Comparative accuracy and readability of three consumer LLMs for interventional pain procedure education

Aliza Perez, DO[1], Sundeep Kahlon, MD[1], Abby Ventura, MD[1], Oscar Dawson, MD[1], Soo Kim, MD[1]

[1] Mount Sinai Hospital

Icahn School of Medicine at Mount Sinai



INTRODUCTION

Patients increasingly use large language model (LLM) chatbots before interventional pain procedures. Prior studies have generally evaluated a single model, a single procedure, or broad pain-related questions. To our knowledge, no prior study has directly compared ChatGPT, Gemini, and Claude across multiple common interventional pain procedures using one standardized framework. We compared the accuracy, completeness, hallucination frequency, readability, and length of patient education generated by these three consumer LLMs.

MATERIAL AND METHODS

In July 2026, ChatGPT, Gemini, and Claude were queried through logged-out sessions on their free consumer interfaces. Each received an identical standardized prompt in a fresh session for five procedures: spinal cord stimulation, lumbar radiofrequency ablation, lumbar epidural steroid injection, lumbar medial branch block, and kyphoplasty, yielding 15 outputs.

Two physicians independently scored each output. Accuracy was assessed across five domains using a 0–3 scale, anchored to society guidance. Completeness was evaluated using a 10-item checklist, and outputs were screened for fabricated content. One rater, blinded to model identity, reviewed de-identified outputs in randomized order; for discordant ratings, the blinded rater's score was used.

Agreement was summarized using percentage agreement, prevalence-adjusted bias-adjusted kappa, and Gwet's AC1. Readability was measured using Flesch-Kincaid grade level against the recommended sixth-grade benchmark. Mount Sinai determined that the project did not constitute human subjects research.

RESULTS

Mean FK grade level was 11.1 (SD 0.9; range, 10.2–13.6); no output met the sixth-grade benchmark or was at or below the eighth-grade level. Mean accuracy was 10.1 of 15, and 70 of 75 domains (93.3%) were rated mostly or fully accurate; none was dangerous or false. Mean completeness was 9.5 of 10, and neither rater identified hallucinated content.

Numerically, ChatGPT had the highest accuracy (10.6/15) and lowest FK grade level (10.5) but produced the longest outputs (1,118 words). Claude had similar accuracy (10.4/15), full completeness (10.0/10), and the shortest outputs (661 words). Gemini had lower accuracy (9.4/15) and completeness (8.6/10) and the highest FK grade level (11.7).

Completeness agreement was 91.3% (PABAK, 0.83; AC1, 0.90); accuracy ratings were within one point in 74 of 75 domains.

Table 1. Readability, response length, accuracy, completeness, and hallucination results by model (n = 15 documents)

Model	Mean FK grade level	Mean length (words)	Documents at ≤6th grade	Analysis accuracy (/15)	Analysis completeness (/10)	Documents with hallucinated content
ChatGPT	10.5	1,118	0/5	10.6	10.0	0/5
Gemini	11.7	694	0/5	9.4	8.6	0/5
Claude	11.2	661	0/5	10.4	10.0	0/5
All documents	**11.1 (SD 0.9)**	**824**	**0/15**	**10.1**	**9.5**	**0/15**

FK, Flesch-Kincaid. Model-level comparisons are descriptive; no inferential statistical testing was performed.

DISCUSSION

This standardized comparison of three LLMs across five interventional pain procedures identified model-level tradeoffs and a shared accessibility deficit. ChatGPT and Claude produced more complete and accurate education than Gemini, while Claude was more concise. However, all models exceeded recommended patient reading levels. These tools may supplement clinician counseling, but professional oversight and readability constraints remain necessary.

Limitations include five procedures, single responses, one evaluation time point, and no assessment of patient comprehension.

REFERENCES

1. Lo Bianco G, Cascella M, Li S, et al. Reliability, accuracy, and comprehensibility of AI-based responses to common patient questions regarding spinal cord stimulation. J Clin Med. 2025;14(5):1453. doi:10.3390/jcm14051453.
2. Olivier T, Ma Z, Patel A, et al. Assessing ChatGPT responses to patient questions on epidural steroid injections: a comparative study of general versus specific queries. Interv Pain Med. 2025;4(2):100592. doi:10.1016/j.inpm.2025.100592.
3. Lo Bianco G, Cascella M, Natoli S, et al. Assessing artificial intelligence-powered responses to common patient questions on radiofrequency ablation and cryoanalgesia for chronic pain. J Clin Med. 2025;14(19):6814. doi:10.3390/jcm14196814.
4. Zhu S, Singh S, Richey B, et al. Can large language models reliably educate patients after kyphoplasty? A clinician-rated comparative study of ChatGPT and Gemini. J Orthop Rep. 2026:100953. doi:10.1016/j.jorep.2026.100953.
5. Weiss BD. Health Literacy: A Manual for Clinicians. Chicago, IL: American Medical Association Foundation; 2003.