Clinical and Experimental Obstetrics & Gynecology

Clin. Exp. Obstet. Gynecol. **2026**; ():
https://doi.org/10.31083/CEOG50327

Original Research

Evaluating the Performance of Large Language Models GPT-4, Claude 3 Sonnet, and Gemini Pro in Recurrent Pregnancy Loss

Han Zhang[1,2,3,†], Chanlin Han[1,2,3,†], Rui Hu[1,2,3], Xiao Zhou[1,2,3], Xuemei Li[1,2,3], Jifan Tan[1,2,3,*]

[1]Reproductive Medicine Center, Shenzhen Maternity and Child Healthcare Hospital, Women and Children's Medical Center, Southern Medical University, 518000 Shenzhen, Guangdong, China
[2]Shenzhen Key Laboratory of Maternal and Child Health and Diseases, 518000 Shenzhen, Guangdong, China
[3]Shenzhen Clinical Research Center for Obstetrics & Gynecology and Reproductive System Diseases, 518000 Shenzhen, Guangdong, China
*Correspondence: tanjifan@alumni.sysu.edu.cn (Jifan Tan)
[†]These authors contributed equally.

Academic Editors: Andrea Tinelli and Michael H. Dahan
Submitted: 23 January 2026 Revised: 27 February 2026 Accepted: 7 April 2026 Published: 23 January 2026

Abstract

Background: The utility of large language models (LLMs) in recurrent pregnancy loss (RPL) consultation and patient education has not yet been systematically investigated. This study evaluated the performance of 3 LLMs (GPT-4, Claude 3 Sonnet, and Gemini Pro) in the field of RPL by assessing accuracy, comprehensiveness, and readability. **Methods:** Two experienced obstetricians and gynecologists developed medical questions based on the 2022 guidelines of the European Society of Human Reproduction and Embryology (ESHRE). The questionnaire included multiple formats, including choice questions (single-answer and multiple-answer) and short-answer questions. Short-answer questions were further categorized as common questions or clinical cases based on the question type, and as prevention, diagnosis, or treatment based on the content. Subsequently, the LLMs-generated answers were graded for accuracy, comprehensiveness, and readability. Choice questions were evaluated for accuracy only, whereas short-answer questions were evaluated for accuracy, comprehensiveness, and readability. Accuracy and comprehensiveness were evaluated using a 5-point Likert scale. Readability was evaluated using the Flesch Reading Ease (FRE) score and the Flesch–Kincaid Grade Level (FKGL). **Results:** Responses to 47 questions generated by LLMs showed that the best-performing model, Claude 3 Sonnet, achieved higher scores in short-answer questions for both accuracy (median score 5.00 [interquartile range (IQR), 4.00–5.00]) and comprehensiveness (median score 5.00 [IQR, 4.13–5.00]). No differences were observed between LLMs in accuracy scores for all choice questions, including single-choice and multiple-choice questions ($p > 0.05$). Regarding readability, the FRE and FKGL scores indicated difficult readability, ranging from college-level to professional-level reading skill. For single LLM, the median accuracy scores did not differ significantly across question types. After targeted, grounded prompting based on specific categories (prevention, diagnosis, and treatment), the accuracy scores of all 3 LLMs were improved (GPT-4, median score 4.00 [IQR, 3.00–5.00] vs. 5.00 [IQR, 5.00–5.00], $p < 0.001$; Claude 3 Sonnet, median score 5.00 [IQR, 4.00–5.00] vs. 5.00 [IQR, 5.00–5.00], $p = 0.001$; Gemini Pro, median score 3.50 [IQR, 2.00–5.00] vs. 5.00 [IQR, 5.00–5.00], $p < 0.001$). The comprehensiveness scores for GPT-4 improved significantly after grounded prompting, whereas Claude 3 Sonnet and Gemini Pro performed worse than baseline, although these differences were not statistically significant. **Conclusions:** Within the field of RPL consultation, Claude 3 Sonnet outperformed GPT-4 and Gemini Pro in terms of accuracy and comprehensiveness of short-answer questions. After targeted, grounded prompting across specific categories (prevention, diagnosis, and treatment), the accuracy scores of all 3 LLMs improved. These findings suggest the potential of LLMs as an important supplementary tool for the current medical system in the field of RPL, supporting improvements in patient management.

Keywords: large language model; artificial intelligence; recurrent miscarriage; recurrent pregnancy loss

1. Introduction

Recurrent pregnancy loss (RPL) is a serious pregnancy disorder affecting approximately 1–5% of women attempting to conceive [1] and is defined as the loss of two or more clinically recognized pregnancies [2]. Previous studies have found that RPL profoundly affects the quality of life of women and their partners and is associated with an increased risk of adverse obstetric outcomes, metabolic syndrome, and psychological disorders such as anxiety and depression in women [3,4]. Therefore, improving the diagnosis, treatment strategies, and preventive management of RPL in couples is essential and holds important clinical and social significance.

Large language models (LLMs) are artificial intelligence (AI) models trained on large-scale datasets to generate natural language text applicable to a wide range of fields [5]. Among them, GPT is the most widely used and has demonstrated strong performance in the medical field, with broad application prospects in clinical decision-making (i.e., risk assessment, diagnosis, treatment selec-

tion), patient engagement (i.e., medical reminders, lifestyle advice), and public health [6,7], contributing to a reduced workload for medical professionals, increased efficiency, and lower costs [8]. While these findings highlight the potential of LLMs as valuable supplementary tools in clinical settings, their accuracy varies considerably across medical domains. This variability underscores the critical importance of model selection, prompt design, and domain-specific evaluation in optimizing the quality of clinical consultation [9,10].

Recently, the potential application of LLMs has attracted considerable attention in the field of obstetrics and gynecology (OBGYN). The study by Grünebaum A et al. [11] demonstrated that LLMs can provide valuable preliminary information across a wide range of topics in OBGYN. Li SW et al. [12] also reported that an LLM outperformed human candidates in an objective structured clinical examination in OBGYN, demonstrating accurate responses to complex and evolving clinical scenarios based on unfamiliar settings within a very short time frame. However, research involving LLMs in OBGYN remains limited, indicating a significant research gap compared with other medical disciplines [13].

To date, in the field of RPL, which causes significant distress to patients, the performance of LLMs for consultation and education remains unexplored and requires evaluation. To the best of our knowledge, this is the first study to evaluate and compare the performance of 3 LLMs (GPT-4, Claude 3 Sonnet, and Gemini Pro) in RPL, aiming to identify their strengths and limitations and to provide a reference for their potential clinical application in RPL.

The significance of this study lies in: (1) systematically evaluating the potential application of LLMs in RPL and providing directions for future research; (2) exploring the feasibility of using LLMs as clinical decision support tools and providing new perspectives for precision medicine in RPL; and (3) promoting the interdisciplinary integration of AI and OBGYN, advancing the development of intelligent healthcare.

2. Materials and Methods

This cross-sectional study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline.

2.1 Data Source and Question Design

Questions were designed in accordance with the 2022 guidelines of the European Society of Human Reproduction and Embryology (ESHRE) on RPL. These evidence-based guidelines provide recommendations on the definition, epidemiology, etiology, diagnosis, treatment, and prognosis of RPL. In this study, we focused on guideline-based recommendations for RPL risk factor assessment, diagnostic investigation, and treatment management. Two types of questions were designed: choice questions (including single-answer and multiple choice) and short-answer questions (categorized as common questions and clinical cases based on the question type; and as prevention, diagnosis, and treatment based on the content), covering a range of difficulty from basic concepts to clinical decision-making. Questions were extracted and structured by two experienced obstetricians and gynecologists into a standardized format to facilitate processing by LLMs. Two rounds of Delphi surveys were then conducted to revise and refine the initial question set. All questions and LLM responses are shown in Table 1 and Supplementary Materials-Appendix.

2.2 LLMs Selection and Grouping

We deliberately selected 3 general-purpose LLMs GPT-4 (OpenAI, accessed May 2024), Claude 3 Sonnet (Anthropic, accessed May 2024), and Gemini Pro (Google, accessed May 2024) rather than medical-specific models for the following reasons: (1) these tools are the most accessible for patients and healthcare providers seeking RPL information; (2) we aimed to evaluate real-world performance as experienced by non-specialist users and general practitioners; and (3) this approach establishes baseline performance for comparison with future medical-specific models. The experiment was divided into three groups:

(1) Single-LLM comparisons: The application of RPL knowledge across different domains was investigated separately for each of the 3 pre-trained LLMs, aiming to reflect their basic performance in the RPL field.

(2) Multi-LLM comparisons: GPT-4, Claude 3 Sonnet, and Gemini Pro were evaluated and compared for accuracy, comprehensiveness, and readability, aiming to identify the most suitable LLM for RPL clinical consultation before AI-grounded prompting.

(3) Pre- and post-grounded prompting of a single LLM: Structured excerpts from the three chapters (risk factors, diagnosis, and treatment) of the 2022 ESHRE guideline were inserted into the prompt as contextual input prior to each query, without modifying model parameters. The protocol was as follows: a new conversation session was initiated for each model; the full chapter of the ESHRE guideline corresponding to the question category was provided as the first message, with the instruction “Based on the following clinical guideline, please answer the subsequent questions”. The test question was then submitted within the same session. Performance changes before and after grounded prompting were compared to explore the potential of different LLMs for continuous optimization in the RPL domain.

2.3 Test and Data Collection

Questions were input into GPT-4, Claude 3 Sonnet, and Gemini Pro in a standardized format. To account for the stochastic nature of LLM outputs and ensure result stability, each question was independently tested 5 times per model, consistent with prior LLM evaluation studies in the

Table 1. Questions of recurrent pregnancy loss (RPL).

Content	Question type		Question
Prevention	Multiple-Choice Questions	Single-Choice Questions	Women should be sensitively informed that the risk of pregnancy loss is lowest in women aged: A. 20 to 25 years; B. 20 to 35 years; C. 30 to 40 years; D. Over 40 years
			Women should be informed that the risk of pregnancy loss rapidly increases after the age of: A. 35 years; B. 40 years; C. 45 years; D. 50 years
			Regarding the association between stress and RPL, which statement is correct? A. Stress is a direct cause of RPL; B. Stress is associated with RPL, but there is no evidence that it is a direct cause; C. Stress has no association with RPL; D. Stress increases the chances of pregnancy
			For couples with RPL, what is maternal obesity or being significantly underweight associated with? A. Obstetric complications; B. Chances of a live birth; C. General health issues; D. All of the above
			For couples with RPL, which recommendation is suggested? A. Maintain low body weight; B. Maintain a healthy normal range BMI; C. Gain weight; D. Make no changes
			What is the relationship between excessive alcohol consumption and RPL? A. A proven risk factor; B. A possible risk factor; C. No relationship; D. Beneficial
			For couples with RPL, what is the correct advice regarding alcohol consumption? A. Increase alcohol intake; B. Limit alcohol consumption; C. Consume specific types of alcohol; D. No advice
		Multiple-Choice Questions	Which statements are correct regarding age and the risk of pregnancy loss? A. The risk of pregnancy loss is lowest in women aged 20 to 35 years; B. The risk of pregnancy loss rapidly increases after the age of 40; C. Stress is a direct cause of RPL; D. There is no evidence that stress is a direct cause of pregnancy loss
			Regarding alcohol consumption and RPL, which statements are correct? A. Excessive alcohol consumption is a possible risk factor for pregnancy loss; B. Excessive alcohol consumption is a proven risk factor for fetal problems; C. Couples with RPL do not need to limit their alcohol consumption; D. Couples with RPL should limit their alcohol consumption
	Short-Answer Questions	Common Questions	What is the definition of RPL?
			What are the behavioral and lifestyle factors in RPL?
			How to prevent RPL through health behavior modifications?
			Does advanced maternal age increase the risk of RPL?
			Is RPL related to maternal weight?
		Clinical Case	I am 42 years old and 10 weeks pregnant. I have had excessive alcohol consumption and twice experienced pregnancy loss before. Is there any association? Should I limit alcohol consumption during pregnancy? What is the safe upper limit for daily alcohol consumption during pregnancy?
			I am overweight and have experienced three pregnancy losses. Is obesity a risk for my recurrent miscarriages? Should I have a weight loss?

Table 1. Continued.

Content	Question type		Question
Diagnosis	Choice Questions	Single-Choice Questions	What is the recommendation for thyroid screening in women with RPL? A. Only if there are symptoms of thyroid disorder; B. Recommended in all cases; C. After two pregnancy losses; D. Not recommended For thyroid screening in women with RPL, which of the following are recommended? A. Thyroid-stimulating hormone (TSH). B. Thyroid peroxidase (TPO) antibodies. C. TSH and Thyroxine (T4). D. TSH and TPO antibodies. For women with RPL, when is screening recommended for antiphospholipid antibodies? A. After the first pregnancy loss; B. After two pregnancy losses; C. After three pregnancy losses; D. Only in the presence of additional risk factors for thrombophilia. For women with RPL, which of the following assessments is not recommended to improve the prognosis of the next pregnancy? A. Assessment of Polycystic Ovary Syndrome (PCOS); B. Fasting insulin and glucose testing; C. Ovarian reserve testing; D. All of the above. For a couple with RPL, what is recommended for genetic analysis of pregnancy tissue following pregnancy loss? A. Subtelomere multiplex ligation-dependent probe amplification (MLPA); B. Array-based Comparative Genomic Hybridization (array-CGH); C. Quantitative fluorescence polymerase chain reaction (QF-PCR); D. karyotyping What is the preferred technique for evaluating the uterus in women with RPL? A. Hysterosalpingography (HSG); B. Sonohysterography (SHG); C. Transvaginal 3D ultrasound (US); D. MRI Which of the following is not recommended for routine testing in women with RPL? A. Human Leukocyte Antigen (HLA) determination; B. Thyroid screening; C. Transvaginal 3D US; D. SHG
		Multiple-Choice Questions	Regarding diagnostic investigations in RPL, which statement is correct? A. They can be tailored based on medical and family history; B. Genetic analysis of pregnancy tissue is mandatory for all patients; C. Genetic analysis of pregnancy tissue is not routinely recommended; D. Parental karyotyping should be performed for all patients. Which of the following investigations can be used for RPL diagnosis? A. Medical and family history analysis; B. Genetic analysis; C. Antiphospholipid antibody screening; D. Thyroid function tests The prognosis for women with RPL is recommended to be based on which factors? A. Maternal age; B. Maternal age at first pregnancy; C. Number of previous pregnancy losses; D. Complete pregnancy history For women with RPL, which tests are recommended? A. Thyroid function screening; B. Ovarian reserve testing; C. Luteal phase insufficiency testing; D. Uterine anatomy assessment
	Short-Answer Questions	Common Questions	For couples with RPL, which imaging examinations should be taken? In the diagnosis of RPL, which is the preferred imaging examination for women? Is there any association between family history and RPL? What kinds of genetic analysis of pregnancy tissue are routinely recommended in the diagnosis of RPL? Are serum immunological tests recommended in the diagnosis of RPL? What kinds of immunological tests should be taken? Should women with RPL take a thrombophilia screening? Should women with RPL screen for metabolic or endocrinological tests? Which tests are recommended?
		Clinical Case	I am 29 years old and have had three pregnancy losses. My husband is an alcoholic and obese. Does my husband's lifestyle promote pregnancy loss? Is it necessary for him to limit consumption and lose weight?

Table 1. Continued.

Content	Question type		Question
Treatment	Choice Questions	Single-Choice Questions	What is the recommendation for women with hereditary thrombophilia and RPL? A. Use antithrombotic prophylaxis; B. Avoid antithrombotic prophylaxis unless for VTE prevention or research; C. Administer aspirin; D. Undergo genetic testing
			What is the recommendation for women with RPL and endometrial polyps? A. Regular monitoring with MRI; B. Surgery increases the chance of a live birth in women with RPL; C. Surgery increases the chance of miscarriage in women with RPL; D. Surgical removal of endometrial polyps is not recommended
			What is the recommendation regarding vaginal progesterone in women with RPL? A. Improves live birth rate in women with 3 or more pregnancy losses and vaginal blood loss in a subsequent pregnancy; B. Not recommended in any case of RPL; C. Only recommended for women with less than 3 pregnancy losses; D. Recommended for all women with RPL
			Which of the following is true about multivitamin supplements consumption in women with RPL? A. Women with RPL should be advised on multivitamin supplements that are safe in pregnancy; B. Multivitamin supplements are not recommended for women with RPL; C. Multivitamin supplements are risk factors for RPL; D. Vitamins A and E can be taken for women with RPL
		Multiple-Choice Questions	What factors are recommended to base the prognosis on for women with RPL? A. Woman's age; B. Complete pregnancy history; C. Number of previous pregnancy losses; D. Live births and their sequence
			For women with RPL, what can prognostic tools (Kolte & Westergaard) be used for? A. Providing an estimate of the subsequent chance of live birth; B. Providing an estimate of adverse pregnancy outcome; C. Providing information on possible treatments; D. All of the above
			Which treatment methods are not recommended for women with RPL? A. Progesterone; B. Lymphocyte immunization therapy; C. Repeated and high doses of Intravenous immunoglobulin (IVIG); D. Glucocorticoids
	Short-Answer Questions	Common Questions	Should women with RPL be treated with heparin or aspirin during pregnancy?
			Should women with uterine abnormalities and RPL be treated with surgery?
		Clinical Case	I am 12 weeks pregnant and have been diagnosed with RPL. In my thyroid function tests, the TSH level was within the normal reference range. But my thyroid antibody is positive. How should I be followed during pregnancy?
			I have been diagnosed with RPL. And I am pregnant again now. My thyroid function tests, and the TSH level were within the normal reference range. But my thyroid antibody is positive. Should I be treated with levothyroxine during pregnancy?
			I am 12 weeks pregnant and want to use multivitamin supplements. I have been diagnosed with RPL. Is it safe for me? What multivitamin supplements should I take?

RPL, recurrent pregnancy loss; BMI, body mass index; TSH, thyroid-stimulating hormone; TPO, thyroid peroxidase; MLPA, multiplex ligation-dependent probe amplification; array-CGH, Array-based Comparative Genomic Hybridization; QF-PCR, quantitative fluorescence polymerase chain reaction; HSG, hysterosalpingography; SHG, sonohysterography; US, ultrasound; MRI, magnetic resonance imaging; HLA, human leukocyte antigen; VTE, venous thromboembolism.

medical field. The temperature parameter was set to 0.7 for all models to balance creativity and consistency. The modal response across 5 runs was used for choice questions, and mean Likert scores across 5 runs were used for short an-

swer questions. All 5 responses per question were scored independently, and the mean score was calculated to yield a single aggregated value per question per model for final analysis. All tests were conducted within the same time frame to control potential interference from external factors and model updates.

2.4 Evaluation of Accuracy, Comprehensiveness, and Readability

The answers generated by GPT-4, Claude 3 Sonnet, and Gemini Pro were scored on 3 dimensions: accuracy, comprehensiveness, and readability. Accuracy refers to the correctness and adherence to clinical knowledge and the 2022 ESHRE guidelines on RPL. Comprehensiveness evaluates the extent to which the response addresses all relevant aspects of the question. Ease of understanding refers to the ease with which a reader can understand and comprehend the content. A 5-point Likert scale was applied to assess accuracy and comprehensiveness. For accuracy: 1 = completely incorrect or contradicts the 2022 ESHRE guideline; 2 = mostly incorrect with only minor elements correct; 3 = partially correct but contains notable errors or omissions; 4 = mostly correct with only minor inaccuracies; and 5 = fully correct and consistent with the guideline. Readability was evaluated using the Flesch Reading Ease (FRE) score and Flesch–Kincaid Grade Level (FKGL). FRE scores text reading complexity on a scale from 0 to 100, with a higher score indicating a lower complexity level. FKGL evaluates readability based on sentence word count and sentence length, with higher grades indicating lower readability. Readability formulas were used to calculate the FRE [14] and FKGL [15].

Two investigators independently scored accuracy and comprehensiveness to ensure the objectivity and consistency of the evaluation. Prior to scoring, investigators underwent standardized training to clarify the specific meaning of the scoring criteria and scales, thereby improving the reliability of scoring. Inter-rater disagreements were resolved through discussion until consensus was reached.

2.5 Statistical Analysis

All statistical analyses were performed using SPSS 26.0 (IBM Corp, Armonk, NY, USA) and GraphPad Prism 8 (San Diego, CA, USA). Scores for accuracy and comprehensiveness of GPT-4, Claude 3 Sonnet, and Gemini Pro were calculated and displayed as descriptive statistics. Data normality was assessed using the Shapiro-Wilk test; given the ordinal nature of Likert-scale data, non-parametric tests were applied. Normally distributed data were presented as mean ± standard deviation (SD), and non-normally distributed data as medians with interquartile range (IQR). The Mann-Whitney U test or Kruskal-Wallis H test was used to compare the differences among the outcomes. Bonferroni correction was explicitly applied to all multiple subgroup comparisons to control the family-wise error rate. The Wilcoxon signed-rank test was used to compare the scores of each index before and after LLM-grounded prompting. A two-sided $p < 0.05$ was considered statistically significant. Inter-rater agreement between two investigators was evaluated using the weighted kappa coefficient.

3. Results

47 questions based on the 2022 ESHRE guidelines across prevention, diagnosis, and treatment were evaluated independently by two authors. Choice questions (single-choice questions and multiple-choice questions) were only evaluated for accuracy, whereas short-answer questions (categorized as common questions and clinical cases based on the question type, or prevention, diagnosis, and treatment based on the content) were evaluated for accuracy, comprehensiveness, and readability.

The Weighted Cohen's kappa coefficient between the two authors was 0.844 for accuracy and 0.675 for comprehensiveness in GPT-4; 0.786 for accuracy and 0.732 for comprehensiveness in Claude 3 Sonnet; and 0.896 for accuracy and 0.713 for comprehensiveness in Gemini Pro. Higher kappa values for accuracy than for comprehensiveness across all models (0.786–0.896 vs. 0.675–0.732) reflect the greater inter-rater objectivity for factual assessment relative to the inherently subjective evaluation of response coverage.

3.1 Performance of the LLMs Before Grounded Prompting

Before grounded prompting, the median accuracy scores for short-answer questions were 3.50 (IQR, 3.00–4.00) for GPT-4, 4.00 (IQR, 3.50–4.88) for Claude 3 Sonnet, and 3.00 (IQR, 2.00–3.88) for Gemini Pro. For GPT-4, the median accuracy scores for common questions and clinical cases were 3.50 (IQR, 2.50–4.00) and 3.50 (IQR, 3.00–4.00), respectively ($p = 0.72$), and for prevention, diagnosis, and treatment were 3.50 (IQR, 3.00–4.00), 2.75 (IQR, 2.13–3.88), and 4.00 (IQR, 3.25–4.25), respectively ($p = 0.12$) (**Supplementary Table 1**). For Claude 3 Sonnet, the median accuracy scores for common questions and clinical cases were 4.00 (IQR, 3.50–4.63) and 4.25 (IQR, 2.50–5.00), respectively ($p = 0.78$), and for prevention, diagnosis, and treatment were 4.50 (IQR, 4.00–5.00), 3.50 (IQR, 3.50–4.38), and 2.50 (IQR, 2.50–4.75), respectively ($p = 0.11$) (**Supplementary Table 2**). For Gemini Pro, the median accuracy scores for common questions and clinical cases were 2.75 (IQR, 1.88–3.63) and 3.25 (IQR, 2.75–4.13), respectively ($p = 0.24$), and for prevention, diagnosis, and treatment were 3.50 (IQR, 3.00–4.00), 2.50 (IQR, 1.63–3.00), and 2.00 (IQR, 1.50–3.50), respectively ($p < 0.05$) (**Supplementary Table 3**).

The median comprehensiveness scores for short-answer questions were 3.25 (IQR, 3.00–4.00) for GPT-4, 5.00 (IQR, 4.13–5.00) for Claude 3 Sonnet, and 4.00 (IQR, 3.00–4.38) for Gemini Pro. For GPT-4, the median comprehensiveness scores for common questions and clinical cases

were 4.00 (IQR, 3.00–4.50) and 3.00 (IQR, 3.00–3.63), respectively (p = 0.27), and for prevention, diagnosis, and treatment were 4.00 (IQR, 3.00–4.50), 3.25 (IQR, 3.00–4.00), and 3.00 (IQR, 3.00–4.25), respectively (p = 0.45) (**Supplementary Table 1**). For Claude 3 Sonnet, the median comprehensiveness scores for common questions and clinical cases were 5.00 (IQR, 4.38–5.00) and 5.00 (IQR, 3.75–5.00), respectively (p = 0.90), and for prevention, diagnosis, and treatment were 5.00 (IQR, 5.00–5.00), 4.75 (IQR, 4.13–5.00), and 4.00 (IQR, 3.25–5.00), respectively (p < 0.05) (**Supplementary Table 2**). For Gemini Pro, the median comprehensiveness scores for common questions and clinical cases were 3.50 (IQR, 3.00–4.13) and 4.00 (IQR, 3.88–4.50), respectively (p = 0.18), and for prevention, diagnosis, and treatment were 4.00 (IQR, 3.50–5.00), 3.50 (IQR, 3.00–4.00), and 4.00 (IQR, 2.50–4.25), respectively (p = 0.25) (**Supplementary Table 3**).

Overall, GPT-4 showed similar median accuracy and comprehensiveness scores across all question types and categories (**Supplementary Table 1**). For Claude 3 Sonnet, accuracy scores were consistent across groups, whereas comprehensiveness scores differed significantly by content categories (**Supplementary Table 2**). Gemini Pro followed a similar pattern, with comparable accuracy scores across question types but significant group differences in accuracy scores across content categories (**Supplementary Table 3**).

3.2 Multi-LLM Comparisons Before Grounded Prompting

GPT-4, Claude 3 Sonnet, and Gemini Pro were independently evaluated for accuracy, comprehensiveness, and readability to identify the most suitable LLM for clinical consultation of RPL prior to grounded prompting. The accuracy scores for all 47 questions are illustrated in Fig. 1A–C.

The median accuracy scores for GPT-4, Claude 3 Sonnet, and Gemini Pro were 4.00 (IQR, 3.00–5.00), 5.00 (IQR, 4.00–5.00), and 3.50 (IQR, 2.00–5.00), respectively, with significant differences observed between groups (p < 0.01) (Fig. 2 A; **Supplementary Tables 1–3**). Among short-answer questions, significant differences were also observed between groups (p = 0.01), (Fig. 2 B; **Supplementary Tables 1–3**). Among different types of short-answer questions, significant differences were observed for common questions (p = 0.02) (Fig. 2 C; **Supplementary Tables 1–3**), but not for clinical cases (p = 0.59) (Fig. 2 D; **Supplementary Tables 1–3**). In all of the above comparisons, the observed differences were attributable to the Claude 3 Sonnet and Gemini Pro groups. Among different content categories of short-answer questions, significant differences were found among the 3 groups for prevention (p = 0.03), specifically between GPT-4 and Claude 3 Sonnet (Fig. 2 E; **Supplementary Tables 1–3**), but not for diagnosis (p = 0.07) (Fig. 2 F; **Supplementary Tables 1–3**) or treatment (p = 0.13) (Fig. 2 G; **Supplementary Tables 1–3**). Overall, significant differences between groups were observed for the accuracy scores of short-answer questions, common questions, and prevention questions (**Supplementary Tables 1–3**).

The median comprehensiveness scores for GPT-4, Claude 3 Sonnet, and Gemini Pro were 3.25 (IQR, 3.00–4.00), 5.00 (IQR, 4.13–5.00), and 4.00 (IQR, 3.00–4.38) respectively (p < 0.001), with significant differences between groups, Significant differences were found between GPT-4 and Claude 3 Sonnet, as well as between Claude 3 Sonnet and Gemini Pro. (Fig. 3 A; **Supplementary Tables 1–3**). For different types of short-answer questions, significant differences were observed for both common questions (p < 0.01) (Fig. 3B; **Supplementary Tables 1–3**) and clinical cases (p = 0.02) (Fig. 3C; **Supplementary Tables 1–3)**. In common questions, significant differences were found between GPT-4 and Claude 3 Sonnet, and between Claude 3 Sonnet and Gemini Pro, but not between GPT-4 and Gemini Pro. In clinical cases, only GPT-4 and Claude 3 Sonnet differed significantly. Among different content categories of short-answer questions, significant differences were found for prevention (p < 0.01) (Fig. 3 D; **Supplementary Tables 1–3**), with significant differences observed only between GPT-4 and Claude 3 Sonnet, but not between GPT-4 and Gemini Pro, or between Claude 3 Sonnet and Gemini Pro. For diagnosis (p < 0.01) (Fig. 3 E; **Supplementary Tables 1–3**), significant differences were observed between GPT-4 and Claude 3 Sonnet, as well as between Claude 3 Sonnet and Gemini Pro. No significant differences were found for treatment (p = 0.48) (Fig. 3 F; **Supplementary Tables 1–3**). Overall, significant differences between groups were observed in comprehensiveness scores across all content categories except treatment.

Regarding readability, the FRE and FKGL scores were 22.35 ± 12.68 and 15.63 ± 2.45 for GPT-4, 24.45 ± 12.44 and 14.83 ± 3.08 for Claude 3 Sonnet, and 26.21 ± 14.11 and 12.95 ± 1.95 for Gemini Pro, with no observed significant differences across LLMs (all p > 0.05). All scores ranged from college to professional reading level (**Supplementary Tables 1–3**). Among choice questions, the accuracy scores for GPT-4, Claude 3 Sonnet, and Gemini Pro were 5.00 (IQR, 3.00–5.00), 5.00 (IQR, 5.00–5.00), and 5.00 (IQR, 2.00–5.00), respectively (p = 0.07) (**Supplementary Fig. 1A**; **Supplementary Tables 1–3**). Among single-choice questions, the corresponding scores were 5.00 (IQR, 4.50–5.00), 5.00 (IQR, 5.00–5.00), and 5.00 (IQR, 1.88–5.00) (p = 0.42) (**Supplementary Fig. 1B**; **Supplementary Tables 1–3**). Among multiple-choice questions, scores were 5.00 (IQR, 3.00–5.00), 5.00 (IQR, 4.50–5.00), and 3.00 (IQR, 2.00–5.00) (p = 0.08) (**Supplementary Fig. 1C**; **Supplementary Tables 1–3**). No significant differences were found across LLMs for any category of choice questions.

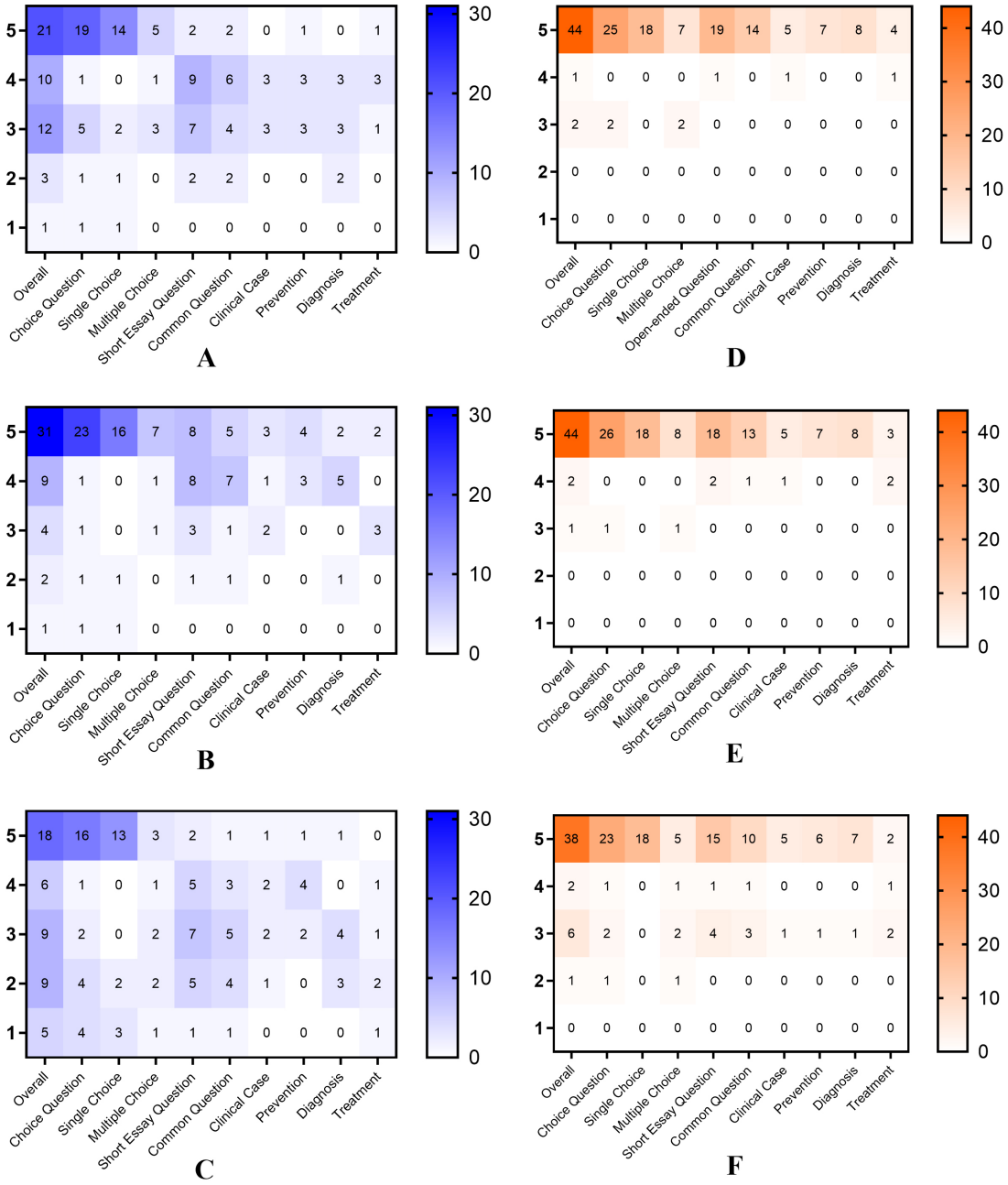

Fig. 1. Heatmap of the accuracy of LLM-generated answers. (A) GPT-4 before grounded prompting. (B) Claude 3 Sonnet before grounded prompting. (C) Gemini Pro before grounded prompting. (D) GPT-4 after grounded prompting. (E) Claude 3 Sonnet after grounded prompting; and (F) Gemini Pro after grounded prompting. LLM, large language model.

3.3 Comparison of LLMs Performance Pre- and Post-Grounded Prompting

GPT-4, Claude 3 Sonnet, and Gemini Pro were independently trained on 3 RPL categories: prevention, diagnosis, and treatment. Following grounded prompting, the models were re-evaluated (**Supplementary Tables 1–3**). Fig. 1D–F illustrates the accuracy and scores across 47 answers for each model. Moreover, we rescored the median FRE and FKGL to evaluate changes in readability after the application of grounded prompting (**Supplementary Tables 1–3**).

Grounded prompting significantly improved accuracy across all 3 models (Fig. 4 A–C). All models converged to a median accuracy of 5.00 (IQR, 5.00–5.00) after prompting. Gemini Pro (Fig. 4C) demonstrated the greatest absolute improvement (median 3.50→5.00, $p < 0.001$), followed by GPT-4 (Fig. 4 A) (median 4.00→5.00, $p < 0.001$), whereas Claude 3 Sonnet (Fig. 4 B) displayed the smallest gain. Notably, this smaller gain for Claude 3 Sonnet occurred despite its superior baseline performance, and the improvement remained statistically significant (median 5.00→5.00, $p = 0.001$). The proportion of questions showing improved

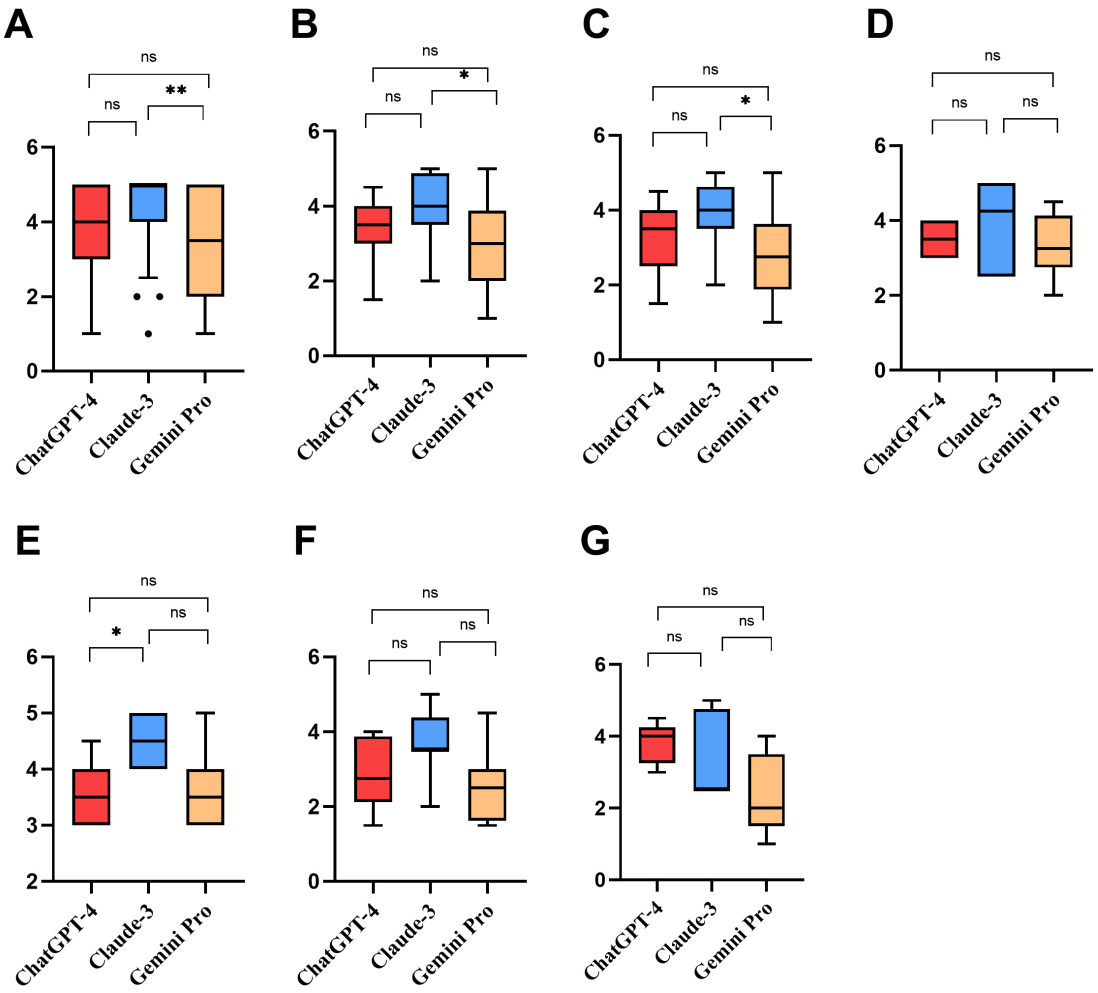

Fig. 2. Comparison of accuracy among 3 LLMs before grounded prompting. (A) Comparison of the median accuracy scores for all questions among the 3 LLMs. (B) Comparison of the median accuracy scores for short-answer questions among the 3 LLMs. (C) Comparison of the median accuracy scores for common questions of short-answer questions among the 3 LLMs. (D) Comparison of the median accuracy scores for clinical cases of short-answer questions among the 3 LLMs. (E) Comparison of the median accuracy scores for short-answer questions related to prevention among the 3 LLMs. (F) Comparison of the median accuracy scores for short-answer questions related to diagnosis among the 3 LLMs; and (G) comparison of the median accuracy scores for short-answer questions related to treatment among the 3 LLMs. Significance levels are indicated as follows: ns, not significant; *: $p < 0.05$; **: $p < 0.01$.

accuracy was highest in GPT-4 and Gemini Pro (53.19% each), with 25 questions improving and none declining for GPT-4, while Gemini Pro showed 25 improvements and 2 declines. In contrast, Claude 3 Sonnet showed a lower improvement rate (40.42%), with 19 questions improving, 26 remaining unchanged, and 2 declining.

As comprehensiveness was evaluated exclusively for short-answer questions, the following section reports comprehensiveness scores pre- and post-grounded prompting (Fig. 4D–F). In contrast to the uniform improvements in accuracy, effects on comprehensiveness diverged significantly across models. GPT-4 (Fig. 4D), which had the lowest baseline comprehensiveness, was the only model to show a statistically significant improvement (median 3.25→4.50, $p = 0.012$). Claude 3 Sonnet (Fig. 4 E) showed a slight, non-significant decline despite its high baseline (median 5.00→4.75, $p = 0.48$). Gemini Pro (Fig. 4 F) exhibited the most pronounced decline, with 11 of 20 questions decreasing and median comprehensiveness falling from 4.00 to 3.00 ($p = 0.39$). Together, these results suggest that grounded prompting reliably enhances accuracy but may constrain response comprehensiveness, particularly in models with stronger baseline comprehensiveness.

4. Discussion

4.1 Main Findings

Our study systematically evaluated the performance of GPT-4, Claude 3 Sonnet, and Gemini Pro regarding accuracy, comprehensiveness, and readability. We observed that all 3 different LLMs demonstrated remarkable perfor-

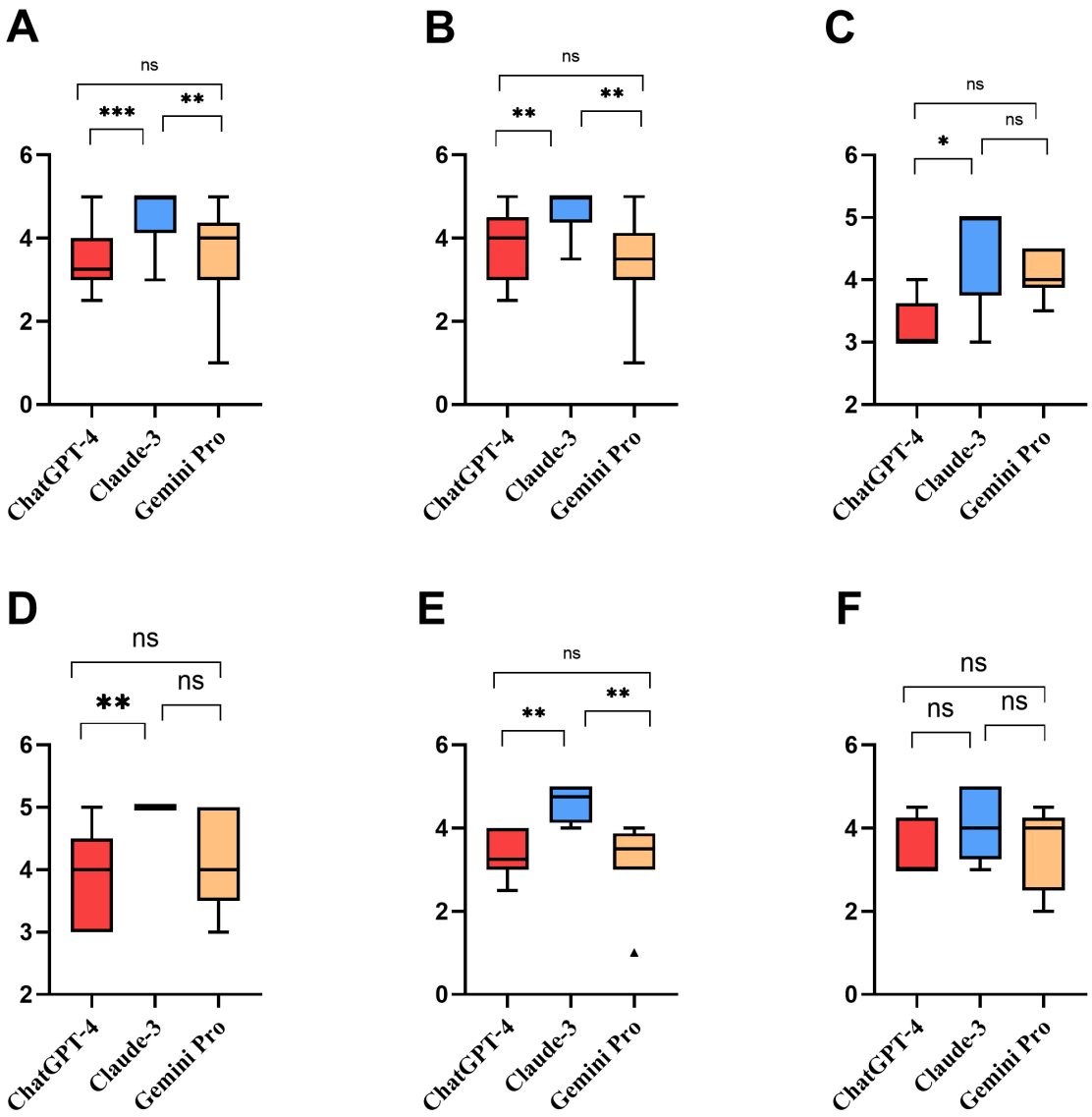

Fig. 3. Comparison of comprehensiveness scores among 3 LLMs before grounded prompting. (A) Comparison of the median comprehensiveness scores for short-answer questions among the 3 LLMs. (B) The comparison of the median comprehensiveness scores for common questions of short-answer questions among the 3 LLMs. (C) The comparison of the median comprehensiveness scores for clinical cases of short-answer questions among the 3 LLMs. (D) The comparison of the median comprehensiveness scores for short-answer questions related to prevention among the 3 LLMs. (E) The comparison of the median comprehensiveness scores for short-answer questions related to diagnosis among the 3 LLMs. (F) The comparison of the median comprehensiveness scores for short-answer questions related to treatment among the 3 LLMs. Significance levels are indicated as: ns, not significant; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

mance in accuracy and comprehensiveness, with Claude 3 Sonnet outperforming GPT-4 and Gemini Pro. To our knowledge, this is the first study to evaluate and compare 3 different LLMs in the field of RPL.

Claude 3 Sonnet outscored GPT-4 and Gemini Pro in short-answer questions, demonstrating superior median accuracy and comprehensiveness, particularly for questions about RPL prevention. These findings reveal performance heterogeneity among the 3 LLMs, with Claude 3 Sonnet exhibiting greater clinical consistency, suggesting its potential as an adjunctive tool for RPL counseling. The superior performance of Claude 3 Sonnet in addressing prevention-related questions also suggests its potential for identifying risk factors and guiding preventive interventions for couples with RPL. However, no differences were found between LLMs in accuracy scores for all choice questions, including single-choice questions and multiple-choice questions ($p > 0.05$). Regarding readability, FRE and FKGL scores indicated difficult readability across all LLMs, ranging from college to professional level. This consistently

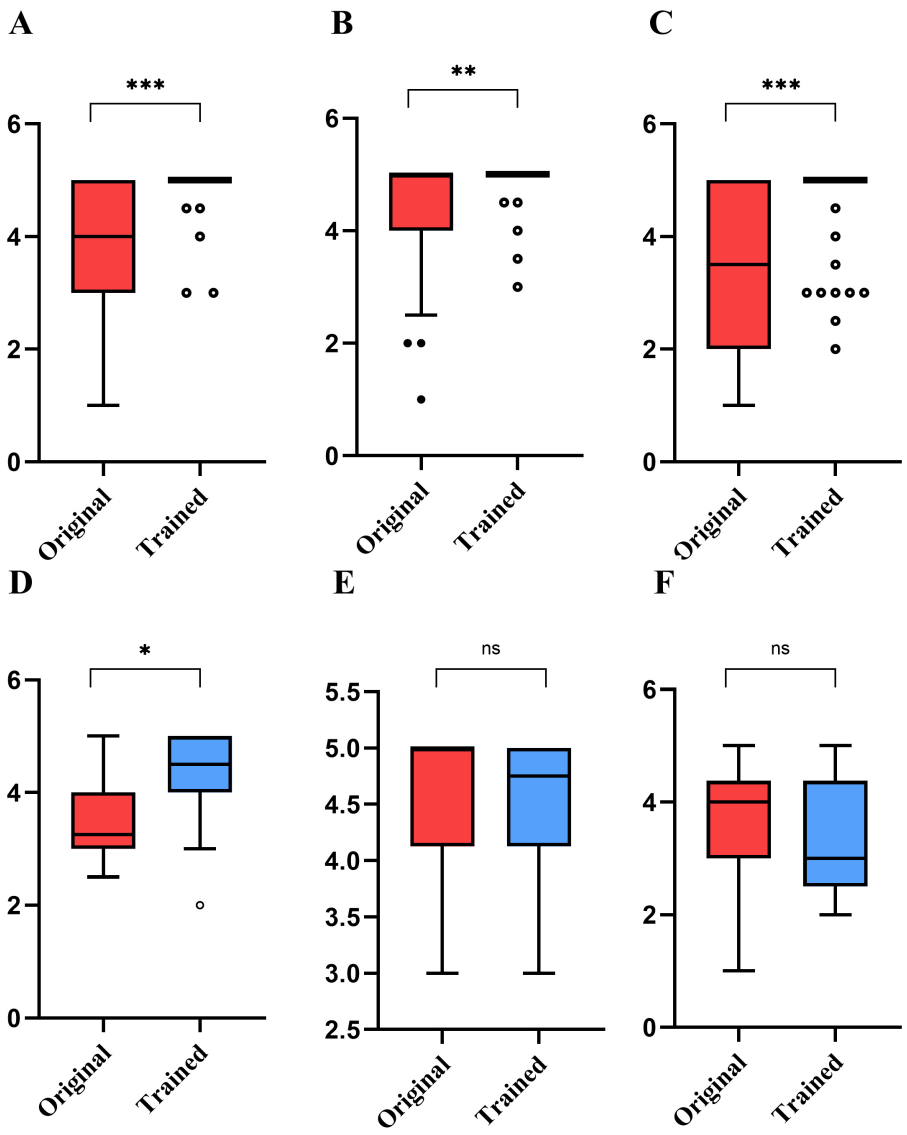

Fig. 4. Comparison of LLM performance based on median accuracy scores before and after grounded prompting. (A) Comparison of the median accuracy scores with GPT-4 pre- and post-grounded prompting. (B) Comparison of the median accuracy scores with Claude 3 Sonnet pre- and post-grounded prompting. (C) Comparison of the median accuracy scores with Gemini Pro pre- and post-grounded prompting. (D) Comparison of the median comprehensiveness scores of short essay questions with GPT-4 pre- and post-grounded prompting. (E) Comparison of the median comprehensiveness scores of short essay questions with Claude 3 Sonnet pre- and post-grounded prompting; and (F) comparison of the median comprehensiveness scores of short essay questions with Gemini Pro pre- and post-grounded prompting. Significance levels are indicated as: ns, not significant; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

poor readability limits their patient-facing utility, highlighting the need for prompt engineering or post-processing strategies to improve accessibility for patients with RPL.

The median accuracy scores for individual LLMs across different question types (common questions and clinical cases) were broadly similar, indicating that LLMs can provide information and consultation support for RPL with broad applicability. After targeted grounded prompting across specific categories (prevention, diagnosis, and treatment), accuracy scores improved for all 3 LLMs, suggesting responsiveness to structured contextual input and potential adaptability to updated ESHRE guidelines. Beyond accuracy and comprehensiveness, future evaluations should incorporate structured safety alignment assessments, including urgency recognition and clinical triage, to ensure that LLMs meet safety standards in RPL care.

4.2 Comparison With Existing Literature

The application and comparison of Claude 3 Sonnet and other LLMs in the medical field have sparked considerable academic interest since their emergence. Our findings suggest that Claude 3 Sonnet emerged as the best-

performing model in RPL among 3 LLMs, consistent with prior studies. Song H et al. [5], reported that, compared with other LLMs, including Bard, GPT-4, and New Bing, Claude performs best in analyzing clinical cases in urolithiasis consultations and education. Similarly, another study reported that, compared with GPT-3.5 and GPT-4, Claude generated higher-quality responses in oncology-related contexts, with superior performance in quality, empathy, and readability [16].

Considering that more than 70% of individuals utilize the internet as their primary source of health information and that false information spreads up to six times faster than factual content online, Menz BD et al. [17] conducted a study evaluating the ability of different LLMs to resist mass generation of health disinformation. This study found that many publicly accessible LLMs, including GPT and Gemini Pro, lacked adequate safeguards against mass content generation, whereas Claude 2 showed robust safeguards against the generation of health-related disinformation. We speculate that, in the field of RPL, the application of Claude 3 Sonnet for medical consultation may not only provide accurate information for couples with RPL but also mitigate the risk of large-scale misinformation generation, thereby avoiding delays in diagnosis or treatment, reducing unnecessary waste of medical resources, and limiting avoidable economic burden.

4.3 Limitations

Our study had several important limitations. First, the low readability of LLM-generated responses may limit their accessibility to the general public. Second, although grounded prompting significantly improved comprehensiveness scores for GPT-4, Claude 3 Sonnet, and Gemini Pro performed no better than, or even marginally worse than, their baseline, suggesting that striking a balance between technical accuracy, comprehensiveness, and accessible language remains a critical challenge for broader clinical adoption of LLMs. Additionally, as the prompting excerpts and test questions were derived from the same guideline chapters, content overlap may have partially inflated post-prompting accuracy scores. Future studies should formally evaluate response generation time and computational complexity across models, conduct sensitivity analyses across a range of temperature settings (e.g., T = 0–0.2) to assess result stability, and employ larger and more diverse question sets, with effect size reporting and per-question score analyses, to better quantify training effects and clarify the observed dissociation between post-prompting accuracy gains and comprehensiveness.

In addition, due to a knowledge gap between machine learning developers (e.g., data scientists) and practitioners (e.g., clinicians), the full utilization of machine learning for clinical data analysis has been hampered [18]. Additionally, advanced data analysis extensions of LLMs should be explored to bridge this gap and provide more robust clinical advice on RPL. Furthermore, LLMs remain prone to hallucination, generating outputs that may be incorrect or misleading [19]. AI tools can also be exploited to mass-produce misinformation, posing direct risks to human life [20,21]. Thus, healthcare professionals must critically evaluate LLM outputs. Navigating the associated legal and ethical landscape is equally essential. Only then can these tools genuinely enhance patient care without compromising clinical expertise [22]. LLMs evolve rapidly, and performance rankings may shift over time. Future evaluations should specify model versions and access dates to ensure reproducibility. Finally, our question set was derived from a single clinical guideline, which limits generalizability. Future studies should incorporate broader and more diverse clinical scenarios to address this limitation.

5. Conclusions

Within the field of RPL consultation, Claude 3 Sonnet outperformed GPT-4 and Gemini Pro in terms of accuracy and comprehensiveness of short-answer questions, particularly in risk factor prevention. After targeted grounded prompting across specific categories (prevention, diagnosis, and treatment), accuracy scores improved for all 3 LLMs. These findings suggest the potential of LLMs as supplementary tools within the current medical system in the field of RPL, promoting the improvement of patient management. However, it should be noted that all 3 LLMs performed suboptimally in terms of readability and failed to improve scores for comprehensiveness after grounded prompting. Medical professionals and patients should recognize these limitations and interpret LLMs-generated information with caution. Overall, this study demonstrates the potential utility of LLMs in the context of RPL and highlights the need for further research to balance technical accuracy, depth of understanding, and public usability.

Availability of Data and Materials

The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Author Contributions

HZ: Conceptualization, Data curation, Writing original draft, Writing review editing. CLH: Data curation, Formal analysis, Writing original draft, Writing review editing. RH, XZ: Data curation, Investigation, Writing original draft, Writing review editing. XML: Data curation, Investigation, Writing original draft, Writing review editing. JFT: Conceptualization, Funding acquisition, Investigation, Methodology, project administration, review editing. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Since the LLMs used in the present study are publicly available applications and no human participants were involved, institutional review board (IRB) approval was not required. Accordingly, an exemption from ethical review was granted by the hospital.

Acknowledgment

Not applicable.

Funding

The study was supported by Sanming Project of Medicine in Shenzhen (No.SZSM202411034), Shenzhen Key Laboratory of Maternal and Child Health and Diseases (ZDSYS20230626091559006) and Shenzhen Clinical Research Center for Obstetrics & Gynecology and Reproductive System Diseases (No.LCYSSQ20220823091401002).

Conflicts of Interest

The authors declare no conflicts of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/CEOG50327.

References

[1] Magnus MC, Wilcox AJ, Morken NH, Weinberg CR, Håberg SE. Role of maternal age and pregnancy history in risk of miscarriage: prospective register based study. BMJ (Clinical Research Ed.). 2019; 364: l869. https://doi.org/10.1136/bmj.l869

[2] ESHRE Guideline Group on RPL, Bender Atik R, Christiansen OB, Elson J, Kolte AM, Lewis S, et al. ESHRE guideline: recurrent pregnancy loss: an update in 2022. Human Reproduction Open. 2023; 2023: hoad002. https://doi.org/10.1093/hropen/hoad002

[3] Bae JH, Jung YM, Lee J, Shivakumar M, Park CW, Park JS, et al. Future risk of metabolic syndrome after recurrent pregnancy loss: a cohort study using UK Biobank. Fertility and Sterility. 2023; 120: 1227–1233. https://doi.org/10.1016/j.fertnstert.2023.09.012

[4] Dimitriadis E, Menkhorst E, Saito S, Kutteh WH, Brosens JJ. Recurrent pregnancy loss. Nature Reviews. Disease Primers. 2020; 6: 98. https://doi.org/10.1038/s41572-020-00228-z

[5] Song H, Xia Y, Luo Z, Liu H, Song Y, Zeng X, et al. Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis. Journal of Medical Systems. 2023; 47: 125. https://doi.org/10.1007/s10916-023-02021-3

[6] Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. Journal of Medical Internet Research. 2023; 25: e50638. https://doi.org/10.2196/50638

[7] Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023; 616: 259–265. https://doi.org/10.1038/s41586-023-05881-4

[8] Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel, Switzerland). 2023; 11: 887. https://doi.org/10.3390/healthcare11060887

[9] Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA. 2025; 333: 319–328. https://doi.org/10.1001/jama.2024.21700

[10] Omar M, Agbareia R, Glicksberg BS, Nadkarni GN, Klang E. Benchmarking the Confidence of Large Language Models in Answering Clinical Questions: Cross-Sectional Evaluation Study. JMIR Medical Informatics. 2025; 13: e66917. https://doi.org/10.2196/66917

[11] Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. American Journal of Obstetrics and Gynecology. 2023; 228: 696–705. https://doi.org/10.1016/j.ajog.2023.03.009

[12] Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. American Journal of Obstetrics and Gynecology. 2023; 229: 172.e1–172.e12. https://doi.org/10.1016/j.ajog.2023.04.020

[13] Levin G, Brezinov Y, Meyer R. Exploring the use of ChatGPT in OBGYN: a bibliometric analysis of the first ChatGPT-related publications. Archives of Gynecology and Obstetrics. 2023; 308: 1785–1789. https://doi.org/10.1007/s00404-023-07081-x

[14] Readability formulas. The Flesch reading ease readability formula. Available at: http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php (Accessed: 4 June 2026).

[15] Readability formulas. The Flesch reading ease readability formula. Available at: https://readabilityformulas.com/the-flesch-kincaid-grade-level-for-digital-content/ (Accessed: 4 June 2026).

[16] Chen D, Parsa R, Hope A, Hannon B, Mak E, Eng L, et al. Physician and Artificial Intelligence Chatbot Responses to Cancer Questions From Social Media. JAMA Oncology. 2024; 10: 956–960. https://doi.org/10.1001/jamaoncol.2024.0836

[17] 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 (Clinical Research Ed.). 2024; 384: e078538. https://doi.org/10.1136/bmj-2023-078538

[18] Tayebi Arasteh S, Han T, Lotfinia M, Kuhl C, Kather JN, Truhn D, et al. Large language models streamline automated machine learning for clinical studies. Nature Communications. 2024; 15: 1603. https://doi.org/10.1038/s41467-024-45879-8

[19] Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digital Medicine. 2023; 6: 120. https://doi.org/10.1038/s41746-023-00873-0

[20] Vinay R, Spitale G, Biller-Andorno N, Germani F. Emotional prompting amplifies disinformation generation in AI large language models. Frontiers in Artificial Intelligence. 2025; 8: 1543603. https://doi.org/10.3389/frai.2025.1543603

[21] Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology. 2023; 307: e230163. https://doi.org/10.1148/radiol.230163

[22] Choudhury A, Chaudhry Z. Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals. Journal of Medical Internet Research. 2024; 26: e56764. https://doi.org/10.2196/56764