J. Health Inform. 2026 Janeiro-Dezembro; 18

Artigo Original

DOI: https://doi.org/10.59681/2175-4411.v18.2026.1570

JHI JOURNAL OF HEALTH INFORMATICS

Role of Large Language Models in Patient Safety in an Expert Evaluation Study

Papel dos Grandes Modelos de Linguagem na Segurança do Paciente em um Estudo de Avaliação por Especialistas

Papel de los grandes modelos lingüísticos en la seguridad del paciente en un estudio de evaluación de expertos

Antonio Valerio Netto[1], Camila de Brito Pontes[2]

ABSTRACT

Objective: This study comparatively evaluates the performance of different Large Language Models (LLMs) as tools to support medical prescription, considering criteria associated with patient safety and clinical applicability, based on a blinded expert evaluation. Methods: Six LLMs answered four questions about contraindications, drug interactions, and dosages. A panel of 34 physicians blindly evaluated 24 responses based on five criteria: consistency, focus, coherence, completeness, and detail. Results: Performance varied according to the criteria and question types; LLM6 showed better completeness and detail, especially in complex cases. Simple questions, such as contraindications, received higher scores, while complex questions showed greater variation. Conclusion: The findings indicate that the use of LLMs to support medical prescription requires careful model selection and consideration of the level of clinical complexity, reinforcing the need for contextual validation before their adoption in healthcare settings.

Keywords: Clinical Decision Support Systems, Artificial Intelligence, Large Language Models

RESUMO

Objetivo: Este estudo avalia comparativamente o desempenho de diferentes Large Language Models (LLMs) como ferramentas de apoio à prescrição médica, considerando critérios associados à segurança do paciente e à aplicabilidade clínica, a partir da avaliação cega de especialistas. Métodos: Seis LLMs responderam quatro questões sobre contraindicações, interações medicamentosas e dosagens. Um painel de 34 médicos avaliou às cegas 24 respostas com base em cinco critérios: consistência, foco, coerência, completude e detalhe. Resultados: O desempenho variou conforme os critérios e tipos de perguntas; o LLM6 teve melhor completude e detalhe, especialmente em casos complexos. Perguntas simples, como contraindicações, tiveram notas mais altas, enquanto as complexas apresentaram maior variação. Conclusão: Os achados indicam que o uso de LLMs no apoio à prescrição médica requer seleção criteriosa do modelo e consideração do nível de complexidade clínica, reforçando a necessidade de validação contextual antes de sua adoção em ambientes assistenciais.

Descritores: Sistemas de Apoio a Decisões Clínicas, Inteligência Artificial, Modelos de Linguagem Grande

RESUMEN

Objetivo: Este estudio evalúa comparativamente el desempeño de diferentes Modelos de Lenguaje Grande (LLMs) como herramientas para apoyar la prescripción médica, considerando criterios asociados con la seguridad del paciente y la aplicabilidad clínica, con base en una evaluación ciega de expertos. Métodos: Seis LLMs respondieron cuatro preguntas sobre contraindicaciones, interacciones medicamentosas y dosis. Un panel de 34 médicos evaluó ciegamente 24 respuestas con base en cinco criterios: consistencia, enfoque, coherencia, integridad y detalle. Resultados: El desempeño varió según los criterios y tipos de preguntas; LLM6 mostró mejor integridad y detalle, especialmente en casos complejos. Las preguntas simples, como contraindicaciones, recibieron puntuaciones más altas, mientras que las preguntas complejas mostraron mayor variación. Conclusión: Los hallazgos indican que el uso de LLMs para apoyar la prescripción médica requiere una cuidadosa selección del modelo y consideración del nivel de complejidad clínica, lo que refuerza la necesidad de validación contextual antes de su adopción en entornos de atención médica.

Descriptores: Sistemas de Apoyo a Decisiones Clínicas, Inteligencia Artificial, Grandes Modelos de Lenguaje.

[1] *Curso de pós-graduação em informática em saúde da UAB/Universidade Federal de São Paulo, São Paulo, SP, Brasil*

[2] *Curso de pós-graduação em informática em saúde da UAB/Universidade Federal de São Paulo, São Paulo, SP, Brasil*

Autor Correspondente: **Antonio Valerio Netto**
e-mail: **avnetto@hotmail.com**

Artigo recebido: 28/11/2025
Aprovado: 29/01/2026

INTRODUCTION

Medication errors are among the most frequent and preventable causes of patient harm(1). According to the National Coordinating Council for Medication Error Reporting and Prevention, a medication error is any preventable event that may cause or lead to inappropriate medication use or patient harm while the medication is under the control of a healthcare professional, patient, or consumer. These errors arise from failures in medication systems and human factors such as fatigue, poor environmental conditions, and staffing shortages that compromise critical steps in the care process, including prescribing, transcription, dispensing, administration, and monitoring. These failures can have serious consequences for patients, leading to severe adverse reactions, permanent disabilities, or, in extreme cases, death(2).

A medical prescription is a written document that provides instructions guiding the patient on medication use during treatment, including dosage, frequency of administration, and duration. Additionally, it must include both the patient's and the healthcare professional's information, as this is crucial for ensuring adherence to the prescribed treatment. Therefore, prescriptions should use clear language that aligns with the patient's level of understanding, presenting information in an organized and easily comprehensible manner, along with legible handwriting. Difficulty in understanding a prescription is one of the main factors contributing to non-adherence to treatment, a critical issue in patient safety, as well as a potential cause of medication errors(3).

Furthermore, medication errors can be associated with aspects of professional practice, healthcare products, and systems. They may occur at various stages, including prescribing, order communication, medication labeling and packaging, as well as in the composition, dispensing, and distribution of pharmaceutical products. Additionally, factors such as improper administration, deficiencies in the education of healthcare professionals and patients, ineffective monitoring, and incorrect medication contribute to these adverse events. Ensuring patient safety requires identifying and mitigating these risks at every stage of the medication process(4).

In response to this issue, the World Health Organization (WHO) launched the Third WHO Global Patient Safety Challenge, titled Medication Without Harm, aiming to reduce serious and preventable medication-related harm by 50% globally between 2017 and 2022(4). By raising awareness and increasing visibility, this initiative emphasizes that many medication administration issues stem from communication failures. Many patients may be unaware of essential information, such as the medication's name, appearance, purpose, correct dosage and frequency, optimal administration time, treatment duration, potential side effects, actions to take if a dose is missed, possible interactions with other drugs or foods, whether the new medication replaces or complements an existing therapy, and difficulties arising from illegible prescriptions(5-6).

Prescription errors not only impact patients' health and the economy but also have serious consequences for the healthcare professionals involved. Those who commit such errors may experience feelings of shame, guilt, and self-doubt, which, in some cases, can contribute to suicidal tendencies. Additionally, the legal and professional repercussions of a prescription error may lead to the revocation of a professional license, compounding the emotional distress already caused by the mistake. Furthermore, these errors can erode patient and family trust while increasing the risk of criminal charges and disciplinary action by professional boards(1).

To address the challenges posed by medication prescription errors and enhance patient safety, this research explores the adoption of Generative Artificial Intelligence tools, leveraging the knowledge generated by Large Language Models (LLMs). These models, trained on vast amounts of internet data using deep learning methods, can generate various forms of content, including text, video, and audio, and respond to user commands. As a result, LLMs hold great potential in healthcare by enabling instant access to medical information, supporting diagnostic processes, and providing insights into potential treatment options(7).

LLMs as aids in drug prescribing

The integration of healthcare data science and artificial intelligence (AI) has transformed process management in the field, influencing everything from data collection and storage to processing, interpretation, and clinical and administrative application. Additionally, AI enables the automation of repetitive tasks, optimizes workflows, supports decision-making, and facilitates personalized treatments. In healthcare education, AI also offers significant advantages, such as generating simulation content and clinical case scenarios to help students enhance their clinical reasoning(7-8-9). Among its various applications, LLMs stand out for their promising role in drug prescribing. These models assist in formulating safer prescriptions and contribute to reducing medical errors, thereby improving patient safety(7).

LLMs are machine learning models based on deep neural networks, designed to perform natural language processing (NLP) tasks such as translation, summarization, and question answering. Trained on vast amounts of textual data, these models can interpret user commands and generate human-like text. In the context of medical prescribing, LLMs, when integrated with clinical decision support systems, assist healthcare professionals in selecting appropriate medications, identifying potential drug interactions, and providing up-to-date information on dosages and side effects. This integration enhances prescription accuracy and helps reduce the incidence of medical errors(10).

Although LLMs hold promise as valuable tools in addressing healthcare demands, analyzing vast volumes of information to respond to user commands, they are also susceptible to generating inaccurate responses or responses based on non-existent data, a phenomenon known as hallucination. This issue poses a significant risk, particularly in drug prescribing, as it can compromise both the quality and safety of prescriptions. Hallucination occurs when models generate information that appears to be grounded in scientific evidence, but upon closer examination, proves to be incorrect or unfounded. As such, this limitation remains one of the key challenges in applying LLMs in healthcare, especially when precise accuracy is crucial to ensuring the efficacy and safety of drug prescriptions(7).

METHODS

This study employed a blinded comparative evaluation design to assess the quality of responses generated by multiple Large Language Models (LLMs) in prescription-related clinical scenarios. The unit of analysis was the individual LLM-generated response. Six LLMs were prompted with four predefined prescription questions, resulting in 24 responses. These responses were independently evaluated by a panel of 34 physicians using predefined quality criteria. The methodological workflow comprised three sequential stages: (i) generation of responses by LLMs under standardized prompting conditions, (ii) blinded expert evaluation using structured scoring criteria, and (iii) quantitative aggregation and comparative analysis of scores.

This research used an online form to collect responses from participants through closed-ended questions. This format allows the study to be classified as qualitative research with a quantitative component or as a mixed-methods approach. The adoption of this format is due to the combination of closed-ended questions typically associated with quantitative research, with the qualitative aim of exploring participants' perceptions and experiences. While the use of closed-ended questions may lean toward a quantitative approach, the focus on understanding the meaning of responses and their contextual interpretation aligns with the qualitative nature of the study(17).

The study involved evaluating prescription-related responses generated by LLMs from a panel of 34 physicians. Participants were recruited by convenience sampling and met the inclusion criteria of being licensed physicians with active clinical practice. The panel comprised professionals from a wide range of medical specialties, including internal medicine, geriatrics, cardiology, pulmonology, psychiatry, pediatrics, nephrology, surgery, and others, ensuring heterogeneity of clinical perspectives. Clinical experience among participants ranged from early-career physicians (approximately three years of practice) to senior physicians with over 20 years of professional experience. There were three physicians with up to three years of experience, 13 between three and seven years, two between 13 and 19 years, and finally, 16 with more than 20 years of experience.

All participating physicians independently evaluated the complete set of 24 anonymized responses generated by the LLMs. The assessments were conducted asynchronously using an online form, with responses presented in random order and without identifying the generating model. No intervention was applied, no personal health or sensitive data was collected, and participants were not exposed to physical, emotional, or social risks. Participation was voluntary, and all assessments were recorded and analyzed anonymously to preserve confidentiality.

This study was approved by the Ethics and Research Committee, under CAAE (Certificate of Presentation of Ethical Appreciation) number 86553225.9.0000.8054 and Opinion Number 7.493.332.

As the study did not involve direct patient participation or personal health data, informed consent was not applicable. However, all participating physicians were fully informed about the study's objectives, and their participation was entirely voluntary.

Six LLMs were selected to generate responses focused on medical prescription. The chosen models were: ChatGPT-4o, Cortex (z_piloto), Claude 3.5 Sonnet, Llama 3.1 405b Instruct, Cohere Command-Nightingale-30B, and Gemini 1.5 Flash. Furthermore, four questions were defined for this research, covering both objective and complex topics such as contraindications, drug interactions, and personalized treatments. These questions were:

- Question 1: "What is the contraindication for ibuprofen?"
- Question 2: "Are there any issues with taking metformin hydrochloride together with valsartan?"
- Question 3: "What are the complications of increasing the daily dosage of metformin hydrochloride?"

- Question 4: “What is the optimal daily dosage of metformin for a 52-year-old male patient who also has high blood pressure?”

The 24 LLM-generated responses were presented to evaluators through an online form. Responses were anonymized and randomized so that evaluators were blinded to both the identity of the LLM and the order of generation. Each physician independently evaluated all responses without discussion or calibration sessions. Prior to evaluation, participants received standardized written instructions defining each assessment criterion but were not provided with reference answers or scoring benchmarks.

The scores assigned were based on five criteria, with each physician scoring between 1 (poor) and 5 (excellent). These criteria were selected to reflect dimensions relevant to clinical usability and patient safety in prescription support, rather than isolated factual accuracy. Together, they capture logical soundness, relevance, completeness, and practical usefulness of the information provided to clinicians. The five criteria are:

- Consistency: Is the answer consistent with other known information?
- Focus: Does the answer directly address the question without unnecessary digressions?
- Coherence: Is the answer logically structured and easy to follow?
- Completeness: Does the answer cover all the important aspects of the question?
- Detail: Does the answer provide enough detail to be useful?

In medical prescribing, “consistency” is essential to ensure that the information provided aligns with best practices and guidelines. “Coherence” is crucial for responses to be clear and logically applicable in clinical practice. “Completeness” ensures that all relevant aspects are considered, preventing omissions that could lead to errors. Additionally, “detail” is vital for ensuring that information is accurate and useful in clinical decision-making. Finally, “Focus” ensures that the information is concise and not dispersed.

In this study, “understanding” of prescription-related information was not treated as a direct outcome variable, but rather as a “multidimensional construct” inferred through expert evaluation. The assessment criteria were selected to capture complementary dimensions that influence the interpretability and clinical usability of prescription guidance. From a patient safety perspective, inadequate understanding may arise not only from incorrect information, but also from fragmented explanations, lack of contextualization, omission of critical elements, or excessive ambiguity. Therefore, higher scores across these criteria were interpreted as proxies for improved comprehensibility and reduced risk of misinterpretation in clinical practice.

The committee of physicians was composed of professionals from various specialties, including clinical medicine, pulmonology, gynecology and obstetrics, anesthesiology, public health, gastroenterology, and neurology. In addition, the physicians had varied levels of clinical experience, ranging from residents with less than three years of practice to seasoned professionals with over 20 years of experience. The goal of the research was to subjectively assess whether the responses generated by the LLMs were adequate for consultation by residents and other physicians, based on the expertise and judgment of each participant. Importantly, the physicians were blinded to which LLM generated each response, ensuring an unbiased evaluation. The intent was for them to assess the quality of each response objectively, without any preconceived notions about the LLM that produced it.

After data collection, the information was tabulated and identified for analysis in the results and discussion. Descriptive statistics (means and variability measures) were calculated for each LLM across criteria and questions. To examine whether physician experience influenced evaluations, inferential analyses were performed using ANOVA when parametric assumptions were met and Kruskal–Wallis tests otherwise. Post-hoc Tukey tests were applied when appropriate.

Selection of Large Language Models

Six LLMs were selected based on three criteria: (i) widespread use or commercial relevance in healthcare-related applications, (ii) architectural diversity (closed-source, open-source, and hybrid platforms), and (iii) public availability at the time of the experiment. All models were accessed between [January/2025] and [March/2025] using their default inference settings, without fine-tuning or retrieval augmentation. The chosen models were: ChatGPT-4o(11), Cortex (z_piloto)(12), Claude 3.5 Sonnet(13), Llama 3.1 405b Instruct(14), Cohere Command-Nightingale-30B(15), and Gemini 1.5 Flash(16).

All LLMs were prompted using identical user-level prompts corresponding to the four prescription-related questions described below. No iterative prompting, follow-up questions, or response refinement was allowed. Each model generated a single response per question. Default generation parameters were used for all models, including temperature and sampling settings, as defined by each platform at the time of access. No external knowledge bases, plugins, or clinical decision support tools were enabled. No LLM was used with RAG (Retrieval-augmented generation).

ChatGPT-4o is one of the most popular language models developed by OpenAI, designed for natural text processing and generation. This version offers faster and more accurate responses, enhancing the user's experience across a variety of applications, from virtual assistants to creative writing tools. Cortex is a platform that integrates multiple LLM models, including the latest versions from OpenAI, Anthropic, Google, Llama, and others. Its architecture supports the creation of specialized agents, enabling more in-depth, qualified, and accurate responses. This flexibility allows Cortex to adapt to different contexts and demands, improving the quality of analyses and the efficiency of cybersecurity, risk management, and compliance processes. In the pilot, a persona called "Medical Assistant" (z_piloto) was created using specific prompt configurations to efficiently process medical information. The "Medical Assistant" was powered by OpenAI's GPT-4o model.

Claude 3.5 Sonnet is an advanced language model developed by Anthropic, released on June 21, 2024. It outperforms both its competitors and previous versions, such as Claude 3 Opus, in assessments of graduate-level reasoning, university-level knowledge, and coding proficiency. Additionally, it offers significant improvements in understanding nuances, humor, and complex instructions, establishing itself as a reference for producing high-quality content with a natural and engaging tone.

Llama 3.1 405B Instruct is a large-scale language model developed by Meta, designed for multilingual dialogue tasks. With 405 billion parameters, it has been fine-tuned to follow instructions, enabling more natural and effective interactions across multiple languages. Compared to other models, it outperforms industry benchmarks, excelling in code generation and complex dialogue tasks. Additionally, it supports the processing of large textual inputs, making it particularly useful for analyzing long documents and generating synthetic data.

Cohere Command-Nightingale-30B is a large-scale language model developed by Cohere. It is an advanced iteration of the Command language model, designed for both text generation and natural language understanding tasks. Named for its size, approximately 30 billion parameters, Nightingale-30B's substantial scale enables it to process and learn from vast amounts of language data. This capability allows the model to generate high-quality text and perform complex natural language processing tasks.

Gemini 1.5 Flash, developed by Google, is designed to be faster and more cost-effective than Gemini 1.5 Pro. It prioritizes efficiency and low latency, making it well--suited for applications that require quick and frequent responses, such as chatbots and document analysis. Despite its optimization for speed, Gemini 1.5 Flash retains a large context window of one million tokens and multimodal capabilities, allowing it to process and analyze large volumes of information in various formats, including text, code, audio, and video.

RESULTS

Based on the tabulated results of the collected responses, some patterns can be identified. For example:

- Most challenging questions:

o Question 4 ("What is the best daily dosage of metformin for a 52-year-old male patient who also has high blood pressure?") was the most challenging, showing greater variability in scores and lower average performance, particularly in the "Completeness" and "Detail" criteria.

o Question 2 ("Is there any issue with taking metformin hydrochloride together with valsartan?") also posed challenges for some LLMs in providing clear and complete answers.

- Criteria with the greatest variability:

o Completeness and Detail showed the largest differences in scores across models, indicating difficulties in covering all relevant aspects and providing detailed information.

- Models with best overall performance:

o LLM6 excelled in most criteria, particularly in "Completeness" and "Detail", showing greater consistency when addressing complex questions.

o LLM3 also achieved high scores in Coherence and Consistency, demonstrating logical reasoning and precision in its answers.

- The impact of physicians' experience:

o More experienced doctors tended to be more rigorous in their assessments, while residents showed greater uniformity in their grades.

Additionally, other patterns were identified and are presented. There is variability in the performance of LLMs across the assessed criteria and the questions asked. Some LLMs excel in certain areas, while others perform more poorly. This may suggest that some LLMs are better suited for specific types of questions, or that their language models are stronger in certain areas. For example, LLM2 and LLM6 seem to score higher overall across multiple criteria and various questions. Regarding the evaluation criteria, it was observed:

- Consistency: Most LLMs performed well on this criterion, indicating that their responses are generally aligned with medical knowledge. The greatest inconsistencies were observed in LLM3, LLM4, and LLM5, particularly in questions 2, 3, and 4.
- Focus: Scores for this criterion vary widely, suggesting that some LLMs have difficulty focusing on the specific question, leading to responses that are sometimes more rambling. LLM1 appears to score better in this regard on almost all questions.
- Coherence: LLMs generally receive good scores for coherence, indicating that responses are logically structured and make sense.
- Completeness and Detail: The scores for this criterion vary more, suggesting that not all LLMs are able to address all important aspects of the question or provide sufficient detail. LLMs such as LLM1, LLM2, and LLM6 seem to perform better in this regard.

DISCUSSION

Regarding the content of the questions, Question 1 about ibuprofen contraindications appears to have generated the most consistent and well-evaluated responses from all LLMs. In contrast, Questions 2, 3, and 4 seem to present more challenges for the LLMs, with evaluations being more variable, particularly in the criteria of Focus, Completeness, and Detail. This suggests that more complex questions require further improvement in LLM capabilities. It is also noticeable that there is variation in the evaluations provided by each individual physician. LLM4 shows significant variation in the physicians' assessments, indicating that some physicians may disagree with the method of response. Similar variations can also be observed in LLM1 and LLM3.

To identify the most challenging questions for the LLMs, we analyzed the average scores and variability in the evaluations provided by the 34 physicians, considering the five assessment criteria: "Consistency", "Focus", "Coherence", "Completeness", and "Detail" (Table 1).

Table 1 - Most challenging questions for LLMs

Question	Difficulty level for LLMs	Clinical and cognitive demands	Interpretation
1	Good	It provided more objective responses based on well-established guidelines, achieving relatively high and consistent scores across all criteria. It was also considered the least challenging.	It was the least challenging, as it was more objective and grounded in widely known information.
2	Moderate	There were more variable scores in Completeness and Detail. This question required LLMs to integrate pharmacological knowledge about drug interactions, which highlighted the limitations of some models.	Moderately challenging, requiring precise knowledge of drug interactions.
3	Variable	The responses needed to address specific side effects and complications, requiring greater detail and clinical context. Some models struggled to provide sufficiently complete responses.	Also challenging, with a focus on the need for detailed information on specific complications.

4	More challenging	It had the lowest scores in Completeness and Detail. It required personalization of the response based on multiple clinical factors (age, comorbidities, specific medication), revealing limitations in the capabilities of LLMs to deal with highly personalized scenarios.	It was the most challenging, due to the need for personalization and the integration of multiple clinical factors.

Source: Prepared by the authors.

To determine which model (LLM1 to LLM6) performed best for each criterion, we conducted an analysis based on the average scores given by the 34 physicians across the four questions. Each criterion is analyzed separately (Table 2).

Table 2 – Performance by criterion

Criterion	Best model	Justification
Consistency	LLM3 e LLM6	LLM3 scored consistently high across all questions, particularly those requiring correlation with established guidelines, such as contraindications (Question 1). LLM6 scored highest in consistency (4.49), indicating that its responses are the most aligned with known medical information.
Focus	LLM2 e LLM1	LLM2 stood out for its direct approach to the questions, avoiding unnecessary digressions. It was particularly effective in objective questions, such as drug interactions (Question 2). LLM1 had the highest mean in focus (4.08), indicating that its responses most directly addressed the question without deviation.
Coherence	LLM3 e LLM6	LLM3 excelled in structuring responses in a logical and fluid manner, enhancing clarity and ease of understanding. LLM6 had the highest average in coherence (4.18), indicating that its responses exhibited the most consistent logical structure.
Completeness	LLM6	LLM6 consistently addressed all aspects of the questions, with particular emphasis on complex issues such as metformin dosage in hypertensive patients (Question 4). LLM6 achieved the highest mean completeness score (3.84), indicating that its responses most thoroughly covered all key aspects of the question.

Detail	LLM6	LLM6 provided responses rich in useful information, receiving particularly high evaluations in the analysis of complications and treatment personalization (Questions 3 and 4). LLM6 achieved the highest average in detail (4.16), indicating that its responses offered sufficient detail to be useful.

Source: Prepared by the authors.

The results indicate that LLM3 and LLM6 were the most consistent and coherent models, demonstrating superior performance in the more challenging criteria, such as “Completeness” and “Detail”. LLM2 and LLM1 excelled in the “Focus” criterion, showing their ability to provide direct and objective answers. It is important to note that the choice of the ideal model will depend on the clinical context and the type of question being addressed.

Based on the evaluated criteria and the performance of the six LLMs, LLM6 emerges as the best choice for the topic of medical prescription. LLM6 stands out due to its ability to provide complete, detailed, and applicable responses to complex clinical scenarios. This choice is supported by its balance across the evaluated criteria and its potential to make significant contributions to clinical practice, particularly in situations requiring in-depth and personalized analysis.

The detailed justification for this choice is provided below:

- Superior performance in “Completeness” and “Detail”: LLM6 consistently scored highest in two crucial criteria for prescribing. These criteria reflect their ability to provide comprehensive answers that cover multiple clinical aspects and present detailed, useful information for the physician.
- Ability to handle complex questions: In challenging scenarios, such as personalizing dosage for patients with comorbidities (Question 4), LLM6 demonstrated a superior ability to integrate information on age, comorbidities, and medications, providing more contextualized recommendations.
- Direct clinical applicability: The responses generated by LLM6 were more closely aligned with the practical needs of medical prescribing, demonstrating precision in addressing drug interactions and treatment complications, as seen in Questions 2 and 3.
- Consistency and reliability: While other models, such as LLM3, excelled in criteria like Consistency and Coherence, LLM6 demonstrated a more robust balance across all criteria, proving to be the most reliable overall.
- Lower variability in assessments: Analysis of the data shows that LLM6 exhibited lower variability in the assessments provided by physicians compared to other LLMs, such as LLM4. This suggests that LLM6 generates responses that are more consistently regarded as high-quality by experts.

During the research, one question piqued the researchers’ curiosity: ‘Is there a relationship between the number of years a doctor has been practicing medicine and the evaluations they provided for each LLM?’ To answer this, a statistical analysis was conducted to determine if there is any correlation between years of medical practice and LLM evaluations. The following steps were taken:

- Data organization: A spreadsheet was created to organize all the evaluations, categorizing the physicians by years of practice.
- Calculation of averages: The average scores for each LLM were calculated for each criterion, within each category of practice duration.
- Statistical analysis: ANOVA was conducted to compare the meaning between the categories. When the assumptions for ANOVA were not met, the nonparametric Kruskal-Wallis test was applied.
- Post-hoc tests: Tukey’s post-hoc test was performed to identify which categories differed significantly when ANOVA or the Kruskal-Wallis test indicated differences.

After analysis, the results revealed that, overall, there was no consistent, statistically significant relationship between years of medical practice and LLM ratings. In other words, the length of physicians’ experience does not appear to generally influence how they evaluate LLM responses. The absence of a consistent statistically significant relationship suggests that clinical experience (measured by years of practice) is not a determining factor in assessing the quality of LLM responses. This may indicate that other factors are more influential in the assessment, such as:

- Medical specialty: A doctor’s area of expertise may influence their perception of the quality of the responses.
- Familiarity with technology: Doctors who are more familiar with technology may assess the responses differently than those with less experience.
- Individual preferences: A physician’s personal preferences may influence their evaluation, regardless of their years of practice.
- Question complexity: The difficulty of the question may have a greater impact on the assessment than the physician’s level of experience.

CONCLUSION

It is important to emphasize that no LLM is perfect, and all have their limitations. LLM6, despite being the best among those evaluated, still requires improvement. The evaluation was based on a specific set of questions and criteria, and LLM performance may vary in different contexts. Therefore, LLMs should be used as supportive tools rather than substitutes for clinical judgment and medical experience.

Trained on vast textual datasets, LLMs can interpret, synthesize, and generate information in natural language with high accuracy. In healthcare, this enables professionals to quickly access insights from medical guidelines, scientific research, and electronic patient histories. This application not only speeds up the decision-making process but also reduces the risk of errors by integrating scientific evidence directly into clinical practice. Another promising application is predictive analytics, which helps identify patients at risk of developing serious conditions, enabling early interventions. LLMs are also being used to assist in interpreting complex tests, translating results into understandable information for both patients and healthcare teams. Furthermore, by processing data from multiple sources, such as electronic health records, wearable devices, and sensors; these models become essential for efficient and personalized patient care management.

However, it is crucial to address the ethical and technical challenges associated with LLMs in healthcare, including data privacy, model transparency, and equitable access to technology. As LLMs continue to evolve, their integration into health data science has the potential to enhance efficiency and quality of care while redefining the future of evidence-based and personalized medicine. Finally, for these technologies to be integrated effectively into clinical practice, LLMs should serve as complementary tools rather than substitutes for human judgment. It is also essential that healthcare professionals understand the limitations of these models and validate the information provided before making critical decisions. The adoption of LLMs must be accompanied by rigorous validation and continuous improvement, prioritizing safety and effectiveness in patient care.

REFERENCES

1. Tariq RA, Vashisht R, Sinha A. Medication Dispensing Errors and Prevention. In: StatPearls, Treasure Island: StatPearls Publishing; [cited 2025 May 14]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519065.
2. National Coordinating Council for Medication Error Reporting and Prevention. About medication errors: What is a medication error?. [cited 2025 Feb 2]. Available from: https://www.nccmerp.org/about-medication-errors.
3. Cruzeta APS, Dourado ACL, Monteiro MTM, Martins RO, Calegario TA, Galato D. Fatores associados à compreensão da prescrição médica no Sistema Único de Saúde de um município do Sul do Brasil. Cien Saude Colet. 2013;18(12):3731-3737.
4. World Health Organization. Medication Without Harm; 2025. [cited 2025 Feb 3]. Available from: https://www.who.int/initiatives/medication-without-harm.
5. Academy of Managed Care Pharmacy. Medication Errors, 2025. [cited 2025 Feb 3]. Available from: https://www.amcp.org/concepts-managed-care-pharmacy/medication-errors.
6. Cohen MR, Smetzer JL. ISMP Medication Error Report Analysis. Hospital pharmacy 2017; 52: 390-393. [cited 2026 Feb 4]. Available from: https://doi.org/10.1177/0018578717715346
7. Shah K., Xu AY, Sharma Y, Daher M, McDonald C, Diebo BG, Daniels AH. Large Language Model Prompting Techniques for Advancement in Clinical Medicine. Journal of Clinical Medicine 2024; 13: 1-12. [cited 2026 Feb 4]. Available from: https://doi.org/10.3390/jcm13175101
8. Netto AV, Berton L, Takahata AK. Ciência de dados e a inteligência artificial na área da saúde. Editora dos Editores; 2021.
9. Netto AV. Ciência de dados em saúde: contribuições e tendências para aplicações. Revista Saúde.com, 2021;(17) 1-5. [cited 2026 Feb 4]. Available from: https://doi.org/10.22481/rsc.v17i3.6290
10. Software & Data: Osford's DrugGPT AI tool enhances medication prescriptions. The Healthcare Technology Report; 2024. [cited 2025 Feb 4]. Available from: https://thehealthcaretechnologyreport.com/oxfords-druggpt-ai-tool-enhances-medication-prescriptions.
11. OpenAI. Hello ChatGPT-4o; 2024. [cited 2025 Feb 8]. Available from: https://openai.com/index/hello-gpt-4o.

12. Cortex. Your Exclusive Corporate AI; 2025. [cited 2025 Feb 8]. Available from: https://sinapse.tech/cortex.
13. Anthropic. Introducing Claude 3.5 Sonnet information; 2024. [cited 2025 Feb 8]. Available from: https://www.anthropic.com/news/claude-3-5-sonnet.
14. Meta. Introducing Llama 3.1: Our most capable models to date; 2024. [cited 2025 Feb 8]. Available from: https://ai.meta.com/blog/meta-llama-3-1.
15. Cohere. The all-in-one platform for private and secure AI; 2024. [cited 2025 Feb 8]. Available from: https://cohere.com/.
16. Google. Gemini models; 2025. [cited 2025 Feb 8]. Available from: https://ai.google.dev/gemini-api/docs/models/gemini.
17. Medeiros CH, Kauark FD, Manhães FC. Metodologia da pesquisa: Guia prático. Via Litterarum, Itabuna; 2010.