GPT-4

Generative Pre-trained Transformer 4 is a multimodal large language model created by OpenAI, and the fourth in its series of GPT foundation models. It was launched on March 14, 2023, and made publicly available via the paid chatbot product ChatGPT Plus, and via OpenAI's API

GPT-4 in neurosurgery

- Specialized Large Language Model Outperforms Neurologists at Complex Diagnosis in Blinded Case-Based Evaluation
- Utilizing Large language models to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation
- AtlasGPT: a language model grounded in neurosurgery with domain-specific data and document retrieval
- Inherent Bias in Large Language Models: A Random Sampling Analysis
- Assessing GPT-4's accuracy in answering clinical pharmacological questions on pain therapy
- Battle of the authors: Comparing neurosurgery articles written by humans and AI
- Artificial intelligence in neurovascular decision-making: a comparative analysis of ChatGPT-4 and multidisciplinary expert recommendations for unruptured intracranial aneurysms
- Retrieval-augmented generation improves precision and trust of a GPT-4 model for emergency radiology diagnosis and classification: a proof-of-concept study

---

GPT 4 can diagnose and triage neurosurgical scenarios at the level of a senior neurosurgical resident. There has been a clear improvement between GPT 3.5 and 4. Likely, the recent updates in internet access and the functionality of ChatGPT will further improve its utility in neurosurgical triage [1)]

GPT-4, an updated language model with additional training parameters, has exhibited exceptional performance on standardized exams. A study examines GPT-4's competence on neurosurgical board-style questions, comparing its performance with medical students and residents, to explore its potential in medical education and clinical decision-making.

GPT-4's performance was examined on 643 Congress of Neurological Surgeons Self-Assessment Neurosurgery Exam (SANS) board-style questions from various neurosurgery subspecialties. Of these, 477 were text-based and 166 contained images. GPT-4 refused to answer 52 questions that contained no text. The remaining 591 questions were inputted into GPT-4, and its performance was evaluated based on first-time responses. Raw scores were analyzed across subspecialties and question types, and then compared to previous findings on Chat Generative pre-trained transformer performance against SANS users, medical students, and neurosurgery residents.

GPT-4 attempted 91.9% of Congress of Neurological Surgeons SANS questions and achieved 76.6% accuracy. The model's accuracy increased to 79.0% for text-only questions. GPT-4 outperformed Chat Generative pre-trained transformer ($P < 0.001$) and scored highest in pain/peripheral nerve (84%) and lowest in spine (73%) categories. It exceeded the performance of medical students (26.3%), neurosurgery residents (61.5%), and the national average of SANS users (69.3%) across all categories.

GPT-4 significantly outperformed medical students, neurosurgery residents, and the national average of SANS users. The mode's accuracy suggests potential applications in educational settings and clinical decision-making, enhancing provider efficiency, and improving patient care [2].

Murphy Lonergan et al. evaluated the performance of LLMs in answering surgical questions relevant to clinical practice and to assess how this performance varies across different surgical specialties. We used the MedMCQA dataset, a large-scale multi-choice question-answer (MCQA) dataset consisting of clinical questions across all areas of medicine. We extracted the relevant 23,035 surgical questions and submitted them to the popular LLMs Generative Pre-trained Transformers (GPT)-3.5 and GPT-4 (OpenAI OpCo, LLC, San Francisco, CA). A Generative Pre-trained Transformer is a large language model that can generate human-like text by predicting subsequent words in a sentence based on the context of the words that come before it. It is pre-trained on a diverse range of texts and can perform a variety of tasks, such as answering questions, without needing task-specific training. The question-answering accuracy of GPT was calculated and compared between the two models and across surgical specialties. Both GPT-3.5 and GPT-4 achieved accuracies of 53.3% and 64.4%, respectively, on surgical questions, showing a statistically significant difference in performance. When compared to their performance on the full MedMCQA dataset, the two models performed differently: GPT-4 performed worse on surgical questions than on the dataset as a whole, while GPT -3.5 showed the opposite pattern. Significant variations in accuracy were also observed across different surgical specialties, with strong performances in anatomy, vascular, and pediatric surgery and worse performances in orthopedics, ENT, and neurosurgery. Large language models exhibit promising capabilities in addressing surgical questions, although the variability in their performance between specialties cannot be ignored. The lower performance of the latest GPT-4 model on surgical questions relative to questions across all medicine highlights the need for targeted improvements and continuous updates to ensure relevance and accuracy in surgical applications. Further research and continuous monitoring of LLM performance in surgical domains are crucial to fully harnessing their potential and mitigating the risks of misinformation [3]

[1] Ward M, Unadkat P, Toscano D, Kashanian A, Lynch DG, Horn AC, D'Amico RS, Mittler M, Baum GR. A Quantitative Assessment of ChatGPT as a Neurosurgical Triaging Tool. Neurosurgery. 2024 Feb 14. doi: 10.1227/neu.0000000000002867. Epub ahead of print. PMID: 38353523.

[2] 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. 2023 Nov;179:e160-e165. doi: 10.1016/j.wneu.2023.08.042. Epub 2023 Aug 18. PMID: 37597659.

[3] Murphy Lonergan R, Curry J, Dhas K, Simmons BI. Stratified Evaluation of GPT's Question Answering in Surgery Reveals Artificial Intelligence (AI) Knowledge Gaps. Cureus. 2023 Nov 14;15(11):e48788. doi: 10.7759/cureus.48788. PMID: 38098921; PMCID: PMC10720372.