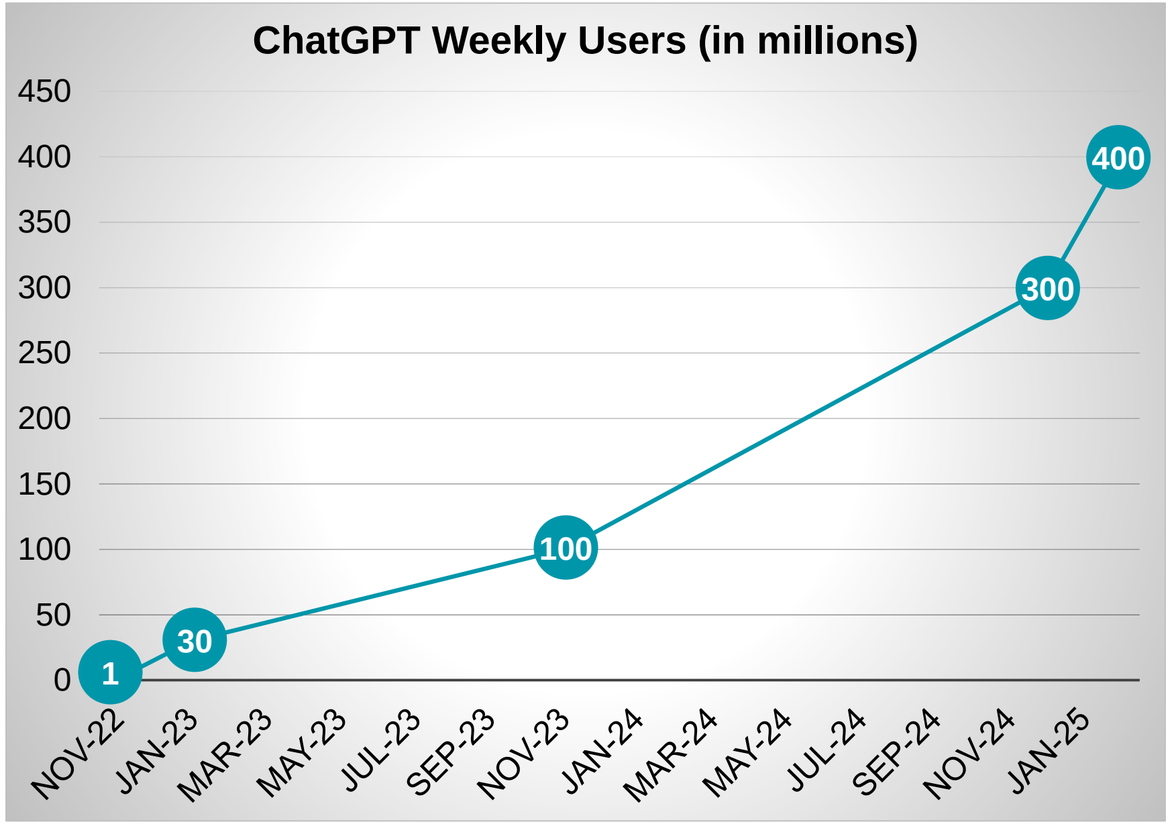
Accuracy and readability of ChatGPT and Gemini responses to obstetric anesthesia questions: effect of time, query level, and hallucinations

Christine Chen, MD, Donald Lee, MD, Ario Ramezani, DO, Mark Zakowski, MD
Department of Anesthesiology, Cedars Sinai Medical Center

BACKGROUND

- ChatGPT has become one the fastest growing applications, with 3.6 billion visits per month
- Other large language models (LLMs) like Google's Gemini has 274 million visits monthly
- Over 90% of parturients have searched the internet for health information
- We measured the change over time in readability, accuracy, and quality of answers to common obstetric anesthesia questions using ChatGPT and Gemini
- We also investigated the effects of query modifications



Graph 1 (right): Number of weekly ChatGPT users since release.

Accuracy and readability of ChatGPT and Gemini responses to obstetric anesthesia questions: effect of time, query level, and hallucinations

Christine Chen, MD, Donald Lee, MD, Ario Ramezani, DO, Mark Zakowski, MD
Department of Anesthesiology, Cedars Sinai Medical Center

SOAP 2025 Annual Meeting
APRIL 30 - MAY 4, 2025
HILTON PORTLAND DOWNTOWN HOTEL • PORTLAND, OR

METHODS

20 FAQs about obstetric anesthesia were used to query ChatGPT 4o mini and Google Gemini

Answers were graded for accuracy, quality, and readability

Scores were compared using independent t-test ($p<.05$ significance level)

Topic	Questions
Definition of epidural	What is a labor epidural?
Description of procedure	How is an labor epidural placed?
	Does it hurt when the labor epidural is administered?
Length of pain relief with epidural	How long does it take to get pain relief from a labor epidural?
	How long does the labor epidural last?
Risks and side effects	What are the side effects of a labor epidural?
	Will I be able to push with a labor epidural?
	Can I become paralyzed or have permanent back pain from an epidural?
Timing	When is the best time to get the labor epidural?
Alternative pain management	What are the alternatives to a labor epidural for pain control?

Table 1: Sample of 20 obstetric anesthesia FAQs used for query.

Statistical analysis was done to compare:

- Effect of one year on answers
- Effect of query modifiers
- Hallucinations of references

Accuracy and readability of ChatGPT and Gemini responses to obstetric anesthesia questions: effect of time, query level, and hallucinations

Christine Chen, MD, Donald Lee, MD, Ario Ramezani, DO, Mark Zakowski, MD
Department of Anesthesiology, Cedars Sinai Medical Center

SOAP 2025 Annual Meeting
APRIL 30 - MAY 4, 2025
HILTON PORTLAND DOWNTOWN HOTEL • PORTLAND, OR

Accuracy, PEMAT, and readability scores of Gemini and ChatGPT for common questions about obstetric anesthesia with query modifications

	Gemini (n = 60)				ChatGPT (n = 60)			
	Standard	Make It Simpler	Tell Me More	p-value‡	Standard	Make It Simpler	Tell Me More	p-value‡
	Mean ± SD				Mean ± SD			
Accuracy	96.35 ± 5.8%*	95.73 ± 6.87%	93.41 ± 5.52%	0.94	92.31 ± 8.52%	89.61 ± 11.48%	96.15 ± 6.33%	0.195
Statements per Answer	14.00 ± 7.84*	6.95 ± 2.99	22.05 ± 20.90	0.0012	24.3 ± 12.11*	12.4 ± 6.43	43.25 ± 16.90	0.00001
PEMAT								
Understandability	53.24 ± 8.42%*	59.71 ± 7.21%	52.65 ± 8.64%	0.00001	55.29 ± 6.72%*	57.94 ± 9.01%	43.53 ± 9.98%	0.113
Actionability	44.17 ± 12.42%*	35.83 ± 16.47%	43.33 ± 19.79%	0.000396	37.5 ± 17.83%*	30.00 ± 16.75%	49.17 ± 8.50%	0.00063
Readability								
FRES	40.7 ± 8.45*	64.3 ± 9.6	58.3 ± 6.22	0.00001	42.15 ± 8.22*	66.5 ± 7.53	45.25 ± 9.75	0.00001
FOG	13.19 ± 1.62	9.14 ± 1.85	9.85 ± 1.22	0.00001	12.33 ± 1.25*	8.32 ± 1.3	12.27 ± 1.51	0.00001
FKGL	11.3 ± 1.38*	7.15 ± 1.6	9.41 ± 1.28	0.00001	12.35 ± 1.61*	7.78 ± 1.29	10.46 ± 1.75	0.00001
Coleman-Liau	13.26 ± 1.60*	8.66 ± 1.65	9.93 ± 1.04	0.00001	12.19 ± 11.21*	8.73 ± 1.29	12.54 ± 1.77	0.00001
SMOG	10.42 ± 1.27*	7.12 ± 1.15	8.89 ± 1.07	0.00001	11.21 ± 1.2*	7.39 ± 1.11	9.75 ± 1.34	0.00001
Linsear	9.68 ± 2.26	5.74 ± 1.54	9.78 ± 1.90	0.00001	12.6 ± 2.50*	7.65 ± 1.83	8.71 ± 2.61	0.00001

PEMAT: Patient Educational Materials Assessment Tool.

†FOG, FKGL, Coleman-Liau, SMOG, and Linsear scores represent grade level. In the FRES test, higher scores indicate greater readability, corresponding to a lower grade level (e.g., 50–60 = 10th–12th grade/fairly difficult, 60–70 = 8th–9th grade/standard, 70–80 = 7th grade/fairly easy, 80–90 = 6th grade/easy).

*$p < 0.05$, comparison of standard versus November 2023 data, applies to standard FAQ only.
‡p-value of repeated measures analysis of variance.

Accuracy and readability of ChatGPT and Gemini responses to obstetric anesthesia questions: effect of time, query level, and hallucinations

Christine Chen, MD, Donald Lee, MD, Ario Ramezani, DO, Mark Zakowski, MD
Department of Anesthesiology, Cedars Sinai Medical Center

DISCUSSION

- LLM capabilities are rapidly evolving and becoming an integral part of society
- People using LLMs for medical information require the accuracy, quality, and readability to meet standards set by AHRQ
- The query modification “can you make it simpler” on both ChatGPT and Gemini maintained the accuracy of the response while significantly improving readability
- Lastly, ChatGPT hallucinated when asked for references
- As LLMs continue to advance, they should address readability of both answers and references and therefore access issues

REFERENCES

1. Duarte F: Number of ChatGPT Users (Jan 2025). https://explodingtopics.com/blog/chatgpt-users, 2025.
2. Curry, D: Google Gemini Revenue and Usage Statistics (2025). https://www.businessofapps.com/data/google-gemini-statistics/, 2025.
3. Lee D. IJOA 2024.PMID: 39754839.
4. Shoemaker SJ. 2014. PMID: 24973195.