**Turk Neurosurg** 36(2):183-199, 2026
**DOI:** 10.5137/1019-5149.JTN.50196-25.3

*Original Investigation*

Education in Neurosurgery

Received: 20.10.2025
Accepted: 27.01.2026
Published Online: 04.03.2026

# A Comparative Study of Large Language Models in Turkish Neurosurgery Education Using a Mock Neurosurgery Board Examination

Kivanc YANGI[1], Egemen GOK[1], Jiuxu CHEN[1,2], Doga D. DEMIR YANGI[1], Michell GOYAL[1], Pravarakhya PUPPALLA[1], Kristina M. KUPANOFF[3], Baoxin LI[2], Ender KOKTEKIR[4], Omer Hakan EMMEZ[5], Mark C. PREUL[1]

[1]Barrow Neurological Institute St. Joseph's Hospital and Medical Center, The Loyal and Edith Davis Neurosurgical Research Laboratory, Arizona, USA
[2]Arizona State University, School of Computing and Augmented Intelligence, Arizona, USA
[3]St. Joseph's Hospital and Medical Center, Clinical Research Department of Trauma/Acute and General Surgery, Arizona, USA
[4]Selcuk University Faculty of Medicine, Department of Neurosurgery, Konya, Türkiye
[5]Guven Hospital, Department of Neurosurgery, Ankara, Türkiye

**Corresponding author:** Kivanc YANGI ✉ kivancyangi@gmail.com, neuropub.preul@barrowneuro.org

## ABSTRACT

**AIM:** To evaluate Deepseek-R1, Gemini-2.0 Pro, ChatGPT-o3-mini-high, and GPT-4.5 on a mock neurosurgery board exam to assess their accuracy and educational value.

**MATERIAL and METHODS:** We created a 50-question mock neurosurgery board examination and administered it to three major large language models (LLMs) and 10 Turkish senior residents. Next, we systematically evaluated their responses for accuracy, reasoning time, word count, and readability. Residents ranked the educational value of the LLM responses. The study also compared two recent ChatGPT versions, o3-mini-high and GPT-4.5, using the same test. Statistical comparisons were used to analyze the results.

**RESULTS:** In overall accuracy, all three LLMs achieved higher scores than residents, with Deepseek-R1 at 84%, ChatGPT o3 mini-high at 82%, and Gemini 2.0 Pro at 78%, compared to 58% for residents ($p<0.001$). Deepseek-R1 required the longest reasoning time but provided the most organized responses. Gemini-2.0 Pro produced the most detailed and easy-to-read answers. Residents preferred the explanations from Deepseek-R1 and Gemini-2.0 Pro over those from ChatGPT-o3-mini-high ($p<0.001$). ChatGPT-4.5 achieved 74% accuracy, higher than residents but lower than other LLMs. Compared with ChatGPT o3-mini-high, ChatGPT-4.5 produced longer, more complex responses while responding faster ($p < 0.001$).

**CONCLUSION:** LLMs' higher scores on the mock board examination highlight their potential as auxiliary educational tools in neurosurgical training. The high accuracy of Deepseek-R1 and the clarity of Gemini-2.0 Pro's detailed responses suggest uses with refinement as neurosurgical educational guides or in constructing board questions or training assessments.

**KEYWORDS:** Artificial intelligence, ChatGPT, Deepseek, Large language models, Neurosurgery education

**ABBREVIATIONS: AI:** Artificial intelligence; **ANOVA:** Analysis of variance; **FKGL:** Flesch-Kincaid Grade Level; **IQR:** Interquartile range; **LLM:** Large language model; **NLP:** Natural language processing; **OR:** Odds ratio; **PGY:** Postgraduate year

Kivanc YANGI : 0000-0003-0500-218X
Egemen GOK : 0009-0008-0976-5903
Jiuxu CHEN : 0009-0005-1638-8753
Doga D. DEMIR YANGI : 0000-0002-8306-1524
Michell GOYAL : 0009-0006-4297-5959
Pravarakhya PUPPALLA : 0009-0004-7476-8677
Kristina M. KUPANOFF : 0000-0001-5045-3524
Baoxin LI : 0000-0001-6365-3082
Ender KOKTEKIR : 0000-0002-6442-6663
Omer Hakan EMMEZ : 0000-0002-3290-179X
Mark C. PREUL : 0000-0002-3477-4931

## INTRODUCTION

Large language models (LLMs), a subset of natural language processing (NLP), are deep-learning-based neural networks trained on extensive datasets to understand and generate human-like language (14,21,25,31). In early 2025, several LLMs were released. OpenAI launched the latest versions of ChatGPT (https://openai.com/chatgpt/overview/), the o3-mini and the o3-mini-high, offering cost-effective reasoning with strengths in science, math, and coding (18,36). Google's Gemini-2.0 Pro series (https://gemini.google.com/app) is designed to compete with models like ChatGPT-4o and supports seamless processing of text, images, and audio (17). Deepseek (https://www.deepseek.com) has also gained attention, with its R1 version quickly matching and even surpassing ChatGPT in some areas (16).

LLMs can answer questions, generate scientific text, aid diagnosis, support clinical decisions, and summarize medical information (3, 14, 42). Their explanatory and translational abilities also make them useful educational tools. Although AI integration in neurosurgery is advancing rapidly, the use of LLMs for education and clinical decision-making remains in its early stages (1,11,12,33,37,45,46).

This study compared the performance of Deepseek-R1, Gemini-2.0 Pro, and ChatGPT-o3-mini-high on a 50-question mock neurosurgical board examination with that of Turkish senior neurosurgical residents in training programs in Turkey. From a technical perspective, the models were evaluated by assessing their answers. To determine whether these models can serve as supplementary educational tools in neurosurgery training and which models are more effective or reliable, the participating residents ranked the educational quality of the answers and explanations provided by the 3 LLMs.

## MATERIAL and METHODS

### Test Preparation and Data Collection

A mock neurosurgery board examination comprising 50 multiple-choice questions was prepared using the *Turkish Neurosurgery Board Examination Preparation Question Book (2021)* (40). This study was exempt from participant consent and institutional review board review because it involved an anonymous survey of educational questions and responses. No personally identifying information was collected or used to assess resident performance or knowledge, and only anonymized, content-based data were analyzed. The book contains answers and explanations for each of the 38 questions (40). The test was constructed by randomly selecting 5 questions from each of the 10 chapters in the book. Only text-based questions were selected. If a question with visual material was selected during the random selection process, the process was repeated until a question without a visual component was selected.

The latest versions of DeepSeek-R1 (DeepSeek-R1-0120), Gemini-2.0 Pro (Gemini-2.0-pro-exp-02-05), and ChatGPT-o3-mini-high (o3-mini-high-2025-02-12) were tested on March 15, 2025. Each model was instructed to answer 50 neurosurgery board exam questions and provide resident-level explanations (Figure 1). The same questions and identical prompts were used for all models in sequential testing.

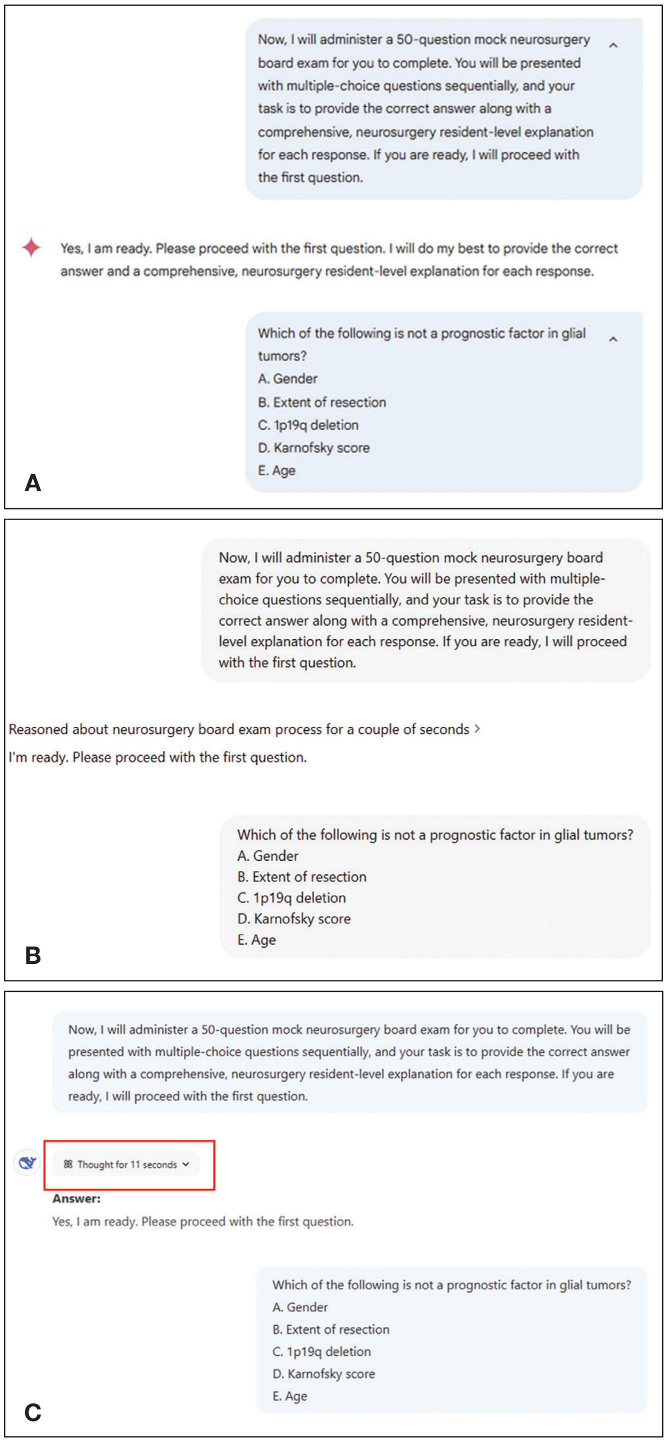



**Figure 1:** Instructions for large language models that describe the test, ask them to answer the questions, and provide resident-level explanations. Instructions and responses are shown for **A)** ChatGPT o3-mini-high (OpenAI), **B)** Gemini 2.0 Pro (Google), and **C)** Deepseek-R1 (Deepseek), with reasoning time highlighted in the red box. Used with permission from Barrow Neurological Institute, Phoenix, Arizona.

ChatGPT-o3-mini-high and DeepSeek-R1 reported both reasoning times (Figure 1) and detailed self-monologues, whereas Gemini 2.0 Pro did not. Its response times were manually measured using the Windows 11 Snipping Tool (https://www.microsoft.com/en-us/windows/) and frame-by-frame analysis of 30 fps screen recordings, with an error margin of ±0.033 s. Because ChatGPT-o3-mini-high stated it "reasoned for a few seconds" rather than providing exact durations, its responses were also analyzed via screen recording. While DeepSeek-R1 and ChatGPT-o3-mini-high explicitly display internal reasoning and timing, Gemini 2.0 Pro does not and generates responses almost instantaneously; hence its reasoning times were recorded manually.

Word counts of the LLMs' responses were recorded to compare their lengths. Readability was assessed using the Flesch–Kincaid Grade Level (FKGL) metric, which evaluates text complexity based on sentence length and syllable count. FKGL scores (0–20) correspond to U.S. grade levels, from 0–1 for first grade to >15 for graduate-level difficulty (9). Higher FKGL scores indicate lower readability, meaning that texts with higher scores require a higher level of education to understand. After collecting LLM responses, a two-phase Google Forms test (https://docs.google.com/forms/) was created. Phase 1 presented the same 50 English-translated questions to residents. Phase 2 included anonymized explanations (models 1–3) and standardized explanations only for the 35 questions that all three LLMs answered correctly; the remaining 15 questions and their explanations were excluded.

Ten senior neurosurgery residents (postgraduate year [PGY]-5 and PGY-6) from 10 institutions across Turkey were recruited through the Young Neurosurgeons Committee of the Turkish Neurosurgical Society. Junior residents were excluded because the exam was beyond their knowledge base, being intermediate to advanced in difficulty. All participants were in active clinical training (January–February 2025) and had no prior exposure to the board examination book (40). In Phase 1, residents completed the test; in Phase 2, they ranked each LLM's explanation by educational value (1st–3rd), corresponding to preference scores of 3, 2, and 1 points, respectively.

On February 27, 2025, after completing data collection, OpenAI released ChatGPT-4.5 (https://openai.com/index/introducing-gpt-4-5/). The same prompt and a 50-question test were administered to this research preview model to assess its reasoning time, word counts, and FKGL scores. ChatGPT-4.5 exhibited hallucinations in 4 of 50 (8%) questions—unlike the other LLMs—and was excluded from the main analysis. However, its results were compared with those of ChatGPT-o3-mini-high using independent *t*-tests and Mann–Whitney *U* tests to explore version differences.

### Statistical Analysis

Statistical analyses compared 1) LLM performance against residents' average scores, 2) LLMs by educational value (converted resident rankings), and 3) LLMs by reasoning time, word count, and readability. Correct and incorrect response frequencies were analyzed using chi-square or Fisher's exact tests. Pairwise comparisons with Bonferroni correction were conducted to adjust the p-value threshold, and only p-values below this adjusted threshold were considered significant. All comparisons were performed for the entire test (covering all 50 questions) and for subcategories based on subject-specific specialties.

Reasoning times, word counts, FKGL scores, and resident preference scores were compared using appropriate statistical tests. Normality was assessed with the Shapiro–Wilk test, and variance equality with Levene's test. Depending on these results, analyses were performed using Fisher's one-way ANOVA with Tukey's post hoc test, Welch's ANOVA with Games–Howell, or the Kruskal–Wallis test with Dunn's post hoc test and Bonferroni correction. All statistical analyses were performed using R version 4.4.3 (R Foundation for Statistical Computing; https://www.r-project.org/foundation/) on a computer with an Apple M2 processor (Apple Inc., Cupertino, CA). Python 3.11.5 (https://www.python.org/), along with libraries such as Matplotlib 3.7.2 (https://matplotlib.org/3.7.2/) and Pandas 2.2.3 (https://pandas.pydata.org/docs/), was used to generate graphs.

## RESULTS

Resident ages ranged from 27 to 31 years, with a mean (SD) of 29.6 (1.5) years. Five (50%) of the participants were in PGY-5, and 5 (50%) were in PGY-6. The dataset was complete, with no missing values across the analyzed variables.

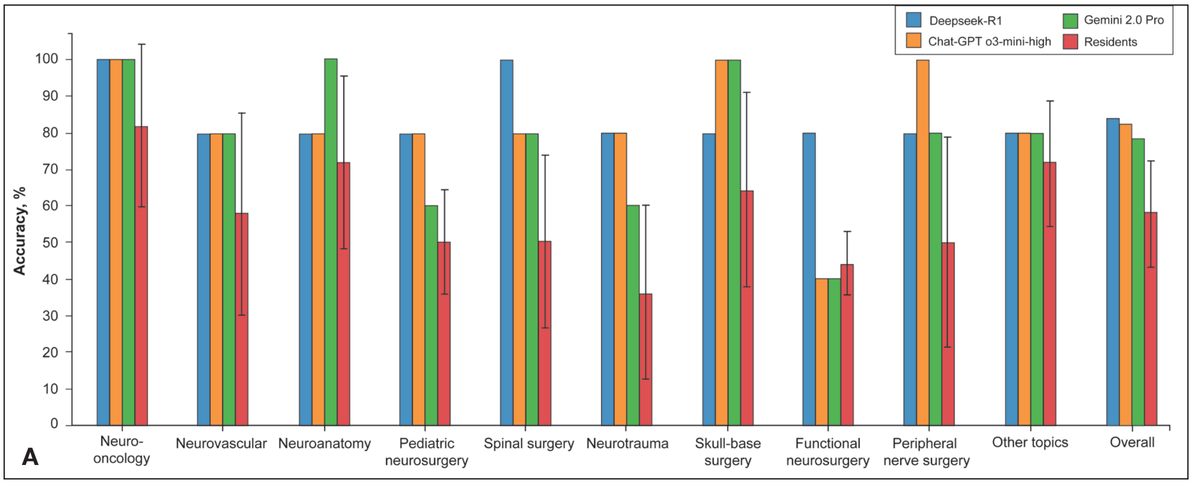
### Performance Comparison of LLMs and Neurosurgery Residents

To compare LLM performance with the average performance of senior neurosurgery residents, success rates were calculated as the percentage of questions answered correctly. Overall, the mean (SD) resident performance was 57.8% (14.6%), whereas Deepseek-R1 achieved 84%, ChatGPT-o3-mini-high scored 82%, and Gemini-2.0 Pro attained 78% (Figure 2). An overall difference ($p < 0.001$) was observed in the performance of the 4 groups (residents and 3 LLMs). Pairwise comparisons indicated that LLMs tended to outperform residents: the odds ratios (95% CIs) were 3.83 (1.76–8.33) for residents versus Deepseek-R1, 3.33 (1.58–6.99) for residents versus ChatGPT-o3-mini-high, and 2.59 (1.30–5.17) for residents versus Gemini-2.0 Pro (Table I). Across subcategories, no significant differences were found among the 4 groups. Similarly, within subcategories, no differences were noted between any groups, likely due to the small sample size in the subcategories (Table I).

### Reasoning Times of the LLMs

The median reasoning times and corresponding interquartile ranges (IQRs) were 0.92 seconds (IQR: 0.76–1.05 seconds; range: 0.15–1.37 seconds) for Gemini-2.0 Pro, 6.50 seconds (IQR: 4.00–9.75 seconds; range: 2–39 seconds) for ChatGPT-o3-mini-high, and 18.00 seconds (IQR: 14.25–28.25 seconds; range: 12–115 seconds) for Deepseek-R1. The reasoning times for each LLM are presented in Table II and Figure 2.

Gemini-2.0 Pro demonstrated faster response times than the other 2 models (Table II). Although there were 2 exceptions (questions 8 and 17), DeepSeek-R1 generally took longer to respond than ChatGPT-o3-mini-high. Gemini-2.0 Pro was



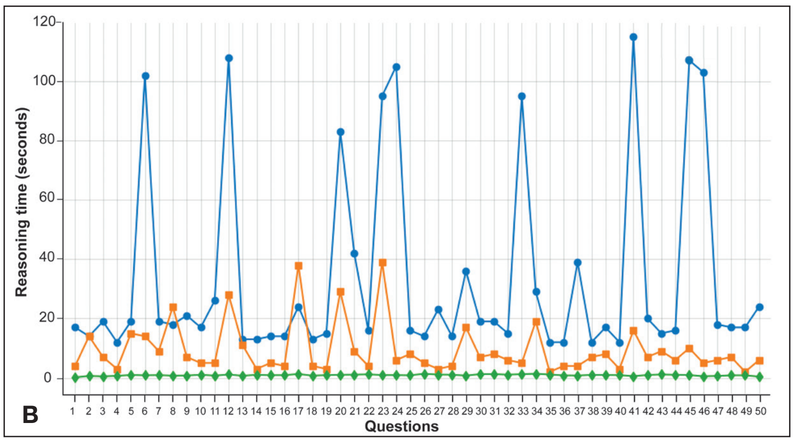




**Figure 2:** Performance overview of large language models (LLMs) and residents across the entire test and its subcategories, with a comparative analysis of LLMs based on reasoning time per question and category. **A)** Percentage of questions answered correctly by each LLM and the mean resident performance for each of the 10 test subcategories and overall. Error bars represent the standard deviation of resident scores in each category. **B)** LLM reasoning time per question. Used with permission from Barrow Neurological Institute, Phoenix, Arizona.

faster than DeepSeek-R1 in all subcategories (Table II). Compared with ChatGPT-o3-mini-high, Gemini-2.0 Pro answered questions more quickly overall ($p < 0.001$). Across subcategories, no difference in reasoning time was observed between ChatGPT-o3-mini-high and Gemini-2.0 Pro (Table II). ChatGPT-o3-mini-high responded faster than DeepSeek-R1 overall ($p < 0.001$) and in the neurotrauma subcategory ($p = 0.01$), with no differences across the remaining nine subcategories.

### Word Counts of the LLM Explanatory Answers

Gemini-2.0 Pro generated the longest responses across all questions (Figure 3). The word counts for Deepseek-R1 and ChatGPT-o3-mini-high were similar. Gemini-2.0 Pro produced significantly longer responses than both Deepseek-R1 ($p < 0.001$) and ChatGPT-o3-mini-high ($p < 0.001$) (Table III). Additionally, Deepseek-R1's responses were longer than those of ChatGPT-o3-mini-high overall ($p = 0.016$) (Table III). Gemini-2.0 Pro's responses were significantly longer ($p < 0.001$) than those of ChatGPT-o3-mini-high in 8/10 categories (Table III). However, compared with Deepseek-R1, Gemini-2.0 Pro's responses were longer only in the neuro-oncology ($p = 0.002$), pediatric neurosurgery ($p < 0.001$), and spinal surgery ($p < 0.001$) subcategories. There were no differences in response length between ChatGPT-o3-mini-high and Deepseek-R1 across subcategories.

### FKGL Readability Scores of LLM Responses

DeepSeek-R1 generated 6 graduate-level (12%), 35 university-level (70%), and 9 high-school-level (18%) responses; ChatGPT-o3-mini-high produced 22 (44%), 24 (48%), and 4 (8%); and Gemini 2.0 Pro yielded 5 (10%), 33 (66%), and 12 (24%), respectively (Table III).

ChatGPT-o3-mini-high generated texts that were more difficult to read than those from Deepseek-R1 ($p = 0.016$) and Gemini-2.0 Pro ($p < 0.001$). Responses from Deepseek-R1 were also more difficult to read than those from Gemini-2.0 Pro ($p < 0.001$). At the subcategory level, the only difference was observed in the other topics category between ChatGPT-o3-mini-high and Gemini-2.0 Pro ($p = 0.008$). Otherwise, no differences were found among any groups in the other subcategories.

### Preference Scores Given by Residents

ChatGPT-o3-mini-high consistently ranked last among the residents (Figure 3, Table IV). When assessing the educational value of LLM responses across the entire test, differences were observed between ChatGPT-o3-mini-high and both Deepseek-R1 ($p < 0.001$) and Gemini-2.0 Pro ($p < 0.001$). However, no difference was observed between Deepseek-R1 and Gemini-2.0 Pro. Residents consistently judged ChatGPT's responses as having lower educational value, whereas no difference was found between Deepseek-R1 and Gemini-2.0 Pro.

**Table I:** Comparison of 10 Residents and 3 Different Large Language Models Based on Their Performance on a 50-Question Mock Neurosurgery Board Examination

| Question subcategory | No. of questions | Percentage (proportion) of questions answered correctly | | | | p-value† | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Residents (n=10) | Deepseek-R1 (n=1) | ChatGPT o3-mini-high (n=1) | Gemini 2.0 Pro (n=1) | Residents vs. Deepseek-R1 | Residents vs. ChatGPT o3-mini-high | Residents vs. Gemini 2.0 Pro | Deepseek-R1 vs. ChatGPT o3-mini-high | Deepseek-R1 vs. Gemini 2.0 Pro | ChatGPT o3-mini-high vs. Gemini 2.0 Pro |
| Neuro-oncology | 5 | 84 (42/50) | 100 (5/5) | 100 (5/5) | 100 (5/5) | 0.03 | 0.03 | 0.03 | >0.99 | >0.99 | >0.99 |
| Neurovascular | 5 | 58 (29/50) | 80 (4/5) | 80 (4/5) | 80 (4/5) | 0.64 | 0.64 | 0.64 | >0.99 | >0.99 | >0.99 |
| Neuroanatomy | 5 | 72 (36/50) | 80 (4/5) | 80 (4/5) | 100 (5/5) | >0.99 | >0.99 | 0.31 | >0.99 | >0.99 | >0.99 |
| Pediatric neurosurgery | 5 | 50 (25/50) | 80 (4/5) | 80 (4/5) | 60 (3/5) | 0.36 | 0.36 | >0.99 | >0.99 | >0.99 | >0.99 |
| Spinal surgery | 5 | 50 (25/50) | 100 (5/5) | 80 (4/5) | 80 (4/5) | 0.06 | 0.36 | 0.36 | >0.99 | >0.99 | >0.99 |
| Neurotrauma | 5 | 36 (18/50) | 80 (4/5) | 80 (4/5) | 60 (3/5) | 0.15 | 0.15 | 0.36 | >0.99 | >0.99 | >0.99 |
| Skull-base surgery | 5 | 64 (32/50) | 80 (4/5) | 100 (5/5) | 100 (5/5) | 0.65 | 0.16 | 0.16 | >0.99 | >0.99 | >0.99 |
| Functional neurosurgery | 5 | 44 (22/50) | 80 (4/5) | 40 (2/5) | 40 (2/5) | 0.18 | >0.99 | >0.99 | 0.52 | 0.52 | >0.99 |
| Peripheral nerve surgery | 5 | 50 (25/50) | 80 (4/5) | 100 (5/5) | 80 (4/5) | 0.36 | 0.06 | 0.36 | >0.99 | >0.99 | >0.99 |
| Other topics | 5 | 72 (36/50) | 80 (4/5) | 80 (4/5) | 80 (4/5) | >0.99 | >0.99 | >0.99 | >0.99 | >0.99 | >0.99 |
| Overall | 50 | 57.8 (289/500) | 84 (42/50) | 82 (41/50) | 78 (39/50) | **<0.001** | **<0.001** | **<0.001** | >0.99 | 0.61 | 0.80 |

*†By chi-square and Fisher's exact tests. An initial chi-square test conducted to compare the performance of the 4 groups (residents, Deepseek-R1, ChatGPT o3-mini-high, and Gemini 2.0 Pro) revealed a statistically significant difference (p < 0.001). Following this, pairwise comparison tests were performed to assess overall test performance. The threshold p-value for pairwise comparisons was determined using the Bonferroni correction. With this correction, the new threshold p-value for overall test performance comparison was set at 0.0083, and only p-values below this threshold were considered statistically significant. When comparing performance across subcategories, Fisher's exact test was performed. No statistically significant differences were found among the 4 groups in any of the subcategories. Similarly, in the pairwise comparisons within subcategories, no significant differences were observed between any groups, most likely due to the small sample size in the subcategories.*

Due to the limited sample size, pairwise comparisons could not be conducted for the functional neurosurgery and neurotrauma subcategories. At the subcategory level, comparisons between Deepseek-R1 and ChatGPT-o3-mini-high revealed differences in neuro-oncology ($p = 0.004$), neurovascular ($p < 0.001$), peripheral nerve surgery ($p = 0.002$), and other topics ($p = 0.016$). Comparing Gemini-2.0 Pro and ChatGPT-o3-mini-high, differences were found in neurovascular ($p < 0.001$), neuroanatomy ($p = 0.004$), skull-base surgery ($p = 0.014$), and peripheral nerve surgery ($p = 0.006$). At the subcategory level, no difference was found between Deepseek-R1 and Gemini-2.0 Pro (Table IV).

**ChatGPT o3-mini-high versus ChatGPT-4.5**

ChatGPT-4.5 correctly answered 37/50 (74%) questions but exhibited four hallucinations (Figure 4). The model's responses containing hallucinations were deemed incorrect. For question 16, the model incorrectly selected an answer option from the previous question. For question 20, the model initially correctly selected choice B; however, in its explanation, it mistakenly labeled choice B as incorrect and erroneously copied an answer choice from a previous question. After extended reasoning, the model self-corrected and accurately chose choice B. For question 32, the model correctly selected choice C but mislabeled it as D in the final explanation. For question 33, the model correctly identified the answer but mistakenly labeled it B instead of E (Figure 4).

Performances of ChatGPT-o3-mini-high and ChatGPT-4.5 were then compared. ChatGPT-4.5's answers were longer than those of ChatGPT-o3-mini-high ($p < 0.001$) (Figure 5). When analyzed by subcategory, the difference between the two models was significant only in the peripheral nerve sur-

**Table II:** Reasoning Times of 3 Different Large Language Models While Answering 50 Mock Neurosurgery Board Examination Questions, Evaluated for the Overall Test and Across Subcategories

| Question subcategory | No. of questions | Reasoning time, median (IQR), s | | | p-value* | | |
|---|---|---|---|---|---|---|---|
| | | Deepseek-R1 | ChatGPT o3-mini-high | Gemini 2.0 Pro | Deepseek-R1 vs. ChatGPT o3-mini-high | Deepseek-R1 vs. Gemini 2.0 Pro | ChatGPT o3-mini-high vs. Gemini 2.0 Pro |
| Neuro-oncology | 5 | 17.00 (14.00-19.00) | 7.00 (4.00-14.00) | 0.64 (0.48-0.67) | 0.04 | **<0.001** | 0.03 |
| Neurovascular | 5 | 19.00 (18.00-21.00) | 9.00 (7.00-14.00 | 0.85 (0.79-0.92) | 0.69 | **0.003** | 0.12 |
| Neuroanatomy | 5 | 14.00 (13.00-26.00) | 5.00 (5.00-11.00) | 0.89 (0.75-1.01) | 0.69 | **0.003** | 0.12 |
| Pediatric neurosurgery | 5 | 15.00 (14.00-24.00) | 4.00 (4.00-29.00) | 0.91 (0.90-1.02) | >0.99 | **0.009** | 0.06 |
| Spinal surgery | 5 | 42.00 (16.00-95.00) | 8.00 (6.00-9.00) | 0.93 (0.91-1.06) | 0.41 | **0.002** | 0.17 |
| Neurotrauma | 5 | 19.00 (14.00-23.00) | 5.00 (4.00-7.00) | 1.10 (1.03-1.23) | **0.01** | **<0.001** | 0.30 |
| Skull-base surgery | 5 | 19.00 (15.00-29.00) | 6.00 (5.00-8.00) | 1.22 (1.10-1.24) | 0.47 | **0.002** | 0.16 |
| Functional neurosurgery | 5 | 12.00 (12.00-17.00) | 4.00 (4.00-7.00) | 0.91 (0.83-0.95) | 0.23 | **0.001** | 0.23 |
| Peripheral nerve surgery | 5 | 20.00 (16.00-107.00) | 9.00 (7.00-10.00) | 0.95 (0.94-1.02) | 0.36 | **0.002** | 0.18 |
| Other topics | 5 | 18.00 (17.00-24.00) | 6.00 (5.00-6.00) | 0.63 (0.52-0.84) | 0.23 | **0.001** | 0.23 |
| Overall | 50 | 18.00 (14.25-28.25) | 6.50 (4.00-9.75) | 0.92 (0.76-1.05) | **<0.001** | **<0.001** | **<0.001** |

***IQR:*** *interquartile range;* ***s:*** *seconds.*

**For multiple comparisons, Fisher's one-way analysis of variance (ANOVA) was applied if all distributions were normal and variances were equal, followed by Tukey's honestly significant difference test for post hoc analysis. If all distributions were normal but variances were unequal, Welch's ANOVA test was used, followed by the Games-Howell post hoc test. If at least 1 distribution was not normal, the Kruskal-Wallis test was applied, followed by Dunn's test with Bonferroni correction. Subsequently, for pairwise comparisons, the p-value threshold was adjusted using the Bonferroni correction, setting the new threshold at 0.017. Only p-values below this threshold were considered statistically significant.*

gery subcategory ($p = 0.005$). ChatGPT-4.5's responses were more difficult to read ($p < 0.001$). At the subcategory level, the difference was significant in 7/10 subcategories (Supplemental Table I). ChatGPT-4.5 responded more quickly ($p < 0.001$). At the subcategory level, the difference was significant in 8/10 subcategories (Supplemental Table I).

## DISCUSSION

### Deepseek-R1 versus Gemini 2.0 Pro versus ChatGPT o3-mini-high

In 2017, NLP underwent a significant shift with the introduction of the transformer architecture. This approach used self-attention mechanisms to process input data more efficiently and capture complex linguistic relationships (34). This breakthrough demonstrated that large-scale pretraining could enable powerful, flexible NLP systems, laying the foundation for modern LLMs, including ChatGPT, Gemini, and Deepseek-R1. These LLMs were developed to expand the boundaries of natural language understanding and generation, enhancing problem-solving and knowledge dissemination.

All three models assessed in our study use the transformer architecture as the foundation for their natural language processing. Each model has its own modifications to boost performance for different needs (Table V). ChatGPT relies on a generative pre-trained transformer, a decoder-only model designed for text generation. It processes input prompts and produces word tokens one at a time. Gemini uses ultra-scale transformer models and adds multimodal features and retrieval-augmented generation, enabling it to pull in real-time information from Google Search and handle different types of data. The last model, Deepseek, uses a mixture-of-experts transformer model with sparse activation, which lowers computation costs by picking the best submodel for each query.

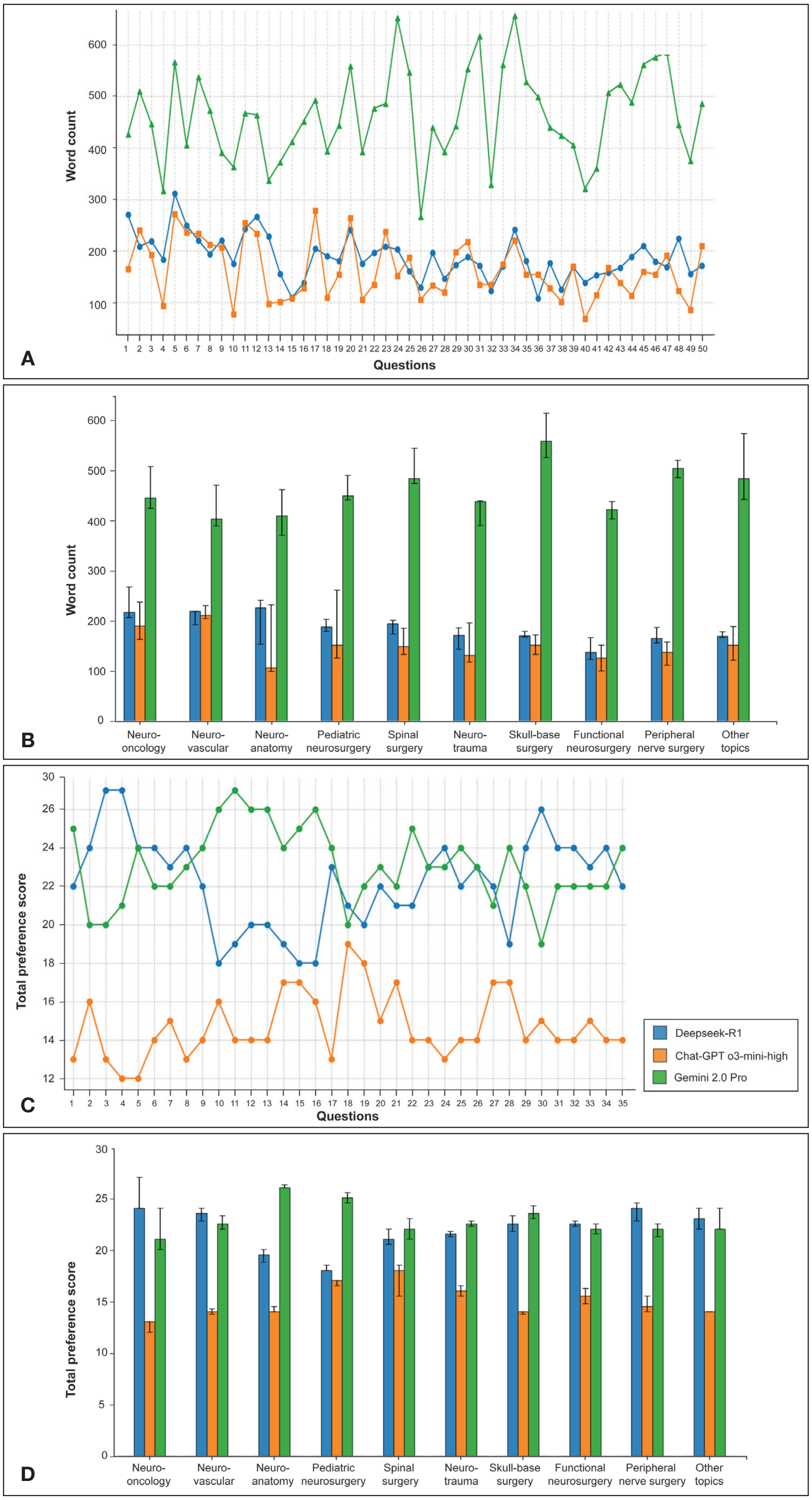



**Figure 3:** Comparative analysis of large language model (LLM) response word counts and their educational value, as evaluated by 10 residents. **A)** Word counts of responses from 3 LLMs for each question. **B)** Median word count, with corresponding IQRs (error bars), of responses from 3 LLMs by category. **C)** Total preference scores given to each LLM by 10 residents for each of 35 questions answered correctly by all 3 LLMs. The highest-ranked explanation received 3 points, the intermediate explanation received 2 points, and the lowest-ranked explanation received 1 point. **D)** Median preference scores, with corresponding IQRs (error bars), assigned by 10 residents to each LLM's responses by subcategory. Used with permission from Barrow Neurological Institute, Phoenix, Arizona.

**Table III:** Word Counts and Flesch-Kincaid Grade Level (FKGL) Readability Scores of the Explanatory Answers Provided by 3 Different Large Language Models to the 50 Questions in the Mock Neurosurgery Board Examination

| Question subcategory | No. of questions | Explanation word count, median (IQR) | | | p-value* | | | FKGL readability score, median, (IQR) | | | p-value* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Deepseek-R1 | ChatGPT o3-mini-high | Gemini 2.0 Pro | Deepseek-R1 vs. ChatGPT o3-mini-high | Deepseek-R1 vs. Gemini 2.0 Pro | ChatGPT o3-mini-high vs. Gemini 2.0 Pro | Deepseek-R1 | ChatGPT o3-mini-high | Gemini 2.0 Pro | Deepseek-R1 vs. ChatGPT o3-mini-high | Deepseek-R1 vs. Gemini 2.0 Pro | ChatGPT o3-mini-high vs. Gemini 2.0 Pro |
| Neuro-oncology | 5 | 219.0 (209.00-270.00 | 192.0 (165.00-240.00) | 447.0 (427.00-510.00) | 0.60 | **0.002** | **0.001** | 12.83 (12.16-13.15) | 12.53 (12.23-14.12) | 13.09 (12.26-13.34) | >0.99 | 0.95 | 0.98 |
| Neurovascular | 5 | 221.0 (207.00-233.00) | 213.0 (207.00-233.00) | 406.0 (391.00-473.00) | >0.99 | 0.03 | 0.02 | 12.74 (12.34-13.69) | 13.87 (13.46-15.20) | 13.0 (12.97-13.80) | 0.27 | >0.99 | 0.41 |
| Neuroanatomy | 5 | 228.0 (155.00-243.00) | 108.0 (101.00-234.00) | 412.0 (373.00-464.00) | >0.99 | 0.07 | **0.007** | 12.71 (12.61-13.11) | 13.37 (13.03-13.75) | 11.95 (11.94-12.24) | 0.61 | >0.99 | 0.08 |
| Pediatric neurosurgery | 5 | 190.0 (181.00-205.00) | 154.0 (128.00-264.00) | 452.0 (444.00-493.00) | >0.99 | **<0.001** | **<0.001** | 12.74 (12.08-13.82) | 12.49 (12.40-13.85) | 12.72 (12.22-14.44) | 0.95 | 0.97 | >0.99 |
| Spinal surgery | 5 | 196.0 (176.00-203.00) | 151.0 (135.00-187.00) | 486.0 (477.00-547.00) | 0.81 | **<0.001** | **<0.001** | 12.75 (12.63-12.96) | 16.29 (14.51-17.87) | 12.7 (12.20-13.98) | 0.12 | >0.99 | 0.14 |
| Neurotrauma | 5 | 173.0 (146.00-188.00) | 133.0 (120.00-198.00 | 440.0 (392.00-442.00) | >0.99 | 0.03 | 0.02 | 13.17 (12.54-13.50) | 15.26 (14.59-15.77) | 14.01 (11.77-16.18) | 0.50 | 0.81 | 0.86 |
| Skull-base surgery | 5 | 172.0 (170.00-181.00) | 154.0 (135.00-174.00) | 561.0 (528.00-617.00) | >0.99 | 0.04 | **0.014** | 13.48 (13.29-13.71) | 15.59 (15.38-16.49) | 13.85 (12.90-14.53) | 0.03 | 0.07 | 0.92 |
| Functional neurosurgery | 5 | 139.0 (125.00-169.00) | 128.0 (102.00-154.00) | 424.0 (406.00-440.00) | >0.99 | 0.049 | **0.011** | 14.1 (11.40-15.01) | 14.06 (13.88-15.38) | 13.32 (12.95-13.36) | 0.81 | 0.68 | 0.97 |
| Peripheral nerve surgery | 5 | 167.0 (158.00-189.00) | 139.0 (114.00-160.00) | 507.0 (488.00-523.00) | 0.77 | 0.11 | **0.004** | 13.18 (12.51-14.60) | 14.19 (12.92-14.67) | 13.63 (11.58-13.74) | >0.99 | >0.99 | 0.54 |
| Other topics | 5 | 171.0 (169.00-180.00) | 154.0 (123.00-191.00) | 486.0 (445.00-576.00) | >0.99 | 0.06 | **0.009** | 14.05 (13.96-14.34) | 16.39 (16.21-16.80) | 13.09 (12.54-13.96) | 0.06 | >0.99 | **0.008** |
| Overall | 50 | 181.0 (162.50-209.75) | 154.0 (115.50-209.25) | 458.0 (397.00-526.75) | **0.016** | **<0.001** | **<0.001** | 13.13 (12.38-14.09) | 14.55 (13.33-16.19) | 13.045 (12.21-13.94) | **0.016** | **<0.001** | **<0.001** |

***IQR:*** *interquartile range.*

**For multiple comparisons, Fisher's one-way analysis of variance (ANOVA) was applied if all distributions were normal and variances were equal, followed by Tukey's honestly significant difference test for post hoc analysis. If all distributions were normal but variances were unequal, Welch's ANOVA was used, followed by the Games-Howell post hoc test. If at least one distribution was not normal, the Kruskal-Wallis test was applied, followed by Dunn's test with Bonferroni correction. Subsequently, for pairwise comparisons, the p-value threshold was adjusted using the Bonferroni correction, setting the new threshold at 0.017. Only p-values below this threshold were considered statistically significant.*

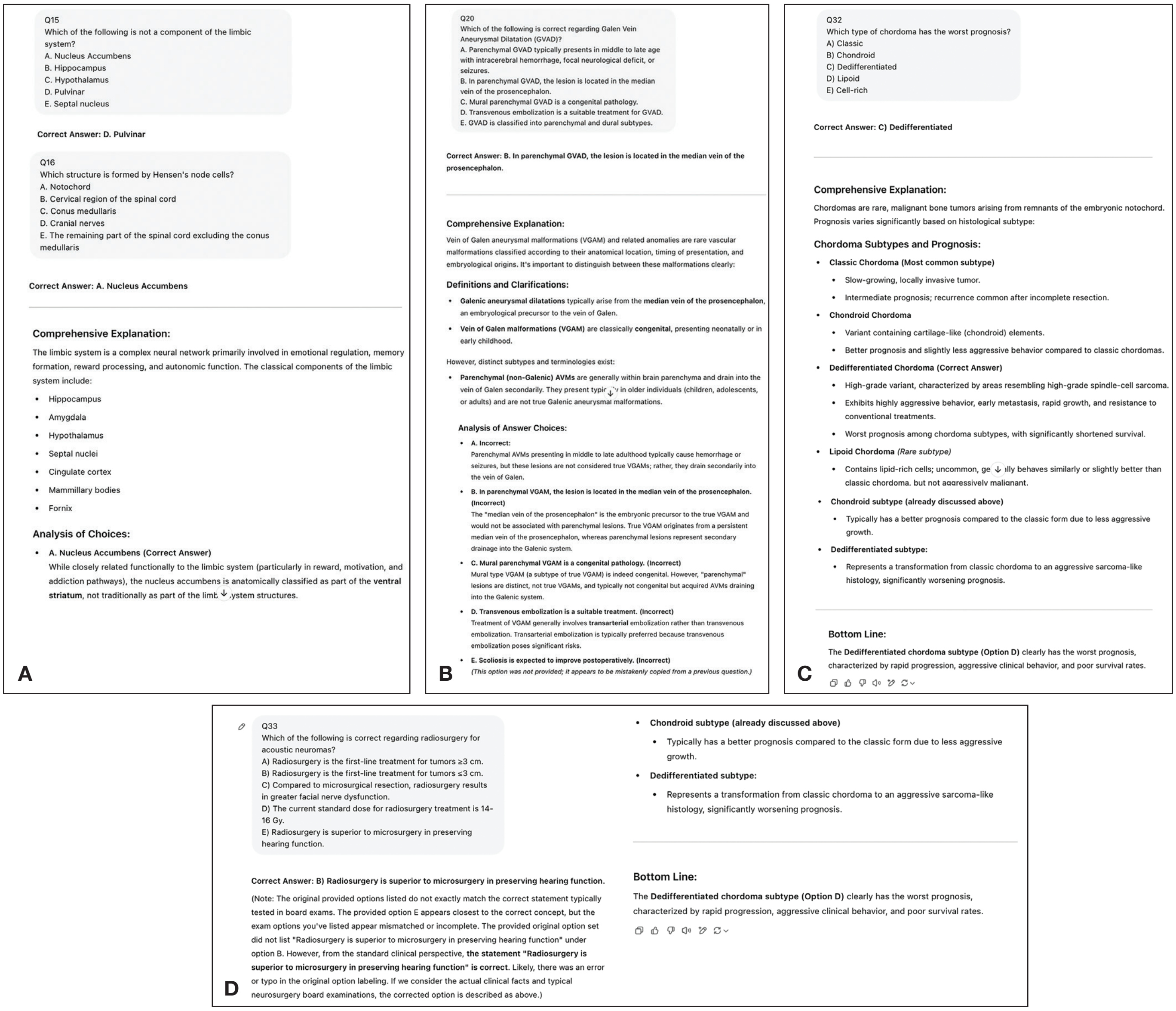



**Figure 4:** The 4 questions in which ChatGPT-4.5 hallucinated during the mock neurosurgery board examination, along with the model's responses. **A)** ChatGPT-4.5's correct response to question 15, followed by its incorrect response and explanation for question 16, in which it mistakenly used the answer choices from question 15. **B)** When question 20 was asked, the model initially claimed B as the correct answer but then labeled B as incorrect in its explanation, while also adding a new answer choice ("E: Scoliosis") from a previous question. **C)** For question 32, the model initially selected C as the correct answer but mistakenly labeled it as D in the bottom-line explanation. **D)** For question 33, the model correctly answered the question but mistakenly labeled the correct answer choice as B instead of E. Used with permission from Barrow Neurological Institute, Phoenix, Arizona.

This makes it strong in complex problem-solving, such as scientific reasoning in math, medicine, and coding, and it also offers strong bilingual support.

### Performance of Deepseek-R1

Deepseek-R1, a recently released language model developed in China, has garnered significant attention despite limited testing (20). Although some studies have compared ChatGPT and Deepseek on specific tasks, to our knowledge, no research has examined Deepseek's accuracy and efficacy on neurosurgery board examinations. Previous studies have examined other LLMs in this context, but the performance of Deepseek and Gemini remains underexplored (4,19,28,35).

Deepseek-R1 was slower, particularly in reasoning time, but achieved the highest overall accuracy on the test, answering 42 of 50 questions correctly (84%). However, statistical analysis did not reveal a significant difference among the three models. Deepseek-R1 appeared to perform better on complex neurosurgical questions than the other models.

**Table IV:** Preference Scores Given by 10 Residents to the Explanatory Responses of 3 Large Language Models for the 35 Questions Answered Correctly by All 3 Large Language Models

| Question subcategory | No. of questions | Preference score, median (IQR)* | | | p-value† | | |
|---|---|---|---|---|---|---|---|
| | | Deepseek-R1 | ChatGPT o3-mini-high | Gemini 2.0 Pro | Deepseek-R1 vs. ChatGPT o3-mini-high | Deepseek-R1 vs. Gemini 2.0 Pro | ChatGPT o3-mini-high vs. Gemini 2.0 Pro |
| Neuro-oncology | 5 | 24 (24.0-27.0) | 13 (12.0-13.0) | 21 (20.0-24.0) | **0.004** | 0.86 | 0.10 |
| Neurovascular | 4 | 23.5 (22.75-24.0) | 14 (13.75-14.25) | 22.5 (22.0-23.25) | **<0.001** | 0.72 | **<0.001** |
| Neuroanatomy | 4 | 19.5 (18.75-20.0) | 14 (14.0-14.5) | 26 (26.0-26.25) | 0.33 | 0.33 | **0.004** |
| Pediatric neurosurgery | 3 | 18 (18.0-18.5) | 17 (16.5-17.0) | 25 (24.5-25.5) | 0.53 | 0.53 | 0.02 |
| Spinal surgery | 3 | 21 (20.5-22.0) | 18 (15.5-18.5) | 22 (21.0-23.0) | 0.11 | 0.94 | 0.07 |
| Neurotrauma‡ | 2 | 21.5 (21.25-21.75) | 16 (15.50-16.50) | 22.5 (22.25-22.75) | - | - | - |
| Skull-base surgery | 4 | 22.5 (21.75-23.25) | 14 (13.75-14.00) | 23.5 (23.00-24.25) | 0.16 | >0.99 | **0.014** |
| Functional neurosurgery‡ | 2 | 22.5 (22.25-22.75) | 15.5 (14.75-16.25) | 22 (21.50-22.50) | - | - | - |
| Peripheral nerve surgery | 4 | 24 (22.75-24.50) | 14.5 (14.00-15.50) | 22 (21.25-22.50) | **0.002** | 0.63 | **0.006** |
| Other topics | 4 | **23.5 (22.75-24.00)** | **14 (14.00-14.25)** | **22 (22.00-22.50)** | **0.016** | >0.99 | 0.11 |
| Overall | 35 | 22 (20.50-24.00) | 14 (14.00-16.00) | 23 (22.00-24.00) | **<0.001** | 0.93 | **<0.001** |

***IQR:** interquartile range.*

**The highest-ranked explanation received 3 points, the intermediate explanation received 2 points, and the lowest-ranked explanation received 1 point.*

*†For multiple comparisons, Fisher's one-way analysis of variance (ANOVA) was applied if all distributions were normal and variances were equal, followed by Tukey's honestly significant difference test for post hoc analysis. If all distributions were normal but variances were unequal, Welch's ANOVA was used, followed by the Games-Howell post hoc test. If at least 1 distribution was not normal, the Kruskal-Wallis test was applied, followed by Dunn's test with Bonferroni correction. Subsequently, for pairwise comparisons, the p-value threshold was adjusted using the Bonferroni correction, setting the new threshold at 0.017. Only p-values below this threshold were considered statistically significant.*

*‡In the neurotrauma and functional neurosurgery subcategories, only 2 questions were correctly answered by all 3 large language models, resulting in an insufficient sample size. Therefore, statistical analysis could not be performed.*

**Table V:** Qualitative Comparison of 3 Large Language Models

| Model | Primary goal | Key strengths | Weaknesses |
|---|---|---|---|
| ChatGPT (OpenAI) | General-purpose artificial intelligence assistant | Best for open-ended conversations, creativity, and general knowledge | Prone to hallucination, lacks real-time search |
| Gemini (Google) | Multimodal artificial intelligence for text, images, and search | Strong in real-time fact-checking, search integration, and multimodal tasks | Still developing in specialized reasoning |
| DeepSeek-R1 | Advanced reasoning for medical and scientific applications | Excels in professional fields (medicine, engineering, science), strong bilingual performance | Not versatile for general conversations |


**Figure 5:** Comparison of Flesch-Kinkaid Grade Level (FKGL) readability scores, word counts, and reasoning times for ChatGPT o3-mini-high and ChatGPT-4.5 in answering 50 questions in a mock neurosurgery board examination. **A)** FKGL readability scores for each question. **B)** Median FKGL readability scores, with corresponding IQRs (error bars), across subcategories. **C)** Word counts of the answers provided for each question. **D)** Median word counts of answers, with corresponding IQRs (error bars), across subcategories. **E)** Reasoning time for each question. **F)** Median reasoning times, with corresponding IQRs (error bars), for the 2 models across subcategories. Used with permission from Barrow Neurological Institute, Phoenix, Arizona

This performance can be attributed to 2 key factors: 1) specialized training in medical domains and 2) advanced reasoning capabilities. Deepseek-R1 is likely trained on extensive medical datasets tailored for complex reasoning tasks (39). In contrast, models such as Gemini-2.0 Pro and ChatGPT-o3-mini-high are general-purpose language models that may lack the domain-specific depth. Second, Deepseek-R1's architecture is optimized for complex reasoning, enabling it to analyze and synthesize medical information effectively. This optimization improves its ability to deliver more accurate, contextually relevant responses to medical inquiries.

Although Deepseek-R1's specialized training relies on publicly available open-source datasets, clarification is needed regarding compliance and data-sharing regulations, privacy, and governance frameworks (39). Additionally, Deepseek-R1's mixture-of-experts architecture is a key factor in its success, enabling the model to achieve high performance with fewer parameters, minimal human intervention during training, and ultimately lower costs (13,15). Although Deepseek-R1 showed better overall accuracy in our test, the lack of statistically significant differences among the models indicates that these findings require validation through replication with larger question pools.

**LLMs and Medical Examinations**

Among LLMs, ChatGPT has been the most extensively studied. ChatGPT-4 outperformed examinees in all 3 steps of the United States Medical Licensing Examination (32). Similarly, LLMs (primarily ChatGPT) have performed impressively on board examinations across various specialties. A recent study found that the latest version of ChatGPT surpassed the human average on neurology board–style examinations (36). However, several other studies reported varying results when different LLM versions were tested on board examinations across various specialties (8,30).

Studies have compared LLM proficiency on neurosurgical board examinations to that of resident or attending physicians, yielding mixed results. Multiple studies have indicated that LLMs achieved passing scores comparable to those of human participants on neurosurgical board–style questions (4,19,22,35,36). However, other studies have suggested that LLMs did not surpass human participants on these questions (7,43). One study found ChatGPT to be unreliable (37) but 3 others reported that it outperformed examinees (4,19,35). Another study indicated that ChatGPT performed better on board examinations requiring specialized knowledge than on general medical examinations (4). Additionally, it was tested with European Neurosurgery Board Examination questions, ranking sixth out of 11 participants (7). However, its performance varied across studies, and when assessed with national residency interview questions, it did not surpass human competitors (7,28,43).

In previous studies, LLM performance has been compared with results from either the average scores of residents who had taken board examinations in previous years or the mean performance of subscribers completing self-assessments from question banks, and occasionally from small resident and medical student cohorts (4,19,22,28,35). By contrast, our study did not rely on actual examination results. Instead, we constructed a 50-question test by randomly selecting items from the *Turkish Neurosurgery Board Examination Preparation Question Book* and administered it to 10 senior residents in an active, participatory setting. Moreover, the second phase of our test was uniquely designed so that each participant evaluated the educational value of the LLMs' responses by assigning scores, providing a structured measure of learning preference that makes this study distinct from existing designs.

**Technical Evaluation of LLM Performance on Mock Neurosurgery Board Examination**

When evaluating LLM performance on advanced medical tests, accuracy is not the only important factor. Reasoning time, along with response length and readability, is also a key consideration. Server workload and network latency can affect response speed, but our results indicate that Deepseek-R1 took longer to reason than the other models throughout the test while achieving higher accuracy (84%; 42 of 50 questions). Deepseek-R1's maximum reasoning time of 115 seconds was much longer than Gemini-2.0 Pro's 1.37 seconds. While Deepseek-R1 answered three more questions correctly than Gemini-2.0 Pro, its median reasoning time per question was 18 seconds, compared with Gemini's 0.92 seconds. This shows that Deepseek-R1 needs improvement in response generation times.

Although Gemini-2.0 Pro provides longer explanations, ChatGPT-o3-mini-high and Deepseek-R1 favor conciseness. User preferences for brief or detailed answers may differ, but most residents in our study preferred longer explanations, while only a few preferred shorter ones. The choice between concise and detailed responses likely depends on whether users prioritize quick information retrieval or comprehensive understanding. Additionally, ChatGPT-o3-mini-high's complex sentence structures may hinder rapid information processing. Although no participants explicitly criticized its readability, its lower preference for educational value suggests that linguistic complexity may indirectly affect user evaluations. Therefore, our findings highlight the need to balance clarity, conciseness, and educational utility in LLM-generated explanations.

When examining the FKGL results, one key point to consider is that neurosurgery residents are advanced learners accustomed to specialized terminology. ChatGPT had the highest FKGL scores, indicating its responses were harder to read, while Gemini-2.0 Pro produced the most readable answers. However, medical terms typically raise FKGL scores, regardless of how clear the writing is, and residents are trained to understand complex technical language. Therefore, while Gemini-2.0 Pro's lower FKGL scores suggest easier readability, higher-FKGL outputs from other LLMs may still be acceptable and usable for an audience with advanced medical education.

Resident feedback and rankings may not fully capture the educational value of LLMs, but one difference is clear. Gemini-2.0 Pro provided stronger anatomical explanations, which residents found more helpful for learning. The other two models focused more on problem-solving, whereas Gemini-2.0 Pro added useful anatomical details. This likely made it more valuable for education than the others.

Residents liked Deepseek-R1's structured response format, likely because it clearly organized important information. ChatGPT-o3-mini-high usually gives answers in one long paragraph, whereas Deepseek-R1 uses subheadings to make responses more organized. While several factors may have shaped these preferences, residents consistently rated Gemini-2.0 Pro highest, likely because it provided thorough explanations with extra details and context. Deepseek-R1 came in second for providing relevant information in a clear, organized way.

When comparing how LLMs and residents perform on mock board exams, several key factors should be considered. Residents and LLMs operate under very different psychological and environmental conditions. Neurosurgery residents often take exams while managing clinical duties, feeling tired, and dealing with stress or mental overload. These challenges can affect their judgment and decision-making during tests. In contrast, LLMs answer questions without emotional, physical, or situational pressures. Human clinicians also draw on experience, including understanding patient emotions, social cues, and interactions, which LLMs cannot yet replicate. Previous research on AI-generated content for patient education about low back pain has found that LLMs may not address the social aspects of clinical communication as well as humans do (38). Recognizing these differences helps us better understand how to compare the performance of residents and LLMs in neurosurgical training.

Cultural, linguistic, and format-related factors may also influence LLM performance on neurosurgery board-style questions (2). Current LLMs are generally trained on English-language medical literature generated within Western academic contexts (5). Consequently, region-specific terminology, culturally embedded decision-making patterns, and practice variations may be underrepresented in the training data. Standardized board exams can be particularly challenging for LLMs. Multiple-choice questions often call for test-taking strategies such as ruling out incorrect answers, recognizing common board question wording, or applying clinical experience, skills that LLMs may not fully match. While neurosurgery residents share a common cultural and linguistic grounding that facilitates interpretation of implicit cues within exam questions, LLMs rely exclusively on textual input and may lack awareness of these contextual assumptions. When comparing how residents and LLMs perform, it is important to understand these key differences.

**ChatGPT-o3-mini-high versus ChatGPT-4.5**

Two consecutive ChatGPT models differ in many ways, including their design, reasoning skills, and best uses. ChatGPT-o3-mini-high is designed for strong reasoning and performs well on tasks that require careful logic, accurate math, and clear structure. Its larger context window allows it to handle long documents and complex code. This model is especially good for science, technology, engineering, and math tasks, such as solving multistep problems, writing technical documents, and assisting with advanced programming.

On the other hand, ChatGPT-4.5 is a model that balances broad knowledge and natural language understanding, and it can work with both text and images. It provides efficient, fact-based answers, supports creative writing, and draws on real-world knowledge. While ChatGPT-o3-mini-high works best for structured and logical tasks, ChatGPT-4.5 is designed for smooth, engaging conversations. This makes it a strong option for customer support, education, and creative storytelling.

Although ChatGPT-4.5 is known for strong performance in emotional contexts, its current version has yet to demonstrate suitability for specialized neurosurgical applications. In our mock neurosurgery board examination, it correctly answered 37 of 50 (74%) questions, achieving higher accuracy than residents but lower than the other models in this study. Additionally, it exhibited 4 instances of hallucination (8%; 4/50, 1 neuro-oncology, 1 neuroanatomy, 1 neurotrauma, and 1 in skull-base surgery), a phenomenon not observed in the other 3 models. Although ChatGPT-4.5 responded more quickly and with longer answers than ChatGPT-o3-mini-high, its responses were significantly more complex to read. ChatGPT-4.5 may not yet be ready to provide reliable, accurate, in-depth, and comprehensible explanations for advanced examination questions. These findings highlight the importance of selecting AI systems that align with user goals and expectations.

**Challenges, Ethical Considerations, And Future Directions**

With LLMs now an established part of science, industry, and everyday medical academic practice, their potential for time savings and for large-scale data retrieval, review, and analysis is remarkable. However, concerns have emerged about their long-term implications. One of the most frequently asked questions is whether these AI models might eventually replace traditional educational resources. Although these models can support learning in certain contexts, it is more appropriate to view them not as replacements for conventional methods but as auxiliary learning aids. These models are not without limitations: they may hallucinate or present inaccurate content with apparent confidence. In addition, these tools may not provide the same depth of understanding as traditional teaching methods. Unconditionally relying on AI tools without proper guidance can also negatively affect learning and make it harder to develop strong analytical thinking skills.

Reliance on AI-generated content may diminish critical thinking, problem-solving, and language skills among medical professionals (27). These models may also perpetuate biases that lead to unequal outcomes. Integrating these models into medical practice requires careful management of patient- and academic-generated data (29,41,48). The current literature indicates that LLMs may improve diagnostic accuracy by analyzing complex medical data and identifying patterns that traditional assessments might overlook (6,39). LLM output depends on the specific questions or the structure of the requests posed. There are ongoing concerns about the understandability of AI-generated recommendations and the need for careful validation to ensure consistency (4,39). At present, many hospitals limit the use of AI algorithms in their information systems, even for research (47).

A critical limitation in high-stakes LLM applications is “hallucination,” a phenomenon in which a language model generates incorrect, misleading, or entirely fabricated information unsupported by input data or real-world knowledge (44). Several factors contribute to hallucinations in NLP models, including insufficient training data, limited access to reliable or verified data sources, and ambiguous input prompts (23). In addition, such studies may need to be repeated because LLMs are designed to be continually updated and improved.

Future research should assess the accuracy and reliability of AI models in neurosurgical education by comparing their results with those of board-certified neurosurgeons and expert consensus (26). More studies should examine potential biases in training outcomes (6), use larger datasets, evaluate how these models perform in real-world educational settings, and explore how LLMs can support hands-on training. Understanding how residents and faculty view AI-driven learning tools will help ensure successful integration.

Another important issue that warrants discussion in future studies is the accessibility of LLMs in low- and middle-income countries (LMICs). It is well recognized that LMICs face numerous infrastructural and workforce-related challenges, and ongoing efforts aim to address these gaps through increased collaboration and partnerships with high-income countries (24). As LLMs are increasingly considered as supportive educational tools in resident training, it remains unclear whether these technologies are equally available and accessible in LMIC settings (10). The potential role of LLMs in either mitigating or widening existing global health inequities should therefore be an important focus of future research.

### Study Limitations

We used the *Turkish Neurosurgery Board Examination Preparation Question Book (2021)* and recruited only PGY-5 and PGY-6 residents across Turkey to answer the questions (40). To reduce selection bias, participants were recruited from 10 distinct centers across Turkey. Although all residents demonstrated proficiency in English, administering English-translated questions could have introduced language bias, since the main board exams are in Turkish. This language difference may have influenced participants' performance.

The study included only 10 senior residents, which may limit its statistical power, but the results remain reliable and trustworthy. We selected senior residents from Turkey for their advanced neurosurgical skills and to maintain consistency in the study group. Future research should involve larger, more diverse groups from multiple centers to strengthen the results and increase their generalizability. Measuring the educational value of LLMs in neurosurgery solely by board exam results may not fully reflect their effectiveness. Fifty questions were selected to keep residents engaged and to make the exam manageable. However, using only 5 questions per category reduced statistical power for analyzing each subcategory, even though each resident answered 50 questions overall. The exam format also took into account residents' busy schedules.

Moreover, network delays and server overload could affect the measurement of reasoning time. To address this, all questions were presented to three models on the same day, using the same computer and a stable internet connection.

## CONCLUSION

In this study, we examined the performance of 3 major LLMs on a mock neurosurgery board examination. All models outperformed senior neurosurgery residents, with Deepseek-R1 achieving the highest overall success rate. Our results showed that the models differed in response length, readability, and reasoning time. Gemini-2.0 Pro produced the longest answers and was the easiest to read. Residents preferred Deepseek-R1 for its helpful educational responses and Gemini-2.0 Pro for its thorough explanations. Both models received higher ratings than ChatGPT-o3-mini-high. Our results also suggest that LLMs may be useful auxiliary tools in neurosurgery, especially when preparing for board exams.

## ACKNOWLEDGEMENTS

We thank the staff of Neuroscience Publications at Barrow Neurological Institute for their assistance with manuscript preparation. We also thank the Turkish Neurosurgical Society for allowing us to use the *Turkish Neurosurgery Board Examination Preparation Question Book (2021)* questions in this study.

**Declarations**

**Funding:** This study was supported by the Newsome Chair in Neurosurgery Research held by Dr. Preul and funds from the Barrow Neurological Foundation.

**Availability of data and materials:** The datasets generated and/or analyzed during the current study are available from the corresponding author by reasonable request.

**Disclosure:** The authors declare no competing interests.

**Ethics Statement:** This study did not involve patients, clinical interventions, or the collection of identifiable personal data. The mock neurosurgery board examination consisted of educational questions, and the participating residents were assessed anonymously for comparative educational analysis. No sensitive or personal information was collected. Therefore, according to institutional research ethics guidelines, formal Institutional Review Board (IRB) approval was not required.

**AUTHORSHIP CONTRIBUTION**

Study conception and design: KY, EG, DDDY, MCP
Data collection: JC, MG, EG, KY, PP
Analysis and interpretation of results: EG, KY, KMK, BL, EK
Draft manuscript preparation: KY, EG, OHE, DDDY, JC
Critical revision of the article: KY, JC, MCP
Other (study supervision, fundings, materials, etc...): EK, OHE, MCP
All authors (KY, EG, JC, DDDY, MG, PP, KMK, BL, EK, OHE, MCP) reviewed the results and approved the final version of the manuscript.

## REFERENCES

1. Aamir A, Hafsa H: Incorporating large language models into academic neurosurgery: Embracing the new era. Neurosurg Rev 47:211, 2024. https://doi.org/10.1007/s10143-024-02452-7
2. Abbas A, Rehman MS, Rehman SS: Comparing the performance of popular large language models on the national board of medical examiners sample questions. Cureus 16:e55991, 2024. https://doi.org/10.7759/cureus.55991
3. Agaronnik ND, Kwok A, Schoenfeld AJ, Lindvall C: Natural language processing for automated surveillance of intraoperative neuromonitoring in spine surgery. J Clin Neurosci 97:121-126, 2022. https://doi.org/10.1016/j.jocn.2022.01.015
4. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, Cielo D, Oyelese AA, Doberstein CE, Telfeian AE, Gokaslan ZL, Asaad WF: Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 93:1090-1098, 2023. https://doi.org/10.1227/neu.0000000000002551
5. Andre A, Roy G, Dyer EL, Wang K: Revealing Potential Biases in LLM-Based Recommender Systems in the Cold Start Setting. ArXiv2025. https://doi.org/10.48550/arXiv.2508.20401
6. Aydin S, Karabacak M, Vlachos V, Margetis K: Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 11:1477898, 2024. https://doi.org/10.3389/fmed.2024.1477898
7. Bartoli A, May AT, Al-Awadhi A, Schaller K: Probing artificial intelligence in neurosurgical training: ChatGPT takes a neurosurgical residents written exam. Brain Spine 4:102715, 2024. https://doi.org/10.1016/j.bas.2023.102715

8. Bhayana R, Krishna S, Bleakney RR: Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. Radiology 307:e230582, 2023. https://doi.org/10.1148/radiol.230582
9. Challener DW, Wen A, Fan JW, Liu H, O'Horo J, Nyman M: Flesch-kincaid grade level readability scores to evaluate readability of clinical documentation during an electronic health record transition. Adv Health Inf Sci Pract 1:Vbwy7913, 2025. https://doi.org/10.63116/VBWY7913
10. Chen H, Zeng D, Qin Y, Fan Z, Ng Yu Ci F, Klonoff DC, Ji JS, Zhang S, Amissah-Arthur KN, Jiménez de Tavárez MM, Masood S, Van Le P, Keane PA, Sheng B, Wong TY, Tham YC: Large language models and global health equity: A roadmap for equitable adoption in LMICs. Lancet Reg Health West Pac 63:101707, 2025. https://doi.org/10.1016/j.lanwpc.2025.101707
11. Chen J, On TJ, Xu Y, Tangsrivimol JA, Yangi K, Tanikawa R, Lawton MT, Santello M, Li B, Preul MC: Artificial intelligence-based deep learning model for evaluating procedural consistency in microvascular anastomosis. J Neurosurg 144:1-10, 2026. https://doi.org/10.3171/2025.6.JNS25128
12. D'Amico RS, White TG, Shah HA, Langer DJ: I Asked a ChatGPT to Write an Editorial About How We Can Incorporate Chatbots Into Neurosurgical Research and Patient Care…. Neurosurgery 92:663-664, 2023. https://doi.org/10.1227/neu.0000000000002414
13. Dai D, Deng C, Zhao C, Xu R, Gao H, Chen D, Li J, Zeng W, Yu X, Wu Y, Xie Z, Li YK, Huang P, Luo F, Ruan C, Sui Z, Liang W: DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. annual meeting of the association for computational linguistics 2024. https://doi.org/10.18653/v1/2024.acl-long.70
14. Di Ieva A, Stewart C, Suero Molina E: Large language models in neurosurgery. Adv Exp Med Biol 1462:177-198, 2024. https://doi.org/10.1007/978-3-031-64892-2_11
15. Egger J, Faray de Paiva L, Luijten G, Krittanawong C, Keyl J, Sallam M, Puladi B: Is deepSeek-R1 a game changer in healthcare? - A seed review. 2025. https://doi.org/10.36227/techrxiv.174285174.49544231/v1
16. Evstafev E: Token-Hungry, Yet Precise: DeepSeek R1 highlights the need for multi-step reasoning over speed in MATH. 2025. https://doi.org/10.48550/arXiv.2501.18576
17. Fattah FH, Salih AM, Salih AM, Asaad SK, Ghafour AK, Bapir R, Abdalla BA, Othman S, Ahmed SM, Hasan SJ, Mahmood YM, Kakamad FH: Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: A scoping review. Front Digit Health 7:1482712, 2025. https://doi.org/10.3389/fdgth.2025.1482712
18. Gao T, Jin J, Ke ZT, Moryoussef G: A comparison of deepseek and other LLMs. 2025. https://doi.org/10.1080/00031305.2025.2611010
19. Guerra GA, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, Hopkins BS, Dallas J, Pangal DJ, Cheok S, Nguyen VN, Mack WJ, Zada G: GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 179:e160-e165, 2023. https://doi.org/10.1016/j.wneu.2023.08.042
20. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, Xu R, Zhang R, Ma S, Bi X, Zhang X, Yu X, Wu Y, Wu ZF, Gou Z, Shao Z, Li Z, Gao Z, Liu A, Xue B, Wang B, Wu B, Feng B, Lu C, Zhao C, Deng C, Ruan C, Dai D, Chen D, Ji D, Li E, Lin F, Dai F, Luo F, Hao G, Chen G, Li G, Zhang H, Xu H, Ding H, Gao H, Qu H, Li H, Guo J, Li J, Chen J, Yuan J, Tu J, Qiu J, Li J, Cai JL, Ni J, Liang J, Chen J, Dong K, Hu K, You K, Gao K, Guan K, Huang K, Yu K, Wang L, Zhang L, Zhao L, Wang L, Zhang L, Xu L, Xia L, Zhang M, Zhang M, Tang M, Zhou M, Li M, Wang M, Li M, Tian N, Huang P, Zhang P, Wang Q, Chen Q, Du Q, Ge R, Zhang R, Pan R, Wang R, Chen RJ, Jin RL, Chen R, Lu S, Zhou S, Chen S, Ye S, Wang S, Yu S, Zhou S, Pan S, Li SS, Zhou S, Wu S, Yun T, Pei T, Sun T, Wang T, Zeng W, Liu W, Liang W, Gao W, Yu W, Zhang W, Xiao WL, An W, Liu X, Wang X, Chen X, Nie X, Cheng X, Liu X, Xie X, Liu X, Yang X, Li X, Su X, Lin X, Li XQ, Jin X, Shen X, Chen X, Sun X, Wang X, Song X, Zhou X, Wang X, Shan X, Li YK, Wang YQ, Wei YX, Zhang Y, Xu Y, Li Y, Zhao Y, Sun Y, Wang Y, Yu Y, Zhang Y, Shi Y, Xiong Y, He Y, Piao Y, Wang Y, Tan Y, Ma Y, Liu Y, Guo Y, Ou Y, Wang Y, Gong Y, Zou Y, He Y, Xiong Y, Luo Y, You Y, Liu Y, Zhou Y, Zhu YX, Huang Y, Li Y, Zheng Y, Zhu Y, Ma Y, Tang Y, Zha Y, Yan Y, Ren ZZ, Ren Z, Sha Z, Fu Z, Xu Z, Xie Z, Zhang Z, Hao Z, Ma Z, Yan Z, Wu Z, Gu Z, Zhu Z, Liu Z, Li Z, Xie Z, Song Z, Pan Z, Huang Z, Xu Z, Zhang Z, Zhang Z: DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 645:633-638, 2025. https://doi.org/10.1038/s41586-025-09422-z
21. Hirschberg J, Manning CD: Advances in natural language processing. Science 349:261-266, 2015. https://doi.org/10.1126/science.aaa8685
22. Hopkins BS, Nguyen VN, Dallas J, Texakalidis P, Yang M, Renn A, Guerra G, Kashif Z, Cheok S, Zada G, Mack WJ: ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. J Neurosurg 139:904-911, 2023. https://doi.org/10.3171/2023.2.JNS23419
23. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ArXiv2023. https://doi.org/10.48550/arXiv.2311.05232
24. Iqbal J, Naseem A, Bashir MA, Yangi K, Bozkurt I, Chaurasia B: Global perspective of neurosurgery practice in lower middle-income countries: challenges, opportunities, and the path forward. Ann Med Surg (Lond) 87:2532-2536, 2025. https://doi.org/10.1097/MS9.0000000000003221
25. Jha AK: The promise of electronic records: around the corner or down the road? Jama 306:880-881, 2011. https://doi.org/10.1001/jama.2011.1219
26. Liu M, Okuhara T, Chang X, Okada H, Kiuchi T: Performance of ChatGPT in medical licensing examinations in countries worldwide: A systematic review and meta-analysis protocol. PLoS One 19:e0312771, 2024. https://doi.org/10.1371/journal.pone.0312771
27. Lucas HC, Upperman JS, Robinson JR: A systematic review of large language models and their implications in medical education. Med Educ 58:1276-1285, 2024. https://doi.org/10.1111/medu.15402

28. Mannam SS, Subtirelu R, Chauhan D, Ahmad HS, Matache IM, Bryan K, Chitta SVK, Bathula SC, Turlip R, Wathen C, Ghenbot Y, Ajmera S, Blue R, Chen HI, Ali ZS, Malhotra N, Srinivasan V, Ozturk AK, Yoon JW: Large language model-based neurosurgical evaluation matrix: A novel scoring criteria to assess the efficacy of ChatGPT as an educational tool for neurosurgery board preparation. World Neurosurg 180:e765-e773, 2023. https://doi.org/10.1016/j.wneu.2023.10.043
29. Mesko B, Topol EJ: The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 6:120, 2023. https://doi.org/10.1038/s41746-023-00873-0
30. Mihalache A, Popovic MM, Muni RH: Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 141:589-597, 2023. https://doi.org/10.1001/jamaophthalmol.2023.1144
31. Mitha S, Schwartz J, Hobensack M, Cato K, Woo K, Smaldone A, Topaz M: Natural language processing of nursing notes: An integrative review. Comput Inform Nurs 41:377-384, 2023. https://doi.org/10.1097/CIN.0000000000000967
32. Nori H, King N, McKinney SM, Carignan D, Horvitz E: Capabilities of GPT-4 on medical challenge problems. ArXiv2023. https://doi.org/10.48550/arXiv.2303.13375
33. Oermann EK, Kondziolka D: On Chatbots and generative artificial intelligence. Neurosurgery 92:665-666, 2023. https://doi.org/10.1227/neu.0000000000002415
34. Rabiner LR: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77:257-286, 1989. https://doi.org/10.1109/5.18626
35. Sahin MC, Sozer A, Kuzucu P, Turkmen T, Sahin MB, Sozer E, Tufek OY, Nernekli K, Emmez H, Celtikci E: Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. Comput Biol Med 169:107807, 2024. https://doi.org/10.1016/j.compbiomed.2023.107807
36. Schubert MC, Wick W, Venkataramani V: Performance of large language models on a neurology board-style examination. JAMA Netw Open 6:e2346721, 2023. https://doi.org/10.1001/jamanetworkopen.2023.46721
37. Sevgi UT, Erol G, Dogruel Y, Sonmez OF, Tubbs RS, Gungor A: The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. Neurosurg Rev 46:86, 2023. https://doi.org/10.1007/s10143-023-01998-2
38. Tabanli A, Demirkiran ND: Comparing ChatGPT 3.5 and 4.0 in low back pain patient education: Addressing strengths, limitations, and psychosocial challenges. World Neurosurgery 196:123755, 2025. https://doi.org/10.1016/j.wneu.2025.123755
39. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, Temsah MH: DeepSeek in healthcare: Revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus 17:e79221, 2025. https://doi.org/10.7759/cureus.79221
40. Türk Nöroşirürji Derneği: Yeterlik Sınavı Hazırlık Soru Kitabı. Türk Nöroşirürji Derneği Yayinlari, 2021, Ankara, Türkiye
41. Ullah E, Parwani A, Baig MM, Singh R: Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. Diagn Pathol 19:43, 2024. https://doi.org/10.1186/s13000-024-01464-7
42. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS: Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 30:1134-1142, 2024. https://doi.org/10.1038/s41591-024-02855-5
43. Williams SC, Starup-Hansen J, Funnell JP, Hanrahan JG, Valetopoulou A, Singh N, Sinha S, Muirhead WR, Marcus HJ: Can ChatGPT outperform a neurosurgical trainee? A prospective comparative study. Br J Neurosurg 39:627-636, 2025. https://doi.org/10.1080/02688697.2024.2308222
44. Xu Z, Jain S, Kankanhalli MS: Hallucination is inevitable: An innate limitation of large language models. ArXiv2024. https://doi.org/10.48550/arXiv.2401.11817
45. Yangi K, Hong J, Gholami AS, On TJ, Reed AG, Puppalla P, Chen J, Calderon Valero CE, Xu Y, Li B, Santello M, Lawton MT, Preul MC: Deep learning in neurosurgery: a systematic literature review with a structured analysis of applications across subspecialties. Front Neurol 16:1532398, 2025. https://doi.org/10.3389/fneur.2025.1532398
46. Yangi K, On TJ, Xu Y, Gholami AS, Hong J, Reed AG, Puppalla P, Chen J, Tangsrivimol JA, Li B, Santello M, Lawton MT, Preul MC: Artificial intelligence integration in surgery through hand and instrument tracking: A systematic literature review. Front Surg 12:1528362, 2025. https://doi.org/10.3389/fsurg.2025.1528362
47. Zhang Y, Rando J, Evtimov I, Chi J, Smith EM, Carlini N, Tramèr F, Ippolito D: Persistent pre-training poisoning of LLMs. ArXiv2024. https://doi.org/10.48550/arXiv.2410.13722
48. Zhui L, Fenghe L, Xuehu W, Qining F, Wei R: Ethical considerations and fundamental principles of large language models in medical education: Viewpoint. J Med Internet Res 26:e60083, 2024. https://doi.org/10.2196/60083

**Supplemental Table I:** Pairwise Comparison of Performances of Chatgpt O3-Mini-High and Chatgpt-4.5 on A 50-Questions Mock Neurosurgery Board Examination

| Question subcategory | No. of Questions | Explanation word count, median (IQR) | | p-value | Readability score, median (IQR) | | p-value | Reasoning time, median (IQR) | | p-value |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **ChatGPT-4.5** | **ChatGPT o3-mini-high** | | **ChatGPT-4.5** | **ChatGPT o-3 mini-high** | | **ChatGPT- 4.5** | **ChatGPT o3-mini-high** | |
| Overall | 50 | 218.50 (199.00-249.50) | 154.00 (115.50-209.25) | **<0.001** | 17.50 (15.75-18.60) | 14.55 (13.33-16.19) | **<0.001** | 1.84 (1.43-1.98) | 6.50 (4.00-9.75) | **<0.001** |
| Neuro-oncology | 5 | 284.0 (220.00-302.00) | 192.0 (165.00-240.00) | 0.08 | 17.3 (15.30-18.80) | 12.53 (12.23-14.12) | **0.013** | 1.29 (1.23-1.46) | 7.00 (4.00-14.00) | **0.045** |
| Neurovascular | 5 | 227.0 (205.00-275.00) | 213.0 (207.00-233.00) | 0.69 | 18.3 (17.80-18.80) | 13.87 (13.46-15.20) | **<0.001** | 1.56 (1.33-1.83) | 9.00 (7.00-14.00) | **0.040** |
| Neuroanatomy | 5 | 223.0 (217.00-238.00) | 108.0 (101.00-234.00) | 0.31 | 15.7 (15.20-18.70) | 13.37 (13.03-13.75) | **0.024** | 1.84 (1.84-2.00) | 5.00 (5.00-11.00) | **0.012** |
| Pediatric neurosurgery | 5 | 223.0 (208.00-248.00) | 154.0 (128.00-264.00) | 0.42 | 15.6 (15.40-15.90) | 12.49 (12.40-13.85) | **0.026** | 1.94 (1.82-2.01) | 4.00 (4.00-29.00) | **0.012** |
| Spinal surgery | 5 | 221.0 (192.00-221.00) | 151.0 (135.00-187.00) | 0.19 | 16.2 (16.10-17.80) | 16.29 (14.51-17.87) | 0.78 | 1.88 (1.81-2.01) | 8.00 (6.00-9.00) | **0.008** |
| Neurotrauma | 5 | 203.0 (175.00-244.00) | 133.0 (120.00-198.00) | 0.22 | 18.9 (18.30-18.90) | 15.26 (14.59-15.77) | **0.008** | 1.83 (1.63-1.90) | 5.00 (4.00-7.00) | 0.10 |
| Skull-base surgery | 5 | 213.0 (190.00-213.00) | 154.0 (135.00-174.00) | 0.21 | 18.0 (17.50-19.80) | 15.59 (15.38-16.49) | **0.036** | 1.92 (1.43-2.00) | 6.00 (5.00-8.00) | 0.10 |
| Functional neurosurgery | 5 | 199.0 (199.00-250.00) | 128.0 (102.00-154.00) | **0.016** | 17.4 (14.50-17.50) | 14.06 (13.88-15.38) | 0.22 | 1.72 (1.42-1.93) | 4.00 (4.00-7.00) | **0.022** |
| Peripheral nerve surgery | 5 | 208.0 (207.00-217.00) | 139.0 (114.00-160.00) | **0.005** | 15.9 (14.20-17.40) | 14.19 (12.92-14.67) | 0.13 | 1.92 (1.33-1.98) | 9.00 (7.00-10.00) | **0.011** |
| Other topics | 5 | 229.0 (158.00-231.00 | 154.0 (123.00-191.00) | 0.11 | 18.3 (17.90-19.30) | 16.39 (16.21-16.80) | **0.039** | 1.92 (1.87-3.18) | 6.00 (5.00-6.00) | **0.035** |