10.1371/journal.pdig.0000198
35. Bommineni VL, Bhagwagar S, Balcarcel D, Davatzikos C, Boyer D. Performance of ChatGPT on the MCAT: The Road to Personalized and Equitable Premedical Learning. Cold Spring Harbor Laboratory; 2023 Mar; doi: 10.1101/2023.03.05.23286533
36. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. *BMC Medical Education* 2024 Feb 14;24(1):143. doi: 10.1186/s12909-024-05125-7
37. Schubert MC, Wick W, Venkataramani V. Performance of Large Language Models on a Neurology Board–Style Examination. *JAMA Network Open* 2023 Dec 7;6(12):e2346721. doi: 10.1001/jamanetworkopen.2023.46721
38. Zhang H, Chen J, Jiang F, Yu F, Chen Z, Li J, Chen G, Wu X, Zhang Z, Xiao Q, Wan X, Wang B, Li H. HuatuoGPT, towards Taming Language Model to Be a Doctor. 2023. Available from: <https://arxiv.org/abs/2305.15075>
39. Liu S, McCoy AB, Wright AP, Carew B, Jenkins JZ, Huang SS, Peterson JF, Steitz B, Wright A. Leveraging Large Language Models for Generating Responses to Patient Messages. 2023. doi: 10.1101/2023.07.14.23292669
40. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. *Scientific Reports* 2023 Oct 19;13(1):17885. doi: 10.1038/s41598-023-45223-y
41. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C-H, Lam JSH, Chen DZ, Goh JHL, Tan MCJ, Sheng B, Cheng C-Y, Koh VTC, Tham Y-C. Benchmarking large language models'

- performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *eBioMedicine* 2023 Sep;95:104770. doi: 10.1016/j.ebiom.2023.104770
42. Farooq M, Hardan S, Zhumbhayeva A, Zheng Y, Nakov P, Zhang K. Understanding Breast Cancer Survival: Using Causality and Language Models on Multi-omics Data. *arXiv*; 2023. Available from: <http://arxiv.org/abs/2305.18410> [accessed Oct 12, 2024]
 43. Fleming SL, Morse K, Kumar A, Chiang C-C, Patel B, Brunskill E, Shah N, Bedi S, Yamaguchi C. Assessing the Potential of USMLE-Like Exam Questions Generated by GPT-4. *medrxiv*; 2023. doi: <https://doi.org/10.1101/2023.04.25.23288588>
 44. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *npj Digital Medicine Springer Science and Business Media LLC*; 2023 Oct;6(1). doi: 10.1038/s41746-023-00939-z
 45. Wu C-K, Chen W-L, Chen H-H. Large Language Models Perform Diagnostic Reasoning. 2023. Available from: <https://arxiv.org/abs/2307.08922>
 46. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, Martin C, Flores MG, Zhang Y, Magoc T, Lipori G, Mitchell DA, Ospina NS, Ahmed MM, Hogan WR, Shenkman EA, Guo Y, Bian J, Wu Y. A study of generative large language model for medical research and healthcare. *npj Digital Medicine Springer Science and Business Media LLC*; 2023 Nov;6(1). doi: 10.1038/s41746-023-00958-w
 47. Oh N, Choi G-S, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Annals of Surgical Treatment and Research XMLink*; 2023;104(5):269. doi: 10.4174/astr.2023.104.5.269
 48. Lehman E, Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, Nadler D, Szolovits P, Johnson A, Alsentzer E. Do We Still Need Clinical Language Models? *arXiv*; 2023. Available from: <http://arxiv.org/abs/2302.08091> [accessed Oct 12, 2024]
 49. Chen Z, Balan MM, Brown K. Language models are few-shot learners for prognostic prediction. *arXiv preprint arXiv:230212692* 2023;
 50. Bumgardner V, Mullen A, Armstrong S, Hickey C, Talbert J. Local large language models for complex structured medical tasks. *arXiv preprint arXiv:230801727* 2023;
 51. Zheng H, Zhu Y, Jiang LY, Cho K, Oermann EK. Making the Most Out of the Limited Context Length: Predictive Power Varies with Clinical Note Type and Note Section. *arXiv*; 2023. Available from: <http://arxiv.org/abs/2307.07051> [accessed Oct 12, 2024]
 52. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, McCoy AB, Sittig DF, Wright A. Assessing the Value of ChatGPT for Clinical Decision Support Optimization. *medrxiv*; 2023. doi: 10.1101/2023.02.21.23286254
 53. Davies NP, Wilson R, Winder MS, Tunster SJ, McVicar K, Thakrar S, Williams J, Reid A. ChatGPT sits the DFPH exam: large language model performance and potential to support public health learning. *BMC Medical Education* 2024 Jan;24(1):57. doi: 10.1186/s12909-024-05042-9

54. Tang R, Han X, Jiang X, Hu X. Does Synthetic Data Generation of LLMs Help Clinical Text Mining? arXiv; 2023. Available from: <http://arxiv.org/abs/2303.04360> [accessed Oct 12, 2024]
55. Peikos G, Symeonidis S, Kasela P, Pasi G. Utilizing ChatGPT to Enhance Clinical Trial Enrollment. arXiv; 2023. Available from: <http://arxiv.org/abs/2306.02077> [accessed Oct 12, 2024]
56. Huemann Z, Lee C, Hu J, Cho SY, Bradshaw TJ. Domain-adapted Large Language Models for Classifying Nuclear Medicine Reports. *Radiology: Artificial Intelligence* 2023 Nov;5(6):e220281. doi: 10.1148/ryai.220281
57. Hulman A, Dollerup OL, Mortensen JF, Fenech ME, Norman K, Støvring H, Hansen TK. ChatGPT- versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center. Kolahi J, editor. *PLOS ONE* 2023 Aug;18(8):e0290773. doi: 10.1371/journal.pone.0290773
58. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, Xu Z, Ding Y, Durrett G, Rousseau JF, Weng C, Peng Y. Evaluating large language models on medical evidence summarization. *npj Digital Medicine* 2023 Aug;6(1):158. doi: 10.1038/s41746-023-00896-7
59. Perlis RH. Research Letter: Application of GPT-4 to select next-step antidepressant treatment in major depression. *medrxiv*; 2023. doi: 10.1101/2023.04.14.23288595
60. Taira K, Itaya T, Hanada A. Performance of the Large Language Model ChatGPT on the National Nurse Examinations in Japan: Evaluation Study. *JMIR Nursing* 2023 Jun;6:e47305. doi: 10.2196/47305
61. King R, Samaan JS, Yeo YH, Mody B, Lombardo DM, Ghashghaei R. Appropriateness of ChatGPT in answering heart failure related questions. *medrxiv*; 2023. doi: <https://doi.org/10.1101/2023.07.07.23292385>
62. Rahimzadeh V, Kostick-Quenet K, Blumenthal Barby J, McGuire AL. Ethics Education for Healthcare Professionals in the Era of ChatGPT and Other Large Language Models: Do We Still Need It? *The American Journal of Bioethics* 2023 Oct;23(10):17–27. doi: 10.1080/15265161.2023.2233358
63. Heston TF. Evaluating Risk Progression in Mental Health Chatbots Using Escalating Prompts. *medrxiv*; 2023. doi: 10.1101/2023.09.10.23295321
64. Marafino BJ, Liu VX. Performance of a large language model (ChatGPT-3.5) for Pooled Cohort Equation estimation of atherosclerotic cardiovascular disease risk. *medrxiv*; 2023. doi: 10.1101/2023.08.11.23293957
65. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, Moningi S, Qian JM, Goldstein M, Harper S, Aerts HJWL, Catalano PJ, Savova GK, Mak RH, Bitterman DS. Large language models to identify social determinants of health in electronic health records. *npj Digital Medicine* 2024 Jan;7(1):6. doi: 10.1038/s41746-023-00970-0
66. Zakka C, Chaurasia A, Shad R, Dalal AR, Kim JL, Moor M, Alexander K, Ashley E, Boyd J, Boyd K, Hirsch K, Langlotz C, Nelson J, Hiesinger W. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. arXiv; 2023. Available from:

<http://arxiv.org/abs/2303.01229> [accessed Oct 12, 2024]

67. Yuan J, Tang R, Jiang X, Hu X. Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching. arXiv; 2023. Available from: <http://arxiv.org/abs/2303.16756> [accessed Oct 12, 2024]
68. Rabbani N, Brown C, Bedgood M, Goldstein RL, Carlson JL, Pageler NM, Morse KE. Evaluation of a Large Language Model to Identify Confidential Content in Adolescent Encounter Notes. medrxiv; 2023. doi: <https://doi.org/10.1101/2023.08.25.23294372>
69. Guo E, Gupta M, Sinha S, Rössler K, Tatagiba M, Akagami R, Al-Mefty O, Sugiyama T, Stieg PE, Pickett GE, De Lotbiniere-Bassett M, Singh R, Lama S, Sutherland GR. neuroGPT-X: Towards an Accountable Expert Opinion Tool for Vestibular Schwannoma. medrxiv; 2023. doi: [10.1101/2023.02.25.23286117](https://doi.org/10.1101/2023.02.25.23286117)
70. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards Interpretable Mental Health Analysis with Large Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 2023. p. 6056–6077. doi: [10.18653/v1/2023.emnlp-main.370](https://doi.org/10.18653/v1/2023.emnlp-main.370)
71. Liu Z, Wu Z, Hu M, Zhao B, Zhao L, Zhang T, Dai H, Chen X, Shen Y, Li S, Li Q, Li X, Murray B, Liu T, Sikora A. PharmacyGPT: The AI Pharmacist. arXiv; 2024. Available from: <http://arxiv.org/abs/2307.10432> [accessed Oct 12, 2024]
72. Kane MJ, King C, Esserman D, Latham NK, Greene EJ, Ganz DA. A compressed large language model embedding dataset of ICD 10 CM descriptions. BMC Bioinformatics 2023 Dec;24(1):482. doi: [10.1186/s12859-023-05597-2](https://doi.org/10.1186/s12859-023-05597-2)
73. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Seifman MA. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. ANZ Journal of Surgery 2024 Feb;94(1–2):68–77. doi: [10.1111/ans.18666](https://doi.org/10.1111/ans.18666)
74. Sarker S, Qian L, Dong X. Medical Data Augmentation via ChatGPT: A Case Study on Medication Identification and Medication Event Classification. arXiv; 2023. Available from: <http://arxiv.org/abs/2306.07297> [accessed Oct 12, 2024]
75. Ma L, Han J, Wang Z, Zhang D. CephGPT-4: An Interactive Multimodal Cephalometric Measurement and Diagnostic System with Visual Large Language Model. arXiv; 2023. Available from: <https://arxiv.org/abs/2307.07518>
76. Lu Y, Srinivasan G, Preum S, Pettus J, Davis M, Greenburg J, Vaickus L, Levy J. Assessing the Impact of Pretraining Domain Relevance on Large Language Models Across Various Pathology Reporting Tasks. 2023. doi: [10.1101/2023.09.10.23295318](https://doi.org/10.1101/2023.09.10.23295318)
77. Chen S, Li Y, Lu S, Van H, Aerts HJWL, Savova GK, Bitterman DS. Evaluating the ChatGPT family of models for biomedical reasoning and classification. Journal of the American Medical Informatics Association 2024 Apr 3;31(4):940–948. doi: [10.1093/jamia/ocad256](https://doi.org/10.1093/jamia/ocad256)
78. Kirk R, Mediratta I, Nalmpantis C, Luketina J, Hambro E, Grefenstette E, Raileanu R. Understanding the Effects of RLHF on LLM Generalisation and Diversity. arXiv; 2023. doi: <https://arxiv.org/abs/2307.10432>

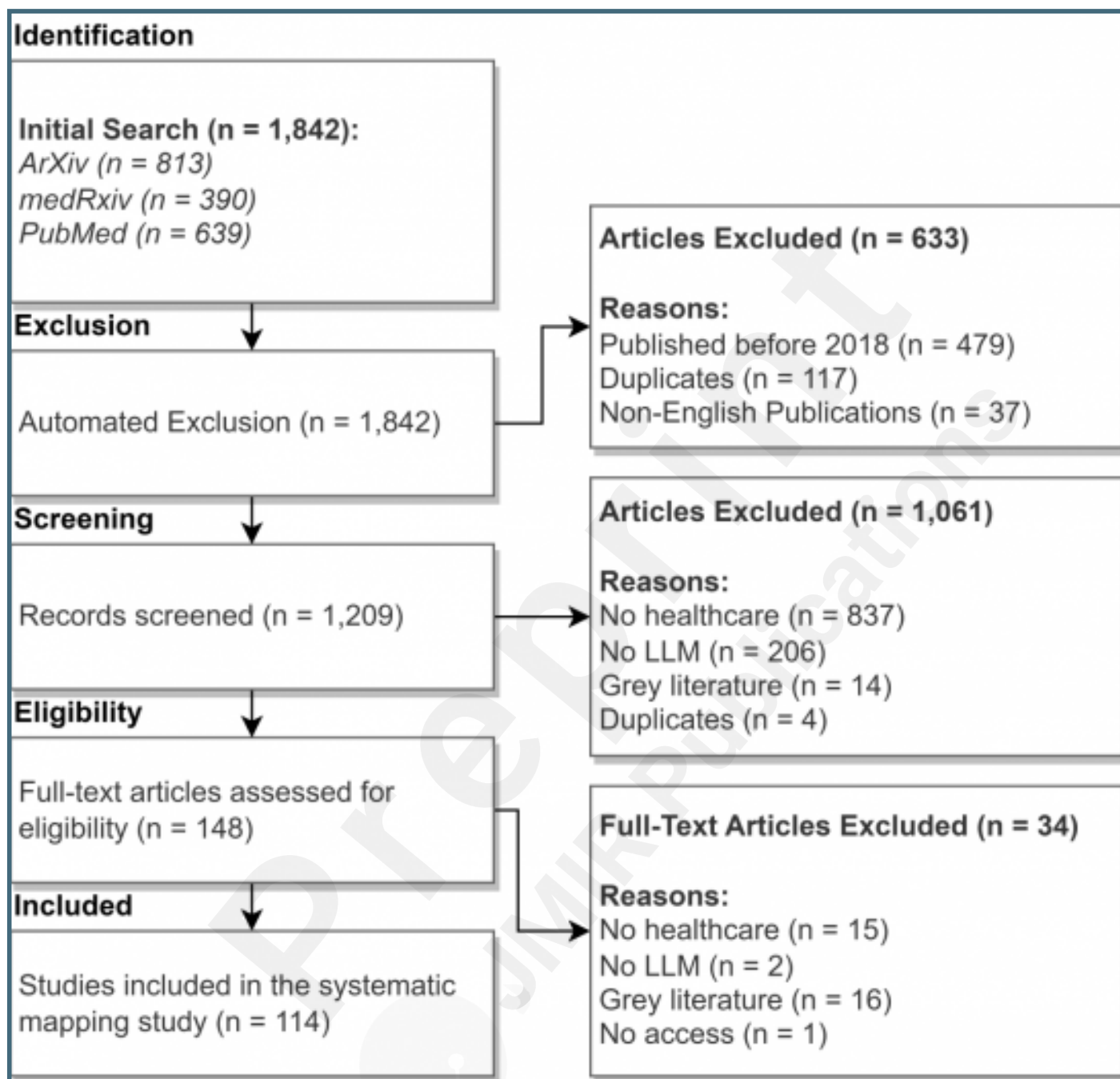
10.48550/ARXIV.2310.06452

79. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, Floridi L. The ethics of AI in health care: A mapping review. *Social Science & Medicine* 2020 Sep;260:113172. doi: 10.1016/j.socscimed.2020.113172
80. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S, Sheller M, Summers RM, Trask A, Xu D, Baust M, Cardoso MJ. The future of digital health with federated learning. *npj Digital Medicine* 2020 Dec;3(1):119. doi: 10.1038/s41746-020-00323-1
81. Buçinca Z, Malaya MB, Gajos KZ. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM Human-Computer Interaction* New York, NY, USA: Association for Computing Machinery; 2021 Apr;5(CSCW1). doi: 10.1145/3449287
82. Xu L, Sanders L, Li K, Chow JC, others. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR cancer JMIR Publications Inc., Toronto, Canada; 2021;7(4):e27850.*
83. von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Walczak M, Pfrommer J, Pick A, Ramamurthy R, Garcke J, Bauckhage C, Schuecker J. Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transaction on Knowledge and Data Engineering* 2021;35(1):614–633. doi: 10.1109/TKDE.2021.3079836
84. Leiser F, Rank S, Schmidt-Kraepelin M, Thiebes S, Sunyaev A. Medical informed machine learning: A scoping review and future research directions. *Artificial Intelligence in Medicine* 2023 Nov;145:102676. doi: 10.1016/j.artmed.2023.102676

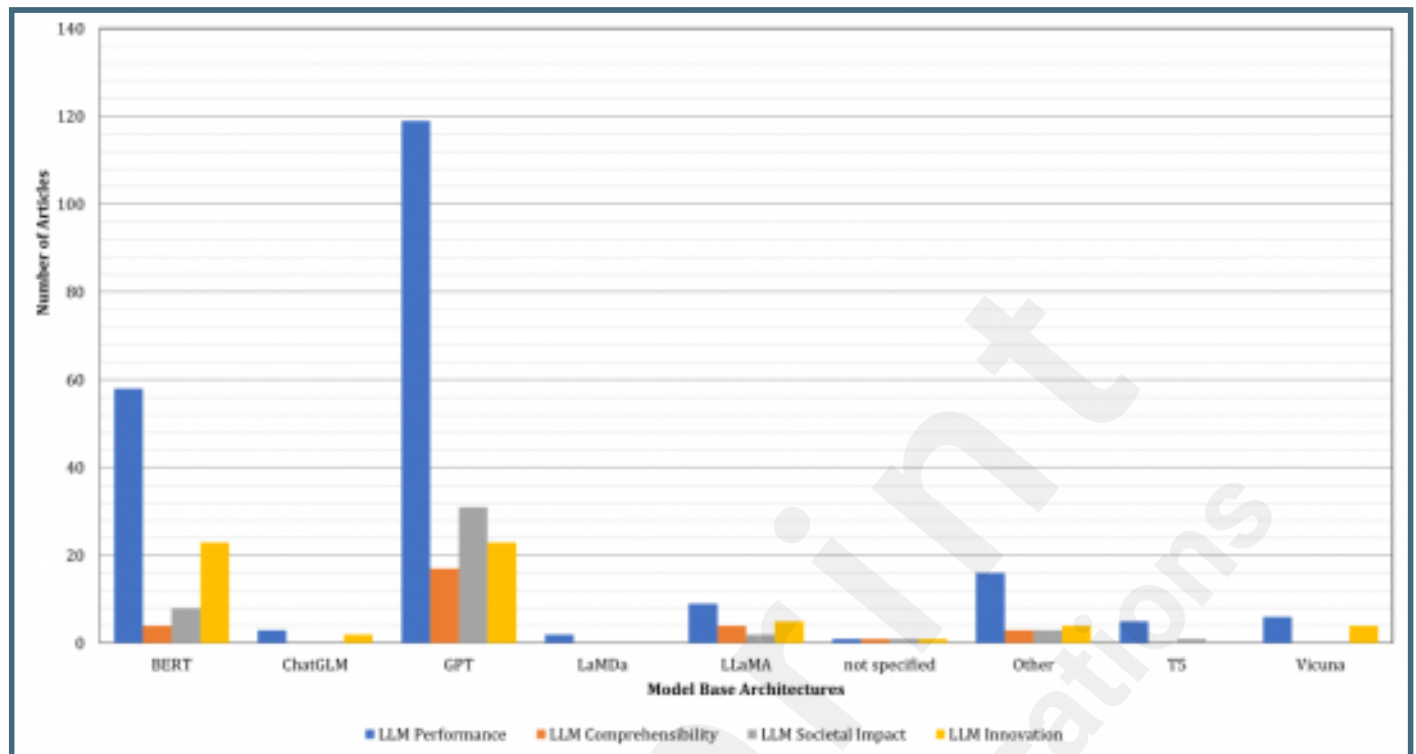
Supplementary Files

Figures

PRISMA chart.



Number of manuscripts per LLM model architectures and research focus.



Multimedia Appendixes

Overview of all included publications along with our coding dimensions and our analysis of each dimension for each research focus.

URL: <http://asset.jmir.pub/assets/5a46bf2631dfa73d232ab1c9fb11727d.xlsx>

Additional Tables and Figures for the analysis of each dimension.

URL: <http://asset.jmir.pub/assets/bb48e72c11a8e151725b21ada66106e1.docx>

PRISMA Checklist.

URL: <http://asset.jmir.pub/assets/c4ee0e965f8d610b8c1b58d44b6ace21.docx>

