Computational and Structural Biotechnology Journal 27 (2025) 1140–1147

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

Artificial intelligence-large language models (AI-LLMs) for reliable and accurate cardiotocography (CTG) interpretation in obstetric practice

Khanisyah Erza Gumilar [a,b,*], Manggala Pasca Wardhana [c], Muhammad Ilham Aldika Akbar [b], Agung Sunarko Putra [d], Dharma Putra Perjuangan Banjarnahor [e], Ryan Saktika Mulyana [f], Ita Fatati [g], Zih-Ying Yu [h], Yu-Cheng Hsu [h,i], Erry Gumilar Dachlan [b], Chien-Hsing Lu [j], Li-Na Liao [h,**], Ming Tan [a,k,***]

[a] Graduate Institute of Biomedical Science, China Medical University, Taichung, Taiwan
[b] Department of Obstetrics and Gynecology, Universitas Airlangga Hospital - Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
[c] Department of Obstetrics and Gynecology, Dr. Soetomo General Hospital - Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
[d] Department of Obstetrics and Gynecology, Dr. Ramelan Naval Hospital, Surabaya, Indonesia
[e] Department of Obstetrics and Gynecology, Dr. Mohamad Soewandhie Hospital, Surabaya, Indonesia
[f] Department of Obstetrics and Gynecology, Department of Obstetrics and Gynecology, Udayana University Hospital, Denpasar, Indonesia
[g] Department of Obstetrics and Gynecology, Bandung Kiwari General Hospital, Bandung, Indonesia
[h] Department of Public Health, China Medical University, Taichung, Taiwan
[i] School of Chinese Medicine, China Medical University, Taichung, Taiwan
[j] Department of Obstetrics and Gynecology, Taichung Veteran General Hospital, Taichung, Taiwan
[k] Institute of Biochemistry and Molecular Biology and Research Center for Cancer Biology, China Medical University, Taichung, Taiwan

ARTICLE INFO

Keywords:
Cardiotocography (CTG)
Artificial intelligence-large language models (AI-LLMs)
ChatGPT
Gemini
Copilot
Fetal monitoring
Obstetrics

ABSTRACT

Background: Accurate cardiotocography (CTG) interpretation is vital for the monitoring of fetal well-being during pregnancy and labor. Advanced artificial intelligence (AI) tools such as AI-large language models (AI-LLMs) may enhance the accuracy of CTG interpretation, but their potential has not been extensively evaluated.

Objective: This study aimed to assess the performance of three AI-LLMs (ChatGPT-4o, Gemini Advanced, and Copilot) in CTG image interpretation, compare their results to those of junior (JHDs) and senior human doctors (SHDs), and evaluate their reliability in clinical decision-making.

Study design: Seven CTG images were interpreted by the three AI-LLMs, five SHDs, and five JHDs, with the evaluations scored by five blinded maternal-fetal medicine experts using a Likert scale for five parameters (relevance, clarity, depth, focus, and coherence). The homogeneity of the expert ratings and group performances were statistically compared.

Results: ChatGPT-4o scored 77.86, outperforming the Gemini Advanced (57.14), Copilot (47.29), and JHDs (61.57). Its performance closely approached that of the SHDs (80.43), with no statistically significant difference between the two ($p > 0.05$). ChatGPT-4o excelled in the depth parameter and was only marginally inferior to the SHDs regarding the other parameters.

Conclusion: ChatGPT-4o demonstrated superior performance among the AI-LLMs, surpassed JHDs in CTG interpretation, and closely matched the performance level of SHDs. AI-LLMs, particularly ChatGPT-4o, are promising tools for assisting obstetricians, improving diagnostic accuracy, and enhancing obstetric patient care.

* Correspondence to: Department of Obstetrics and Gynecology, Hospital of Universitas Airlangga Faculty of Medicine, Universitas Airlangga, Jl. Dharmahusada Permai, Mulyorejo, Surabaya, Jawa Timur 60115, Indonesia.
** Correspondence to: Department of Public Health, China Medical University, No. 100, Section 1, Jingmao Rd, Beitun Dist, Taichung City 406040, Taiwan.
*** Correspondence to: Institute of Biochemistry and Molecular Biology, Graduate Institute of Biomedical Sciences, China Medical University (Taiwan), No. 100, Section 1, Jingmao Rd, Beitun Dist, Taichung City, 406040, Taiwan.
E-mail addresses: khanisyah@fk.unair.ac.id (K.E. Gumilar), linaliao@mail.cmu.edu.tw (L.-N. Liao), mingtan@mail.cmu.edu.tw (M. Tan).

https://doi.org/10.1016/j.csbj.2025.03.026
Received 16 December 2024; Received in revised form 5 March 2025; Accepted 13 March 2025
Available online 18 March 2025

1. Introduction

Cardiotocography (CTG) is a vital monitoring method used in modern obstetrics to assess fetal condition during labor [1]. It monitors fetal heart rate and uterine contractions simultaneously, providing crucial information regarding fetal well-being. It serves as a valuable monitoring tool during pregnancy and labor because it can detect fetal heart rate deceleration [2], hypoxia [3], and excessive contractions [4], all of which affect the safety of the mother and baby. CTG can reduce labor complications, morbidity, and infant mortality [5]. Although its use has become standard, interpreting CTG results requires specialized training, experience, and expertise. A mistaken interpretation or delayed response can significantly impact both mother and fetus. Therefore, there is an urgent need for technological enhancement of the accuracy

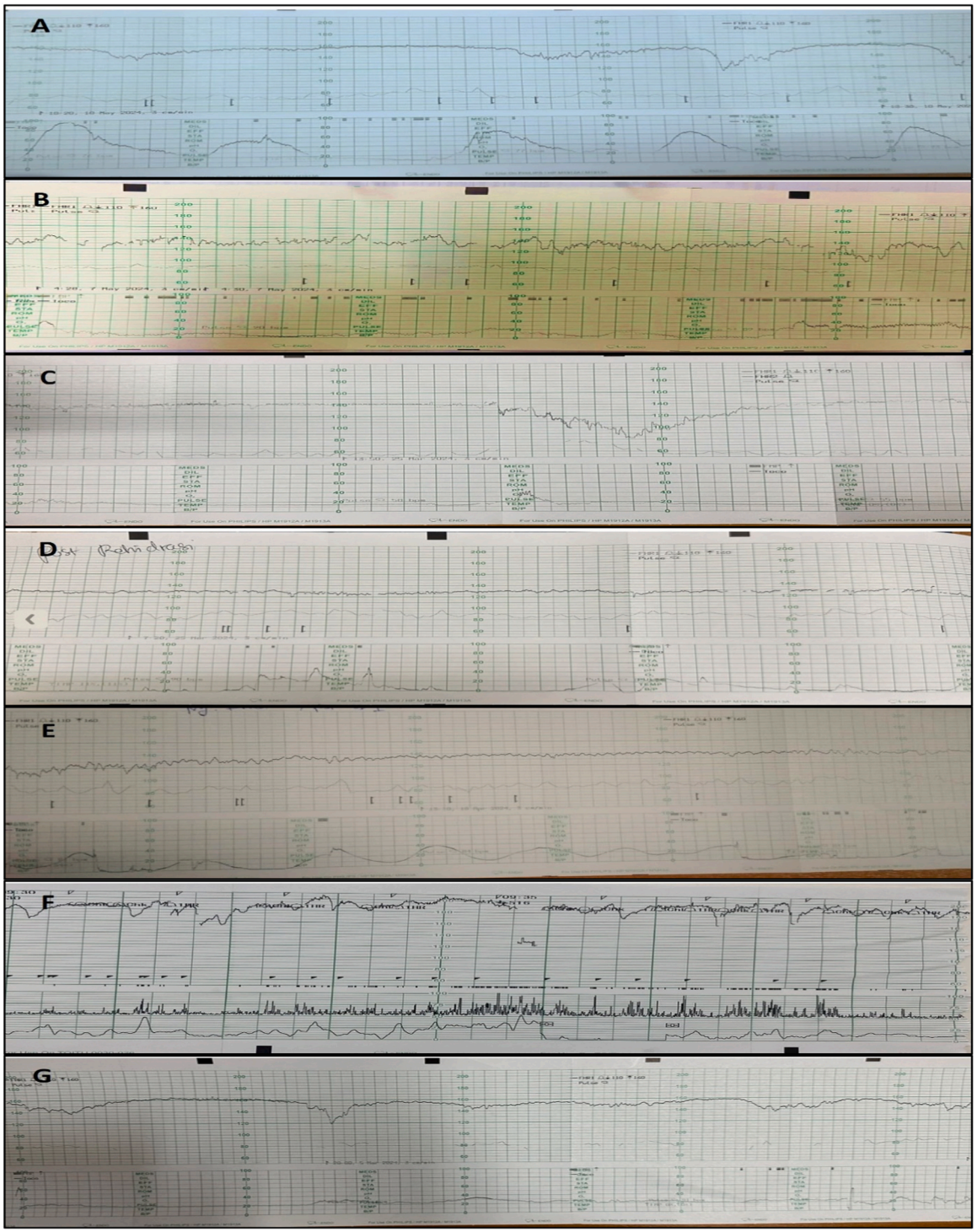

Fig. 1. CTG images used in this study. A. A 37-week pregnancy with preeclampsia and gestational diabetes undergoing misoprostol induction showing regular uterine contractions; B. A 38-week pregnancy with PROM without accompanying contractions; C. First stage of labor in the active phase experiencing secondary arrest and decelerations; D. CTG condition after intrauterine resuscitation showing low variability; E. Post-due date pregnancy with PROM undergoing induction. The induction was eventually stopped due to meconium-stained amniotic fluid; F. A pregnant patient with severe symptomatic pneumonia and fever for 3 days before arriving at the hospital; and G. Misoprostol induction in a post-date pregnancy at 41 weeks. After one series of administration, there was no progress in labor.

and efficiency of CTG interpretation.

Artificial intelligence-large language model (AI-LLM) development in recent years has opened new possibilities in various fields, including medicine. AI-LLMs can quickly and accurately analyze large volumes of data and learn from patterns and trends that humans may not easily detect [6]. In the context of CTG, AI-LLMs can provide rapid and accurate initial analyses, and assist medical professionals in making timely and data-driven decisions.

The use of AI-LLMs in CTG interpretation offers several advantages. The models can be trained on large historical CTG datasets, enabling them to accurately identify the patterns associated with specific risks. They are accessible any time and place, and provide consistent support without the fatigue or biases that can affect human interpretation. AI-LLMs can also learn and improve with continuous increases in data availability [7]. With new interactions and interpretations, the algorithms of these models can be refined, with increasing reliability and efficacy over time. Furthermore, integrating AI-LLMs into healthcare systems can reduce the workload of the medical staff, allowing them to focus on other aspects of patient care.

In an increasingly digital world, implementing AI-LLMs in CTG interpretation can represent a significant innovation in obstetric and childbirth practices by providing fast, accurate, and consistent analyses, thereby improving the safety of mothers and babies during labor. This study aimed to assess the accuracy of three different AI-LLMs at interpreting CTG images representing diverse patient conditions, and to determine these models are reliable for assisting doctors. By combining medical expertise with advanced technology, we aimed to develop improved and more efficient solutions to ensure safe and optimal childbirth.

2. Materials and methods

2.1. Materials

We investigated three AI-LLMs in this study: ChatGPT-4o (https://chatgpt.com; referred to as CG4o), Gemini Advanced 1.5 Pro (https://gemini.google.com/app; referred to as Gem Adv), and Copilot (https://copilot.microsoft.com/). The examination was conducted on August 17, 2024, within 24 h for each model. To ensure consistency, all models were tested using uniform default settings, with the temperature parameter set to 1 and top_p ranging from 0.9 to 1. A new task was initiated for each image input and new command.

The CTG images used in this study were obtained from the Department of Obstetrics and Gynecology, Universitas Airlangga Hospital, Surabaya, Indonesia (ethical approval from Universitas Airlangga Hospital no.156/KEP/2024; protocol number UA-02-24216). CTG-1 (Fig. 1A) was obtained from a 37-week pregnant patient with preeclampsia and gestational diabetes undergoing induction with misoprostol. CTG-2 (Fig. 1B) is related to a 38-week pregnant patient with a history of premature rupture of membranes (PROM). CTG-3 (Fig. 1C) was obtained from a 37-week and 5-day pregnant patient in active labor with observed decelerations. CTG-4 (Fig. 1D) was derived from a 39-week pregnant patient who underwent intrauterine resuscitation after a previous CTG indicated category-2 or suspicious nonstress test result. CTG-5 (Fig. 1E) was obtained from a 41-week pregnant patient with PROM who underwent induction. CTG-6 (Fig. 1F) related to a 37-week pregnant patient who presented with a 3-day history of fever and was diagnosed with a severe respiratory tract infection. CTG-7 (Fig. 1G) was obtained from a 41-week pregnant patient undergoing misoprostol induction.

We recruited 10 obstetricians as participants for the comparative assessment. Five physicians with over 10 years of experience in obstetrics were categorized as senior human doctors (SHDs), while the remaining five physicians with under 5 years of obstetrics experience were classed as junior human doctors (JHDs). Both groups of doctors were tested using the same CTG images, with an interpretation time limit of 10 minutes. Additionally, we enlisted five maternal-fetal medicine (MFM) specialists to evaluate all interpretations derived from the three AI-LLMs and two groups of doctors.

2.2. Study design

We evaluated the performance of the three AI-LLMs (referred to as models): CG4o, GemAdv, and Copilot, when presented with the seven CTG images of the different patients. Both groups of doctors (SHDs and JHDs) were tested using the same questions as those given to the AI-LLMs. The CTG images were presented to each model along with a specific prompt, and the interpretations were recorded (Fig. 2). For instance, we provided the case scenario and CTG image, followed by a prompt sentence requesting interpretation and analysis. The complete case scenarios and prompts used are provided in Supplementary Material (Supp.1). These interpretations were reviewed by a team of five MFM (referred to as raters). To ensure impartiality, the responses from the models were coded and randomized before being blindly assessed by the raters. To analyze the model outputs, we used assessed five parameters: relevance, clarity, depth focus, and coherence (Table 1) [8–12], using a five-point Likert scale (Table 2)[13,14].

2.3. Statistical analysis

The performances of the three models and human doctors in the interpretation of the seven CTG images were evaluated. Five raters evaluated all of the responses generated by the three models and human doctors using five parameters: relevance, clarity, depth, focus, and coherence. A five-point Likert scale was employed for scoring, with a score of 5 indicating superior performance. Subsequently, the scores were linearly converted to a 0–100 scale based on a previous study [15]. To ascertain the extent of inter-rater consistency (homogeneity), Pearson and Spearman correlation coefficients were employed. A one-way analysis of variance (ANOVA) with Scheffé's post-hoc analysis was used to investigate differences in mean scores among raters. A multiple linear regression model was constructed to evaluate the performance of both the models and human doctors. This accounted for the potential influence of subjectivity among raters and the varying degrees of complexity observed among image sets. All statistical analyses were conducted using SAS software (version 9.4; SAS Institute, Cary, NC, USA), with the significance level set at 0.05.

3. Results

3.1. Moderate to high homogeneity among rater scores

To assess the consistency of CTG interpretation response rater scores, we utilized a heat map to illustrate the homogeneity index among the five MFM experts, and analyzed the results using Pearson (Fig. 3A) and Spearman correlation coefficients (Fig. 3B). Overall, the raters exhibited a relatively high level of consensus in assessing responses, with the homogeneity index ranging from moderate to high. Specifically, Pearson and Spearman correlation coefficients ranged from 0.54 to 0.91 and from 0.36 to 0.94, respectively.

To further validate the reliability of the scoring process, a one-way ANOVA test with Scheffé's post-hoc analysis was used to examine the total scores for all responses (Fig. 3C). The variation among raters was not significant ($p > 0.05$). Collectively, this indicates that the raters had consistent views regarding CTG interpretation responses, suggesting that the evaluated results were dependable.

3.2. AI-LLMs showed variable performances in CTG image interpretation

CG4o showed the best performance with a score of 77.86 (Fig. 4A), compared to GemAdv and Copilot with scores of 57.14 and 47.29, respectively. These results indicate significant variations in CTG

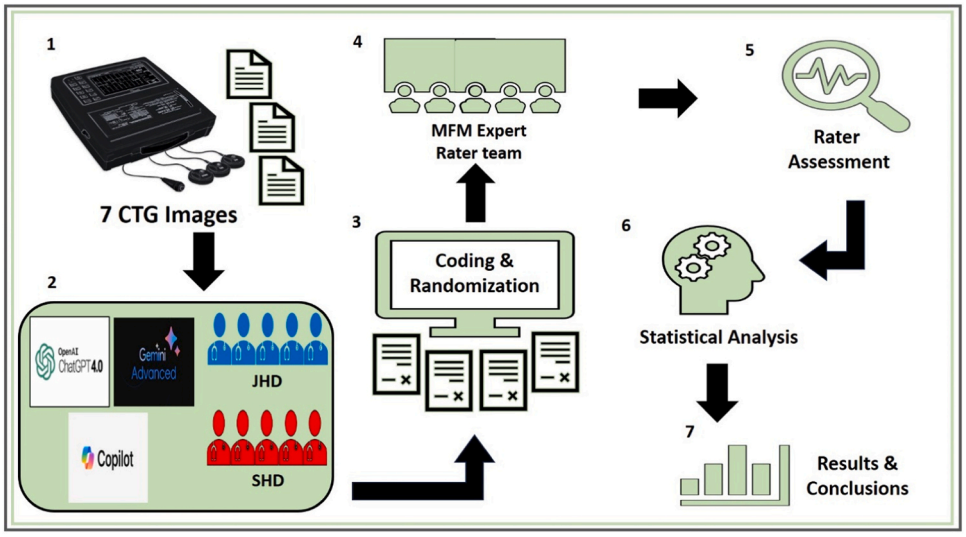

Fig. 2. Study design. Seven CTG images representing diverse labor cases (1) were tested on three groups (AI-LLMs, JHDs, and SHDs) (2). The interpretations from these three groups were assigned specific codes and randomized (3) for subsequent evaluation by MFM experts (raters) (4). The raters assessed the CTG interpretations from various entities using agreed-upon parameters (5). The results were statistically evaluated and compared (6), and presented as figures and conclusions (7).

Table 1
Assessment parameters.

Definition & Parameters
Relevance: The response is closely related or appropriate to the issue
Clarity: Clear, easy to understand, and free from ambiguity
Depth: The answer provides detailed and specific information, not just a general or superficial answer
Focus: Contains the main points or keywords expected
Coherence: All parts of the answer work together in a logical and structured way, with no conflicting parts

Table 2
Likert scale rating and descriptions.

Score	Rating	Description
1	Very Poor	The performance is unsatisfactory and unacceptable.
2	Poor	The performance does not meet established standards and reflects a need for significant improvement.
3	Average	The performance meets the minimum standard; however, it falls short of excellence and lacks any notable distinction.
4	Above Average	The performance is good, consistently meeting the expected standard with reliable quality and efficiency.
5	Outstanding	The performance goes beyond expectations, showing exceptional quality, creativity, or innovation.

interpretation performance among different AI-LLM systems, with CG4o showing a superior performance.

In the evaluation of human doctor performance, SHDs achieved the highest score of 80.43, while JHDs scored 61.57. JHDs scored lower than SHDs and CG4o, but still outperformed GemAdv and Copilot. Of note, CG4o performance (77.86) was only slightly lower than that of the SHDs (80.43), with no statistically significant difference between the two, highlighting the success of CG4o in CTG interpretation.

There was a statistically significant difference ($p < 0.001$) in performance between SHDs, and JHDs, Copilot, and GemAdv, further indicating that SHDs possess significantly superior expertise and experience in CTG interpretation. Among AI-LLM systems, CG4o also showed a statistically significant difference compared to GemAdv and Copilot ($p < 0.001$), reaffirming its superiority. JHDs showed a significant difference ($p < 0.001$) in performance compared to Copilot, suggesting that human expertise, even at the junior level, still surpasses that of some AI models in complex clinical tasks. The responses from the models and human doctors were

further evaluated by five expert raters who used a five-point Likert scale for assessing relevance, clarity, depth, focus, and coherence (Fig. 4B). The SHDs consistently ranked the highest across all five parameters. This indicates that they provide more comprehensive CTG interpretations than AI-LLMs and JHDs. JHDs demonstrated a competitive performance across most parameters. Although they did not achieve the same level as SHDs, they consistently performed better than GemAdv and Copilot, particularly in terms of coherence and clarity.

When comparing all three AI-LLM models with human doctors, statistically significant differences ($p < 0.001$) were observed in the pentagram (Fig. 4B), suggesting a clear distinction between the tested groups. CG4o excelled in all assessment parameters compared to GemAdv and Copilot, and even surpassed SHDs regarding the depth parameter.

4. Discussion

To the best of our knowledge, this is the first study to assess the CTG interpretation ability of different AI-LLMs. CTG is a crucial method for monitoring fetal condition during pregnancy and labor. It helps prevent complications that are potentially harmful to both the mother and fetus, and can be used to detect conditions such as hypoxia, fetal heart rate deceleration, and uterine rupture. However, interpreting CTG requires the specialized expertise and experience of a physician.

The participation of two groups of doctors (SHDs and JHDs) in the present study allowed us to examine the performance of AI-LLMs compared to human doctors. The CTG image interpretations provided by the AI-LLMs and human doctors were evaluated by five obstetric experts. To ensure inter-rater consistency, three statistical analyses were performed to measure the homogeneity of the assessments (Fig. 3A–C).

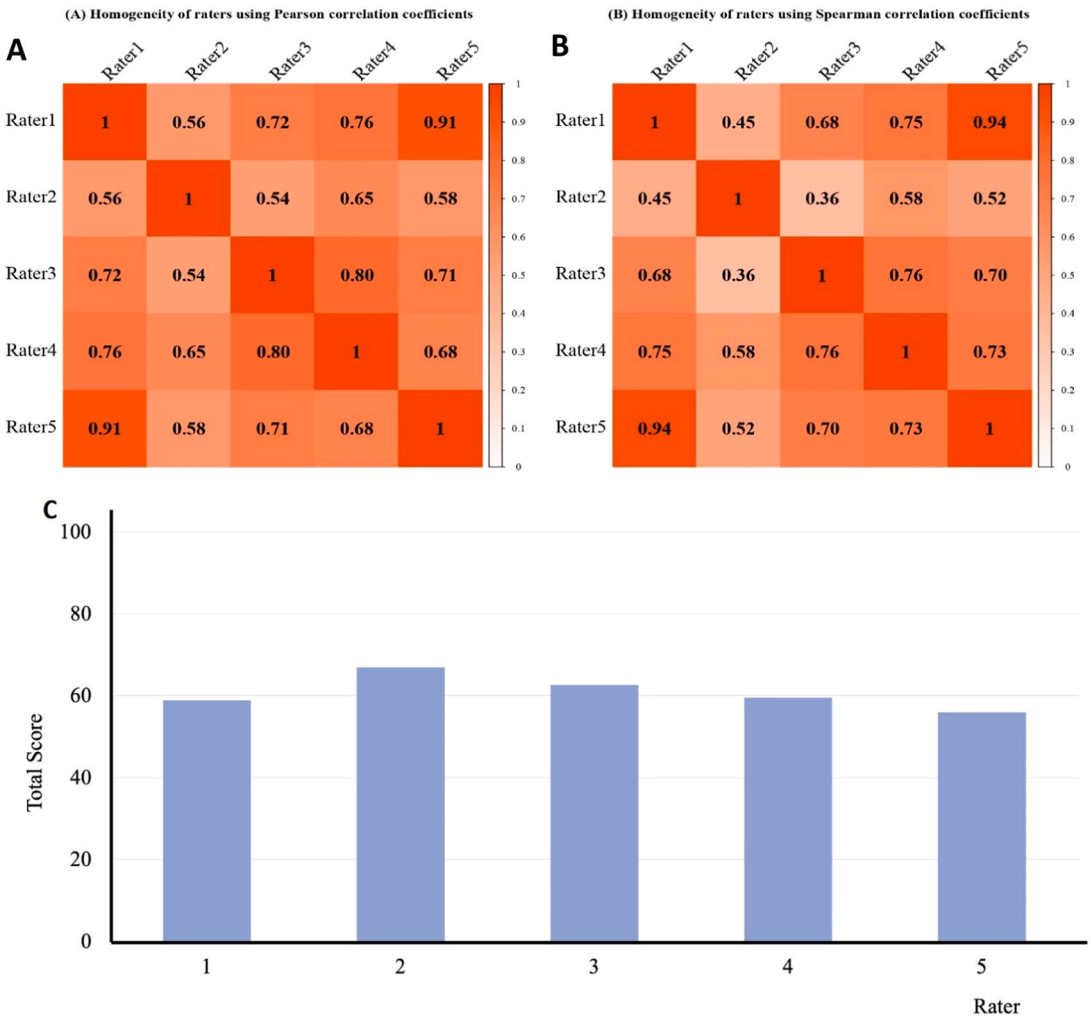

Fig. 3. Homogeneity scores among raters. Pearson correlation coefficients test (A) and Spearman correlation coefficients (B) revealing moderate to high homogeneity (0.54–0.91 and 0.36–0.94, respectively). One-way ANOVA test with Scheffe's post-hoc analysis showed that there was no significant difference among the five raters (C).

There was a moderate to high homogeneity index among raters, indicating high agreement in their perception of the quality of AI-LLM CTG interpretations. Although there were some variations in their ratings, they were not statistically significant (Fig. 3C), suggesting that their judgments were objective and consistent. Moreover, to ensure impartiality, the model responses were coded, randomized, and blindly assessed by the raters. This further enhanced the confidence in the reliability of the AI-LLM evaluations, indicating that the assessments were not significantly influenced by potential rater bias.

We evaluated the three AI-LLMs based on five key parameters: relevance, clarity, depth, focus, and coherence. The CG4o outperformed GemAdv and Copilot in CTG image interpretation (Fig. 4A). Among the three models, CG4o consistently achieved the highest scores across all evaluation criteria, indicating its superior accuracy and reliability in CTG analysis (Fig. 4B). These findings suggest that CG4o provides the most consistent and effective performance compared with the other models assessed.

Of note, CG4o performed better than JHDs and closely approached the performance ability of SHDs, with no significant differences observed between CG4o and SHDs. These results indicate that CG4o has significant potential for CGT interpretation. However, the expert raters agreed that SHDs remained superior to the AI-LMMs in terms of relevance, coherence, and focus. This emphasizes the need for AI-LLM systems to be validated and supervised by CTG image interpreters.

The inferior performances of GemAdv and Copilot also highlight the need for further improvement in developing algorithms capable of achieving an accuracy comparable to that of human doctors, particularly in highly complex contexts, such as maternal health. However, CG4o can be used to assist less experienced doctors in CTG interpretation. Improved diagnostic accuracy based on CTG results will ultimately lead to better patient care, while minimizing treatment errors and complications. Given the performance of CG4o, which was almost as good as that of SHDs, and even exceed them regarding the depth parameter, there are numerous prospects for its use in patient care.

The application of AI-LLMs in CTG interpretation is a novel area of study that has not yet been explored. Machine learning and deep learning have been utilized as supporting tools for CTG analysis, significantly enhancing fetal monitoring by improving accuracy and clinical applicability. For instance, a study employing recursive feature elimination and Bayesian optimization for CTG classification achieved a 97.20 % accuracy rate [16]. Similarly, an AI-driven framework, the Tabular Prior Data Fitted Network, demonstrated a 97.91 % accuracy rate with a low computational complexity, making it highly suitable for real-time clinical applications [17]. Additionally, convolutional neural networks shave shown robust performance in detecting late deceleration and fetal distress in CTG [18]. These advancements in AI-driven fetal monitoring have significantly improved obstetric decision making and maternal-fetal outcomes. However, the models are limited in terms of

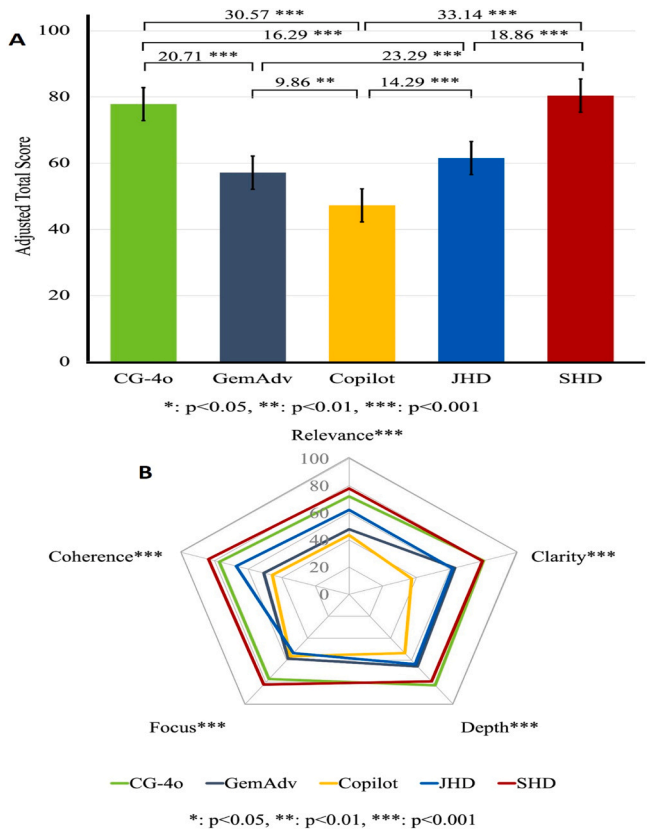

Fig. 4. Ability of AI-LLMs and human doctors to interpret CTG images. (A) CG4o scores significantly better in CTG interpretation than the other two AI-LLMs. (B) CG4o shows significant superiority in the five parameters (relevance, clarity, depth, focus, and coherence) used for interpretation of quality assessment.

text comprehension and natural language processing. In contrast, AI-LLMs encompass a broader range of capabilities and leverage vast datasets to understand, generate, and predict text with greater accuracy [19].

In the present study, AI-LLMs demonstrated the ability to interpret CTG findings in patients undergoing labor. The consistently high performance of CG4o across the five key evaluation parameters suggests that it can assist clinicians in making better-informed decisions, potentially reducing diagnostic errors and improving patient outcomes. Because CTG interpretation requires the specialized expertise and experience of a physician, AI-LLM utilization is particularly valuable in settings where expert interpretation may not be readily available, thereby ensuring a high standard of care across different healthcare environments. The advantage of AI-LLMs over machine learning and deep learning is their ability to provide explanations in easily understandable languages. AI-LLMs can also enable two-way communication and provide healthcare professionals with patient management recommendations.

Beyond CTG analysis, the broader application of AI-LLMs in the medical field has shown varying levels of proficiency across key areas such as answer accuracy, clinical interpretation, reference generation, and image analysis [20–23]. Although previous research has highlighted their high accuracy in responding to medical queries, integrating these models into clinical workflows remains a challenge. Studies indicate that AI-LLMs often struggle to strictly follow explicit instructions, and are highly sensitive to both the amount and sequence of input data [24]. These limitations hinder their seamless deployment in structured medical environments, necessitating further advancements in terms of model robustness and contextual adaptability.

The accuracy of LLMs in interpreting complex medical information varies according to the specialization. Models specifically designed for medical applications tend to outperform general-purpose AI-LLMs in understanding clinical details [25]. However, one persistent challenge is the generation of reliable references because some models produce incorrect or unverifiable sources [26]. Specialized medical AI-LLMs mitigate this issue by leveraging the peer-reviewed scientific literature to enhance credibility. In medical imaging, AI models have demonstrated strong capabilities for analyzing modalities such as magnetic resonance imaging and computed tomography [27]. Nevertheless, rigorous validation is required to ensure that their reliability matches that of radiologists. Although AI-LLMs have potential for supporting medical professionals, their widespread adoption depends on continuous advancements in accuracy, interpretability, and reliability.

In summary, CG4o demonstrated outstanding performance in CGT interpretation, making it an excellent tool for supporting human doctors in this task. Its accuracy was comparable to that of SHDs, and it significantly outperformed the other models. However, it is important to emphasize that the role of the doctor as a validator and supervisor remains crucial because the AI-LLM does not operate autonomously and may provide imperfect answers.

5. Study limitations

Despite these promising results, this study has several limitations. First, the evaluation was conducted using a relatively small sample size of CTG images. This may limit the generalizability of the findings because AI-LLM may perform differently when exposed to more extensive and diverse data. Additionally, potential bias may arise from the selection of evaluation parameters. While the five parameters of relevance, clarity, depth, focus, and coherence were carefully designed and tested in previous studies [12,15,28], they may not encompass all the critical aspects of effective CTG interpretation. Notably, this study did not assess the ability of AI-LLM to detect rare but clinically significant patterns, or its adaptability to evolving clinical guidelines.

Another important consideration is the evolving nature of AI-LLM, the performance of which may fluctuate owing to updates, retraining, or algorithmic shifts, posing challenges to their consistent reliability in medical applications. Furthermore, the lack of transparency regarding the training data raises concerns about potential data leakage, and limits the ability to comprehensively assess biases or knowledge gaps. These factors underscore the need for specialized models that are fine-tuned for medical tasks, such as MEDGemini, which may be better suited to addressing domain-specific challenges than general-purpose models such as Gemini. Future research should prioritize the development and validation of tailored models to enhance the accuracy, interpretability, and clinical applicability of AI-driven CTG analyses.

6. Conclusion

This study evaluated the performance of three AI-LLMs (CG4o, GemAdv, and Copilot) in CTG image interpretation, and compared them with those of SHDs and JHDs. CG4o outperformed the other AI models, closely approaching the performance of SHDs, and surpassing it with regards to the depth parameter. However, SHDs still exhibited the best performance in terms of relevance, coherence, and focus, emphasizing that human expertise remains crucial. These findings suggest that AI-LLMs, especially CG4o, have the potential to assist doctors, particularly less experienced doctors, in CTG interpretation. However, their use should be supervised by specialists to ensure patient safety and accuracy. This study highlights the potential of AI-LLM utilization for the enhancement of obstetric care quality, but also underscores the need for further research to ensure the reliability of these models in more complex clinical scenarios. In conclusion, the integration of advanced AI-LLMs such as CG4o into clinical practice can yield substantial benefits in the field of obstetrics, especially fetal monitoring in labor. The

accurate and consistent application of AI-LLMs has the potential to improve clinical outcomes and operational efficiency, thereby improving patient outcomes, and avoiding morbidity and mortality.

Funding

This research was partly funded by China Medical University Ying-Tsai Scholar Fund CMU109-YT-04, CMU Internal Funds (CMU112-IP-01 and CMU113-MF-56), and a NSTC grant 113-2314-B-039-067 (to MT). KEG is a recipient of an Elite Program Scholarship from the Taiwan Ministry of Education.

CRediT authorship contribution statement

Erry Gumilar Dachlan: Writing – review & editing, Validation. **Yu-Cheng Hsu:** Software, Formal analysis. **Zih-Ying Yu:** Software, Formal analysis. **Ita Fatati:** Validation, Investigation. **Muhammad Ilham Aldika Akbar:** Validation, Investigation. **Ming Tan:** Writing – review & editing, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization. **Manggala Pasca Wardhana:** Validation, Investigation. **Li-Na Liao:** Writing – review & editing, Visualization, Software, Formal analysis, Conceptualization. **Khanisyah Erza Gumilar:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Chien-Hsing Lu:** Writing – review & editing. **Ryan Saktika Mulyana:** Validation, Investigation. **Dharma Putra Perjuangan Banjarnahor:** Validation, Investigation. **Agung Sunarko Putra:** Validation, Investigation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this study, we used ChatGPT-4o and Grammarly to edit and proofread the manuscript to improve its readability. After using these tools and services, the authors reviewed, verified, and edited the content as required. The authors take full responsibility for the contents of this manuscript.

Acknowledgments

The authors are grateful for our collaborators, Ach S. Faridzi, MD (Dr. Mohammad Zyn General Hospital, Sampang, Indonesia), Gallaran Matu, MD (Dr. Soemarno Sosroatmodjo General Hospital, Bulungan, Indonesia), Nurlaella I. Nusi, MD (Dr. Mohamad Soewandhie General Hospital, Surabaya, Indonesia), Roziana, MD (Dr. Zainoel Abidin General Hospital, Banda Aceh, Indonesia), and Yuni Setiawaty, MD (Dr. Wahidin Sudiro Husodo General Hospital, Mojokerto, Indonesia). We also thank Bayu A. S. Putra, MD, Prita A. Malinda, MD, Ni Nyoman A. R. Pradnyani, MD, Rizqy Rahmatyah, MD, and Royan M. Varendra, MD (Department of Obstetrics and Gynecology, Faculty of Medicine, Airlangga University, Surabaya, Indonesia).

Competing interests

None of the authors declare competing interests.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.03.026.

Data availability

We have ensured that all data required are included in the Supplementary file. Exceptions are raw values provided by individual doctors, which can be provided upon request.

References

[1] Grivell RM, Alfirevic Z, Gyte GM, Devane D. Antenatal cardiotocography for fetal assessment. Cochrane Database Syst Rev 2015;2015(9):Cd007863.
[2] Caning MM, Thisted DLA, Amer-Wählin I, Laier GH, Krebs L. Interobserver agreement in analysis of cardiotocograms recorded during trial of labor after cesarean. J Matern Fetal Neonatal Med 2019;32(22):3778–83.
[3] Cömert Z, Şengür A, Budak Ü, Kocamaz AF. Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models. Health Inf Sci Syst 2019;7(1):17.
[4] Tarvonen M, Sainio S, Hämäläinen E, Hiilesmaa V, Andersson S, Teramo K. Saltatory pattern of fetal heart rate during labor is a sign of fetal hypoxia. Neonatology 2020;117(1):111–7.
[5] Ranaei-Zamani N, David AL, Siassakos D, Dadhwal V, Aughwane R, Russell-Buckland J, et al. Saving babies and families from preventable harm: a review of the current state of fetoplacental monitoring and emerging opportunities. npj Women's Health 2024;2(10).
[6] Hartka T. The American Journal of Emergency Medicine's policy on large language model usage in manuscript preparation: balancing innovation and responsibility. Am J Emerg Med 2024;82:105–6.
[7] Rios-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. Front Med (Lausanne) 2024;11:1380148.
[8] Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. J Am Coll Radiol 2023.
[9] Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023;307(5):e230922.
[10] Wu T, He S, Liu J, Sun S, Liu K, Han Q-L, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Autom Sin 2023;10(5):1122–36.
[11] Bhardwaz S, Kumar J. An extensive comparative analysis of Chatbot Technologies - ChatGPT, Google BARD and Microsoft Bing. 2023 2nd Int Conf Appl Artif Intell Comput (ICAAIC) 2023:673–9.
[12] Gumilar KE, Indraprasta BR, Hsu Y-C, Yu Z-Y, Chen H, Irawan B, et al. Disparities in medical recommendations from AI-based chatbots across different countries/regions. Sci Rep 2024;14(1).
[13] Sikander B, Baker JJ, Deveci CD, Lund L, Rosenberg J. ChatGPT-4 and human researchers are equal in writing scientific introduction sections: a blinded, randomized, non-inferiority controlled study. Cureus 2023;15(11):e49019.
[14] Veras M, Dyer JO, Rooney M, Barros Silva PG, Rutherford D, Kairy D. Usability and efficacy of artificial intelligence chatbots (ChatGPT) for health sciences students: protocol for a crossover randomized controlled trial. JMIR Res Protoc 2023;12:e51873.
[15] Gumilar KE, Indraprasta BR, Faridzi AS, Wibowo BM, Herlambang A, Rahestyningtyas E, et al. Assessment of large language models (LLMs) in decision-making support for gynecologic oncology. Comput Struct Biotechnol J 2024.
[16] Hardalaç F, Akmal H, Ayturan K, Acharya UR, Tan R-S. A pragmatic approach to fetal monitoring via cardiotocography using feature elimination and hyperparameter optimization. Interdiscip Sci: Comput Life Sci 2024;16(4):882–906.
[17] Alzakari SA, Aldrees A, Umer M, Cascone L, Innab N, Ashraf I. Artificial intelligence-driven predictive framework for early detection of still birth. SLAS Technol 2024;29(6):100203.
[18] Sato I, Hirono Y, Shima E, Yamamoto H, Yoshihara K, Kai C, et al. Comparison and verification of detection accuracy for late deceleration with and without uterine contractions signals using convolutional neural networks. Front Physiol 2025;16:1525266.
[19] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689.
[20] Akyon SH, Akyon FC, Camyar AS, Hızlı F, Sari T, Hızlı Ş. Evaluating the capabilities of generative AI tools in understanding medical papers: qualitative study. JMIR Med Inf 2024;12:e59258.
[21] Menz BD, Kuderer NM, Bacchi S, Modi ND, Chin-Yee B, Hu T, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. Bmj 2024;384:e078538.
[22] Rydzewski NR, Dinakaran D, Zhao SG, Ruppin E, Turkbey B, Citrin DE, et al. Comparative evaluation of LLMs in clinical oncology. Nejm ai 2024;1(5).
[23] Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy course: Comparative analysis. Clin Anat 2025;38(2):200–10.
[24] Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024;30(9):2613–22.
[25] He K, Mao R, Lin Q, Ruan Y, Lan X, Feng M, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Inf Fusion 2025;118.
[26] Fleurence RL, Bian J, Wang X, Xu H, Dawoud D, Higashi M, et al. Generative artificial intelligence for health technology assessment: opportunities, challenges,

and policy considerations: an ISPOR working group report. Value Health 2025;28 (2):175–83.
[27] Paudyal R, Shah AD, Akin O, Do RKG, Konar AS, Hatzoglou V, et al. Artificial intelligence in CT and MR imaging for oncological applications. Cancers (Basel) 2023;15(9).
[28] Gumilar KE, Ariani G, Wiratama PA, Rimbun R, Yuliawati TH, Chen H, et al. Assess the capabilities of AI-based large language models (AI-LLMs) in interpreting histopathological slides and scientific figures: performance evaluation study. JMIR Prepr 2024.