Cureus
Part of SPRINGER NATURE

Open Access Review Article

Understanding Large Language Models in Healthcare: A Guide to Clinical Implementation and Interpreting Publications

Review began 03/26/2025
Review ended 04/14/2025
Published 04/16/2025

DOI: 10.7759/cureus.82397

Julia Maslinski [1, 2], Rachel Grasfield [3, 1], Raghav Awasthi [4, 1], Shreya Mishra [5, 1], Dwarikanath Mahapatra [1], Piyush Mathur [6, 1, 7]

1. Artificial Intelligence, BrainXAI ReSearch, BrainX LLC, Cleveland, USA 2. Integrative Neuroscience, Binghamton University, Binghamton, USA 3. Medicine and Health Sciences, University of Iowa, Iowa City, USA 4. Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, USA 5. Data Science, Lerner Research Institute, Cleveland Clinic, Cleveland, USA 6. Anesthesiology and Perioperative Medicine, Cleveland Clinic, Cleveland, USA 7. Anesthesiology, Case Western Reserve University School of Medicine, Cleveland, USA

Corresponding author: Piyush Mathur, pmathurmd@gmail.com

Abstract

Large language models (LLMs) have generated excitement and interest in their capability to impact various facets of healthcare delivery. However, the rapidly expanding literature on LLMs presents challenges in understanding recent work, associated terminology, and potential applications for healthcare professionals. In this review, we discuss the development and evolution of LLMs, especially in healthcare. We provide a description of the key terminologies associated with LLMs to improve the understanding of these terms and their application context in healthcare. Evaluation of the experiments and research related to LLMs is fundamentally important for clinicians. Thus, we provide a description of evaluation methodologies used in LLM research. Lastly, through illustrative examples of research in application of LLMs in healthcare, we showcase the opportunities to leverage this state-of-the-art artificial intelligence (AI) technique with considerations for clinical and administrative adoption both by the patients and healthcare professionals. Through this review, we hope to equip the healthcare professionals with the knowledge they need to understand LLM in healthcare research.

Categories: Medical Education, Healthcare Technology
Keywords: artificial intelligence in medicine, deep learning artificial intelligence, health professional's education, large language models (llms), large language models (llms) in medicine, research

Introduction And Background

The number of publications related to artificial intelligence (AI) in healthcare using large language models (LLMs) is increasing across all aspects of healthcare [1-4]. This is further substantiated by the emergence of journals solely dedicated to AI in healthcare, which actively support research and publications related to LLMs in healthcare [5]. We have observed a similar ‘infodemic’ for image-based AI in healthcare research using deep learning in the fields of radiology, gastroenterology, and cardiology [6].

Since the arrival of generative pre-trained transformer (GPT), there is increasing research work being done to understand where and how these powerful models can be applied in healthcare [7]. Over the last couple of years, there has been a rapid growth in not only generic LLMs, but also in numerous healthcare-specific LLMs trained on healthcare data, including electronic health record data (Figure *1*) [8]. New frameworks for LLM implementation, such as Langchain, are being developed and refined at a rapid pace. Healthcare-specific tasks, such as summarization of clinical notes, ambient documentation of clinician-patient interaction, and patient question answering, are being developed and trialed. As the use of LLMs increases, their pitfalls and challenges in utilization, ranging from inaccuracies, hallucination generation, bias, and cost, are being realized. Increasingly, research is also focusing on understanding these issues in a better way and how to mitigate them [9].

With the excitement about the use of LLMs, large amounts of resources are being allocated for their rapid development and integration in healthcare.This will likely support and promote continued LLM research and implementation in healthcare at a rapid pace. It is important that as the number of publications related to this new and evolving technology in healthcare grows, clinicians understand the basics of LLMs to be able to read and interpret publications related to this domain. This primer aims to educate clinicians and encourage them to participate in research related to LLM applications in healthcare.

How to cite this article
Maslinski J, Grasfield R, Awasthi R, et al. (April 16, 2025) Understanding Large Language Models in Healthcare: A Guide to Clinical Implementation and Interpreting Publications. Cureus 17(4): e82397. DOI 10.7759/cureus.82397

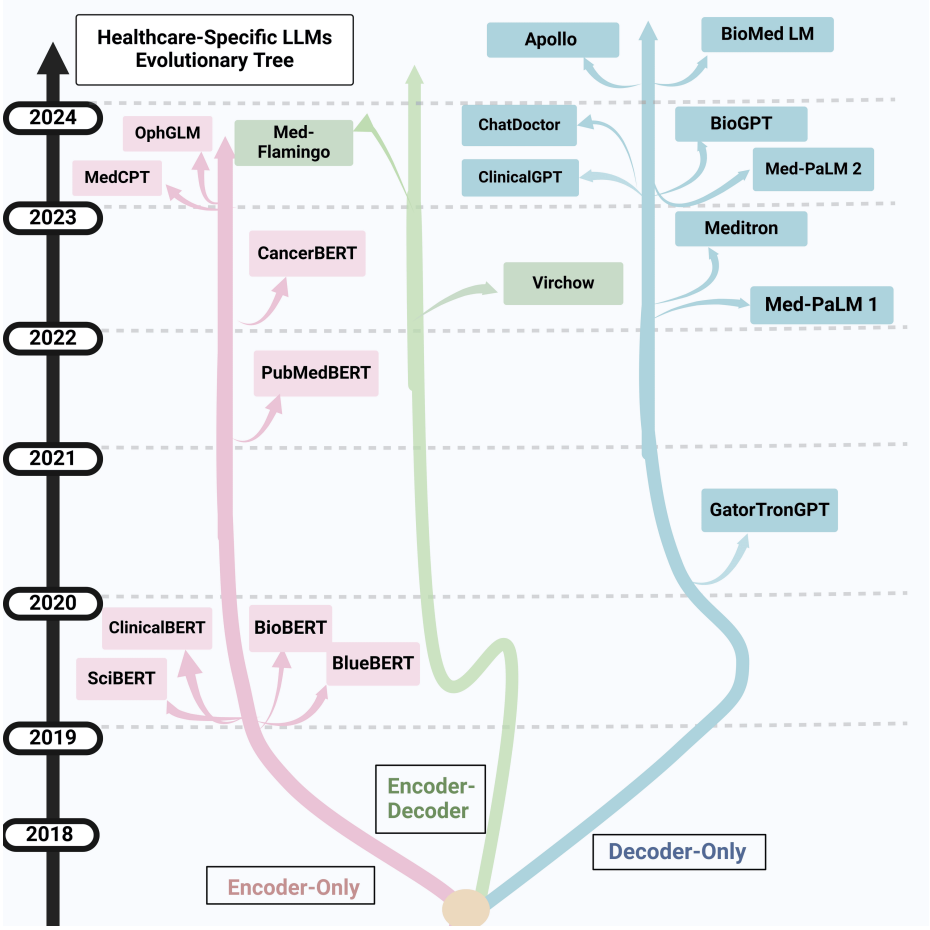

FIGURE 1: Healthcare LLM development over the years

Figure credit: Julia Maslinski

LLM: Large language model

Review

LLM basics

LLMs are a type of AI model (Table *1*) that are developed with advanced deep-learning neural network techniques. Most LLMs architecture is inspired by the transformer architecture, which consists of two key blocks: the encoder and the decoder [10]. Both the encoder and decoder are multilayered neural networks that use self-attention mechanisms and process data in a feed-forward manner. The encoder processes the data and encodes it to generate the desired output such as in classification of diseases. The decoder takes the encoded data and decodes it for the desired output such as text summarization. LLMs can be classified into three categories based on the encoder and decoder. The encoder-only models were proposed in the bidirectional encoder representations from transformers (BERT) paper and later adapted for healthcare applications with models such as BioBERT, PubMedBERT, ClinicalBERT, and others [11-14]. Models such as BERT can learn contextual representations of words and sentences, encoding the relationships between symptoms and diseases (Table *1*). As in ClinicalBERT, readmission prediction can be performed by training the models on discharge summaries and physician notes in the intensive care unit. The decoder-only models include only the decoder part of the transformer, such as GPT models, including healthcare models like ClinicalGPT and BioGPT [15-17]. These models have been shown to perform well at tasks such as medical question and answering in examination scenarios [16]. The encoder-decoder models incorporate both the encoder and decoder, such as Med-Flamingo and Virchow [18,19]. Models such as Med-Flamingo can extend to generative medical question answering abilities to include multimodal data, such as a combination of images and text, to address visual question and answering in examination problems [18].

Terms	Explanation
LLMs	A type of AI model trained on large amounts of data to generate language responses to user input.
Transformer Architecture	A type of neural network model that is used in natural language processing to develop LLMs.
BERT	A transformer architecture-based AI generative model that many LLMs are based on. It processes text from both directions, left-to-right and right-to left, allowing it to more efficiently grasp the context.
GPT	A transformer architecture-based AI generative model that many LLMs are based on. GPT is trained for natural language-processing tasks that can use previous words to predict the next words and form coherent sentences. It generates text sequentially rather than bidirectionally, but is usually trained on more data than BERT.
PLMs	A framework LLM that can be used in their development, as they are already trained on large data sets. PLMs can be customized to fit the uses of the specific LLM being developed.
Fine-Tuning	Training an LLM on new data and adjusting the parameters to ensure it is well-suited for all types of inputs to deliver the desired output. Training on domain-specific data allows specialization of general baseline PLMs.
RLHF	Optimization method for LLMs where the model is updated based on positive and negative human feedback to improve the model over time. This is a supervised learning method as human-labeled data is used to train the model.
RM	A part of RLHF where human-labeled data, usually a ranking of best to worst response from a sample of responses, trains the RM. This model is then used to predict which response the labeler would rank as best, which is then used as a reward function to improve the PLM.
PPO	Another part of RLHF where the RM provides a reward for the output and this reward updates the final LLM.
Prompt	User input that the LLM generates a response to, based on patterns it learned during the training process.
Prompt Engineering	An optimization method for LLMs in which the inputs are designed and formatted in a way the LLM can understand to produce a desired output.
Zero Shot	A learning method for LLMs in which they can generalize and respond to inputs they have never seen before. It is one of the key features of LLMs that make them so useful to all types of inputs and data.
Temperature	A randomness factor of LLMs that can be adjusted by the user. A low temperature means a less varied response, while a high temperature allows a more creative and diverse output to the prompt.
RAG	Optimizes LLM output by integrating data from external sources before producing a response. This allows the model to have access to current,referenced and reliable data beyond its original training set.
LLM Hallucination	The LLM generates grammatically correct and coherent sentences, but they have factually incorrect information.

TABLE 1: Terms related to LLMs and their explanations

LLM: Large language model; AI: Artificial intelligence; BERT: Bidirectional encoder representations from transformers; GPT: Generative pre-trained transformer; PLM: Pre-trained language model; RLHF: Reinforcement learning from human feedback; RM: Reward model; PPO: Proximal policy optimization; RAG: Retrieval augmented generation

LLMs are trained on extensive datasets, including text from articles, books, and websites, aiming to generate human-like responses [20]. By learning from vast amounts of data, LLMs recognize patterns and respond to human input in ways that can perform tasks such as language translation, summarization, answering questions, and text generation [21]. In contrast, traditional deep learning models use far fewer parameters than LLMs and are usually specialized for one task such as determining stroke risk based on patient information input. The recently developed LLMs have billions of parameters, are trained on vast amounts of data, utilize massive computational resources, have broad applicability, and can adapt and update responses in real-time based on user input [21]. The real-time adaptation and broad-use aspect of LLMs that are not present in other traditional deep-learning algorithms make their applicability much greater, especially for healthcare.

Development of LLMs

LLMs are developed using two methods, both of which have the same final steps. The first method is pre-training LLMs from a defined set of data (Figure *2*). After the objective of the model is identified and an extensive amount of data is available for training, the model can be developed [22]. Next, the model is

trained on the data, which may take days or weeks depending on the size of the data and computing resources.

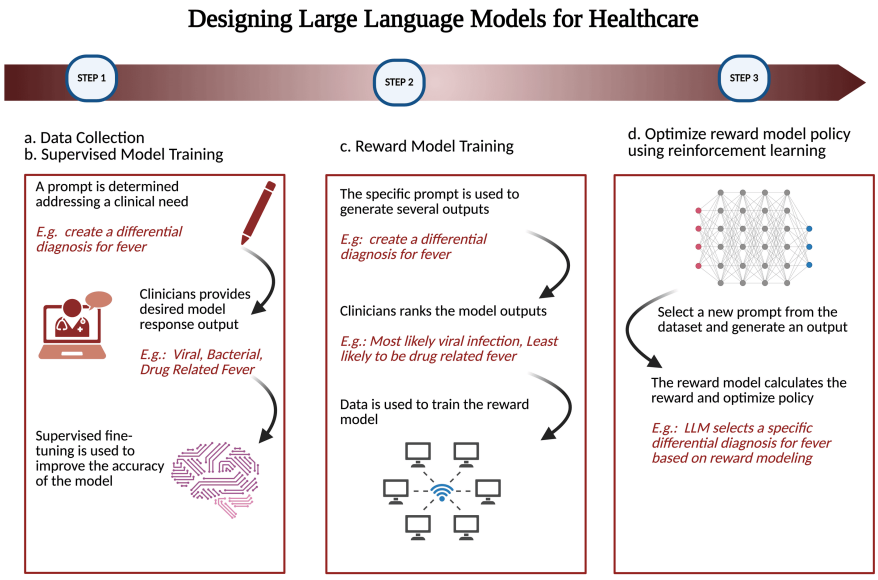

FIGURE 2: Flowchart for designing and developing healthcare-specific LLM

Figure Credit: Rachel Grasfield and Raghav Awasthi

LLM: Large language model

The second method involves using existing pre-trained language models (PLMs), which is less resource-intensive as it uses an already-existing model as the base for the new LLM. With a PLM, the steps of obtaining extensive training data and selecting an architecture for the model can be skipped. To develop a model with this method, one must first find a well-suited PLM by considering both the model size and the training data that was used. It is important that the training data is diverse and useful for the model being developed. The chosen PLM should fit the intended uses of the LLM being developed.

Once the model is developed using either of the methods previously described, validation data should be used to validate the model's performance on data other than the training data. This step is crucial for both pre-training LLMs and using PLMs to ensure the model responds well to never-before-seen data and is able to generalize to generate responses based only on the training data it was exposed to. The model should be modified accordingly and even re-trained on more data if it is not performing to the expectations. Further, the model must be fine-tuned to ensure it is responding correctly based on user input. The most common method used for fine-tuning is reinforcement learning from human feedback (RLHF). RHLF is a supervised learning method as human-labeled data is used to train the model (Figure *2*). The method involves a team of humans tasked with preferencing the responses provided by the model. For a given prompt, a human labeler will indicate the desired response, and this is used to fine-tune the model using supervised learning. Next, a given prompt and a few outputs will be sampled, for which the labeler must rank the best to worst response, which trains the reward model (RM) (Figure *2*). The trained RM is then used to predict which response the labeler would rank as best. It is then used as a reward function to fine-tune the baseline model, such as proximal policy optimization (PPO), another reinforcement learning algorithm where the RM provides a reward for the output, and this reward updates the LLM [23]. RLHF is used to optimize the finalized LLM over time through positive and negative human feedback to constantly improve the model over time.

Healthcare-specific LLM development

While many of the popular general-purpose LLMs, such as GPT, are valuable for a broad range of tasks, they are not specifically developed for use in clinical medicine. However, there are numerous emerging LLMs created specifically for clinical purposes (Figure *1*). GatorTronGPT is an example of such a model that was trained on 277 billion words of mixed clinical and English text, exhibiting state-of-the-art performance in biomedical datasets [24]. GatorTronGPT was developed specifically to generate and comprehend clinical text, which is useful in medical documentation [8]. Similarly, ClinicalBERT was successful in predicting 30-day hospital readmission using only clinical notes from the first few days and discharge summaries [25]. Another example of a healthcare-specific LLM is Google's Med-PaLM, whose performance on the United

States Medical Licensing Examination (USMLE) style questions had an accuracy exceeding 67% on the MedQA dataset [2]. Med-PaLM is designed to answer medical questions with high accuracy [2]. LLMs specifically developed for healthcare tasks such as generating clinical notes, predicting readmission, and answering medical questions are extremely useful and are likely to be increasingly prevalent in clinical practice soon.

Data Type and Needs for LLMs

To successfully develop an LLM, extensive amounts of data is necessary. A diverse dataset taken from articles, websites, and books allows training of the LLM for optimal performance. The type of data necessary depends on the intended uses of the LLM being developed. If the LLM is being developed for a specific task, it should be trained and optimized with a vast amount of data from that specific topic. For example, to train a healthcare focused LLM to help perform clinical tasks such as ClinicalGPT, training on electronic health record data would be ideal [16].

To optimize the quality of the training data gathered, deduplication methods are often used. Deduplication allows sampling of only unique instances of data, so repeats are removed. Furthermore, filtering of the training data involves analysis of the data quality, which leads to removing some of the unneeded data in training the model. Finally, a remixing stage is utilized to ensure the diversity and broad applicability of the dataset. After this stage, data may be added to fill in underrepresented domains noticed during this stage.

Scaling laws in the development of LLMs refer to the relationship between model size, amount of training data being used, and the computational resources required for the LLM to be trained. These laws suggest that the LLMs' performance improves with increases in these three factors: model size, data amount, and computational power [26]. The larger the model size, the more likely it is able to perform complex language-related tasks. Similarly, the larger and more diverse training dataset, the better the LLM is likely to perform on a wide-range of inputs. Following this, the larger the size of the LLM, the more likely it needs more computational power to train, but in turn, leads to a better model overall. Large amounts of high-quality and diverse datasets are key to accurate LLM functioning and development, and it is what distinguishes their performance from other AI models.

Validation and Assessment of LLM Performance

Validation and assessment of LLMs can be defined as an evaluation that checks various dimensions of output such as accuracy, fluency, relevance, bias, coherence, potential harm, and hallucinations (Table *2*). Overall evaluation of LLMs can be categorized into two categories: 1) automated evaluation, which includes mathematical formulas or model-based approaches; and 2) human evaluation, where humans manually assess the LLMs' output [27].

Evaluation Indicators	Quantitative Metrics	Explanation
Task Success Rate	Accuracy	Accuracy is a measure to calculate the percentage of the correct predictions by the models.
Fluency	Perplexity score	Perplexity measures the amount of randomness in the generated text by LLM.
Relevance	ROUGE score	A set of metrics used to evaluate how well a generated text such as a summary matches the reference text.
Biasness	LLMBI	LLMBI is the weight-adjusted sums of various dimensions of bias, such as gender, race, etc.
Coherence	Coh-Metrix	Coh-Metrix is an exhaustive computational tool that evaluates the cohesion of text using over 200 measures.
Hurtfulness	HONEST	HONEST is a scoring evaluation system designed to assess the harmfulness of sentence completions generated by language models.
Hallucinations	SelfCheckGPT	SelfCheckGPT is a method to detect the hallucination of LLMs.

TABLE 2: Key indicators and their corresponding metrics for evaluation[28-34]

ROGUE: Recall-Oriented Understudy for Gisting Evaluation; LLMBI: Large Language Model Bias Index; HONEST: Hurtfulness of Language Model Sentence Completion; GPT: Generative pre-trained transformer

Automated Evaluation

Automatic evaluation of natural language processing (NLP) systems, in general, is designed to be task-specific . For example, the Bilingual Evaluation Understudy (BLEU) score is more appropriate for language translation, the Recall-Oriented Understudy for Gisting Evaluation (ROGUE) score is more relevant for text summarization, and the perplexity score is used to evaluate the quality of text generation (Table *2*) [35]. For text classification, confusion matrix-based scores, such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC), are utilized. These metrics have different interpretations. For instance, higher BLEU and ROUGE scores indicate better model performance, while lower perplexity suggests better performance [30]. However, we understand that LLMs like GPT, Large Language Model Meta AI (LLaMA), BigScience Large Open-Science Open-Access Multilingual Language Model (BLOOM), Falcon, etc., are capable of performing multiple tasks such as language translation, text summarization, sentiment analysis, and text classification [36-38]. Therefore, benchmarks like SuperGLUE, the lm-eval Python package by BIG-Bench, and MT-Benchmark, have been developed [39-41]. These benchmarks provide multiple task-specific datasets and metrics to evaluate LLMs and benchmark them. This benchmarking helps in comparing models on different tasks and making better choices of models for different use cases. Bias, leakage of private data, and toxicity are additional key factors to consider when evaluating LLMs.

Human Evaluation

Despite the development of these benchmarks and new automatic evaluation techniques, it is widely reported that while models may perform well on these benchmark datasets, they often fail to perform equally well when faced with user-specific datasets [42]. This is commonly due to the lack of information in the data on which the models are trained. For example, the instruction-tuning datasets of OpenAI models are unknown (not open-source) [43]. Also, automatic evaluations techniques are not robust to various adversarial attacks, including lexical overlap and factuality errors. Previous claims suggest that existing evaluation metrics are not robust against even simple perturbations and may disagree with scores assigned by humans to the perturbed output [44,45]. Furthermore, there is very little correlation between these metrics and human judgment [46]. Considering the challenges associated with the automatic evaluation of LLMs, human evaluation is one of the key methods to robustly assess LLMs. Human evaluation has been recently conducted in recent LLM publications and remains the gold standard method to evaluate LLM performance, particularly in fields like healthcare [47-49]. In a review of 142 publications, researchers demonstrated significant gaps in variability, generalizability and reliability of human evaluations of LLM outputs [49]. To addresses this gap, the authors have proposed QUEST, as a comprehensive and practical framework for human evaluation of LLMs [49]. Similarly, a robust set of metrics and a web application, named HumanELY, has also been recently proposed to enhance the standardization and efficacy of human evaluation [48].

Human evaluation of LLM outputs has its own set of challenges. It is resource intensive, and has a lack of guidance on issues such as: framing of the questions, number of samples needed, number of human annotators to minimize inter-annotator agreement (IAA), and who should perform the evaluation. To overcome challenges associated with human evaluation, recent research has explored how LLMs themselves can be utilized to perform human evaluation [48]. With the continuous development of diverse evaluation methods, the hope is to have more robust and reliable validation and assessment approaches for LLMs, thereby building trustworthy LLM-based systems.

Considerations for clinical implementation

Clinical

LLMs have the capacity to be implemented in every stage of clinical diagnosis: diagnosis, management, and prognosis (Figure *3*). Many studies have tried GPT to assist with diagnosis, although with variable results [50-52]. In a study of 100 patient medical records, ChatGPT was used to diagnose knee or hip arthrosis and recommend treatment plans such as pain management and surgery [53]. While the model did not perform perfectly, there was an 83% agreement between the ChatGPT and physicians on diagnosis. Treatment plans for pain approached statistical significance, and quality of life plan recommendations were statistically significant, indicating that ChatGPT could be a future diagnostic tool in analyzing patient records to prepare a diagnosis and offer treatment plans [53]. Similarly, in a study of patients with Glioma, ChatGPT responses were assessed for adjuvant therapy decision-making [54]. Experts who formed the tumor board rated the responses as ‘good’ for the ChatGPT recommendations related to treatment recommendations and therapy regimen. Similar to what the authors concluded in this study, many others have recommended a human-in-the-loop approach for complex medical decision-making with these model responses serving as useful aids [54]. LLMs have also been trialed by clinicians other than physicians for overall care delivery of patients too. For example, using ChatGPT to identify patients with type II diabetes, the model was able to suggest diet plans that matched with a nutritionists’ recommendation [55].

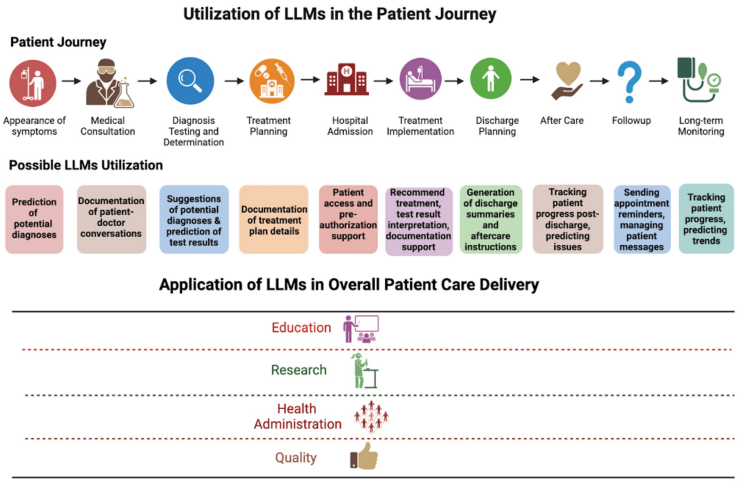

FIGURE 3: Opportunities and applications of LLMs in patient-care delivery

Figure Credit: Shreya Mishra

LLM: Large language model

Administrative

LLMs can play an important role in quickly and accurately completing various administrative tasks that are currently completed by clinicians. Clinical note documentation and summarization has been one of the key opportunities identified for use of LLM, which might help decrease clinician burnout. Use of AI scribes in some early trials has shown to be very promising as rated by physicians who used it for 10 weeks across many different specialties and many different locations [56]. Similarly, for other tasks performed by clinicians, such as writing prescriptions, ChatGPT was trained on textbook information on International Classification of Diseases (ICD) codes and rehabilitative pharmacy following a case study of a stroke [57]. It was able to accurately complete the required prescriptions and outline rehabilitation goals, as well as complete specialized therapeutic information. While only tested on a single case study, this supports a potential larger use of LLMs in writing and sending prescriptions to pharmacies, or completing medical plans and treatment goals for medical records [57].

LLMs have the ability to improve clinician functioning by maximizing the area of medicine they find most meaningful. They may be a useful tool in relieving some of this stress by providing compassionate, empathetic responses to medical questions from patients, which physicians can then review before sending. In one study, the AI Chatbot responding to patient messages was found to be significantly higher quality regarding empathy (“bedside manner”) than the physician [58].

LLMs' role in education has been one of the key focus areas in the early trials, especially using GPT. From medical school education to training specialists such as neurosurgeons and nursing education, many early studies have tried LLMs [59-61]. Many of the early studies have also compared the performance of GPT on various medical examinations, sometimes even correlating them with clinician performance [62,63]. These are early explorations in the use of LLMs in healthcare, which, over the next few years, are likely to be continuously experimented upon and will have the use cases defined and refined.

Research in medicine is very labor-intensive and inefficient. Many early use cases of LLMs in the same have been recently explored including, evidence synthesis, scientific writing, and patient record matching for clinical trial enrollment, amongst many others [64-67]. Interestingly, while many scientific journals have restricted the use of LLMs for scientific writing, the New England Journal of Medicine (NEJM) AI has allowed their use as long as authors take complete responsibility for the content and provide transparency [5]. Use of LLMs in research is a very promising area that is likely to bring in efficiencies at scale.

Patient Use

Several studies to date have demonstrated the ability of ChatGPT to serve as a bridge between physicians and patients by providing tailored responses and educational materials that best fit the patient’s level of

medical literacy [68-70]. This provides a potential resource for both parties, as LLMs can serve a supportive role in providing patients with useful resources and answers to their questions, while also allowing physicians to communicate their understanding of diagnoses, prognoses, and commonly asked questions in a way that maximizes patient understanding.

ChatGPT has demonstrated that it can provide accurate and appropriate answers to the questions about sleep apnea from a patient [69]. Prompts to ChatGPT can decrease the response grade level, improving readability while providing an acceptable level of medical information [69]. This supports the theory that ChatGPT can improve medical literacy in patients by providing accurate responses to their medical questions within the frame of educational answers that best fit their needs [69]. An area of medical education that may not be accessible to the general public is the information provided in medical journals, which may surpass the reading level or comfort of patients. For example, in the field of urology, ChatGPT has been shown to be able to read and analyze academic papers and provide a summary that fits the patient's level of understanding [70]. While promising, limitations such as hallucinations, accuracy, and readability, of such applications still needs further evaluation [49].

Implementing LLMs in clinical practice

While studies have shown the potential benefits of LLM implementation as a clinical decision-making tool in a research framework, the reality of actual implementation may pose more challenges. Implementation of LLMs in healthcare facilities requires a strategy around managing the following key aspects of deployment: 1) data storage, 2) LLM deployment, and 3) healthcare professional interface. Data storage, especially of electronic medical record (EMR) data, for most hospitals is managed using on-premise servers or cloud services [71]. LLM deployment can occur on a server/cloud within the EMR or externally, which interacts with the stored data. Delivery of LLM output is most likely to occur through EMRs or other devices including mobile. All these require significant additional resources, including bandwidth for data transmission, cybersecurity applications, and trained professionals for managing them.

Additionally, multidisciplinary teams that can guide the implementation strategy, including ethical concerns, legal issues, and resource allocation, similar to other AI model deployments, are important [72]. The type of information stored, both onsite or in the cloud, may contain sensitive and identifying information. Patient healthcare data is protected by medical professionals and regulators using the Health Insurance Portability and Accountability Act (HIPAA) standards. This is a significant issue, considering that chatbots are not immune to data breaches and leaks, and have the capacity to share personal information with those who can manipulate the algorithm. Additionally, chatbots have been theorized to evolve such that physicians can be manipulated into including patient information in their queries, which may be sold to third parties [73].

Measuring and Monitoring Clinical Effect

The purpose of deploying LLMs in a clinical environment is to support clinician decision-making to deliver high-quality care. Measuring the performance of LLMs, both on technical and clinical benchmarks, is essential. While the field of LLM operations (LLMOps) is nascent, it is important to incorporate it in the strategy for LLM deployments [74]. Establishing mechanisms for monitoring model and pipeline lineage, with alerts for detecting model drift and identifying potential malicious user behavior, is critical from a technical point of view. From a clinical point of view, establishing key clinical metrics for clinical decision support delivery, adoption, and feedback need to be established.

Evaluating and Interpreting LLM-Based Publications and Contextualizing the Results

Similar to any other AI model, training data forms the basis of the likely output and performance of the model. One of the key aspects of understanding research related to LLMs is to understand the data they are trained on. There are many models that are trained on generic Internet-extracted data, which may or may not perform well on specific healthcare tasks. LLMs trained on PubMed, patient-clinician interaction data, or electronic health record data are likely to exhibit varying performance across different tasks.

It is also important to understand the differences in performances of these LLMs for different technical tasks such as data summarization, question answering, and data classification. Based on the training data and model architecture, LLMs might behave differently for these tasks. It is important to understand the base performance of these using task-based evaluation metrics, the appropriateness of their use, or alternatives. Many of these models are not static either, and their versions are updated with new data or learning architecture.

Most of the models researched in healthcare publications so far have been used "out of the box" with limited prompt engineering. It is important to understand if any significant modifications were made or can be made to impact model output, such as fine-tuning which might change the performance from that of the base model. Further, determining how these models have been used by including techniques like retrieval

augmented generation (RAG) is important.

Understanding the evaluation and performance metrics used in researching these LLMs is crucial. While technical metrics might be helpful, comprehensive and standardized evaluation methods, including those for human evaluation, are also important. Ultimately, understanding the performance of these LLMs in healthcare applications is necessary. While many have compared the LLM with human performance, understanding the goals of the research and the impact on patient care delivery is key. LLMs might outperform or underperform in some areas of healthcare, but these outcomes must be interpreted cautiously, in the context of precision care delivery and the generalizability of results across various patient populations.

To ensure transparency and standardization in reporting of LLM model for individual prognosis or diagnosis, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-LLM (TRIPOD-LLM) is one of the first guideline with a checklist of 19 main items and 50 subitems [75]. This guideline builds on the prior reporting statement of TRIPOD + AI [75]. Evaluation frameworks, such as HumanELY and QUEST, also provide additional checklists that can be used to report human evaluation of LLM model outputs [48,49]. Standardization of methodology in research using evaluation frameworks and reporting guidelines are likely to improve the quality of publications related to LLM in healthcare in the future.

Conclusions

Availability of state-of-the-art AI models, such as LLMs, is very exciting and spurs large-scale research for their application in healthcare. These models provide us with an unprecedented opportunity to improve all aspects of healthcare, including clinical and administrative tasks. It is important for clinicians to understand the basics of these models as well as their adoption methods and pitfalls to be able to interpret this important research work and related publications.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Piyush Mathur, Julia Maslinski, Rachel Grasfield, Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra

Acquisition, analysis, or interpretation of data: Piyush Mathur, Julia Maslinski, Rachel Grasfield, Raghav Awasthi, Shreya Mishra

Drafting of the manuscript: Piyush Mathur, Julia Maslinski, Rachel Grasfield, Raghav Awasthi, Shreya Mishra

Critical review of the manuscript for important intellectual content: Piyush Mathur, Rachel Grasfield, Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra

Supervision: Piyush Mathur, Julia Maslinski, Rachel Grasfield, Raghav Awasthi, Shreya Mishra

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

No artificial intelligence tools were used in preparation of this manuscript.

References

1. Clusmann J, Kolbinger FR, Muti HS, et al.: The future landscape of large language models in medicine . Commun Med (Lond). 2023, 3:141. 10.1038/s43856-023-00370-1
2. Singhal K, Azizi S, Tu T, et al.: Large language models encode clinical knowledge . Nature. 2023, 620:172-80. 10.1038/s41586-023-06291-2
3. Awasthi R, Mishra S, Grasfield R, et al.: Artificial intelligence in healthcare: 2023 year in review . medRxiv. 2024, 10.1101/2024.02.28.24303482

4. Shah NH, Entwistle D, Pfeffer MA: Creation and adoption of large language models in medicine . JAMA. 2023, 330:866-9. 10.1001/jama.2023.14217
5. Koller D, Beam A, Manrai A, et al.: Why we support and encourage the use of large language models . NEJM AI. 2023, 1:10.1056/AIe2300128
6. Awasthi R, Mishra S, Cywinski JB, Maheshwari K, Khanna AK, Papay FA, Mathur P: Quantitative and qualitative evaluation of the recent artificial intelligence in healthcare publications using deep-learning. medRxiv. 2023, 10.1101/2022.12.31.22284092
7. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, Rizzo C: ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023, 11:1166120. 10.3389/fpubh.2023.1166120
8. Peng C, Yang X, Chen A, et al.: A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023, 6:210. 10.1038/s41746-023-00958-w
9. Yang JF, Jin HY, Tang RX, et al. : Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. arXiv. 2023, 10.48550/arXiv.2304.13712
10. Zhang Y, Liu C, Liu M, Liu T, Lin H, Huang CB, Ning L: Attention is all you need: utilizing attention in AI-enabled drug discovery. Brief Bioinform. 2023, 25: 10.1093/bib/bbad467
11. Devlin J, Chang MW, Lee K, Toutanova K: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. 2019, 10.48550/arXiv.1810.04805
12. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020, 36:1234-40. 10.1093/bioinformatics/btz682
13. Gu Y, Tinn R, Cheng H, et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. 2021, 3:1-23. 10.1145/3458754
14. Alsentzer E, Murphy J, Boag W, Weng WH, Di J, Naumann T, McDermott M: Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop. Rumshisky A, Roberts K, Bethard S, Naumann T (ed): Association for Computational Linguistics, Minneapolis; 2019. 72-8. 10.18653/v1/W19-1909
15. Brown T, Mann B, Ryder N, et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems 33. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (ed): Curran Associates Inc., New York; 2020. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html:1877-901.
16. Wang G, Yang G, Du Z, Fan L, Li X: ClinicalGPT: Large language models finetuned with diverse medical data and comprehensive evaluation. arXiv. 2023, 10.48550/arXiv.2306.09968
17. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon HF, Liu TY: BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings Bioinf. 2022, 23:bbac409. 10.1093/bib/bbac409
18. Moor M, Huang Q, Wu S, et al.: Med-Flamingo: a multimodal medical few-shot learner . arXiv. 2023, 10.48550/arXiv.2307.15189
19. Vorontsov E, Bozkurt A, Casson A, et al.: Virchow: a million-slide digital pathology foundation model. arXiv. 2023, 10.48550/arXiv.2309.07778
20. Blank IA: What are large language models supposed to model?. Trends Cogn Sci. 2023, 27:987-9. 10.1016/j.tics.2023.08.006
21. Meskó B, Topol EJ: The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023, 6:120. 10.1038/s41746-023-00873-0
22. Huang Y, Xu J, Lai J, et al.: Advancing transformer architecture in long-context large language models: a comprehensive survey. arXiv. 2023, 10.48550/arXiv.2311.12351
23. Ouyang L, Wu J, Jiang X, et al.: Training language models to follow instructions with human feedback . arXiv. 2022, 10.48550/arXiv.2203.02155
24. Yang X, Chen A, PourNejatian N, et al.: A large language model for electronic health records . NPJ Digit Med. 2022, 5:194. 10.1038/s41746-022-00742-2
25. Huang K, Altosaar J, Ranganath R: ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. 2019, 10.48550/arXiv.1904.05342
26. Bi X, Chen DL, Chen GT, et al. : DeepSeek LLM: scaling open-source language models with longtermism . arXiv. 2024, 10.48550/arXiv.2401.02954
27. Sottana A, Liang B, Zou K, Yuan Z: Evaluation metrics in the era of GPT- 4: reliably evaluating large language models on sequence to sequence tasks. arXiv. 2023, 10.48550/arXiv.2310.13800
28. The Handbook of Computational Linguistics and Natural Language Processing . Clark A, Fox C, Lappin S (ed): Blackwell Publishing Ltd, Malden; 2010. 10.1002/9781444324044
29. Jelinek F, Mercer RL, Bahl LR, Baker JK: Perplexity—a measure of the difficulty of speech recognition tasks . J Acoust Soc Am. 2005, 62:S63. 10.1121/1.2016299
30. Lin CY: ROUGE: a package for automatic evaluation of summaries . Text Summarization Branches. Association for Computational Linguistics, Barcelona; 2004. 74-81.
31. Oketunji AF, Anas M, Saina D: Large Language Model (LLM) Bias Index—LLMBI . Zenodo. 2023, 10.5281/zenodo.10441700
32. McNamara DS, Graesser AC, McCarthy PM, Cai ZQ: Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, Cambridge, UK; 2014. 10.1017/CBO9780511894664
33. Nozza D, Bianchi F, Hovy D: HONEST: measuring hurtful sentence completion in language models . Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tur D, Beltagy I, Bethard S, Cotterell R, Chakraborty T, Zhou YC (ed): Association for Computational Linguistics, 2021. 2398-406. 10.18653/v1/2021.naacl-main.191
34. Manakul P, Liusie A, Gales MJ: SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. arXiv. 2023, 10.48550/arXiv.2303.08896
35. Papineni K, Roukos S, Ward T, Zhu WJ: BLEU: a method for automatic evaluation of machine translation . Proceedings of the 40th Annual Meeting on Association for Computational Linguistics . Isabelle P, Charniak

E, Lin DK (ed): Association for Computational Linguistics, Philadelphia; 2002. 311-18. 10.3115/1073083.1073135
36. Touvron H, Lavril T, Izacard G, et al. : LLaMA: open and efficient foundation language models. arXiv. 2023, 10.48550/arXiv.2302.13971
37. Le Scao T, Fan A, Akiki C, et al. : BLOOM: a 176B-parameter open-access multilingual language model . arXiv. 2022, 10.48550/arXiv.2211.05100
38. Almazrouei E, Alobeidli H, Alshamsi A, et al.: The Falcon series of open language models . arXiv. 2023, 10.48550/arXiv.2311.16867
39. Sarlin PE, DeTone D, Malisiewicz T, Rabinovich A, et al.: SuperGlue: learning feature matching with graph neural networks. arXiv. 2019, 10.48550/arXiv.1911.11763
40. Srivastava A, Rastogi A, Rao A, et al.: Beyond the imitation game: quantifying and extrapolating the capabilities of language models. arXiv. 2022, 10.48550/arXiv.2206.04615
41. Bai G, Liu J, Bu X, et al.: MT-Bench-101: a fine-grained benchmark for evaluating large language models in multi-turn dialogues. arXiv. 2024, 10.48550/arXiv.2402.14762
42. Liang P, Bommasani R, Lee T, et al. : Holistic Evaluation of Language Models . arXiv. 2023, 10.48550/arXiv.2211.09110
43. Laskar MT, Bari MS, Rahman M, et al. : A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. arXiv. 2023, 10.48550/arXiv.2305.18486
44. Chen Y, Eger S: MENLI: robust evaluation metrics from natural language inference . Trans Assoc Comput Ling. 2023, 11:804-25. 10.1162/tacl_a_00576
45. Sai AB, Dixit T, Sheth DY, et al.: Perturbation checklists for evaluating NLG evaluation metrics . arXiv. 2021, 10.48550/arXiv.2109.05771
46. Moramarco F, Korfiatis AP, Perera M, et al.: Human evaluation and correlation with automatic metrics in consultation note generation. arXiv. 2022, 10.48550/arXiv.2204.00447
47. Reddy S: Evaluating large language models for use in healthcare: a framework for translational value assessment. Inf Med Unlocked. 2023, 41:101304. 10.1016/j.imu.2023.101304
48. Awasthi R, Mishra S, Mahapatra D, et al. : HumanELY: human evaluation of LLM yield, using a novel web-based evaluation tool. medRxiv. 2024, 10.1101/2023.12.22.23300458
49. Tam TY, Sivarajkumar S, Kapoor S, et al.: A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024, 7:258. 10.1038/s41746-024-01258-7
50. Suhag A, Kidd J, McGath M, et al.: ChatGPT: a pioneering approach to complex prenatal differential diagnosis. Am J Obstet Gynecol MFM. 2023, 5:101029. 10.1016/j.ajogmf.2023.101029
51. Wei Q, Cui Y, Wei B, Cheng Q, Xu X: Evaluating the performance of ChatGPT in differential diagnosis of neurodevelopmental disorders: a pediatricians-machine comparison. Psychiatry Res. 2023, 327:115351. 10.1016/j.psychres.2023.115351
52. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J: Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun. 2024, 15:2050. 10.1038/s41467-024-46411-8
53. Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J, Holzapfel DE: Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol. 2023, 24:61. 10.1186/s10195-023-00740-4
54. Haemmerli J, Sveikata L, Nouri A, et al.: ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board?. BMJ Health Care Inform. 2023, 30:e100775. 10.1136/bmjhci-2023-100775
55. Sun H, Zhang K, Lan W, et al.: An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res. 2023, 25:e51300. 10.2196/51300
56. Tierney AA, Gayre G, Hoberman B, et al.: Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal Innov Care Deliv. 2024, 5: 10.1056/CAT.23.0404
57. Zhang L, Tashiro S, Mukaino M, Yamada S: Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. J Rehabil Med. 2023, 55:jrm13373. 10.2340/jrm.v55.13373
58. Ayers JW, Poliak A, Dredze M, et al.: Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023, 183:589-96. 10.1001/jamainternmed.2023.1838
59. Seetharaman R: Revolutionizing medical education: can ChatGPT boost subjective learning and expression? . J Med Syst. 2023, 47:61. 10.1007/s10916-023-01957-w
60. Sevgi UT, Erol G, Doğruel Y, Sönmez OF, Tubbs RS, Güngor A: The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. Neurosurg Rev. 2023, 46:86. 10.1007/s10143-023-01998-2
61. Seney V, Desroches ML, Schuler MS: Using ChatGPT to teach enhanced clinical judgment in nursing education. Nurse Educ. 2023, 48:124. 10.1097/NNE.0000000000001383
62. Wang X, Gong Z, Wang G, et al.: ChatGPT performs on the Chinese National Medical Licensing Examination . J Med Syst. 2023, 47:86. 10.1007/s10916-023-01961-0
63. Saad A, Iyengar KP, Kurisunkal V, Botchu R: Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. Surgeon. 2023, 21:263-6. 10.1016/j.surge.2023.07.001
64. Roberts RH, Ali SR, Hutchings HA, Dobbs TD, Whitaker IS: Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. BMJ Health Care Inform. 2023, 30:e100830. 10.1136/bmjhci-2023-100830
65. Salvagno M, Taccone FS, Gerli AG: Can artificial intelligence help for scientific writing? . Crit Care. 2023, 27:75. 10.1186/s13054-023-04380-2
66. Rozencwajg S, Kantor E: Elevating scientific writing with ChatGPT: a guide for reviewers, editors... and authors. Anaesth Crit Care Pain Med. 2023, 42:101209. 10.1016/j.accpm.2023.101209
67. Gupta SK, Basu A, Nievas M, et al. PRISM: Patient Records Interpretation for Semantic Clinical Trial Matching using Large Language Models . (2024). Accessed: 04/28/2024: http://arxiv.org/abs/2404.15549.

68. Krittanawong C, Virk HU, Kaplin SL, Wang Z, Sharma S, Jneid H: Assessing the potential of ChatGPT for patient education in cardiac catheterization care. JACC Cardiovasc Interv. 2023, 16:1551-2. 10.1016/j.jcin.2023.04.042
69. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS: Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med. 2023, 19:1989-95. 10.5664/jcsm.10728
70. Eppler MB, Ganjavi C, Knudsen JE, et al.: Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson's summaries. Urol Pract. 2023, 10:436-43. 10.1097/UPJ.0000000000000428
71. Liu Y, Chen PH, Krause J, Peng L: How to read articles that use machine learning: users' guides to the medical literature. JAMA. 2019, 322:1806-16. 10.1001/jama.2019.16489
72. Younis HA, Eisa TA, Nasser M, et al.: A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. Diagnostics (Basel). 2024, 14:109. 10.3390/diagnostics14010109
73. Marks M, Haupt CE: AI chatbots, health privacy, and challenges to HIPAA compliance. JAMA. 2023, 330:309-10. 10.1001/jama.2023.9458
74. Aryan A, Nain AK, McMahon A, et al.: The costly dilemma: generalization, evaluation and cost-optimal deployment of large language models. arXiv. 2023, 10.48550/arXiv.2308.08061
75. Gallifant J, Afshar M, Ameen S, et al.: The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025, 31:60-9. 10.1038/s41591-024-03425-5