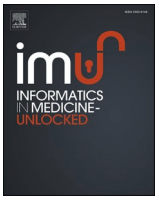
Informatics in Medicine Unlocked 41 (2023) 101304

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Evaluating large language models for use in healthcare: A framework for translational value assessment

Sandeep Reddy

School of Medicine, Deakin University, Waurn Ponds, VIC, 3216, Australia

A B S T R A C T

The recent focus on Large Language Models (LLMs) has yielded unprecedented discussion of their potential use in various domains, including healthcare. While showing considerable potential in performing human-capable tasks, LLMs have also demonstrated significant drawbacks, including generating misinformation, falsifying data, and contributing to plagiarism. These aspects are generally concerning but can be more severe in the context of healthcare. As LLMs are explored for utility in healthcare, including generating discharge summaries, interpreting medical records and providing medical advice, it is necessary to ensure safeguards around their use in healthcare. Notably, there must be an evaluation process that assesses LLMs for their natural language processing performance and their translational value. Complementing this assessment, a governance layer can ensure accountability and public confidence in such models. Such an evaluation framework is discussed and presented in this paper.

1. Introduction

Many healthcare systems and their delivery models revolve around episode-based care instead of preventative ones [1]. This approach has led to high costs in delivering healthcare, inefficiencies, and an inability to service the rising demand for quality and safe healthcare. Digital health interventions, including AI, have been identified as solutions to these problems. In this regard, Artificial intelligence (AI) has demonstrated significant potential and use in addressing many entrenched healthcare planning and delivery issues [2]. Of the many categories of AI, Natural language processing (NLP) has become an essential technology in aiding clinicians and healthcare administrators in transcription, documentation analysis and summary generation [3–5].

In the early years of NLP research, they were centred on what is known as the rationalist approach, while statistical NLP, which takes an empiricist approach, became the dominant school of thought in the 1990s [6]. Statistical NLP assumes that a significant degree of latent semantic knowledge resides in text corpora. Numerical representations of language, called word representations or embeddings, are necessary to encode this knowledge. Statistical language models, which are probability distributions of sequences of words, can be created using word representations to model the probability of word occurrence in sentences. N-gram models are a simple form of language model commonly used in text mining, among others. Using word representations to encode the semantics of words in a language, statistical language models can be created to model the probability of word occurrence in sentences [6].

Language models are probability distributions of sequences of words that are useful for problems that require predicting the next word in a sequence given the previous words [7,8]. Large Language Models (LLM) are AI neural network models that can carry on a dialogue. These models have made significant strides in size and capability in recent years and are versatile and can perform many different language tasks. They are already being used in various fields, such as journalism, advertising, writing, and programming [9,10]. The most popular LLM currently is 'Chat Generative Pre-trained Transformer (ChatGPT)', a 175-billion--parameter autoregressive language models model designed by OpenAI, with the ability to generate human-like responses to user prompts [11]. The response to prompts is generated by deep learning algorithms trained on vast amounts of data. ChatGPT is a descendant of the fine-tuned version of GPT-3.5, trained on articles, websites, books, and written conversations derived from the web [9]. Adding both supervised and reinforcement learning techniques to train ChatGPT has enhanced its ability to perform competently on various tasks [12].

LLM models are generally developed and work as outlined in Fig. 1 [7,13,14]. First raw text (input) is provided to the language model, then input text is down into individual tokens (words, subwords, or characters), which the language model can then process. Each token is then represented as a high-dimensional vector, which captures its semantic meaning and contextual information. This embedding is learned through training the language model on large amounts of text. The embedding vectors are then passed through a series of transformer layers. Each transformer layer consists of multi-head self-attention and feedforward neural networks, which allow the model to capture

E-mail address: Sandeep.reddy@deakin.edu.au.

https://doi.org/10.1016/j.imu.2023.101304
Received 9 May 2023; Received in revised form 26 June 2023; Accepted 1 July 2023
Available online 3 July 2023

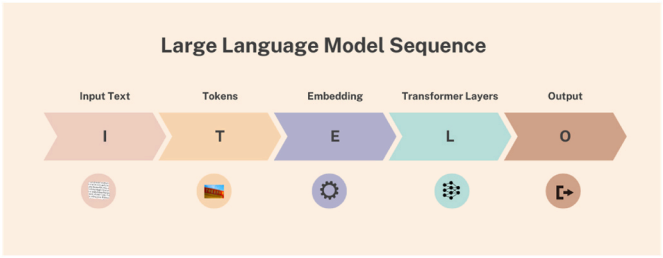


Fig. 1. LLMs architecture.

complex dependencies and relationships between words. Finally, the output of the transformer layers is passed through a linear layer, which produces a probability distribution over the vocabulary of the language model. This distribution can be used to predict the likelihood of different words or sequences of words following the input text.

2. LLMs in healthcare

AI and healthcare have a symbiotic relationship that can improve healthcare delivery and patient outcomes and reduce healthcare costs [2,15] (see Fig. 2). In a medical context, LLMs would follow the same process in their development as outlined in Fig. 1. However, in this instance, the input would be electronic health record (EHR) notes, radiology reports, or other medical documentation [16,17]. There would be an added step of pre-processing data to remove patient-identifying information, correct spelling or grammar errors, and handle medical terminology or abbreviations. The output is a probability distribution over the vocabulary of the language model. In a clinical context, this distribution can predict diagnoses, suggest treatment options, or provide other clinical decision support. Overall, this architecture allows large language models to process and understand medical text data, providing valuable insights and support for clinical decision-making [18]. In this context, LLM's potential applications are limitless. We have already seen LLMs applications to generate discharge summaries, clinical concept extraction, answer medical questions, interpret electronic health records, and generate medical articles [9, 16–19].

3. Ethical and other concerns about LLMs

The development of pre-trained LLMs like ChatGPT has revolutionized the natural language processing (NLP) field and opened new possibilities for generating medically relevant content [10,18]. However, their use has raised ethical concerns about the potential spread of misinformation, misinterpretation, plagiarism and questions about authorship [12,20]. The potential spread of misinformation can entail significant societal hazards. LLMs' ability to generate plausible sounding but incorrect or nonsensical answers highlights the ethical challenges of using them to provide medical advice. Using LLMs for scholarly publishing has opened new possibilities but also raised ethical concerns related to plagiarism, authorship, and the potential spread of misinformation [20]. LLMs have limitations that are known to produce incorrect or biased output. These errors might be recycled and amplified as LLM outputs can be used to train future model iterations [12]. This raises concerns about the integrity of the scientific record and the potential for false information to be used in future research or health policy decisions. The developers of LLMs have set up guardrails to minimize the risks, but users have found ways around them [21].

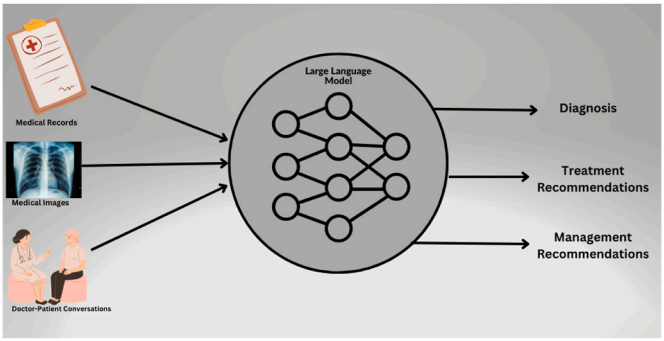


Fig. 2. LLMs use in healthcare.

4. An evaluation framework for LLM application in healthcare

The potential of artificial intelligence (AI) to revolutionize healthcare delivery is widely acknowledged [2]. However, limited assessments have found that many AI systems have fallen short of their translational goals due to intrinsic inadequacies only assessed after deployment [22]. The early rollout of ChatGPT has spawned competitors, potentially rendering the issues with LLMs a far-reaching problem [23]. Evaluation frameworks must evolve to assess LLMs for their safety and quality used in healthcare. While LLMs have shown impressive performance in modelling source code, leading to AI-based programming assistance, some of the most substantial performing models, like ChatGPT and PaLM 2 are not publicly available. This can limit transparency around the model's architecture and its output, thus limiting ability of users to mitigate biases and hallucinations [24]. Several pre-trained language models are publicly available, but their performance and impact on modelling and training design decisions still need to be determined.

This paper presents a conceptual evaluation framework to assess the performance of large language models and explore an appropriate governance and monitoring mechanism for their use in healthcare. Specific NLP metrics have been proposed to assess the performance of LLMs, as outlined in Table 1.

However, while assessing the NLP performance of LLMs, the above evaluation metrics do not assess the models' functional, utility, and ethical aspects as they apply to healthcare. Therefore, additional layers that assess the translational and governance aspects of LLMs in healthcare are required. However, one does not have to commence from scratch in devising such a framework. Considerable work has been undertaken in recent years to develop and promote various evaluation and governance frameworks for AI models in healthcare [22,30]. It is practical to draw upon such frameworks and customise them to evaluate LLM applications in healthcare. This paper presents two such

Table 1
Current evaluation metrics for language models [25–29].

Metric	Description
Perplexity	A common evaluation metric used in natural language processing (NLP) to measure the effectiveness of language models. It measures how well the model predicts the probability distribution of a test dataset. A lower perplexity value indicates better performance.
BLEU	The Bilingual Evaluation Understudy (BLEU) score is a metric used to evaluate the quality of machine translation output by comparing it to one or more reference translations. It ranges from 0 to 1, with 1 indicating perfect translation.
ROUGE	The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score is a family of metrics used for evaluating automatic summarization and machine translation systems. It measures the overlap between the generated summary and the reference summaries
F1 Score	A measure of a model's accuracy, combining precision and recall. It is commonly used in binary classification tasks to evaluate the performance of a model on a given dataset.
Human evaluation	The effectiveness of a language model like ChatGPT is best evaluated by humans who can judge the quality of the generated text in terms of its fluency, coherence, and relevance to the task at hand.

frameworks and a modified framework incorporating critical components of them in addition to NLP metrics.

In 2021, an international team of medical researchers and data scientists developed the TEHAI (Translational Evaluation of Healthcare AI) framework to introduce a multi-stage and comprehensive evaluation framework that assessed AI models beyond regulatory and reporting requirements [22]. This framework emphasises the translational value and draws upon the principles of translational research and health technology assessment. TEHAI includes three main components (capability, utility, and adoption), with fifteen subcomponents identified based on a critical review of related literature and frameworks/guidelines covering AI in healthcare reporting and evaluation. The components and subcomponents are designed to assess various aspects of AI systems at different development and deployment stages. The capability component assesses the intrinsic technical capability of the AI system to perform its expected purpose by reviewing key aspects of how the AI system was developed. The utility component evaluates the usability of the AI system across different dimensions, including contextual relevance, safety, ethical considerations, and efficiency. The adoption component appraises the translational value of current AI systems by evaluating key elements that demonstrate the model's adoption in real-life settings. TEHAI provides a standardised approach to evaluating the translational aspects of AI systems in healthcare, which can support or contradict the use of a specific AI tool in each healthcare setting. The framework can be used at various stages of development and deployment, providing a comprehensive yet practical instrument to assess AI's functional, utility, and ethical aspects [22]. Full details of the framework components and questions are presented in the appendix (Fig. 3).

To complement the translational assessment, a layer of governance is added to the evaluation framework. The governance layer is essential to ensure oversight and accountability when LLMs are developed and deployed in healthcare environments. While general-purpose governance models have been available for some time [31,32], a healthcare specific and practical governance model that covers the nuances of the application of AI in healthcare are few. Specialised governance models for healthcare ensure aspects such as bio-medical ethics, medico-legal and patient safety are adequately assessed and monitored. One such specialised framework is 'The Governance Model for AI in Healthcare (GMAIH)', which consists of four main components: fairness, transparency, trustworthiness, and accountability [30]. These aspects are outlined below.

5. Fairness [30]

The use of AI in healthcare requires appropriate and representative training datasets to avoid biases, inaccurate predictions, medical errors, and discrimination. To ensure fairness in data collection and utilisation, a data governance panel comprising AI developers, patient and target group representatives, clinical experts, and individuals with relevant ethical and legal expertise is proposed. The panel will review datasets and algorithms used to develop LLMs to ensure they conform to the principle of justice and do not lead to health inequities, discrimination, or unfair allocation of resources.

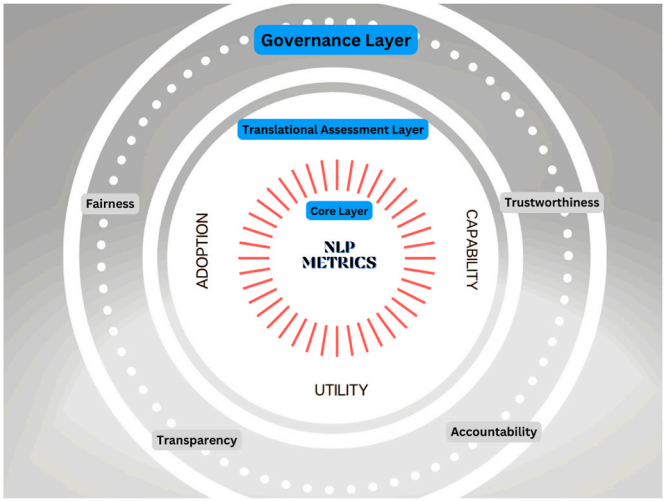


Fig. 3. A suggested evaluation framework for LLMs in Healthcare.

6. Transparency [30]

The interpretability and explainability of AI models used in medical imaging analysis and clinical risk prediction is paramount in healthcare. Limited transparency and explainability can reduce trustworthiness and impair validation of clinical recommendations. Therefore, appropriate governance emphasises ongoing or continual explainability and interpretable frameworks to enhance the decision-making process.

7. Trustworthiness [30]

Clinicians need to understand the causality of medical conditions and the methods and models employed to support the decision-making process. In addition to explainability, the potential autonomous functioning of AI applications and potential unintended consequences must be considered. The trustworthiness of AI models can be enhanced by ensuring data privacy, security, and confidentiality.

8. Accountability [30]

Accountability is critical in ensuring that AI applications in healthcare are used responsibly and ethically. Clear policies, procedures, and regulations should be in place to ensure compliance with legal and ethical standards. Thus, it is proposed that healthcare institutions and governmental bodies develop normative standards for the application of AI in healthcare to inform the design and deployment of AI models.

The framework can be accompanied by a scoring system to allow for a quantitative assessment and a meaningful evaluation of the LLM's relevance to the use case. It is beyond the scope of this paper to outline a detailed scoring mechanism for each component here but meaningful guidance for scoring the translational assessment layer can be found in Reddy et al. [22]. For the governance layer, equal weightage is recommended for the four components with a score range aligning to the translational assessment layer be accorded.

The layered assessment with healthcare aligned components makes this framework comprehensive. The framework has the potential to guide developers, healthcare organizations and other stakeholders to ensure the responsible and ethical use of LLMs in healthcare. This aspect is crucial in ensuring patient safety, quality of care, and public trust in these models. A comprehensive assessment of LLMs can be achieved by incorporating translational and governance elements in addition to NLP metrics. We can also ensure that LLMs are safe, transparent, trustworthy, equitable and accountable as they get increasingly used in healthcare.

9. Conclusion

LLMs have shown significant promise in the recent period in performing human-capable tasks [10,17,18]. Their utility in healthcare has also been demonstrated. Nevertheless, their drawbacks, like generating misinformation, have also been seen [12,20]. If these models were to be adopted in healthcare, there must be an appropriate governance process around their utilisation. In this paper, a comprehensive assessment and governance approach are presented. By adopting such an approach, LLM's potential benefits in healthcare delivery, better patient outcomes, and reduced healthcare costs can be appropriately realised. While LLMs are ready to play in healthcare, we need a referee to ensure they play well.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sandeep Reddy reports a relationship with Medical Artificial Intelligence Pty Ltd that includes: equity or stocks.

Acknowledgements

None.

Appendix

TEHAI Components [22].

Component
1. Capability
1.1. Objective
This subcomponent assesses whether the system has a clear objective i.e., stated contribution to a specific healthcare field. This subcomponent is scored on a scale of how clearly the objective is articulated
1.2. Dataset Source and Integrity
An AI system is only as good as the data it was derived from. If the training data does not reflect the intended purpose, the model predictions are likely to be useless or even harmful. This subcomponent evaluates the source of the data and the integrity of datasets used for training and testing the AI system including an appraisal of the representation of the target population in the data, coverage, accuracy and consistency of data collection processes and transparency of datasets. This subcomponent is scored on a scale of how well the dataset is described, how well the datasets fit with the ultimate objective and use case, and how credible/reliable the data source is. The subcomponent also considers when new data is acquired to train an embedded model that appropriate checks are undertaken to ensure integrity and alignment of data to previously used data
1.3. Internal Validity
An internally valid model will be able to predict health outcomes reliably and accurately within a pre-defined set of data resources that were used wholly or partially when training the model. This includes the classical concept of goodness-of-fit, but also cross-validation schemes that derive training and tests sets from the same sources of data. Scoring is based on the size of the training data set with respect to the health care challenge, the diversity of the data to ensure good modelling coverage, and whether the statistical performance of the model (e.g., classification) is high enough to satisfy the requirements of clinical usefulness.
1.4. External Validity
To qualify as external validation, we require that the external data used to assess AI system performance must come from substantially distinct external source that did NOT contribute any data towards model training. Examples of external data sources include independent hospitals, institutions or research groups that were not part of the model construction team or a substantial temporal difference between the training and validation data collections. The scoring is based on the size and diversity of the external data (if any) and how well the external data characteristics fit with the intended care recipients under the study objective.
1.5. Performance Metrics
Performance metrics refers to mathematical formulas that are used for assessing how well an AI model predicts clinical or other health outcomes from the data. If the metrics are chosen poorly, it is not possible to assess the accuracy of the models reliably. Furthermore, specific metrics have biases, which means the use of multiple metrics is recommended for robust conclusions. This subcomponent examines whether performance measures relevant to the model and the results stated in the study are presented. These performance metrics can be classification or regression or qualitative metrics. This subcomponent is scored on a scale of how well the performance metrics fit the study and how reliable they are likely to be considering the nature of the health care challenge.
1.6. Use Case
This subcomponent is seeking justification for the use of AI for the health need as opposed to other statistical or analytical methods. This tests if the application has considered the relevance and fit of the AI to the particular healthcare domain it is being applied to. This subcomponent is scored on a scale of how well the use case is stated.
2. Utility
2.1. Generalizability and Contextualization
The context of an AI application is defined here as the match between the model performance, expected features, characteristics of the training data and the overall objective. In particular, biases or exacerbation of disparities due to underrepresentation or inappropriate representation due to the availability of datasets used both in training and validation can have an adverse effect on the real-world utility of an AI model. This subcomponent is scored based on how well it is expected to perform on the specific groups of people it is most intended for.
2.2. Safety and Quality
It is critical that AI models being deployed in healthcare, especially in clinical environments, are assessed for their safety and quality. Appropriate consideration should be paid to the presence of ongoing monitoring mechanisms in the study, such as adequate clinical governance that will provide a systematic approach to maintaining and improving the safety and quality of care within a healthcare setting. This subcomponent is scored based on the strength of the safety and quality process and how likely it is to ensure safety and quality when AI is applied in the real-world.
2.3. Transparency
This subcomponent assesses the extent to which model functionality and architecture is described in the study and the extent to which decisions reached by the algorithm are understandable (i.e., black box or interpretable). Important elements are the overall model structure, the individual model components, the learning algorithm, and how the specific solution is reached by the algorithm. This subcomponent is scored on a scale of how transparent, interpretable and reproducible the AI models are, given the information available.
2.4. Privacy
This subcomponent covers personal privacy, data protection and security. This subcomponent is ethically relevant to the concept of autonomy/self-determination, the right to control access to and use of personal information, and the consent processes used to authorize data uses. This subcomponent is scored on the extent of consideration of privacy aspects including consent by study subjects, the strength of data security and data life cycle throughout the study itself and consideration for future protection if deployed in the real-world.

(*continued on next page*)

(*continued*)

Component
2.5. Non-Maleficence This subcomponent refers to the identification of actual and potential harms caused by the AI and actions to avoid foreseeable or unintentional harms. Harms to individuals may be physical, psychological, emotional, economic. Harms may affect systems/organizations, infrastructure and social wellbeing. This subcomponent is scored on the extent to which potential harms of the AI are identified, quantified and the measures taken to avoid harms and reduce risk.
3. Adoption
3.1. Use in a Healthcare Setting As discussed earlier, many AI systems have been developed in controlled environments or in-silico, but there is a need to assess for evidence of use in real world environments and integration of new AI models with existing information systems. This subcomponent is scored according to the extent to which the model has been adopted by and integrated into 'real world' healthcare services i.e., healthcare settings beyond the test site. This subcomponent also considers the applicability of the system to end-users, both clinicians and administrators, and the beneficiaries of the system, patients as part of the evaluation.
3.2. Technical Integration This subcomponent evaluates how well the AI systems integrate with existing clinical/administrative workflows outside of the development setting, and their performance in such situations. In addition, the subcomponent includes reporting of integration even if the model performs poorly. This subcomponent is scored on a scale of how well the integration aspects of the model are anticipated and if specific steps to facilitate practical integration have been taken.
3.3. Number of Services Many AI in healthcare studies are based on single site use without evidence of wider testing or validation. In this subcomponent, we review reporting of wider use. This subcomponent is scored on a scale of how well the use of the model across multiple healthcare organizations is described.
3.4. Alignment with Domain This category considers how much of information about the alignment and relevance of the AI system to the healthcare domain and its likely long-term acceptance are reported. In other words, the model is assessing the benefits of the AI model to the particular medical domain the model is being applied to. This again relates to the translational aspects of the AI model. This subcomponent is scored on a scale of how well the benefits of the AI model to the medical domain are articulated.

References

[1] Lewis SJ, Leeder SR. Why health reform? Med J Aust 2009;191(5):270–2.
[2] Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J R Soc Med 2019;112(1):22–8.
[3] Zhou B, Yang G, Shi Z, Ma S. Natural Language Processing for Smart Healthcare. IEEE Rev Biomed Eng 2022. https://doi.org/10.1109/RBME.2022.3210270.
[4] Edirippulige S, Gong S, Hathurusinghe M, Jhetam S, Kirk J, Lao H, et al. Medical students' perceptions and expectations regarding digital health education and training: a qualitative study. J Telemed Telecare 2022;28(4):258–65.
[5] Chen JS, Baxter SL. Applications of natural language processing in ophthalmology: present and future. Front Med 2022;9:906554.
[6] Gruetzemacher R, Paradice D. Deep transfer learning & beyond: transformer Language Models in information systems research. ACM Comput Surv 2022;54 (10s):1–35.
[7] Sejnowski TJ. Large language models and the reverse turing test. Neural Comput 2023;35(3):309–42.
[8] Mars M. From word embeddings to pre-trained Language Models: a state-of-the-art walkthrough. Appl Sci 2022;12(17).
[9] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of Large Language Models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312.
[10] Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature 2023;614(7947):214–6.
[11] De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120.
[12] The Lancet Digital H. ChatGPT: friend or foe? Lancet Digit Health 2023 Mar;5(3). https://doi.org/10.1016/S2589-7500(23)00023-7. e102.
[13] Chen M, Tworek J, Jun H, Yuan Q, HpdO Pinto, Kaplan J, et al. Evaluating large language models trained on code. 2021 Jul 7. arXiv preprint arXiv:210703374.
[14] Chen SF, Beeferman D, Rosenfeld R. Evaluation Metrics For Language Models [Internet]. Carnegie Mellon University; 2018 [cited 2023Jul3]. Available from: https://kilthub.cmu.edu/articles/journal_contribution/Evaluation_Metrics_For_Language_Models/6605324/1.
[15] Reddy S. Artificial intelligence and healthcare—why they need each other? Journal of Hospital Management and Health Policy 2020;5:9.
[16] Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022;5(1):194.
[17] Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107–8. https://doi.org/10.1016/S2589-7500(23)00021-3.
[18] Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are zero-shot clinical information extractors. 2022 May 25. arXiv preprint arXiv:220512689.
[19] Dagan A, Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digital Health 2023;2(2).
[20] Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health 2023 Mar;5(3):e105–6. https://doi.org/10.1016/S2589-7500(23)00019-5.
[21] Taylor J. ChatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards. The Guardian 2023 Mar 7. Available from: https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards?CMP=share_btn_tw.
[22] Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform 2021;28(1).
[23] Hart R. ChatGPT's biggest competition: here are the companies working on rival AI chatbots. Forbes 2023 Feb 23. Available from: https://www.forbes.com/sites/roberthart/2023/02/23/chatgpts-biggest-competition-here-are-the-companies-working-on-rival-ai-chatbots/.
[24] Wang D-Q, Feng L-Y, Ye J-G, Zou J-G, Zheng Y-F. Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. MedComm – Future Medicine 2023;2(2):e43.
[25] Józefowicz R, Vinyals O, Schuster M, Shazeer NM, Wu Y. Exploring the limits of language modeling. ArXiv. 2016;abs 2016:02410.
[26] Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics; 2002. p. 311–8.
[27] ROUGE: a package for automatic evaluation of summaries. In: Lin C-Y, editor. Annual meeting of the association for computational linguistics; 2004.
[28] Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. ArXiv. 2011;abs 2010:16061.
[29] Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. 2019 Apr 22. arXiv preprint arXiv:1904.09751.
[30] Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inf Assoc 2020;27(3):491–7.
[31] University of Oxford. AI governance: a research agenda. Oxford: Centre for the Governance of AI & Future of Humanity Institute; 2018.
[32] Daly A, Hagendorff T, Li H, Mann M, Marda V, Wagner B, Wang WW, Witteborn S. Artificial Intelligence, Governance and Ethics: Global Perspectives. In: The Chinese University of Hong Kong Faculty of Law Research Paper No. 2019-15, University of Hong Kong Faculty of Law Research Paper No. 2019/033; 2019 Jul 4. Available from: https://ssrn.com/abstract=3414805.