Ann Med Res

Current issue list available at Ann Med Res

Annals of Medical Research

journal page: annalsmedres.org

A comparative analysis of large language models in digital content on painless labor: A systematic evaluation of readability, reliability, quality, and accuracy

Ahmet Ridvan Dogan[a], *, Havva Kocayigit[b], Ayca Tas Tuna[b]

[a]Sakarya Training and Research Hospital, Department of Anesthesiology and Reanimation, Sakarya, Türkiye
[b]Sakarya University, Faculty of Medicine, Department of Anesthesiology and Reanimation, Sakarya, Türkiye
*Corresponding author: a.ridvandogan@gmail.com (Ahmet Ridvan Dogan)

MAIN POINTS

- Large language models, such as ChatGPT, Gemini, and DeepSeek, are increasingly used for patient education and health communication.
- The readability and reliability of AI-generated medical information vary widely across platforms.
- This study provides the first systematic comparison of these three models in the context of information on painless labor.
- It identifies distinct strengths and weaknesses: ChatGPT excels in readability, Gemini excels in quality and reliability, and DeepSeek exhibits variable performance.
- The findings emphasize the need for verifiable references and audience-specific optimization in AI-based health communication.

Cite this article as: Dogan AR, Kocayigit H, Tas Tuna A. A comparative analysis of large language models in digital content on painless labor: A systematic evaluation of readability, reliability, quality, and accuracy. *Ann Med Res*. 2026;33(6):267–276. doi: 10.5455/annalsmedres.2025.11.346.

ABSTRACT

Aim: Painless labor is an important concern in obstetric care, with both medical and psychosocial dimensions. With the increasing use of artificial intelligence (AI) in health communication, evaluating the quality of content generated by large language models (LLMs) is essential. This study aimed to compare AI-generated content on "painless labor" produced by three LLMs---ChatGPT (Chat Generative Pre-trained Transformer), Gemini, and DeepSeek---in terms of readability, reliability, content quality, and medical accuracy.

Materials and Methods: A total of 270 texts were generated using 30 frequently searched keywords on three dates in May 2025 via the free versions of the three LLMs. Readability was assessed with six validated indices. Reliability and quality were evaluated using the Modified DISCERN tool, the JAMA (Journal of the American Medical Association) Benchmark Criteria, the Ensuring Quality Information for Patients (EQIP)-36, and the Global Quality Score. The medical accuracy of 90 texts was assessed by an obstetric anesthesia expert.

Results: ChatGPT texts were the most readable and were written at lower grade levels, but scored lower in reliability and quality. Gemini ranked highest for both reliability and content quality, although it produced more complex language. DeepSeek showed variable performance. No significant differences were found in medical accuracy or content completeness. Gemini also demonstrated the most consistent performance across all time points.

Conclusion: LLMs vary substantially in how they present medical content. ChatGPT achieved higher readability scores, indicating simpler language structures, whereas Gemini demonstrated higher reliability and content quality metrics. However, the absence of source citations across all models raises concerns about the verifiability of content, highlighting the need for critical oversight in healthcare applications.

Keywords: Painless labor, Large language models, Readability, Health literacy, Patient education, Obstetric anesthesia

Received: Nov 10, 2025 **Accepted:** Jan 05, 2026 **Available Online:** Jun 25, 2026

INTRODUCTION

Painless labor aims to enhance the childbirth experience by reducing pain through pharmacological or non-pharmacological methods [1]. Expectations and anxiety during pregnancy influence delivery preferences, including cesarean rates [2].

In recent years, patients increasingly search online for labor analgesia options—such as epidural or spinal techniques—prior to consulting healthcare professionals [3,4]. However, online content varies significantly in readability and reliability [5].

Artificial intelligence (AI)-based large language models (LLMs), including ChatGPT (Chat Generative Pre-trained Transformer), Gemini, and DeepSeek, have transformed access to health information by providing conversational, simplified answers to complex queries [6]. Despite their growing use, the scientific accuracy, readability, and reliability of AI-generated content remain largely unexamined [7]. Especially for individuals with limited health literacy, this raises concerns about misinformation and content

quality [8].

Readability is a key parameter for digital health communication. Leading institutions recommend that patient materials be understandable to readers with basic literacy [9]. Individuals with low health literacy often struggle with technical texts, which may negatively affect decision-making [10]. In this context, painless labor serves as an ideal topic for evaluating LLMs in terms of readability, reliability, quality, and medical accuracy. Unlike many general medical topics, painless labor is a preference-sensitive and emotionally charged clinical context where preconsultation information may directly influence patients' attitudes and decision-making. Therefore, evaluating AI-generated content in this domain addresses not only informational quality but also potential clinical and psychosocial implications.

ChatGPT, Gemini, and DeepSeek differ in architecture, data sources, and citation capabilities [11–13]. These distinctions justify a comparative analysis of their responses to a sensitive and frequently searched medical topic. Beyond readability, the reliability of content (e.g., source usage, authorship, currency) and its quality (relevance, objectivity) are critical for evaluating health information [14]. This study aims to systematically evaluate texts generated by these LLMs on the topic of painless labor across four domains: readability, reliability, quality, and medical accuracy.

MATERIALS AND METHODS

Study design and ethical approval

This cross-sectional, comparative study was conducted in May 2025 using experimental content analysis. No human participants were involved, and only publicly accessible outputs from AI models were used; thus, ethical approval was not required.

Keyword selection and data collection

Browser history and cookies were cleared, and all queries were performed in incognito mode to avoid personalization bias. Google Trends was used to identify relevant keywords by searching for the terms "epidural delivery," "painless labor," and "painless delivery." After filtering to remove irrelevant results, 30 keywords were selected.

Each keyword was submitted to three AI models—ChatGPT (GPT-3.5, OpenAI), Gemini (Gemini 1.5, Google Bard), and DeepSeek (DeepSeek-V3)—on three dates: May 11, 18, and 25, 2025. All queries were conducted using new browser sessions in incognito mode. All queries for each predefined date were conducted in a single session within the same time window to minimize temporal variability. DeepSeek required account registration; therefore, a separate account was created. In total, 270 response texts were collected (3 models × 3 dates × 30 keywords).

Responses were saved in Microsoft Word 2016 after removing formatting elements, such as lists and bullet points. Full dataset was archived at: https://archive.org/details/all_answers_LLMs.

Readability analysis

Each response was uploaded to https://readabilityformulas.com, and analyzed using six readability indices: Automated Readability Index (ARI), Flesch Reading Ease Score (FRES), Gunning Fog Index (GFO), Flesch–Kincaid Grade Level (FKGL), Simple Measure of Gobbledygook (SMOG), and FORCAST Readability Formula (FORCAST) (Table 1). Word and sentence counts were also recorded. All texts were analyzed by the same researcher using a standardized method and were subsequently entered into the statistical dataset.

Reliability assessment

Two independent researchers evaluated the texts using two validated tools. The Modified DISCERN tool (0–5 points) assessed source credibility, balance, and clarity [15]. The JAMA (Journal of the American Medical Association) Benchmark Criteria (0–4 points) evaluated authorship, attribution, update date, and information transparency [16]. In the event of disagreement, a third researcher resolved any discrepancies.

Quality assessment

Text quality was assessed using the Ensuring Quality Information for Patients (EQIP)-36 scale and the Global Quality Score (GQS) [15]. EQIP-36 includes 36 binary items, and scores were categorized as high (76–100%), moderate (51–75%), low (26–50%), or poor (0–25%). GQS is a 1–5 scale reflecting overall usefulness to patients. Discrepancies were resolved through third-party consensus. Researchers used standardized forms and reviewed the responses in different orders to minimize bias.

Medical accuracy assessment

A random sample of 90 texts (30 per model) was selected using a Python-based algorithm. All identifying metadata were removed, and evaluations were conducted blindly by a specialist in obstetric anesthesia. Each text was scored 1–10 for scientific accuracy and 1–10 for content comprehensiveness, following prior methodologies [17]. Obvious medical errors and fabricated references were identified through qualitative assessment.

Statistical analysis

Analyses were performed using IBM SPSS v25.0. Descriptive statistics were reported as mean ± standard deviation, median (minimum–maximum), and frequency (%). The distribution of the data was assessed using the Shapiro–Wilk test. Kruskal–Wallis tests were used for group comparisons, followed by Bonferroni-corrected Mann–Whitney U tests when

https://doi.org/10.5455/annalsmedres.2025.11.346

appropriate. Chi-square tests were applied to categorical variables. The Friedman test was used to assess variation in performance across dates. A p-value < 0.05 was considered statistically significant, with adjustments for multiple comparisons where necessary.

RESULTS

A total of 30 English keywords frequently associated with "epidural delivery," "painless labor," and "painless delivery" were identified using Google Trends. Repetitive or semantically overlapping terms were excluded, and the most contextually relevant expressions were retained (Table 2).

Each keyword was queried using three AI models—ChatGPT, Gemini, and DeepSeek—on three separate dates (May 11, 18, and 25, 2025), resulting in a total of 270 response texts.

Comparison of readability indices

Significant differences were found for all readability indices, word counts, and sentence counts across the three models (all p<0.001) (Table 3). Gemini and DeepSeek had significantly higher ARI scores than ChatGPT (p<0.01). For FRES, ChatGPT scored highest, followed by DeepSeek and Gemini (p<0.05 to p<0.001). In GFI, Gemini had the highest scores; there was no difference between ChatGPT and DeepSeek.

FKGL scores were ranked as follows: Gemini > ChatGPT > DeepSeek (all p<0.001). Gemini also had the highest SMOG scores, while DeepSeek had the highest FORCAST scores (all p<0.001). Gemini produced significantly longer texts, in terms of both word and sentence counts, than the other two models. DeepSeek responses were also longer than ChatGPT's in terms of sentence count (all adjusted p<0.05).

In grade-level classifications, Gemini's content was consistently grouped into higher difficulty levels by ARI, GFI, FKGL, and SMOG (Figure 1). ChatGPT exhibited the lowest difficulty levels across most indices. FORCAST scores indicated that DeepSeek had higher complexity than both Gemini and ChatGPT (p<0.01). Detailed pairwise comparisons are provided in Supplementary Table 1.

Comparison of reliability and quality assessments

Significant differences between the three AI models were observed across all four evaluation metrics (all p<0.001). ChatGPT had significantly lower Modified-DISCERN scores than DeepSeek and Gemini, both of which showed similar reliability levels. In the JAMA Benchmark assessment, DeepSeek ranked highest, followed by Gemini; ChatGPT scored significantly lower than both. Gemini achieved the highest EQIP quality scores, significantly outperforming ChatGPT and DeepSeek. In GQS evaluations, both Gemini and DeepSeek scored significantly higher than ChatGPT, with no difference between them (Figure 2). Detailed pairwise comparisons are provided in Supplementary Table 2.

Time-based analyses

Temporal analyses using the Friedman test revealed some model-specific trends across the three data collection dates. For ChatGPT, both word and sentence counts significantly increased over time (p<0.01). ARI, FKGL, and SMOG scores rose between May 11 and May 18, then slightly decreased on May 25, indicating fluctuations in complexity. FRES scores dipped on May 18 and rebounded on May 25, while FORCAST scores gradually declined (p<0.01). No significant changes were found in reliability or quality scores across the three dates. In DeepSeek, only EQIP scores increased significantly over time (p<0.05), while other metrics remained stable. Gemini showed no significant changes in any metric, indicating consistent performance across all time points (Table 4).

Comparison of medical accuracy assessments

Median scores for both scientific accuracy and content completeness were similar across the three models, with no statistically significant differences (Table 5). All models demonstrated comparable medical accuracy.

DISCUSSION

This study systematically compared the content generated on the topic of "painless labor" by three popular large language models (ChatGPT, Gemini, and DeepSeek) across metrics of readability, reliability, quality, and medical accuracy. The findings demonstrated statistically significant differences among the models regarding readability, reliability, and quality. Beyond confirming previously reported performance patterns of large language models, the present study extends the existing literature by contextualizing these differences in painless labor, a preference-sensitive, emotionally charged clinical scenario. In addition, the time-based evaluation was designed as an exploratory assessment of output stability, highlighting consistency differences between models rather than serving as a primary outcome measure.

In terms of readability, ChatGPT produced texts with lower grade levels and greater ease of comprehension. In contrast, Gemini and DeepSeek generated longer texts with higher technical content, requiring more advanced reading levels. Regarding reliability, ChatGPT received the lowest scores, whereas DeepSeek and Gemini received higher scores. Regarding quality metrics, Gemini produced the most balanced and structured content for patient information, whereas DeepSeek showed competitive performance on certain quality criteria but occasionally yielded variable results in content consistency.

In time-based analyses, the Gemini model was distinguished by producing consistent output, whereas ChatGPT exhibited a more variable performance; however, the observed fluctuations were small in magnitude and are unlikely to be clinically or practically significant. In assessments of medical accuracy, no significant differences were observed among the three models; all produced similar scores for accuracy and content

https://doi.org/10.5455/annalsmedres.2025.11.346

Table 1. Readability indices used in the study, their formulas, and conceptual descriptions.

Index Name	Formula	Description
ARI (Automated Readability Index)	4.71 × (characters/word) + 0.5 × (words/sentence) -- 21.43	Based on average character count and sentence length; provides an approximate U.S. educational grade level.
FRES (Flesch Reading Ease Score)	206.835 -- 1.015 × (words/sentence) -- 84.6 × (syllables/word)	Higher scores indicate easier readability (range: 0--100).
GFO (Gunning Fog Index)	0.4 × [(words/sentence) + % complex words]	Estimates the number of years of education required to understand the text.
FKGL (Flesch-Kincaid Grade Level)	0.39 × (words/sentence) + 11.8 × (syllables/word) -- 15.59	Indicates the educational grade level needed to comprehend the text.
SMOG (Simple Measure of Gobbledygook)	1.0430 × √(number of polysyllabic words × (30/number of sentences)) + 3.1291	Recommended especially for health literacy; more reliable for texts exceeding 100 words.
FORCAST	20 -- (number of single-syllable words per 10 words)	Focuses solely on word level, not sentence structure; particularly suitable for documents other than prose text.

Complex Word: A word consisting of three or more syllables. Grade Level: The educational level required to comprehend the text. For example, "8th grade level" means the text can be understood by an 8th-grade student.

Table 2. List of the 30 most relevant English keywords associated with the terms "epidural delivery," "painless labor," and "painless delivery" according to Google Trends data (2004--May 2025). (Source: Google Trends, https://trends.google.com).

Keywords		
birth with epidural	epidural for normal delivery	painless delivery
birth without epidural	epidural labor	painless delivery cost
can contractions be painless	epidural painless delivery	painless delivery side effects
delivery pain	epidural pregnancy	painless labor
epidural anesthesia	epidural side effects	painless normal delivery
epidural birth	labor analgesia	painless pregnancy
epidural delivery	labor pain	what is epidural
epidural during labor	labor with epidural	what is epidural birth
epidural for birth	painless birth	what is epidural delivery
epidural for labor	painless contractions	what is painless delivery

Table 3. Comparison of readability indices, word count, and sentence count across artificial intelligence models.

Readability Index	ChatGPT (Median, (Min--Max))	DeepSeek (Median, (Min--Max))	Gemini (Median, (Min--Max))	p *	ChatGPT vs DeepSeek p †	ChatGPT vs Gemini p †	DeepSeek vs Gemini p †
ARI	12.71 (8.24--17.19)	13.33 (10.08--21.87)	13.34 (10.67--16.93)	**0.001**	**<0.001**	**0.002**	0.674
FRE	42.69 (27.85--60.67)	39.64 (2.40--54.84)	36.79 (19.16--52.02)	**<0.001**	<0.032	**<0.001**	**0.005**
GFI	9.79 (6.62--14.91)	10.08 (7.49--14.22)	10.49 (8.48--12.99)	**<0.001**	<0.166	**<0.001**	**0.001**
FKGL	11.70 (7.85--16.01)	10.23 (7.84--16.00)	12.52 (9.96--15.78)	**<0.001**	**<0.001**	**<0.001**	**<0.001**
SMOG	13.02 (8.69--16.18)	11.12 (9.35--15.05)	13.51 (11.57--16.26)	**<0.001**	**<0.001**	**<0.001**	**<0.001**
FORCAST	11.58 (10.30--14.11)	12.29 (10.77--14.13)	11.89 (10.73--12.93)	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Word Count	332.50 (82--775)	290.50 (134--424)	526.50 (159--1144)	**<0.001**	0.069	**<0.001**	**<0.001**
Sentence Count	18.00 (2--55)	28.00 (9--47)	30.00 (7--57)	**<0.001**	**<0.001**	**<0.001**	**0.004**

* Overall group comparisons were performed using the Kruskal–Wallis test. † For measures with significant differences, pairwise comparisons were conducted using the Bonferroni-corrected Mann–Whitney U test (significance level $p < 0.0167$). Abbreviations: ARI = Automated Readability Index; FRES = Flesch Reading Ease Score; GFI = Gunning Fog Index; FKGL = Flesch-Kincaid Grade Level; SMOG = Simple Measure of Gobbledygook; FORCAST = FORCAST Readability Formula.

https://doi.org/10.5455/annalsmedres.2025.11.346

Table 4. Comparison of temporal changes in readability, reliability, and quality metrics of artificial intelligence models (May 11–25, 2025).

Metric	ChatGPT (Median (min-max))			DeepSeek (Median (min-max))			Gemini (Median (min-max))		
	11.May.25	18.May.25	25.May.25	11.May.25	18.May.25	25.May.25	11.May.25	18.May.25	25.May.25
Word Count	210 (110–416)	315 (115–517)	385 (82–775)	278 (162–420)	284 (134–424)	267 (154–404)	520 (219–822)	528 (159–1144)	523 (164–796)
p *		**<0.01**			0.611			0.909	
Sentence Count	16 (2–28)	16 (6–29)	21 (4–55)	25 (12–37)	28 (9–43)	25 (12–47)	30 (12–48)	30 (9–57)	31 (7–50)
p *		**0.003**			0.57			0.995	
ARI Score	11.42 (8.40–14.56)	13.77 (10.23–17.19)	12.65 (8.24–15.28)	14.09 (10.08–19.77)	13.67 (10.29–18.07)	14.23 (11.53–21.87)	13.65 (10.67–16.93)	13.74 (11.46–16.18)	13.15 (10.95–16.80)
p *		**<0.01**			0.671			0.225	
FRES Score	42.83 (28.26–55.58)	39.50 (27.85–49.99)	43.61 (34.61–60.67)	38.74 (14.27–54.44)	39.55 (13.05–54.84)	38.17 (2.40–51.41)	35.28 (19.16–52.02)	35.56 (20.89–44.98)	37.86 (20.81–49.51)
p *		**0.024**			0.86			0.266	
Gunning Fog Score	9.85 (7.39–14.91)	9.92 (6.62–11.47)	9.62 (6.71–12.56)	10.14 (7.68–13.37)	10.12 (7.49–13.00)	9.99 (7.89–14.22)	10.61 (8.48–12.99)	10.65 (8.78–12.62)	10.47 (8.91–12.88)
p *		0.682			0.887			0.742	
FKGL Score	10.74 (8.38–13.87)	12.76 (10.21–16.01)	11.77 (7.85–14.10)	10.71 (7.84–14.60)	10.42 (8.11–15.14)	10.66 (8.36–16.00)	12.71 (9.96–15.78)	12.75 (11.16–15.32)	12.30 (10.04–15.73)
p *		**<0.01**			0.788			0.301	
SMOG Grade	11.81 (8.69–15.90)	13.71 (11.36–16.18)	13.10 (9.97–15.32)	11.51 (9.44–14.55)	11.27 (9.35–14.39)	11.25 (9.40–15.05)	13.64 (11.57–15.56)	13.74 (12.24–15.90)	13.45 (11.94–16.26)
p *		**<0.01**			0.656			0.49	
FORCAST Score	11.85 (10.54–14.11)	11.46 (10.30–12.24)	11.38 (10.51–12.14)	12.31 (11.26–14.13)	12.25 (10.77–13.17)	12.29 (11.45–13.87)	11.99 (10.73–12.93)	11.87 (11.08–12.69)	11.81 (10.99–12.57)
p *		**0.002**			0.888			0.301	
M-DISCERN	1.90 (1–3)	1.93 (1–3)	2.00 (1–3)	2.87 (2–4)	3.03 (2–4)	3.30 (2–4)	2.80 (2–4)	3.20 (2–4)	3.13 (2–4)
p *		0.898			0.126			0.12	
JAMA	0.50 (0–1)	0.47 (0–1)	0.37 (0–1)	1.37 (1–2)	1.63 (1–2)	1.43 (1–2)	1.00 (0–2)	1.13 (0–2)	0.93 (0–2)
p *		0.566			0.1			0.604	
EQIP (%)	53.7% (30.6–75.0%)	57.3% (33.3–77.8%)	54.4% (33.3–77.8%)	54.3% (33.3–69.4%)	58.6% (33.3–80.6%)	61.8% (38.9–77.8%)	63.4% (47.2–83.3%)	63.7% (47.2–86.1%)	63.4% (44.4–83.3%)
p *		0.479			**0.033**			0.99	
GQS	3.03 (2–4)	2.83 (2–4)	3.00 (2–4)	3.27 (2–4)	3.53 (2–4)	3.20 (2–4)	3.37 (2–4)	3.50 (2–4)	3.27 (2–4)
p *		0.583			0.125			0.378	

* Friedman test. Abbreviations: ARI: Automated Readability Index; FRES: Flesch Reading Ease Score; GFI: Gunning Fog Index; FKGL: Flesch-Kincaid Grade Level; SMOG: Simple Measure of Gobbledygook; FORCAST: FORCAST Readability Formula; M-DISCERN: Modified DISCERN; JAMA: Journal of the American Medical Association ;EQIP: Ensuring Quality Information for Patients; GQS: Global Quality Score.

Table 5. Comparison of medical accuracy and content completeness metrics among artificial ıntelligence models.

Metric	ChatGPT (Median, (Min–Max))	DeepSeek (Median, (Min-Max))	Gemini (Median, (Min–Max))	p *
Accuracy	8 (5-9)	8 (5-8)	8 (5-9)	0.765
Completeness	7 (5-8)	7 (5-7)	7 (5-8)	0.938

* Overall group comparisons were performed using the Kruskal–Wallis test.

completeness. Overall, the study highlighted notable differences among models in the technical sophistication, comprehensibility, reliability, and accuracy of their AI-generated content.

The ChatGPT model produced the simplest texts and at the lowest grade levels in terms of readability, a finding consistent with previous studies reporting that ChatGPT achieves the highest readability scores in patient information materials [18,19]. However, those studies also found ChatGPT inadequate in content comprehensiveness and guidance, noting that its simplified language reduced the depth of information. Similarly, in our study, although ChatGPT produced simpler

https://doi.org/10.5455/annalsmedres.2025.11.346

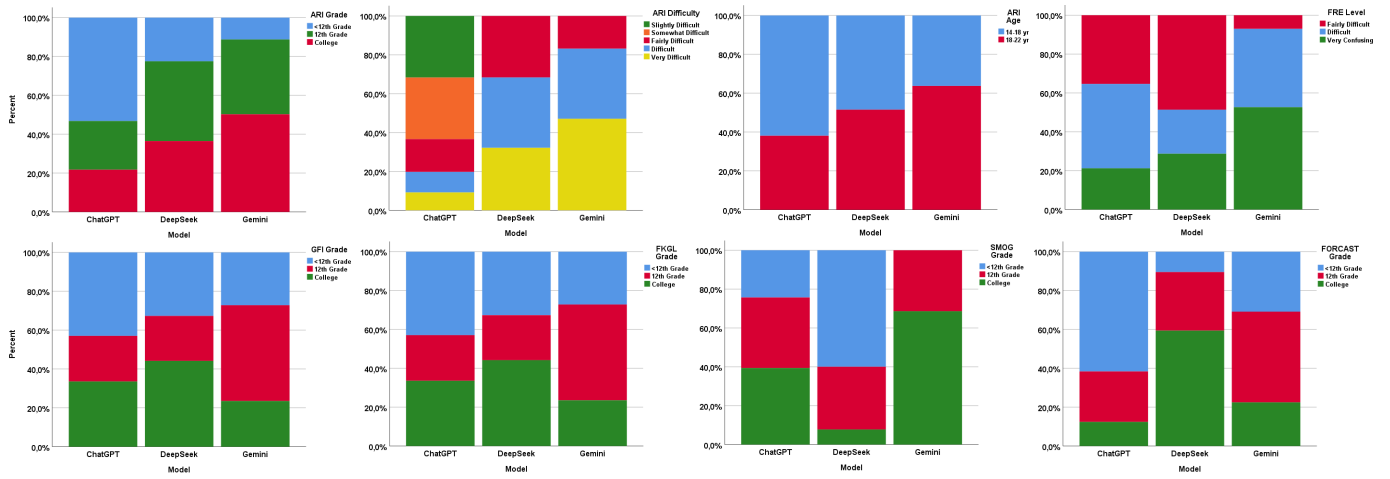

Figure 1. Grade-level classifications of texts generated by ChatGPT, DeepSeek, and Gemini across readability indices (ARI, ARI Difficulty, ARI Age, FRES, GFI, FKGL, SMOG, FORCAST). Bars represent the percentage distribution of responses in each grade-level category. Significant differences were observed across models for all indices except certain ARI subcategories (Chi-suqare tests, p<0.01). Detailed pairwise comparisons are provided in Supplementary Table 1.

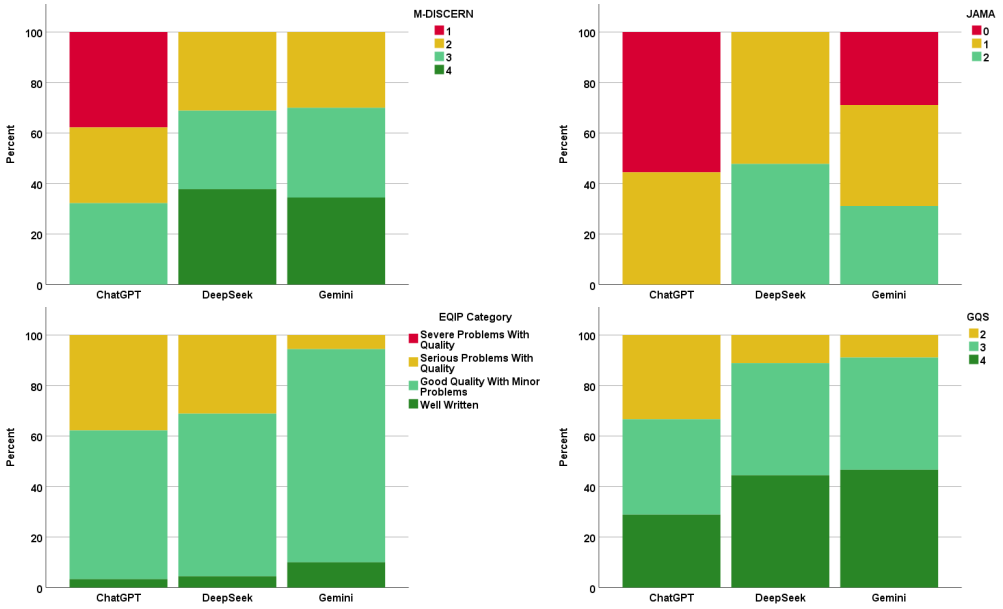

Figure 2. Reliability and quality assessments of AI-generated texts across models. Stacked bar charts show the percentage distribution of scores for M-DISCERN (top left), JAMA Benchmark Criteria (top right), EQIP-36 (bottom left), and Global Quality Score (GQS, bottom right) among ChatGPT, DeepSeek, and Gemini. Significant differences were observed across all four evaluation metrics (Chi-square tests, p<0.001). Detailed pairwise comparisons are provided in Supplementary Table 2.

texts, it scored significantly lower on reliability and quality metrics. This reflects the challenge of balancing simplification of information delivery with preservation of clinical value in the content. Notably, despite these relative differences, the absolute readability levels of outputs across all evaluated models predominantly corresponded to college-grade reading levels (FKGL and SMOG > 12), which remain well above the recommended $\leq 8^{th}$ grade level for patient education materials.

In our study, the Gemini model achieved the highest scores on reliability (JAMA, M-DISCERN) and quality (EQIP, GQS) metrics. This aligns with previous studies emphasizing Gemini's capacity to produce patient-friendly and balanced content However, those studies also reported that Gemini tends to produce higher Gunning Fog and SMOG scores, indicating that the texts are structurally more complex, which may limit comprehensibility. Similarly, in our study, Gemini generated content at the highest grade levels, particularly according to SMOG and FKGL indices, thereby reducing readability. This discrepancy may stem from Gemini's tendency to produce more comprehensive, explanatory content, which may be less accessible to individuals with limited health literacy.

Findings related to the DeepSeek model revealed a more het-

https://doi.org/10.5455/annalsmedres.2025.11.346

erogeneous profile, consistent with both the literature and the internal analyses of our study. Previous studies have reported that DeepSeek offers more technical content and provides more frequent reference links but has been criticized for higher hallucination rates and content inconsistencies [20,21]. In our study, DeepSeek demonstrated moderate performance in terms of readability but produced more variable results in quality scores compared to both ChatGPT and Gemini. Possible reasons include the model's open-source structure, incomplete optimization of its document-linking capabilities, and lack of training specifically focused on structured health-related content.

In this study, the medical accuracy and completeness scores of the three models were similar; however, the lack of source citations by any of the models limited the verifiability of the content. This finding aligns with previous reports indicating that although ChatGPT often demonstrates high content accuracy, it lacks transparency in referencing and occasionally includes misleading information [22]. Similarly, earlier studies have reported that while GPT-4 can maintain accuracy and comprehensiveness even in simplified responses, models like Bard tend to experience content loss during simplification [19]. In our study, the Gemini and DeepSeek models produced accuracy scores comparable to ChatGPT, but, like ChatGPT, did not provide sources to support clinically critical information. The absence of references poses a significant risk to patient safety, as texts may appear accurate but remain unverifiable. The findings underscore that large language models still require careful oversight for generating medical content and are not suitable as standalone tools for clinical decision support. It should be noted that instruments such as the Modified DISCERN and JAMA Benchmark Criteria were originally developed to evaluate static, authored health websites rather than conversational AI-generated outputs. Therefore, lower reliability scores may partly reflect limitations in transparency and attribution inherent to current general-purpose LLMs, rather than definitive deficits in the factual correctness of the information provided.

Previous studies focusing on obstetric anesthesia have demonstrated that large language models have significant potential for patient education, although this potential varies substantially with respect to content quality, source transparency, and readability. In a study on epidural analgesia during labor, ChatGPT was reported to perform well in terms of accuracy (97.7%) and content quality; however, the texts were generally written above the 11th-grade reading level [23]. This finding aligns with our study, in which ChatGPT, despite achieving high accuracy scores, offered only a relative advantage in readability. Another study found that both ChatGPT and Gemini provided similar accuracy rates (87–85%) when responding to frequently asked questions about childbirth; however, the Gemini model was considered superior in terms of using simpler language and delivering more actionable content [5]. In contrast, our study found that Gemini produced longer and more complex texts while achieving the highest quality scores. This difference may be attributable to the nature of the questions used in our study, which were neither directly instructive nor action-oriented, and to Gemini's tendency to offer more comprehensive explanations. In a comparative study involving American Society of Anesthesiologists (ASA) materials, Gemini outperformed ChatGPT in terms of accuracy and comprehensiveness; however, both models were found to produce more challenging texts in terms of readability compared to ASA content [17]. Similarly, in our study, the texts generated by Gemini and DeepSeek were classified at significantly higher grade levels than those recommended for health literacy. Another analysis focusing on pregnancy-related questions showed that the Gemini model achieved a higher rate of acceptable answers (83%) compared to ChatGPT (58%) and also demonstrated significantly fewer errors in source citation [24]. However, in our study, none of the evaluated models provided references, resulting in equivalent limitations regarding verifiability. In a comparative analysis focusing on obstetric pathologies, ChatGPT was found to perform significantly better than Gemini in terms of understandability and actionability [25]. Conversely, in our study, the Gemini model achieved the highest scores on the GQS and EQIP metrics, while ChatGPT exhibited lower reliability scores. This discrepancy may be due to the specific patient scenarios and prompts used in different studies, which can directly influence the text structure and content strategies of the models. Overall, consistent with findings from other studies in the obstetric field, our study highlights that, while the technical accuracy of these models is generally high, significant differences exist among models in user-centered aspects such as presentation style, readability, and content organization.

These findings from the literature largely align with the main results of our study. However, previous research has often focused on broader pregnancy-related topics rather than on specific themes such as labor analgesia. The unique aspect of our study lies in analyzing the content-generation performance of AI models on the topic of painless labor—a subject with both medical and socio-cultural implications—and in systematically revealing differences in performance through a large dataset that also includes temporal consistency assessments.

The model-based differences observed in this study are related not only to the content output but also to the structural and functional parameters governing how this output is produced. The tendency of ChatGPT to generate texts that are more readable but with lower reliability and content quality likely stems from how the model balances its prioritization of simplification against information density. Gemini's propensity to produce more technical, longer, and more instructive content may be explained by the diversity of its training data and differences in its source pool [26]. Athough the DeepSeek model demonstrated performance similar to Gemini in terms of readability in our study, it produced more variable and at

https://doi.org/10.5455/annalsmedres.2025.11.346

times superficial responses regarding content coherence and clarity of guidance. This may be attributable to the model's open-source nature, which allows continuous updates, but which has not yet undergone sufficient structured evaluation in the health domain [27]. Additionally, by structuring the evaluation around model–time–text interactions, rather than isolated outputs, this study adopted a multilayered analytical approach that allowed for a more nuanced understanding of model behavior within clinical communication contexts.

Moreover, the differences among models are related to the prompt-processing strategies and response-generation architectures they employ. For example, the tendency of ChatGPT's free versions to produce shorter and more summarized content, due to token limits, may lead to information gaps and superficial narratives [28]. Conversely, models like Gemini, which generate longer responses, may increase information coverage but adversely affect readability scores. Additionally, previous studies have shown that some models are more prone to "hallucination" behavior, incorporating non-existent sources or unverified information [11]. Although these phenomena were not quantitatively assessed in the present study, they represent well-documented concerns in the literature that may influence the perceived reliability of AI-generated health information. Another factor explaining these differences lies in the scope and currency of the datasets used during model training. Models such as ChatGPT-3.5, trained only on data up to 2021, may present certain contemporary medical practices incompletely or inaccurately, whereas newer LLMs like Gemini, with access to broader and more updated data pools, may offer enhanced content coverage [26]. However, if this advantage is not effectively integrated into the model architecture, it can result in texts that lack readability and user-friendliness.

The use of AI-generated texts in sensitive areas such as labor analgesia depends not only on technical accuracy but also on comprehensibility. Our study found that content lacking high readability remains common, potentially hindering information access for individuals with low health literacy [10]. Information that is insufficiently understood can negatively influence patients' decisions, particularly in clinical areas such as labor, where anxiety levels are high. While some models produce technically accurate yet complex content, potentially complicating patient education, overly simplified, superficial responses may create a false sense of confidence among users. Both situations underscore that, beyond accuracy, the manner in which information is presented is equally critical. Previous studies have also highlighted that while online peer-to-peer interactions can enhance access to information and support, they may simultaneously carry risks when content promotes harmful behaviors, underscoring the importance of careful moderation and presentation of health information [29].

Limitations

This study provides a systematic and multidimensional evaluation of large language models on the clinically and socially significant topic of labor analgesia. However, certain limitations should be acknowledged. First, all analyses were based on the free versions of the models; therefore, the results may not reflect the performance of premium versions, such as GPT-4. Second, textual outputs in English were evaluated, limiting generalizability to other languages and to multimedia capabilities. Third, the accuracy assessment was performed by a single expert on a limited subset of responses, which may introduce subjectivity and limit statistical generalizability. Although the texts were randomly selected and assessed in a blinded manner, the absence of multiple independent evaluators and of a standardized accuracy checklist limits findings on medical accuracy. Accordingly, the accuracy results should be interpreted as exploratory and hypothesis-generating rather than definitive. Additionally, the study employed a relative model-to-model comparison design and did not benchmark against established gold-standard patient education materials, which limits the interpretation of their absolute adequacy for clinical use. Furthermore, the use of single-turn, keyword-based prompts does not reflect the iterative and conversational nature of real-world LLM use, potentially underrepresenting the performance of models optimized for multi-turn dialogue. Lastly, evaluations did not include end-user perspectives; how patients perceive and interpret the content remains unknown.

CONCLUSION

This study demonstrates that large language models differ substantially in the informational content they generate about labor analgesia. While ChatGPT offers the highest readability and may therefore be more suitable for patient education, Gemini provides superior reliability and content quality, making it more appropriate for professional or information-dense contexts. DeepSeek demonstrates variable performance, indicating the need for further refinement.

Importantly, none of the models provided source citations, a critical limitation that undermines verifiability and clinical trust. Optimizing LLM outputs for target audiences and establishing mechanisms for transparent referencing will be essential for their safe integration into healthcare communication. Future research should incorporate real patient perspectives, premium model versions, and multilingual analyses to better assess the practical implications of AI-generated medical information.

Acknowledgements: The authors would like to thank Ali İrfan Doğan, an undergraduate student in Management Information Systems at Esenyurt University, for his valuable technical support during the data collection and archiving phases of this study.

https://doi.org/10.5455/annalsmedres.2025.11.346

Ethics Committee Approval: This study did not involve human participants, identifiable personal data, or animals, and therefore did not require approval by an institutional ethics committee.

Informed Consent: This study does not involve human participants, identifiable personal data, or animals.

Peer-review: Externally peer-reviewed.

Conflict of Interest: The authors declare that they have no competing interests related to the content of this article.

Author Contributions: Conception: A.R.D; Design: A.R.D, H.K; Supervision: H.K, A.T.T; Materials: A.R.D; Data Collection or Processing: A.R.D, H.K, A.T.T; Analysis or Interpretation: A.R.D, H.K, A.T.T; Literature Review: A.R.D, A.T.T; Writing: A.R.D; Critical Review: H.K, A.T.T.

Financial Disclosure: No external funding was received for this study.

Artificial Intelligence Disclosure: The authors declare that an artificial intelligence–based tool (ChatGPT, OpenAI, version 3.5) was used solely for language editing and improving the clarity and readability of the manuscript. The AI tool did not contribute to the study design, data analysis, interpretation of results, or scientific conclusions. All content was critically reviewed and approved by the authors, who take full responsibility for the manuscript.

REFERENCES

1. Nori W, Kassim MAK, Helmi ZR, et al. Non-Pharmacological Pain Management in Labor: A Systematic Review. *J Clin Med.* 2023;12(23):7203. doi: 10.3390/jcm12237203.
2. Fuglenes D, Aas E, Botten G, Øian P, Kristiansen IS. Why do some pregnant women prefer cesarean? The influence of parity, delivery experiences, and fear. *Am J Obstet Gynecol.* 2011;205(1):45.e1-45.e9. doi: 10.1016/j.ajog.2011.03.043.
3. Gjestvang C, Haakstad LAH. Navigating Pregnancy: Information Sources and Lifestyle Behavior Choices—A Narrative Review. *J Pregnancy.* 2024;2024(1):4040825. doi: 10.1155/2024/4040825.
4. Murphy J, Vaughn J, Gelber K, Geller A, Zakowski M. Readability, content, quality and accuracy assessment of internet-based patient education materials relating to labor analgesia. *Int J Obstet Anesth.* 2019;39:82–7. doi: 10.1016/j.ijoa.2019.01.003.
5. Lee D, Brown M, Hammond J, Zakowski M. Readability, quality and accuracy of generative artificial intelligence chatbots for commonly asked questions about labor epidurals: a comparison of ChatGPT and Bard. *Int J Obstet Anesth.* 2025;61:104317. doi: 10.1016/j.ijoa.2024.104317.
6. Ayers JW, Zhu Z, Poliak A, et al. Evaluating Artificial Intelligence Responses to Public Health Questions. *JAMA Netw Open.* 2023;6(6):e2317517. doi: 10.1001/jamanetworkopen.2023.17517.
7. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel).* 2023;11(6):887. doi: 10.3390/healthcare11060887.
8. Murugesu L, Damman OC, Timmermans DRM, et al. Health literate-sensitive shared decision-making in maternity care: needs for support among maternity care professionals in the Netherlands. *BMC Pregnancy Childbirth.* 2023;23(1):594. doi: 10.1186/s12884-023-05915-9.
9. Kripalani S, Weiss BD. Teaching about health literacy and clear communication. *J Gen Intern Med.* 2006;21(8):888–90. doi: 10.1111/j.1525-1497.2006.00543.x.
10. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low Health Literacy and Health Outcomes: An Updated Systematic Review. *Ann Intern Med.* 2011;155(2):97–107. doi: 10.7326/0003-4819-155-2-201107190-00005.
11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198.
12. Islam R, Ahmed I. Gemini-the most powerful LLM: Myth or Truth. 2024 5th Inf Commun Technol Conf ICTC. Nanjing, China, 2024, pp. 303-308. doi: 10.1109/ICTC61510.2024.10602253.
13. Peng Y, Malin BA, Rousseau JF, et al. From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare. *J Biomed Inform.* 2025;163:104791. doi: 10.1016/j.jbi.2025.104791.
14. Li Y, Ouyang H, Lin G, Peng Y, Yao J, Chen Y. Evaluation of the measurement properties of online health information quality assessment tools: A systematic review. *Int J Nurs Sci.* 2025;12(2):130–6. doi: 10.1016/j.ijnss.2025.02.015.
15. Ozdemir ZM, Yavuz SA, Surmelioglu DG. YouTube as an information source in deep margin elevation: Reliability, accuracy and quality analysis. *PLoS One.* 2025;20(2):e0318568. doi: 10.1371/journal.pone.0318568.
16. Silberg WM, Lundberg GD, Musacchio RA. Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet: Caveant Lector et Viewor—Let the Reader and Viewer Beware. *JAMA.* 1997;277(15):1244–5. doi: 10.1001/jama.1997.03540390074039.
17. Gondode PG, Singh R, Mehta S, Singh S, Kumar S, Nayak SS. Artificial intelligence chatbots versus traditional medical resources for patient education on "Labor Epidurals": an evaluation of accuracy, emotional tone, and readability. *Int J Obstet Anesth.* 2025;61:104302. doi: 10.1016/j.ijoa.2024.104302.
18. Demir M, Bedir N, Boğa D, et al. Comparison of Artificial Intelligence Responses About Penile Prosthesis implantation in Terms of Quality and Readability: A Methodological and Validity Study on ChatGPT, Gemini, and DeepSeek. *J Reconstr Urol.* 2025;15(1):21–6. doi: 10.5336/urology.2025-109438.
19. Srinivasan N, Samaan JS, Rajeev ND, Kanu MU, Yeo YH, Samakar K. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. *Surg Endosc.* 2024;38(5):2522–32. doi: 10.1007/s00464-024-10720-2.
20. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. *Front Med.* 2024;11:1477898. doi: 10.3389/fmed.2024.1477898.
21. Zhou J, Cheng Y, He S, Chen Y, Chen H. Large Language Models for Transforming Healthcare: A Perspective on DeepSeek-R1. *MedComm – Future Med.* 2025;4(2):e70021. doi: 10.1002/mef2.70021.
22. Rahimli Ocakoglu S, Coskun B. The Emerging Role of AI in Patient Education: A Comparative Analysis of the Accuracy of Large Language Models for Pelvic Organ Prolapse. *Med Princ Pract.* 2024;33(4):330–7. doi: 10.1159/000538538.
23. Tan CW, Chan JCY, Chan JJI, Nagarajan S, Sng BL. Information about labor epidural analgesia: an updated evaluation on the readability, accuracy, and quality of ChatGPT responses incorporating patient preferences and complex clinical scenarios. *Int J Obstet Anesth.* 2025;63:104688. doi: 10.1016/j.ijoa.2025.104688.
24. Khromchenko K, Shaikh S, Singh M, Vurture G, Rana RA, Baum JD. ChatGPT-3.5 Versus Google Bard: Which Large Language Model Responds Best to Commonly Asked Pregnancy Questions? *Cureus.* 2024;16(7):e65543. doi: 10.7759/cureus.65543.
25. Khemlani A, Singavarapu J, Vasa R, et al. Comparing the Accuracy of Large Language Models(LLM)in Trending Obstetrical Topics. *Open Access J Gynecol Obstet.* 2025;7(1):18–22. doi: 10.22259/2638-5244.0701003.
26. Rane N, Choudhary S, Rane J. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. *J Appl Artif Intell.* 2024;5(1):69–93. doi: 10.48185/jaai.v5i1.1052.

27. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in Healthcare: Revealing Opportunities and Steering Challenges of a New Open-Source Artificial Intelligence Frontier. *Cureus.* 2025;17(2):e79221. doi: 10.7759/cureus.79221.
28. Roberts RH, Ali SR, Hutchings HA, Dobbs TD, Whitaker IS. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. *BMJ Health Care Inform.* 2023;30(1):e100830. doi: 10.1136/bmjhci-2023-100830.
29. Ferraro G, Loo Gee B, Ji S, Salvador-Carulla, L. Lightme: analysing language in internet support groups for mental health. *Health Inf Sci Syst.* 2020;8(34):1-10. doi: 10.1007/s13755-020-00115-7.

https://doi.org/10.5455/annalsmedres.2025.11.346