Article

Not peer-reviewed version

An Overview of Medical Knowledge Evaluation of Large Language Models: An Endeavor Toward a Standardized Evaluation and Reporting Guideline

Omid Kohandel Gargari and Gholamreza Habibi *

Posted Date: 9 January 2025

doi: 10.20944/preprints202501.0699.v1

Keywords: LLM; reporting guidline; linguistic statistics

 doi:10.20944/preprints202501.0699.v1

Article

An Overview of Medical Knowledge Evaluation of Large Language Models: An Endeavor Toward a Standardized Evaluation and Reporting Guideline

Omid Kohandel Gargari and Gholamreza Habibi *

Farzan Artificial Intelligence Team, Farzan Clinical Research Institute, Tehran, Iran

* Correspondence: GholamReza (GR) Habibi, Email: farzan.ai.research@gmail.com; Address: Farzan Clinical Research Instituted, Siyami Dead End, Tohid Sq, Satarkhan Str, Tehran, Iran; Phone: +982166597463-4

Abstract: Large language models (LLMs) have increasingly been recognized for their potential to revolutionize various aspects of healthcare, including diagnosis and treatment planning. However, the complexity of evaluating these models, particularly in the medical domain, has led to a lack of standardization in assessment methodologies. This study, conducted by the Farzan Clinical Research Institute, aims to establish a standardized evaluation framework for medical LLMs by proposing specific checklists for multiple-choice questions (MCQs), question-answering tasks, and case scenarios. The study demonstrates that MCQs provide a straightforward means to assess model accuracy, while the proposed confusion matrix helps identify potential biases in model choice. For question-answering tasks, the study emphasizes the importance of evaluating dimensions like relevancy, similarity, coherence, fluency, and factuality, ensuring that LLM responses meet clinical expectations. In case scenarios, the dual focus on accuracy and reasoning allows for a nuanced understanding of LLMs' diagnostic processes. The study also highlights the importance of model coverage, reproducibility, and the need for tailored evaluation methods to match study characteristics. The proposed checklists and methodologies aim to facilitate consistent and reliable assessments of LLM performance in medical tasks, paving the way for their integration into clinical practice. Future research should refine these methods and explore their application in real-world settings to enhance the utility of LLMs in medicine.

Keywords: LLM; reporting guidline; linguistic statistics

1. Introduction

Large language models have shown a significant potential to impact different realms of science. One of the most promising areas is medicine[1,2]. Performance evaluation is one of the most critical stages in the development of any artificial intelligence model. For simpler machine learning models that produce outputs in the form of qualitative or quantitative variables, specific metrics such as Accuracy, Recall and etc. are reported. However, language models, due to their inherent nature, require a more complex evaluation process. This evaluation process also varies depending on the specific task at hand. This evaluation could be divided into two categories automated numeric linguistics metrics and human evaluation[3], evaluation of scientific aspect of responses and models' knowledge lies in the hands of human evaluator. There are several methods suggested for evaluation of medical LLMs or LLMs fine-tuned on medical data. Methods introduced for models' evaluation include multiple choice questions (MCQs), case scenarios and medical question answers[4]. Currently each study uses its own knowledge evaluation checklist causing a heterogenous reporting among studies. While there is some reporting guidelines suggested for prediction models[5,6] need for a standard evaluation and reporting system is sensed in the realm of LLMs in medicine.

In the following we will cover detail regarding each method. Finally suggesting an evaluation checklist for each method, these checklists are result of multiple discussion and literature review

 doi:10.20944/preprints202501.0699.v1

sessions of medical and computer science researchers at Farzan Clinical Research Institute. It is crucial to note that selection of evaluation methods should be based on characteristics of the models, for instance if a model is designed for summarization tasks it should not be tested for other tasks. In this study we focus on models that are capable of reasoning and text generation.

2. Linguistic Analysis

Linguistic analysis, also known as computational linguistics [5], is an aspect of NLP involves the systematic study and interpretation of text generated. This discipline combines insights from linguistics, mathematics, and artificial intelligence to understand and manipulate language generated by the model in a meaningful way. Linguistic analysis covers a wide array of techniques and methodologies aimed at extracting semantic, syntactic, and pragmatic information from text. In this section we delve into various methods used for linguistic analysis, including similarity measurement, statistical evaluation, and token frequency and distribution.

2.1. Passage Ranking

Passage Ranking or Document Ranking involves the process of sorting or ranking documents based on their relevance to a given query. One common evaluation metric used in Passage Ranking tasks is Mean Reciprocal Rank (MRR) [6], which assesses the effectiveness of the ranking by considering the reciprocal of the rank of the first relevant document.

In this task, MRR is a key evaluation metric that quantifies the effectiveness of the ranking by considering the reciprocal of the rank of the first relevant document [7] which can be defined as:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

where |Q| is the total number of queries and ranki is the rank of the first relevant document for query i.

The MRR score ranges from 0 to 1, with higher values indicating better performance. A higher MRR signifies that, on average, the relevant documents are ranked higher in the list, indicating a more effective ranking system.

The fundamental concept behind Passage Ranking models is to assess the relevance of each document in a collection to a specific query [8]. These models typically take as input a query and a set of documents, and their output consists of scores indicating the degree of relevance of each document to the query. The documents are then sorted or ranked based on these scores, with the most relevant ones appearing at the top of the list.

One common approach to Passage Ranking is through the use of neural network-based models, such as BERT (Bidirectional Encoder Representations from Transformers) [9]. These models leverage pre-trained language representations to understand the context of the query and the content of the documents, allowing them to produce accurate relevance scores.

Equations used in Passage Ranking models often include similarity measures between the query and each document, such as cosine similarity or dot product, which quantify the degree of similarity or relevance between the query and the document embeddings.

2.2. Semantic Textual Similarity

Pieces of text, such as sentences or paragraphs, align with each other [10]. The goal is to measure the degree of similarity in meaning between these text segments. In practical terms, STS models compare a given source sentence with a list of other sentences and return scores that indicate how similar each sentence is to the source. This process helps in understanding the extent to which two texts convey the same or similar information.

This metric enable machines to understand and process text in a way that aligns more closely with human interpretation. Unlike syntactic similarity, which focuses on the surface form of the text

 doi:10.20944/preprints202501.0699.v1

(e.g., word overlap), semantic similarity delves into the underlying meaning. This distinction is required for applications that require understanding context and nuance rather than just matching keywords. Several approaches can be employed to measure semantic similarity between texts, ranging from traditional methods using linguistic resources to leveraging deep learning models.

Some of the methods are defined below:

2.2.1. Word Embeddings

Represent words as dense vectors in a continuous vector space. Methods like Word2Vec [11], GloVe [12], and FastText [13] capture semantic relationships between words. Text similarity is computed by aggregating word vectors (e.g., averaging) and then measuring the similarity between the aggregated vectors.

2.2.2. Contextualized Embeddings

Utilize models like BERT (Bidirectional Encoder Representations from Transformers) to generate context-sensitive embeddings for words and sentences. These models capture the meaning of words in their specific context, resulting in more accurate similarity measures. Sentence Transformers extend the concept of word embeddings to sentences. Models such as Sentence-BERT (SBERT) [14] and Universal Sentence Encoder (USE) [15] provide embeddings for entire sentences, allowing for efficient similarity computation.

2.2.3. Evaluation Metrics

To evaluate the performance of semantic similarity models, several metrics are commonly used:

1 .Pearson Correlation Coefficient [16–18]: Measures the linear correlation between the predicted similarity scores and the ground truth scores.

2 .Spearman's Rank Correlation Coefficient [16,18]: Assesses the monotonic relationship between the predicted scores and the ground truth scores, comparing the rank orders rather than the raw values.

2.3. Cosine Similarity

Cosine similarity is a mathematical tool for quantifying the similarity between two vectors within a multi-dimensional space [19]. In the assessment of outputs from LLMs, it is used for semantic assessment between the generated text and a given reference text [20,21]. It calculates the cosine of the angle between the embedding vectors of the given texts.

The main idea is to represent textual data as vectors in the multi-dimensional embedding space, where each dimension corresponds to a feature or aspect of the text. By comparing the angles between these vectors, we can determine how similar or dissimilar the corresponding texts are in terms of their semantic content. Cosine similarity isn't influenced by vector size; it's determined solely by the angle between them. A cosine similarity of 1 signifies vectors that are almost identical (angle of 0 degrees), while 0 indicates perpendicular vectors. A value of -1 suggests vectors pointing in opposite directions.

In this process, we give identical prompts to various models, asking them to generate content on specific topics. Next, we compare the vector embeddings of their outputs with those of reference texts. The cosine similarity score, which falls between 0 and 1, reflects how closely the generated text aligns with the reference text. A score nearing 1 suggests a strong semantic resemblance between the two texts[16,22,23]. Given two vectors A and B, the Cosine Similarity *similarity(A, B)* is computed as follows:

$$similarity(A, B) = \frac{A.B}{||A|.|B||}$$

where A·B denotes the dot product of vectors A and B, ‖A‖ and ‖B‖ represent the Euclidean norms of vectors A and B respectively.

 doi:10.20944/preprints202501.0699.v1

2.4. Rouge Score

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score [24] is a set of metrics used for evaluating summary quality by comparing overlaps in n-grams, word sequences, and word pairs between system-generated and reference summaries. N-grams are contiguous sequences of 'n' items from a given sample of text or speech. These items can be phonemes, syllables, letters, words, or base pairs according to the application. In our context, n-grams typically refer to sequences of words.

This type of evaluation consists of multiple metrics including:

1 .ROUGE-N: ROUGE-N measures the overlap of n-grams between the candidate text and the reference text. Common values are ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-3 (trigrams).
Example: If the reference text contains "the cat sat on the mat" and the generated text contains "the cat is on the mat", the ROUGE-1 score would be calculated based on the overlap of unigrams like "the", "cat", "on", and "mat."
2 .ROUGE-L: ROUGE-L measures the longest common subsequence (LCS) between the candidate and reference summaries. It takes into account sentence-level structure similarity.
Example: For the reference summary "the cat sat on the mat" and the candidate summary "the cat is on the mat", the LCS would be "the cat on the mat."
3 .ROUGE-W: ROUGE-W is a weighted version of ROUGE-L that gives more importance to consecutive matches than to non-consecutive matches.
Example: If the reference summary is "the cat sat on the mat" and the candidate summary is "the cat sat quickly on the mat", ROUGE-W will score the consecutive "the cat sat on the mat" higher than non-consecutive matches.
4 .ROUGE-S: ROUGE-S, or ROUGE-Skip-Bigram, considers skip-bigram matches between the candidate and reference summaries. Skip-bigrams are any pair of words in their order of appearance, allowing for gaps.
Example: For the reference summary "the cat sat on the mat" and the candidate summary "the cat on mat", the skip-bigrams like "the on", "cat mat" will be considered.

This metric proves its worth particularly in evaluating machine-generated summaries for extensive texts. Imagine a scenario where we need to assess the effectiveness of an ML model designed to produce such summaries. Here, a top-notch summary would be one that effectively retains the key information from the source text. It's vital to note that while a high score achieved by the text-summarization algorithm indicates its ability to preserve essential details, this alone doesn't guarantee the overall quality of the text. Despite maintaining critical information, the algorithm might inadvertently introduce toxic or biased content. In essence, it could overlook additional quality-related constraints we expect the generated output to meet.

Hence, evaluating the quality of generated text presents a complex challenge, shaped by the specific requirements of the application at hand.

2.4.1. Calculation of ROUGE Scores

ROUGE scores can be calculated in different forms, focusing on precision, recall, and Fmeasure:
-Recall:

$$Recall = \frac{Number\ of\ overlapping\ n-grams}{Total\ number\ of\ n-grams\ in\ the\ reference\ summary} \quad (20)$$

-Precision:

$$Percision = \frac{Number\ of\ overlapping\ n-grams}{Total\ number\ of\ n-grams\ in\ the\ candidate\ summary} \quad (21)$$

 doi:10.20944/preprints202501.0699.v1

F1-Score:

$$F1 - score = \frac{2 \times Percision \times Recall}{Percision + Recall}$$

(22)

Example

Given the reference summary ”the cat sat on the mat” and the candidate summary ”the cat is on the mat:”

•ROUGE-1 (unigram): Count overlaps like ”the”, ”cat”, ”on”, and ”mat.”

•ROUGE-2 (bigram): Count overlaps like ”the cat”, ”on the”, and ”the mat.”

•ROUGE-L: Identify the longest common subsequence ”the cat on the mat.”

• ROUGE-S: Consider skip-bigrams like ”the on”, ”cat mat”.

3. Human Evaluation

3.1. Multiple Choice Questions

MCQs are a relatively reliable method for medical LLMs knowledge evaluation, since for each question there is a binary outcome, correct or incorrect answer, making it easier to calculate numeric metrics (Accuracy percent). Theses question could be general, for general LLMs, or specific if the model is designed for a certain topic or task. MCQ exams such as United States Medical Licensing Exam (USMLE) are a promising source for evaluation and are being frequently used. There are also multiple datasets which could be used for this purpose. For instance, medical multiple choice question answering dataset or MedMCQA is a very famous dataset containing 194,000 multiple choice medical question on several topics and subjects. Other medical MCQ datasets are summarized in Table 1.

Table 1. list of some medical multiple choice questions datasets.

Name	Description	Reference
MedQA	collected from the professional medical board exams. It covers three languages: English, simplified Chinese, and traditional Chinese, and contains 12,723, 34,251, and 14,123 questions for the three languages	[7]
PubmedQA	Collected from PubMed abstracts. The task of PubMedQA is to answer research questions with yes/no/maybe. PubMedQA has 1k expert-annotated, 61.2k unlabeled and 211.3k artificially generated QA instances	[8]
MedMCQA	MedMCQA has More than 194k high-quality AIIMS & NEET PG entrance exam MCQs covering 2.4k healthcare topics and 21 medical subjects are collected with an average token length of 12.77 and high topical diversity.	[9]

The evaluation of a model's responses to MCQs involves two primary areas: accuracy and reasoning. Accuracy is determined by the number of correct answers given by the model, typically calculated by dividing the number of correct answers by the total number of questions. For questions with multiple correct options, careful prompt engineering is necessary. We recommend analyzing the distribution of selected options; one approach is to design confusion matrices based on the various choices (Figure 1). This can help identify any potential bias in favor of selecting certain options.

 doi:10.20944/preprints202501.0699.v1

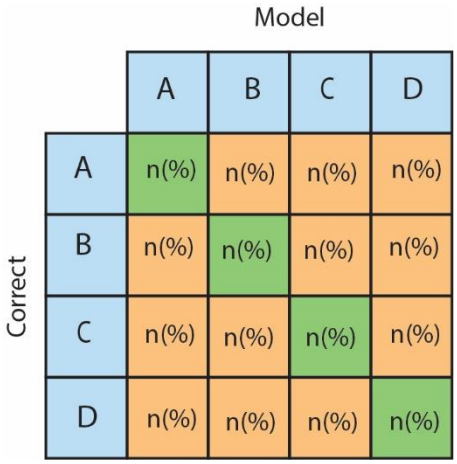

Figure 1. Sample of a confusion matrix for multiple choice questions.

Models could also be asked to provide the reasoning behind their answer. Sometimes models are unable to follow instructions and provide only one answer and also sometimes models fail to select the correct answer although they had the correct reasoning. To evaluate the reasoning following scoring checklist is suggested (Table 2).

Table 2. Multiple choice questions reasoning evaluation checklist.

N	Question
1	The number of questions in which the correct option is selected and reasoning is in line with model's choice
2	The number of questions in which the correct option is mentioned in the description, but no option is not selected
3	The number of questions in which the correct option is mentioned in the description, but the chosen option is different.
4	The number of questions in which the correct option is selected but description does not provide robust reasoning.
5	Incorrect choice and reasoning

3.2. Question-Answers

Evaluating LLMs often involves assessing their question-answering capabilities, which are crucial for determining if their responses align with user expectations. However, while evaluating LLMs typically involves some form of question answering, few datasets and studies focus exclusively on this aspect. Instead, most resources are designed to assess various other skills of LLMs[4]. By question answering we aim to focus on short and long answer questions. Domains of models' responses evaluation include:

1. Relevancy: Defined as models' ability to properly understand the question and provide relevant response regardless of answers correctness
2. Similarity: by similarity we aim to compare model's response to ground truth or expected answer. This test could also be used in case models is fine-tuned on question-and-answer dataset and the aim is to compare model's response to the response available in fine tuning dataset.
3. Coherence: this domain aims to check whether response follows a logical flow with an introduction, body and conclusion.
4. Fluency: assesses whether answers are linguistically fluent and easy to understand.
5. Factuality: at this domain the aim is to compare model's response to reliable scientific resources.

 doi:10.20944/preprints202501.0699.v1

A suggested evaluation checklist is provided in Table 3

Table 3. Evaluation checklist for qustion answers.

Relevancy	
1	In how many questions was the answer related to the question regardless of the correctness?
Similarity	
1	The points in the ground truth have been mentioned and also some correct additional items have been added to it.
2	The points in the ground truth have been mentioned
3	The key points are incompletely mentioned
4	The key points are not mentioned
Coherence	
1	In how many questions does the model address additional and irrelevant details?
2	In how many questions does the tone remain constant during the answer?
3	In how many questions are ironies and non-scientific literature used during the answer?
4	In some questions, the structure of the answer is logical (introduction, body and clear conclusion).
Fluency	
1	How many answers were understood in one reading?
2	How many answers do you think were fluent?
3	How many answers included repeating items?
4	In how many questions are punctuation marks used correctly?
5	In how many questions are words from other languages used inappropriately?
6	In how many questions is the length of sentences and paragraphs appropriate?
Factuality	
1	The number of questions that have incorrect scientific information. (compared with reliable sources)

3.3. Case Scenarios

Case study researches are types of research which provide LLMs with a set of case scenarios and ask them to analyze cases and provide a diagnosis or evaluation of the case. These types of research have 5 main steps after research design: 1. Case scenarios preparation 2. Model selection and deployment. 3. Prompt design 4. Running model 5. Response evaluation. In this section we mainly focus on final step which is evaluation of model's response. As mentioned before due to growing evidence on this type of studies design of a standard reporting guideline must subjected for future research.

Case diagnosis evaluation consists of two main domains: Accuracy and reasoning. Unlike previous methods we have a scoring system. For accuracy it is assessed that how close is model's diagnosis to actual diagnosis. Next domain evaluates the step-by-step reasoning of the model. Case diagnosis evaluation scoring checklist is available in Table 4.

 doi:10.20944/preprints202501.0699.v1

Table 4. Case diagnosis evaluation checklist.

Accuracy		
N	**Question**	**Mark**
1	Correct diagnosis	3 Points
2	Correct diagnosis of the disease without mentioning details	2 Points
3	Correct diagnosis of the general category of the disease	1 Point
4	Incorrect diagnosis	0 Point
Reasoning		
1	The reasoning includes all the important and diagnostic signs and symptoms of the case and points to the correct signs outside the case.	4 Points
2	The reasoning includes all the important and diagnostic signs and symptoms of the case	3 Points
3	The reasoning considers most of the symptoms of the case	2 Points
4	The reasoning does not consider most of the symptoms of the case	1 Point
5	Incorrect reasoning	0 Point

Figure 2 illustrates a graphical abstract of evaluation methods.

4. Discussion and Conclusion

The integration of large language models (LLMs) in the medical field promises to revolutionize various aspects of healthcare, from diagnosis to treatment planning. Our study aimed to establish a standardized evaluation framework for LLMs applied to medical tasks, addressing the heterogeneity in current evaluation practices.

Our study suggest that multiple choice questions (MCQs) offer a straightforward and reliable method for assessing the knowledge of medical LLMs. By using MCQs sourced from reputable exams like the USMLE and datasets like MedMCQA, researchers can quantify the accuracy of LLMs in a controlled environment. The confusion matrix approach, as demonstrated in our study, provides deeper insights into potential biases and errors in model choice, enhancing the understanding of model behavior in decision-making processes.

For question-answering tasks, our study highlights the importance of evaluating multiple dimensions such as relevancy, similarity, coherence, fluency, and factuality. Each of these dimensions plays a crucial role in ensuring that the responses generated by LLMs align with clinical expectations and scientific accuracy. The evaluation checklist we propose allows for a comprehensive assessment of these dimensions, ensuring that LLMs not only provide accurate information but also deliver it in a coherent and fluent manner.

In the context of case scenarios, the dual focus on accuracy and reasoning in our proposed scoring system allows for a nuanced understanding of how well LLMs can replicate human-like reasoning in medical diagnoses. The detailed scoring checklist ensures that models are evaluated

 doi:10.20944/preprints202501.0699.v1

based on their ability to recognize and interpret critical signs and symptoms accurately, thereby mimicking the diagnostic process of medical professionals.

There are some points that although not directly mentioned in our checklists but could be assessed using our method. For instance, the need to assess the coverage of models across various medical topics. If the evaluation questions or cases are from different topics this step is crucial to determine if the model maintains homogenous performance across different subjects. For this purpose, we suggest comparing checklists results between topics. The reproducibility of the model is another critical factor, which can be tested using prompts like "Are you sure?" after the model provides its answer. This allows for recalculating the accuracy and further assessing the model's consistency and reliability[10].

Different evaluation methods in past research have highlighted the diverse ways to assess LLMs. For example, Barile et al.'s study provided pediatric cases to GPT for differential diagnosis, categorizing the outcomes as 'Correct', 'Incorrect', and 'Did not fully capture diagnosis'. This type of evaluation offers a straightforward comparison between model responses and actual labels, avoiding complex subjective analysis[11]. Conversely, Duey et al. assessed ChatGPT's recommendations for thromboembolic prophylaxis in spine surgery through accuracy metrics, as well as categories like over-conclusiveness, supplementary information, and incompleteness. Such diverse approaches demonstrate the breadth of methodologies and the importance of matching evaluation strategies to the specific characteristics of each study[12].

The study by Taira et al. highlights the performance of ChatGPT in the Japanese National Nurse Examinations, demonstrating the model's ability to meet or nearly meet the passing criteria, especially in 2019. Their methodology assessed ChatGPT's responses to various question formats using two distinct prompt types—simple-choice and situation-setup. This approach, complemented by statistical analyses like chi-square and Cochran-Armitage trend tests, emphasized the importance of tailored evaluation methodologies for LLMs in medical contexts, ensuring thorough and contextually relevant assessments[13]. Similarly, the study by Kung et al. investigated ChatGPT's performance on the United States Medical Licensing Exam (USMLE). The model performed at or near the passing threshold for all three exams without specialized training or reinforcement. By evaluating accuracy, concordance, and insight in explanations and employing statistical analyses and rigorous adjudication, the study provided a nuanced assessment of ChatGPT's capabilities in medical education and clinical decision-making. This research underscores the need for transparency and explainability in AI's role in healthcare[14].

Cascella et al. explored the feasibility of ChatGPT in clinical and research scenarios within healthcare, focusing on support for clinical practice, scientific production, potential misuse in medicine and research, and reasoning about public health topics. Their findings highlighted ChatGPT's ability to structure medical notes and summarize scientific information effectively. However, concerns about potential misuse, such as fabricating research data or generating plausible but incorrect analyses, were significant. The study also noted ChatGPT's limitations in understanding complex medical relationships and the risk of "hallucinating" believable but inaccurate answers, emphasizing the need for clear guidelines and ethical considerations in AI deployment in healthcare[15]. Yeo et al. evaluated ChatGPT's performance in answering questions about cirrhosis and hepatocellular carcinoma. They graded ChatGPT's responses using a system that categorized them as comprehensive, correct but inadequate, mixed, or incorrect. The study also assessed reproducibility by entering each question twice into ChatGPT and compared the responses to published knowledge questionnaires based on AASLD guidelines. Additionally, the study examined ChatGPT's ability to respond to questions about quality measures in cirrhosis management and its capacity to provide emotional support to patients and caregivers[16]. Hermann et al. assessed the accuracy of ChatGPT in responding to cervical cancer-related questions, with the study querying ChatGPT with 64 questions. Two Gynecologic Oncologists scored the responses, revealing that ChatGPT provided "correct and comprehensive" answers to 53.1% of the questions. While effective in answering questions about cervical cancer prevention and QOL, ChatGPT's accuracy in diagnosis

 doi:10.20944/preprints202501.0699.v1

and treatment was less reliable, indicating the need for further development before it can be recommended as a reliable resource for patients or Gynecologists[17].

This study has several limitations that should be considered. One key limitation is the lack of a comprehensive safety evaluation of the LLMs in clinical settings. While we focused on performance metrics like accuracy, coherence, and reasoning, the safety implications of deploying these models in real-world medical scenarios were not fully explored. This includes potential risks of misinformation, misdiagnosis, and the ethical implications of model responses. Additionally, our study did not account for the variability in model performance across different demographics or the impact of data biases, which are crucial for ensuring equitable and reliable medical advice. These limitations highlight the need for further research to address these critical aspects and ensure the safe and effective use of LLMs in medicine.

Overall, our study underscores the need for standardized evaluation and reporting systems for medical LLMs. By adopting the proposed checklists and methodologies, researchers and practitioners can ensure consistent and reliable assessments of LLM performance. This will ultimately facilitate the development of LLMs that are not only technically proficient but also aligned with the intricate needs of medical practice.

Future research should focus on refining these evaluation methods, considering the rapid advancements in AI and the increasing complexity of medical LLM applications. Additionally, exploring the integration of these evaluation frameworks in real-world clinical settings will be vital in understanding and enhancing the practical utility of LLMs in medicine.

References

1. Gargari, O.K., et al., *Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo.* BMJ Evidence-Based Medicine, 2024. **29**(1): p. 69-70.
2. Horiuchi, D., et al., *Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases.* Neuroradiology, 2024. **66**(1): p. 73-79.
3. Ganesan, K., *Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.* arXiv preprint arXiv:1803.01937, 2018.
4. Guo, Z., et al., *Evaluating large language models: A comprehensive survey.* arXiv preprint arXiv:2310.19736, 2023.
5. Liu, X., et al., *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension.* Nature Medicine, 2020. **26**(9): p. 1364-1374.
6. Collins, G.S., et al., *TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods.* Bmj, 2024. **385**: p. e078378.
7. Jin, D., et al., *What disease does this patient have? a large-scale open domain question answering dataset from medical exams.* Applied Sciences, 2021. **11**(14): p. 6421.
8. Jin, Q., et al. *PubMedQA: A Dataset for Biomedical Research Question Answering*. 2019. Hong Kong, China: Association for Computational Linguistics.
9. Pal, A., L.K. Umapathi, and M. Sankarasubbu, *MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*, in *Proceedings of the Conference on Health, Inference, and Learning*, F. Gerardo, et al., Editors. 2022, PMLR: Proceedings of Machine Learning Research. p. 248--260.
10. Brin, D., et al., *Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments.* Sci Rep, 2023. **13**(1): p. 16492.
11. Barile, J., et al., *Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies.* JAMA Pediatr, 2024. **178**(3): p. 313-315.
12. Duey, A.H., et al., *Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations.* Spine J, 2023. **23**(11): p. 1684-1691.
13. Taira, K., T. Itaya, and A. Hanada, *Performance of the Large Language Model ChatGPT on the National Nurse Examinations in Japan: Evaluation Study.* JMIR Nurs, 2023. **6**: p. e47305.
14. Kung, T.H., et al., *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.* PLOS Digit Health, 2023. **2**(2): p. e0000198.

 doi:10.20944/preprints202501.0699.v1

15. Cascella, M., et al., *Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios.* J Med Syst, 2023. **47**(1): p. 33.
16. Yeo, Y.H., et al., *Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma.* Clin Mol Hepatol, 2023. **29**(3): p. 721-732.
17. Hermann, C.E., et al., *Let's chat about cervical cancer: Assessing the accuracy of ChatGPT responses to cervical cancer questions.* Gynecol Oncol, 2023. **179**: p. 164-168.