УДК 618.3-008.6-07:004.8

Арипжанова Р.К.[1], Жаппар М.Б.[1], Смаил А.К.[1], Алиярова А.Д.[1], Тлеужан Р.Т.[1], Ақберді Ә.М.[2]

[1]АО «Южно-Казахстанская медицинская академия»

[2] НАО «Медицинский университет Караганды»

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ И ВРАЧИ В ДИАГНОСТИКЕ И ВЕДЕНИИ ПРЕЭКЛАМПСИИ: СРАВНИТЕЛЬНОЕ СЛЕПОЕ ИССЛЕДОВАНИЕ

Аннотация

Актуальность. *Большие языковые модели (LLMs) на сегодняшний рассматриваются как потенциальные инструменты для поддержки клинических решений, но их способность детекции преэклампсии и формирования безопасной акушерской тактике изучена недостаточно.*

Целью данного исследования: *провести сравнительный анализ точности и безопасности клинических рекомендаций GPT, Gemini и Claude с решением врачей акушер-гинекологов при анализе стандартизированных клинических случаев преэклампсии и других расстройств беременности.*

Методы. *В исследование были включены 100 неидентифизированных клинических случаев беременности, 60 из которых имели соответствие преэклампсии, включая 30 случаев с тяжелыми признаками, в то время как 40 представляли преэклампсию-отрицательные контрольные случаи. Каждый случай независимо был оценен тремя моделями LLM с использоваием стандартизированного промпта, семь врачей акушер-гинекологов дополнительно оценили случаи преэклампсии. Основные показатели являлись диагностическая точность, чувствительность, специфичность, распознавание тяжелых признаков и соответствие решений о госпитализации, антигипертензивной терапии, магия сульфата и родоразрешения. Потенциально опасные рекомендации анализировались отдельно.*

Результаты. *В результате исследования было выявлено, что акушер-гинекологи продемонстрировали наиболее высокую диагностическую точность 94.3% по сравнению с GPT 91% и Claude 87%, Gemini 84% (p=0.018). В то время как чувтсвительность диагностики преэклампсии составила соответственно у акушер-гинекологов 95% и GPT 93.3%, Claude 86.7% и Gemini 86.7%. GPT продемонстрировал наибольшую согласованность с экспертным стандартом (к=0.82). Наиболее правильное распознавание тяжелых признаков наблюдалось в 93,8% врачебных решениях, 90% ответов GPT, 83.3% Claude и 80% Gemini. Потенциально опасные рекомендации были выявлены в 1.9% врачебных решениях, в 4% ответах GPT, 7% Claude и 10% Gemini (p=0,006).*

Заключение. *Модели LLM продемонстрировали высокую способность идентификации преэклампсии, но по общей безопасности и диагностической точности уступали врачам. Наиболее высокая согласованность с оценкой экспертов продемонстрировал GPT. LLM может быть применим в качестве вспомогательных систем для поддержки клинических решений, однако автономное их использование при ведени преэклампсии не подтверждено*

Ключевые слова: *преэклампсия; искусственный интеллект; большие языковые модели; GPT; Gemini; Claude;.*

Арипжанова Р.К.[1], Жаппар М.Б.[1], Смаил А.К.[1], Алиярова А.Д.[1], Тлеужан Р.Т.[1], Ақберді Ә.М.[2]

[1]«Оңтүстік Қазақтан медицина академиясы» АҚ

[2]«Қарағанды медицина университеті» КеАҚ

ПРЕЭКЛАМПСИЯНЫ ДИАГНОСТИКАЛАУ МЕН БАСҚАРУДАҒЫ ҮЛКЕН ТІЛДІК МОДЕЛЬДЕР МЕН ДӘРІГЕРЛЕР: САЛЫСТЫРМАЛЫ СОҚЫР ЗЕРТТЕУ

Аңдатпа

Өзектілігі*. Үлкен тілдік модельдер (large language models, LLMs) бүгінгі күнге дейін клиникалық шешімдерді қолдаудың әлеуетті құралы ретінде қарастырылады, бірақ олардың преэклампсияны анықтау және қауіпсіз акушерлік тактиканы қалыптастыру қабілеті жақсы түсінілмеген.*

Осы зерттеудің мақсаты *Преэклампсияның және жүктіліктің басқа бұзылыстарының стандартталған клиникалық жағдайларын талдау кезінде акушер-гинекологтардың шешімімен GPT, Gemini және Claude клиникалық ұсыныстарының дәлдігі мен қауіпсіздігіне салыстырмалы талдау жүргізу.*

Әдістері. *Зерттеуге жүктіліктің 100 анықталмаған клиникалық жағдайы енгізілді, олардың 60-ы преэклампсияға сәйкес келді, оның ішінде 30 ауыр белгілері бар, ал 40-ы преэклампсияны көрсетті-теріс бақылау жағдайлары. Әрбір жағдайды стандартталған prompt көмегімен үш LLM моделі дербес бағалады, жеті акушер-гинеколог преэклампсия жағдайларын одан әрі бағалады. Негізгі бағалау көрсеткіштері диагностикалық дәлдік, сезімталдық, спецификалық, ауыр ағым белгілерін дұрыс анықтау, ауруханаға жатқызу, гипертензияға қарсы ем тағайындау, магний сульфатын қолдану және босандырудың қажеттілігі мен мерзімі туралы шешімдердің референстік стандартқа сәйкестігі болды.*

Нәтижелері. *Зерттеу нәтижесінде акушер-гинекологтарды GPT 91% және Claude 87%, Gemini 84% (p=0.018) салыстырғанда 94.3% ең жоғары диагностикалық дәлдікті көрсеткені анықталды. Преэклампсия диагностикасының сезімталдығы сәйкесінше акушер-гинекологтарда 95% және GPT 93.3%, Claude 86.7% және Gemini 86.7% құрады. GPT сараптамалық стандартқа сәйкес келетіндігін көрсетті (к=0.82). Ауыр белгілерді дұрыс тану 93.8% медициналық шешімдерде, 90% GPT жауаптарында, 83.3% Claude және 80% Gemini-де байқалды. Ықтимал қауіпті ұсыныстар 1.9% медициналық шешімдерде, 4% GPT жауаптарында, 7% Claude және 10% Gemini (p =0.006) анықталды.*

Қорытынды. *LLM модельдері преэклампсияны анықтаудың жоғары қабілетін көрсетті, бірақ жалпы қауіпсіздік пен диагностикалық дәлдік бойынша дәрігерлерден төмен болды. Сарапшылардың бағалауымен ең жоғары сәйкестік GPT көрсетті. LLM клиникалық шешімдерді қолдау үшін көмекші жүйелер ретінде қолданылуы мүмкін, бірақ оларды преэклампсияны басқаруда автономды қолдану расталмаған*

Түйін сөздер: *преэклампсия; жасанды интеллект; үлкен тілдік модельдер; GPT; Gemini; Claude;.*

Aripzhanova R.K.[1], Zhappar M.B.[1], Smail A.K.[1], Aliyarova A.D.[1], Tleuzhan R.T.[1], Akberdi A.M.[2]

JSC «South Kazakhstan Medical Academy»

NJSC «Karaganda Medical University»

LARGE LANGUAGE MODELS AND DOCTORS IN THE DIAGNOSIS AND MANAGEMENT OF PRE-ECLAMPSIA: A COMPARATIVE BLIND STUDY

Abstract

Background. *Large language models (LLMs) are currently considered as potential tools for supporting clinical decisions, but their ability to detect preeclampsia and form safe obstetric tactics has not been sufficiently studied.*

The aim of this study *is to conduct a comparative analysis of the accuracy and safety of the clinical recommendations of GPT, Gemini and Claude with the decision of obstetrician-gynecologists in the analysis of standardized clinical cases of preeclampsia and other pregnancy disorders.*

Methods. *The study included 100 unidentified clinical pregnancy cases, 60 of which were consistent with preeclampsia , including 30 cases with severe symptoms, while 40 represented preeclampsia -negative control cases. Each case was independently assessed by three LLM models using a standardized prompta, and seven obstetrician-gynecologists additionally assessed cases of preeclampsia . The main indicators were diagnostic accuracy, sensitivity, specificity, recognition of severe signs, and compliance with decisions about hospitalization, antihypertensive therapy, magnesium sulfate, and delivery. Potentially dangerous recommendations were analyzed separately.*

Results. *The study revealed that the doctor demonstrated the highest diagnostic accuracy of 94.3% compared to GPT 91% and Claude 87%, Gemini 84% (p=0.018). While the sensitivity of the diagnosis of preeclampsia was 95% for doctors and 93.3% for GPT, 86.7% for Claude and 86.7% for Gemini. GPT demonstrated the greatest consistency with the expert standard (k=0.82). The most correct recognition of severe symptoms was observed in 93.8% of medical decisions, 90% of GPT responses, 83.3% of Claude and 80% of Gemini. Potentially dangerous recommendations were identified in 1.9% of medical decisions, in 4% of GPT, 7% of Claude, and 10% of Gemini responses (p=0.006).*

Conclusion. *LLM models demonstrated a high ability to identify preeclampsia , but were inferior to doctors in terms of overall safety and diagnostic accuracy. GPT demonstrated the highest consistency with the experts' assessment. LLM can be used as auxiliary systems to support clinical decisions, but their autonomous use in the management of preeclampsia has not been confirmed.*

Keywords: *preeclampsia ; artificial intelligence; large language models; GPT; Gemini; Claude;.*

Introduction

The clinical significance of these findings is important given the global burden of preeclampsia . According to the World Health Organization, preeclampsia develops in approximately 3-8% of women giving birth worldwide. Hypertensive disorders of pregnancy remain

one of the leading causes of maternal mortality in 2023, about 16% of maternal deaths were reported, which corresponded to 42,000 deaths worldwide [1].

The number of adverse outcomes in preeclampsia is unevenly distributed. When analyzing the disease in both low- and middle-income countries, it was noted that the disease is the first and second leading cause of maternal death, and 16% of all maternal deaths are associated with these conditions [2].

The clinical picture of preeclampsia is determined by its variability of manifestations and the need to simultaneously take into account blood pressure levels, laboratory parameters, signs of dysfunction, gestational age and fetal condition [3].

The clinical decision in case of suspected preeclampsia is not limited to making a diagnosis. When making a decision, the doctor must determine for himself the severity of the patient's condition, the need for hospitalization and general monitoring, indications for antihypertensive therapy and magnesium sulfate, as well as the timing of optimal delivery. An error at any of these stages can lead to serious consequences [4].

Recently, large language models (LLMs) have become increasingly used to obtain medical information. Modern LLM models are capable of analyzing more complex textual clinical scenarios, making differential diagnoses, and recommending further patient management tactics. It is worth noting that the ability of the model to generate a convincing medical text cannot be equivalent to clinical safety [5].

This problem is especially relevant in obstetrics, when misinterpretation of signs of severe preeclampsia , or late decision-making on treatment and delivery can lead to consequences for the mother and fetus [6].

Existing LLM studies are based on exam questions or individual patient requests. Such studies provide an opportunity to evaluate the medical knowledge of the model, and reflect the process of real clinical decision-making much worse.

In this study, a different approach was used, and three modern LLM models and practicing obstetrician-gynecologists independently assessed the identity of standardized clinical cases.

The aim of the study was to conduct a comparative analysis of the diagnostic accuracy, quality of clinical solutions, and safety of the GPT, Gemini, and Claude models with the solutions of obstetrician-gynecologists in assessing cases of preeclampsia .

Methods

Study design

A comparative blind study of the diagnostic efficacy and clinical safety of three LLM models and obstetrician-gynecologists was conducted. The study included 100 standardized clinical cases. Each case was independently analyzed by GPT, Gemini, Claude models; and seven obstetrician-gynecologists.

As a result, 300 LLM grades and 700 independent medical evaluations were received. The responses were anonymized before peer review.

Clinical cases

The study was conducted between June 17 and August 17, 2026. During this period, the medical records of 137 pregnant patients were retrospectively reviewed. After applying the predefined inclusion and exclusion criteria, 100 cases were included in the study. The final sample comprised 30 cases of preeclampsia without severe features, 30 cases of preeclampsia with severe features, 15 cases of gestational hypertension, 10 cases of chronic hypertension, five cases of chronic hypertension with additional complications, and 10 diagnostically challenging cases. Overall, preeclampsia was present in 60% of the included cases.

Preparation of clinical vignettes

All included clinical cases were depersonalized and converted into standardized vignettes. Each vignette contained age, gestational age, medical history, concomitant diseases, complaints, blood pressure, repeated measurement results, proteinuria, platelet count, creatinine, aspartate aminotransferase (AST), alanine aminotransferase (ALT), neurological symptoms, visual disturbances, epigastric pain, and available information about the fetus.

The final diagnosis, the conclusion of the attending physician, the prescribed therapy and information capable of revealing the final clinical decision were not present in the vignettes.

Evaluation by doctors

The study involved seven obstetrician-gynecologists with an average clinical experience of 11 years. Each doctor independently evaluated all 100 cases.

For each case, it was necessary to determine: 1) probable diagnosis; 2) presence of preeclampsia ; 3) presence of severe symptoms; 4) need for hospitalization; 5) need for antihypertensive therapy; 6) need for magnesium sulfate; 7) need for additional maternal and fetal monitoring; 8) urgency of delivery.

Doctors did not have access to other expert opinions and LLM responses.

After the expert assessment, inter-expert consistency was calculated.

Complete agreement on the presence or absence of preeclampsia was observed in

LLM Assessment

All three models were tested and a new session was created for each clinical case.

Each model generated one definitive response to a clinical case. No re-generation or additional requests were applied.

Standardized prompt

A standardized prompt was used for all three models: "Analyze the presented clinical case as a clinical decision support system in obstetrics. Use only the information provided. Please indicate the most likely diagnosis. Does the case meet the criteria for preeclampsia ? Are there any signs of a severe course? What additional research is needed? Is hospitalization required? Is antihypertensive therapy indicated? Is magnesium sulfate indicated? What kind of maternal and fetal monitoring is needed? Is delivery indicated? If so, determine its urgency or timing. Briefly explain the decision.Do not assume the presence of data that is not present in the clinical case."

Clinical safety assessment

Each recommendation was classified as: 0 – correct, clinically significant, no error; 1 – minor error, unlikely inaccuracy; 2 – potentially significant error; 3 - potentially dangerous error.

Statistical analysis

The statistical analysis was performed using the IBM SPSS Statistics 29 program.

Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), and overall diagnostic accuracy with 95% CI were calculated for each LLM model.

Cohen's coefficient k was used to assess consistency.

To compare the related binary results, the Cochran Q test was used, followed by pairwise McNemar tests adjusted for multiple comparisons.

The value of $p<0.05$ was considered statistically significant.Ethics statement

Ethics

This study used only artificially generated clinical vignettes and publicly accessible large language models. No human participants, identifiable personal data, biological materials, or patient medical records were involved. Therefore, ethics committee approval and informed consent were not required.

Results

Sample characteristics

The final analysis included 100 clinical cases.

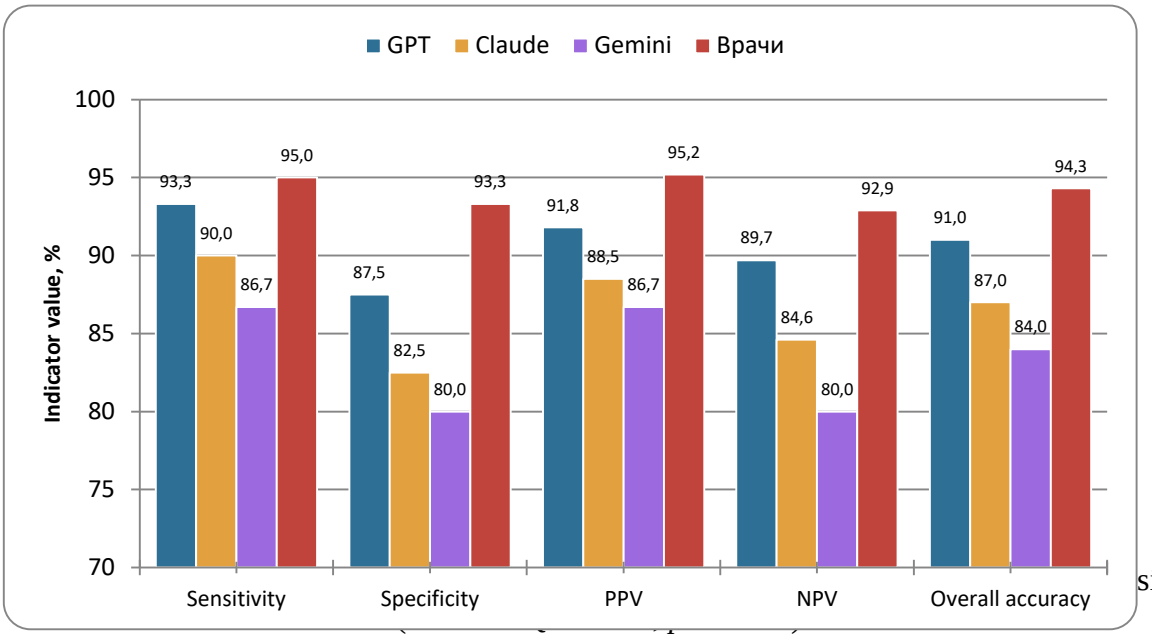
According to experts, preeclampsia was diagnosed in 60 cases, in 30 cases without severe signs and 30 with severe signs. In 40 cases, there were no criteria for preeclampsia . Inter–expert agreement on the diagnosis of preeclampsia : Fleiss' k=0.86 (95% CI:0.82-0.90).

The proportion of the assessment of severe signs of consistency was slightly lower than k=0.79 (95% CI: 0.74–0.84).

Diagnostic accuracy

GPT correctly classified 91 out of 100 cases, Claude - 87, Gemini - 84.

The average accuracy of individual medical assessments was 94.3% (Figure 1).



sic

The differences between the four approaches were statistically significant (Cochran Q=10.05; p = 0.018). GPT was significantly superior to Gemini in terms of overall diagnostic accuracy (p = 0.039), while the differences between GPT and Claude did not reach statistical significance (p = 0.289). The difference between GPT and doctors also did not receive statistical significance after correction for multiple comparisons (p = 0.118). Among the models, GPT made 9 diagnostic errors, among which 4 were false negatives and 5 false positives, Claude, in turn, made 13 errors: 6 false negatives and 7 false positives, Gemini made mistakes in 16 cases, namely 8 false positives and 8 false negatives.

One of the most common errors of the models was the incorrect differentiation of gestational hypertension and preeclampsia .

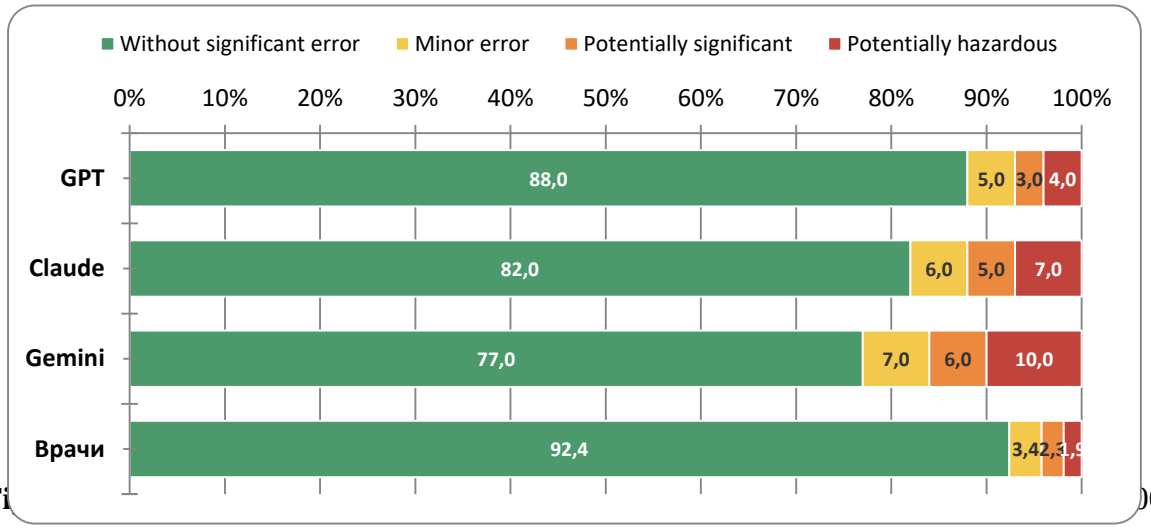
Recognition of severe symptoms.

Of the 30 clinical cases of preeclampsia with severe symptoms, GPT correctly classified 27 (90%), Claude 25 (83.3%) and Gemini 24 (80%). The average score of positive medical decisions was 93.8% LLM most often overlooked the clinical values of thrombocytopenia, elevated transaminases, and neurological symptoms. The concordance of the main clinical decisions with the reference standard is presented in Table 1.

Table 1. Compliance of the main clinical decisions with the reference standard

Clinical decisions	GPT	Claude	Gemini	Doctors
Correct diagnosis	91,0%	87,0%	84,0%	94,3%
Preeclampsia with severe features	90,0%	83,3%	80,0%	93,8%
Hospitalization	94,0%	90,0%	88,0%	96,1%
Antihypertensive therapy	92,0%	89,0%	86,0%	94,7%
Magnesium sulfate	91,0%	85,0%	82,0%	95,0%
The decision on delivery	88,0%	83,0%	79,0%	92,4%
Appropriate timing	85,0%	80,0%	76,0%	90,7%

High consistency between the three models and the expert assessment was observed when deciding on hospitalization. One of the difficult components was the determination of the delivery time ($p < 0.001$). Potentially dangerous errors were identified in 4 out of 100 GPT responses, 7 Claude responses, and 10 Gemini responses. Among 700 medical decisions, 13 (1.9%) were classified as potentially dangerous The differences between the groups were statistically significant ($p = 0.006$) (Figure 2).



Fi)06).

Among the 21 potentially dangerous recommendations of LLM models, the most common were incorrect timing of delivery (n=8; 38.1%), omission of indications for magnesium sulfate (n=6; 28.6%), underestimation of the need for urgent antihypertensive therapy (n=4; 19.0%), and unjustified recommendation of outpatient follow-up (n=3; 14.3%),

The main problem of LLM is not so much in establishing a diagnosis, but in making individual clinical decisions after correctly recognizing diseases.

Discussion

In this comparative study, LLM models showed a relatively high ability to diagnose preeclampsia , but their effectiveness differed both between models and individual components of clinical solutions. GPT demonstrated diagnostic accuracy among the 91% models, while Claude and Gemini showed poor accuracy of 87% and 84%. Consistency with the expert standard has reached k=0.82.

At the same time, the average diagnostic accuracy of doctors was 94.3%, which is a high indicator among all the compared groups.

The findings are consistent with studies that conducted one of the most closely designed studies and compared ChatGPT, DeepSeek, and Gemini with a clinical assessment of exposed cases of suspected severe preeclampsia . In this study, ChatGPT showed high diagnostic effectiveness among the LLM tested [7].

The higher efficiency of GPT is partly due to differences in the design and structure of the sample. The study mainly considered cases of suspected severe preeclampsia , while our sample included the widest clinical spectrum. Our study assessed not only the presence of the disease, but also the sequence of further clinical decisions. A direct comparison of the absolute values of sensitivity and specificity between studies requires caution, and the similarity of the ranking of models enhances the reliability of the observed pattern [8-9].

Our results are consistent with studies where ChatGPT, Gemini DeepSeek evaluated five simulated obstetric cases, including preeclampsia . All models are able to correctly identify the key clinical problem. Scientists have found a noticeable variability in the quality of responses between the models. The plausible clinical reasoning of LLMs is not a sufficient basis for their independent application in obstetric practice. These data are consistent with our observation [10].

In the ChatGPT study, the clinical conclusion of doctors regarding preeclampsia was most closely reproduced, consistency remained moderate (k≈0.60), while sensitivity was below 85% [11].

When compared with the results of the study by Hirosawa et al. When evaluating the diagnostic and therapeutic capabilities of ChatGPT in obstetrics. The study included 30 clinical cases out of six. ChatGPT correctly identified the diagnosis and tactics in 27 out of 30 cases [12].

Larger studies have demonstrated that LLM accuracy is universal for all obstetric and gynecological tasks. In the PERFORM study, 8 LLMs were compared with 24 resident obstetrician-gynecologists in 60 standardized clinical scenarios, thus the diagnostic accuracy of LLMs was 73.75% (95% CI 69.64–77.49), while that of resident physicians was 65.35% (95% CI 62.85–67.76; P<0.001). which is consistent with our research [13].

Our results are complemented by data from Psilopatis et al., which evaluated ChatGPT-3.5, DeepSeek-V2, and Gemini 1.5 in five different simulated scenarios, with responses evaluated by four doctors on a five-point Global Quality Score scale. In the preeclampsia scenarios, the average GQS was 3 for ChatGPT, 4 for DeepSeek, and 3 for Gemini [14-15].

The data obtained also confirm the pronounced inter-model variability. In our study, Claude occupied an intermediate position between GPT and Gemini: the correct diagnosis was established in 87 out of 100 cases, severe symptoms were recognized in 25 out of 30 cases (83.3%), the correct decision on delivery was formed in 83 out of 100 cases, and its correct urgency was determined in 80 out of 100 cases. At the same time, potentially dangerous recommendations were found in 7 out of 100 responses from Claude. For Gemini, the corresponding indicators were lower: 84 correct diagnoses out of 100, 24 correctly recognized cases of severe preeclampsia out of 30, 79 correct decisions about delivery out of 100 and 76 correct decisions about its urgency; 10 out of 100 recommendations were considered potentially dangerous. Thus, a fairly consistent gradient of GPT > Claude > Gemini was observed at almost all major stages of the clinical decision [16-18].

In our study, this is especially evident when comparing GPT and Gemini. The difference in overall diagnostic accuracy was 7 percentage points — 91 correct classifications out of 100 for GPT versus 84 out of 100 for Gemini, while potentially dangerous recommendations were found in 4 and 10 cases, respectively. Consequently, the difference between the models concerned not only the formal diagnostic accuracy, but also the potential safety of clinical recommendations. This confirms the need to evaluate each LLM separately, rather than extrapolating the results of one model to the entire class of generative artificial intelligence [19-20].

Limitations

Our study has several limitations. Firstly, standardized clinical vignettes were used, which may not fully reproduce the dynamic process of medical care. Secondly, LLM performance refers to the models and versions used during the study period, which may change after updating the systems. Thirdly, the assessment by obstetrician-gynecologists is not an absolute standard of assessment.

Conclusion

GPT, Claude, and Gemini demonstrated high but unequal effectiveness in the diagnosis and management of preeclampsia . GPT showed the highest diagnostic accuracy among the LLMs studied and the highest consistency with the expert reference standard.

At the same time, all three models were inferior to doctors in recognizing individual severe conditions and making difficult decisions regarding treatment and delivery. Potentially dangerous errors were more common in LLM responses than in the decisions of obstetrician-gynecologists.

Thus, modern LLMs may be valuable as auxiliary tools to support clinical decisions in preeclampsia , but the results obtained do not support their autonomous use. The evaluation of medical LLMs should be based not only on general diagnostic accuracy, but also on an analysis of specific clinical decisions and potential harms.

Author contribution

Conceptualization: RA, MZ, AS, ADA, RT, and AAk; Investigation: RA, MZ, AS, ADA, and RT; **Supervision:** RA, MZ, and AS**; Methodology:** RA, MZ, AS, ADA, and RT; **Interpretation of data:** RA, MZ, AS, ADA, and RT; **Writing—review and editing:** RA, MZ, AS, ADA, and RT. All authors have read and approved the final version of the manuscript and accept responsibility for the integrity and accuracy of the work.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability: The de-identified data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to institutional approval and applicable ethical and confidentiality restrictions.

Conflict of interest: The authors declare that they have no conflicts of interest.

Disclaimer: The authors state that no part of this report is copied or published elsewhere in whole or in part.

References

1. Vera-Ponce VJ, Loayza-Castro JA, Ballena-Caicedo J, Valladolid-Sandoval LAM, Zuzunaga-Montoya FE et al. Global prevalence of preeclampsia , eclampsia, and HELLP syndrome: a systematic review and meta-analysis. *Front Reprod Health* 2025;7:1706009. https://doi.org/10.3389/frph.2025.1706009

2. Ngene NC, Moodley J. Preventing maternal morbidity and mortality from preeclampsia and eclampsia particularly in low- and middle-income countries. *Best Pract Res Clin Obstet Gynaecol* 2024 94:102473. https://doi.org/10.1016/j.bpobgyn.2024.102473

3. Sharma DD, Chandresh NR, Javed A, Girgis P, Zeeshan M et al. The Management of Preeclampsia : A Comprehensive Review of Current Practices and Future Directions. Cureus. 2024;16(1):e51512. https://doi.org/10.7759/cureus

4. Gestational Hypertension and Preeclampsia : ACOG Practice Bulletin, Number 222. *Obstet Gynecol* 2020;135(6):e237-e260. https://doi.org/10.1097/AOG.0000000000003891

5. 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 Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198

6. Namagembe I, Karavadra B, Kazibwe L, Rujumba J, Kiwanuka N et al. Implementation challenges in preeclampsia care: perspectives from health care professionals in urban Uganda. *AJOG Glob Rep* 2024;4(2):100348. https://doi.org/10.1016/j.xagr.2024.100348

7. Zou L, Zhang P, Jiang YQ, Wang XW, Yan XJ et al. ChatGPT-4o, Gemini Advanced and DeepSeek R1 in preoperative decision-making for thyroid surgery: a comparative assessment with human surgeons. *Front Oncol* 2025;15:1590230. https://doi.org/10.3389/fonc.2025.1590230

8. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y et al. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res* 2024;26:e60807. https://doi.org/10.2196/60807

9. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K et al. Evaluating ChatGPT-4's Accuracy in Identifying Final Diagnoses Within Differential Diagnoses Compared With Those of Physicians: Experimental Study for Diagnostic Cases. *JMIR Form Res* 2024;8:e59267. https://doi.org/10.2196/59267.

10. Del Monte F, Barolo R, Circhetta M, Delmonaco AG, Castagno E et al. Diagnostic efficacy of large language models in the pediatric emergency department: a pilot study. *Front Digit Health* 2025;7:1624786. https://doi.org/10.3389/fdgth.2025.1624786 .

11. Kutbi D, Abou-Bakr E, Haidar HM. Evaluating the Accuracy of Medical Information Generated by ChatGPT and Gemini and Its Alignment With International Clinical Guidelines From the Surviving Sepsis Campaign: Comparative Study. *JMIR Form Res* 2025;9:e84251. https://doi.org/10.2196/84251

12. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K et al. ChatGPT-Generated Differential Diagnosis Lists for Complex Case-Derived Clinical Vignettes: Diagnostic Accuracy Evaluation. *JMIR Med Inform* 2023;11:e48808. https://doi.org/10.2196/48808

13. Martinelli C, Giordano A, Carnevale V, Burk SR, Porto L et al. The PERFORM Study: Artificial Intelligence Versus Human Residents in Cross-Sectional Obstetrics-Gynecology Scenarios Across Languages and Time Constraints. *Mayo Clin Proc Digit Health* 2025;3(2):100206. https://doi.org/10.1016/j.mcpdig.2025.100206

14. Psilopatis I, Heindl F, Cupisti S, Fischer U, Kohlmann V et al. The role of artificial intelligence in gynecologic and obstetric emergencies. *Eur J Obstet Gynecol Reprod Biol* 2025 306:94-100. https://doi.org/10.1016/j.ejogrb.2025.01.007

15. Psilopatis I, Bader S, Krueckel A, Kehl S, Beckmann MW et al. Can Chat-GPT read and understand guidelines? An example using the S2k guideline intrauterine growth restriction of the German Society for Gynecology and Obstetrics. *Arch Gynecol Obstet* 2024;310(5):2425-2437. https://doi.org/10.1007/s00404-024-07667-z

16. West M, Alsaidi A, Siddiqi R, Sayyed F, Counts R et al. Artificial intelligence in obstetrics and gynecology: Evaluating ChatGPT and Google Gemini in answering patient questions. *Int J Gynaecol Obstet* 2026;173(1):328-335. https://doi.org/10.1002/ijgo.70622

17. Harada Y, Suzuki T, Harada T, Sakamoto T, Ishizuka K, et al. Performance evaluation of ChatGPT in detecting diagnostic errors and their contributing factors: an analysis of 545 case reports of diagnostic errors. *BMJ Open Qual* 2024;13(2):e002654. https://doi.org/10.1136/bmjoq-2023-002654

18. Arslan HÇ, Arslan K, Gün M, Karkın PO, Arslanoğlu T. Evaluation of ChatGPT-5 responses in obstetric and gynecological emergencies: concordance, readability, and clinical reliability. *BMC Emerg Med* 2025;25(1):220. https://doi.org/10.1186/s12873-025-01396-5

19. Omar M, Sorin V, Collins JD, Reich D, Freeman R, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun Med (Lond)* 2025;5(1):330. https://doi.org/10.1038/s43856-025-01021-3

20. Draelos RL, Afreen S, Blasko B, Brazile TL, Chase N, et al. Large language models provide unsafe answers to patient-posed medical questions. *NPJ Digit Med* 2026;9(1):241. https://doi.org/10.1038/s41746-026-02428-5