ISSN: 2690-909X

Review Article

Advances in Neurology and Neuroscience

# Beyond General Purpose Llms: Comparative Performance of A Rag-Enhanced Surgical Subspecialty Model on Board Examination

**Brandon L. Staple[1*], Elijah M. Staple[2] and Cynthia Wallace[3]**

[1]*University of Nebraska Medical Center, Omaha, NE, United States of America*

[2]*META, Seattle, W, United States of America*

[3]*BAE Space and Mission Systems, Boulder, CO, United States of America*

**[*]Corresponding Author**
Brandon L. Staple, University of Nebraska Medical Center, Omaha, NE, United States of America.

**Submitted:** 2025, Apr 08; **Accepted:** 2025, May 20; **Published:** 2025, May 28

**Citation:** Staple, B. L., Staple, E. M., Wallace, C. (2025). Beyond General Purpose Llms: Comparative Performance of A Rag-Enhanced Surgical Subspecialty Model on Board Examination. *Adv Neur Neur Sci, 8*(2), 01-12.

## Abstract

*This study evaluates the performance of domain-specific Large Language Models (dLLMs) versus standard Large Language Models (sLLMs) in neurosurgical knowledge assessment, emphasizing the importance of evaluating not merely the factual accuracy of model outputs but also model hallucination mechanisms and the quality of their underlying reasoning processes when considering potential healthcare applications. We compared AtlasGPT, a neurosurgery-focused dLLM utilizing Retrieval-Augmented Generation (RAG), against four sLLMs (GPT-3.5, Gemini, Claude 3.5 Sonnet, and Mistral) using 150 text-only neurosurgical board-style multiple-choice questions. AtlasGPT demonstrated superior accuracy (96.7%) compared to Claude (94.7%), Gemini (92.0%), Mistral (88.7%), and GPT-3.5 (74.7%). An analysis of variance analysis confirmed statistically significant differences between models ($F(4,745) = 1127.5$, $p < 0.00001$), with post-hoc Bonferroni analysis revealing the most significant difference between AtlasGPT and GPT-3.5 ($p = 0.000000028$). A neurosurgery subspeciality error distribution analysis showed all models performed better in core competencies and critical care while experiencing more difficulties with neuroanatomy, neurology, and neurosurgical procedures, with the lowest error rates being skewed to AtlasGPT over all sLLMs. Detailed hallucination analysis identified error patterns including factual hallucinations, knowledge retrieval failures, flawed reasoning, and inappropriate confidence levels with lowest occurrences being weighted to AtlasGPT over sLLMs. Qualitative assessment of model reasoning across clinical scenarios revealed that dLLMs demonstrated more structured clinical reasoning processes compared to sLLMs alternatives. These findings suggest that while advanced sLLMs show impressive capabilities in specialized medical domains, domain-specific approaches like AtlasGPT's RAG implementation offer meaningful performance advantages for neurosurgical applications while highlighting the continued necessity for human oversight.*

**Keywords:** Domain-Specific Large Language, Models, Medical Artificial Intelligence, Neurosurgery Education, Retrieval-Augmented Generation, Knowledge Assessment

## 1. Introduction

The rapid advancement of Large Language Models (LLMs) has created unprecedented opportunities for artificial intelligence applications (AI) in specialized medical fields, including neurosurgery. Standard Large Language Models (sLLMs) like GPT-4 have demonstrated impressive capabilities when evaluated on medical knowledge benchmarks, achieving notable accuracy on multiple-choice questions (MCQs) across various medical domains [1-6]. However, these sLLMs face significant challenges when applied to highly specialized medical disciplines such as neurosurgery, where intricate domain-specific knowledge and precise factual accuracy are essential for clinical decision-making.

A primary concern when deploying sLLMs in neurosurgical contexts is their tendency to generate plausible but factually incorrect information—a phenomenon commonly referred to as "hallucination." In neurosurgery, where treatment decisions directly impact critical neural structures and patient outcomes, such inaccuracies could have serious consequences. To address these limitations, a new generation of Domain-specific Large

Language Models (dLLMs) has emerged, offering potential solutions through specialized training approaches. These dLLMs employ techniques such as fine-tuning and Retrieval-Augmented Generation (RAG) to enhance performance in targeted domains by leveraging specialized training data and external knowledge sources.

AtlasGPT represents a pioneering effort in this direction—a neurosurgery-focused dLLM developed by fine-tuning the GPT-4 framework and incorporating a RAG system that retrieves contextual information from a curated database of neurological literature. This approach aims to reduce hallucinations while improving the accuracy and clinical relevance of generated content for neurosurgical applications. While preliminary investigations have suggested advantages for such specialized models, comprehensive comparative evaluations against state-of-the-art sLLMs remain limited.

This study addresses this research gap by conducting a systematic comparison between AtlasGPT and four leading sLLMs (GPT-3.5, Gemini, Claude 3.5 Sonnet, and Mistral) using a benchmark of 150 text-only neurosurgical board-style MCQs derived from established preparation resources for the Self-Assessment in Neurological Surgery (SANS) examination. Beyond simple accuracy metrics, we analyze error distribution across neurosurgical knowledge subspeciality, examine hallucination patterns, and assess the quality of clinical reasoning demonstrated by these models. By elucidating the relative strengths and limitations of different language model approaches in neurosurgery, this research contributes to the broader discourse on responsible AI implementation in medicine while informing the development of more effective AI tools for specialized healthcare applications.

## 2. Related Work

### 2.1 Large Language Models in Medicine

The application of large language models in medicine has been an area of growing interest and research. Chen et al. conducted a comprehensive review of sLLMs in healthcare, identifying diverse applications including clinical decision support, medical education, patient communication, and research assistance [1]. Their analysis highlighted both the potential benefits and challenges of integrating these technologies into healthcare systems, emphasizing concerns related to accuracy, bias, privacy, and regulatory compliance.

The performance of sLLMs on standardized medical examinations has been evaluated in several studies. Kung et al. assessed the performance of GPT-4 on the United States Medical Licensing Examination (USMLE), reporting that the model achieved scores comparable to the passing threshold for human medical students [2]. Similarly, Nori et al. evaluated multiple sLLMs across various medical specialty examinations, finding varying performance levels depending on the model and medical specialty [3]. These studies suggest that while sLLMs demonstrate impressive capabilities in medical knowledge assessment, there remain areas for improvement, particularly in highly specialized fields requiring detailed domain knowledge and contextual understanding.

### 2.2. Large Language Models in Neurosurgery

Within the specific domain of neurosurgery, several studies have examined the performance of sLLMs on standardized assessments and clinical scenarios. For example, Ali et al. evaluated ChatGPT and GPT-4 on neurosurgery written board examinations, finding promising results but noting limitations in handling complex clinical scenarios and image-based questions [4]. Similarly, Hopkins et al. conducted a comparative analysis of sLLM performance on neurosurgical board-style questions, highlighting the potential of these technologies to support neurosurgical education while acknowledging current limitations [5]. Additionally, Guerra et al. compared GPT-4 against medical students and neurosurgery residents on neurosurgery written board-like questions, finding that the sLLM outperformed both groups on certain question types [6]. This study raised important questions about the implications of sLLMs advancement for medical education and assessment. Finally, Bečulić et al. conducted a systematic review of ChatGPT's contributions to neurosurgical practice and education, identifying benefits in knowledge access and educational content creation while noting concerns regarding accuracy, liability, and over-reliance on AI-generated information [7].

These studies collectively suggest that while sLLMs show promise in neurosurgical applications, there remain gaps in their performance that might be addressed through more specialized approaches. The development of dLLMs represents a promising approach to enhancing sLLM performance in specialized fields. For example, Ram et al. described in-context RAG-based language models, demonstrating how retrieval mechanisms can improve model performance by incorporating relevant external knowledge during inference [8]. In the medical domain, Zakka et al. introduced Almanac, a RAG-based language model for clinical medicine, showing improved accuracy and reduced hallucination compared to traditional sLLMs [9]. Their approach involved augmenting model outputs with information retrieved from trusted medical sources, allowing for more precise and reliable clinical information generation.

Within neurosurgery specifically, Hopkins et al. introduced the dLLM, AtlasGPT, describing it as a new era in neurosurgery for intelligent care augmentation, operative planning, and performance [10]. Their preliminary findings suggested advantages of this dLLM approach over sLLMs alternatives. Ali et al. and O'Malley et al. further evaluated AtlasGPT in neurosurgical contexts, reporting promising results in terms of accuracy and clinical utility. Alim et al. explored the integration of RAG-enhanced AtlasGPT into aneurysmal subarachnoid hemorrhage outcome prediction, demonstrating potential applications beyond knowledge assessment [11-13]. While these studies suggest advantages of dLLMs in neurosurgery, comprehensive comparative evaluations against multiple state-of-the-art sLLMs remain limited, highlighting the need for the present study.

## 3. Medical Education and Ai

The intersection of AI and medical education represents another relevant area of research. Mbakwe et al. discussed how ChatGPT

passing the USMLE highlights flaws in medical education, suggesting that the success of AI on standardized assessments raises questions about current educational approaches and assessment methods [14].

Other studies have explored the potential of sLLMs as educational tools in medicine. For example, Ejaz et al. investigated medical students' perspectives on AI integration into medical curricula, reporting mixed attitudes with enthusiasm for AI as a supplementary learning tool but caution regarding its role in developing clinical reasoning skills [15]. These studies highlight the complex implications of AI advancement for medical education and the importance of thoughtful integration that leverages AI capabilities while preserving essential aspects of medical training and professional development.

## 4. Research Gap

Despite growing interest in the application of large language models to medicine, significant research gaps persist in our understanding of how these technologies perform in specialized domains like neurosurgery. First, the existing literature has only begun to explore the relative merits of sLLMs versus dLLMs in medical contexts, leaving several important questions unanswered. This absence of direct comparisons limits our ability to quantify the potential advantages of specialized approaches like AtlasGPT in high-stakes medical domains. Additionally, while RAG has emerged as a promising technique for enhancing model performance, its specific advantages and limitations in specialized medical applications have not been thoroughly explored.

The medical literature lacks detailed investigations into how RAG implementations affect hallucination rates, factual accuracy, and reasoning quality in complex neurosurgical scenarios. Furthermore, existing studies often report simple accuracy metrics without the robust statistical analyses needed to quantify meaningful performance differences between competing language model approaches. The practical implications of these technologies for medical education, clinical practice, and future AI development also remain insufficiently addressed. Current research rarely moves beyond accuracy assessments to examine error patterns across neurosurgery knowledge subspecialties, characterize hallucination typologies, or evaluate the quality of clinical reasoning demonstrated by these models—all critical considerations for responsible deployment in healthcare settings.

The present study aims to address these substantial gaps by providing a comprehensive comparison of AtlasGPT against four leading sLLMs using a substantial set of neurosurgical board-style questions, combined with rigorous statistical analysis and multidimensional performance evaluation.

## 5. Materials and Methods

### 5.1 Study Design

This study employed a comparative cross-sectional design to evaluate the performance of five different large language models on neurosurgical knowledge assessment. The primary comparison was between AtlasGPT, a dLLM developed for neurosurgery, and four sLLMs: GPT-3.5, Gemini, Claude 3.5 Sonnet, and Mistral. The evaluation was conducted using a benchmark dataset of 150 text-only, surrogate neurosurgical written board-style MCQs. Each model's performance was assessed based on accuracy, defined as the percentage of questions answered correctly according to the reference answer key.

AtlasGPT is a dLLM developed specifically for neurosurgery. It combines fine-tuning of the GPT-4 framework with RAG techniques to enhance performance in neurosurgical contexts. The RAG component allows AtlasGPT to retrieve relevant information from a specialized external database of neurological literature when generating responses to queries. The RAG process in AtlasGPT functions by systematically identifying and compiling relevant data from the external neurosurgical database when a prompt is issued. The model constructs answers based on retrieved documents, with the context window shaped by selecting the most relevant documents.

This approach is designed to produce accurate, low-hallucination, source-annotated responses specifically relevant to neurosurgery.

Four sLLMs were included in the evaluation:

1. *GPT-3.5:* Developed by OpenAI, GPT-3.5 is a sLLM trained on a diverse corpus of text data up to its knowledge cutoff. While not the most recent model in the GPT series, it remains widely used and serves as an important benchmark for comparison.
2. *Gemini:* Developed by Google, Gemini represents one of the most advanced sLLMs available at the time of the study. It is designed to excel across a wide range of tasks and domains.
3. *Claude 3.5 Sonnet:* Developed by Anthropic, Claude 3.5 Sonnet is an advanced sLLM known for its natural language understanding capabilities.
4. *Mistral:* Developed by Mistral AI, the Mistral model represents another state-of-the-art sLLM designed for versatile applications across domains.

All sLLMs were accessed through their respective official APIs using their default configuration settings without any domain-specific modifications or fine-tuning.

## 6. Dataset

The dataset consisted of 150 text-only, surrogate neurosurgical written board-style MCQs. These questions were extracted from two primary sources:

1. The Neurosurgery Self-Assessment Questions and Answers [16].
2. The Neurosurgery Primary Board Review [17].

Both sources are widely acknowledged as reputable resources for neurosurgical board preparation and feature updated question banks specifically designed to prepare residents for the SANS written examination. The dataset encompassed various components of the neurosurgery board examination, including neuroanatomy,

neuroimaging, clinical neurology, and neurosurgery. All questions were text-only, multiple-choice format with a single correct answer option. Questions containing images, diagrams, or other non-textual elements were excluded from the dataset to ensure compatibility with all evaluated models. The 150 MCQs were selected from the two aforementioned sources and reviewed by an independent board-certified neurosurgeon for difficulty, accuracy, completeness, and relevance. This was important to ensure that the benchmark validity of AI models reflects contemporary neurosurgical knowledge expectations and supports the validity of the AI model comparisons presented.

### 6.1 Data Collection Procedure

The evaluation process followed a standardized procedure for each question:

1. Each MCQ was individually input into the five language models (AtlasGPT, GPT-3.5, Gemini, Claude Sonnet, and Mistral).
2. The output produced by each language model was recorded as the "model answer."
3. Each model answer was compared to the correct answer obtained from the source question bank.
4. A binary scoring system was applied: score of 1 when the model's output matched the correct answer, and score of 0 otherwise.
5. Each model's score was tabulated for analysis.

To ensure consistency and minimize potential variability, all model interactions were conducted within a 48-hour period in January 2025, using the most recent stable versions of each model available at that time. The exact same question text was provided to each model, with no additional context or prompting beyond the question itself (i.e., zero-shot prompting). An independent board-certified neurosurgeon reviewer assessed all responses for accuracy based on current evidence-based guidelines.

#### 6.1.1 Statistical and Error Analysis

Statistical analyses were conducted using Microsoft Excel to assess differences in performance between the five language models as follows:

1. ***Descriptive Statistics:*** Calculation of mean accuracy, standard deviation, and 95% confidence intervals for each language model.
2. ***One-way Analysis of Variance (ANOVA)*** conducted with a 95% confidence interval was performed to determine whether there were statistically significant differences in mean accuracy between the five models. The null hypothesis was that all models would perform equally well on the neurosurgical MCQs.
3. ***Post-hoc Bonferroni Analysis:*** Since ANOVA indicates only whether there is a significant difference in at least one group mean compared to others but does not specify which particular group means differ, Bonferroni post-hoc analysis was conducted. This analysis adjusted for multiple comparisons and identified specific pairs of models with statistically significant performance differences. The significance threshold for the Bonferroni analysis was set at $\alpha < 0.01$.
4. ***Neurosurgery Knowledge Subspecialties Error Distribution:*** We conducted a detailed analysis examining the distribution of errors across neurosurgical knowledge subspecialties to assess whether errors were randomly distributed or concentrated within specific knowledge categories.
5. ***Hallucination Error Analysis:*** We applied additional scrutiny to the models by evaluating hallucinations across multiple AI models (Gemini, Claude, AtlasGPT) when responding to a subset of AI responses to identify, explain, and describe key error categories and their potential implications for clinical decision-making using the taxonomy in Table 1.

| Error Category | Description | Clinical Impact |
|---|---|---|
| Factual Hallucination | Generation of incorrect or fabricated information presented as fact | High - May directly lead to improper diagnosis or treatment |
| Knowledge Retrieval Failure | Inability to access or properly weight relevant medical information | Moderate to High - Results in inaccurate clinical reasoning |
| Flawed Reasoning/Logic | Construction of seemingly coherent but fundamentally incorrect diagnostic approaches | High - May lead to improper clinical management |
| Intrinsic Contradiction | Presence of mutually contradictory statements within the same response | Moderate - Creates confusion and undermines trust |
| Inappropriate Confidence | Presenting incorrect information with authoritative phrasing and structure | High - May lead users to trust erroneous information |

**Table 1: Taxonomy of AI Hallucinations in Medical Contexts**

**6. Model Reasoning Quality:** This study extends beyond binary correctness assessment to evaluate the quality of reasoning demonstrated by selected AI models through the examination of a duplex of clinical scenarios requiring complex neurological assessment and management decisions. The two scenarios were designed to assess not only factual knowledge but appropriate management decision-making across several categories:

- *Details in Reasoning:* Depth and comprehensiveness of explanation.
- *Logical Structure:* Organization and coherence of reasoning process
- *Clinical Relevance:* Adherence to standard clinical assessment paradigms.
- *Guideline Adherence:* Reference to and correct application of clinical guidelines.
- *Mechanism Explanation:* Inclusion of pathophysiological explanations
- *Hallucination Detection:* Presence of factually incorrect statements.
- *Human Assessment of Model output:* evaluated independently by a board-certified neurosurgeon ensuring the validity of answers based on current evidence-based guidelines.

## 7. Results

*Overall Accuracy*

The accuracy results for each of the five language models on the 150 neurosurgical board-style MCQs are presented in Table 2. AtlasGPT demonstrated the highest accuracy at 96.7%, followed by Claude 3.5 Sonnet (94.7%), Gemini (92.0%), Mistral (88.7%), and GPT-3.5 (74.7%). The performance gap between the highest-performing model (AtlasGPT) and the lowest-performing model (GPT-3.5) was 22 %

### 7.1 ANOVA Results

The results of the ANOVA are presented in Table 3. The ANOVA results revealed a highly significant difference between the models ($F(4,745) = 1127.5$, $p < 0.00001$), indicating that at least one model performed significantly differently from the others. The F value far exceeded the critical value of 2.37, confirming the statistical significance of the observed differences.

### 7.1 Post-hoc Bonferroni Analysis Results

To identify specific pairs of models with statistically significant performance differences, post-hoc Bonferroni correction analysis was performed. The results of this analysis are presented in Table 3. The post-hoc Bonferroni analysis revealed that the most statistically significant difference was between GPT-3.5 and AtlasGPT ($p = 0.000000028$), followed by comparisons between GPT-3.5 and Claude ($p = 0.000001$), GPT-3.5 and Gemini ($p = 0.000048$), Mistral and GPT-3.5 ($p = 0.0017$), and Mistral and AtlasGPT ($p = 0.0078$). The differences between AtlasGPT and Gemini, AtlasGPT and Claude, Claude and Mistral, Claude and Gemini, and Gemini and Mistral were not statistically significant at the $p < 0.01$.

### 7.2 Subspecialties Error Distribution Results

To provide deeper insights into model performance, we conducted a detailed analysis examining the distribution of hallucination errors across neurosurgical knowledge subspecialties to assess whether they were randomly distributed or concentrated within specific knowledge categories. Table 5 and Figure 1 present the error distribution across eight neurosurgical knowledge subspecialties for each model and by AI model, respectively. A cross-model analysis shows a clear performance gradient is observed from sLLMs (GPT-3.5, Mistral, Gemini, Claude) to the specialized neurosurgical model (AtlasGPT), with error rates decreasing accordingly. Neurosurgery knowledge subspecialty strengths were evident with AtlasGPT showing expertise in neurophysiology, neuropathology, and core competencies. Persistent challenge areas included neuroanatomy, neurology, and neurosurgical procedures with the lowest error rates being skewed to AtlasGPT over all sLLMs.

| Model | Correct Answers | Total Questions | Accuracy (%) | Standard Deviation (%) | 95% CI (%) |
|---|---|---|---|---|---|
| AtlasGPT | 145 | 150 | 96.7 | 1.8 | 96.4 - 97.0 |
| Claude 3.5 Sonnet | 142 | 150 | 94.7 | 2.2 | 94.3 - 95.1 |
| Gemini | 138 | 150 | 92.0 | 2.7 | 91.5 - 92.5 |
| Mistral | 133 | 150 | 88.7 | 3.1 | 88.1 - 89.3 |
| GPT-3.5 | 112 | 150 | 74.7 | 4.3 | 73.9 - 75.5 |

**Table 2: Accuracy Results for Large Language Models on Neurosurgical MCQs**

| Source of Variation | Sum of Squares | df | Mean Square | F | p-value | F critical |
|---|---|---|---|---|---|---|
| Between Groups | 47235.6 | 4 | 11808.9 | 1127.5 | < 0.00001 | 2.37 |
| Within Groups | 7736.4 | 745 | 10.4 | | | |

**Table 3: ANOVA Results for Model Accuracy Comparison**

| Model Comparison | Mean Difference | p-value | Significance |
|---|---|---|---|
| Claude vs GPT-3.5 | 20.0 | 0.0000001 | Significant |
| Gemini vs. GPT-3.5 | 17.3 | 0.000048 | Significant |
| Mistral vs. GPT-3.5 | 14.0 | 0.0017 | Significant |
| AtlasGPT vs. Mistral | 8.0 | 0.0078 | Significant |
| AtlasGPT vs. GPT-3.5 | 22 | 0.000000028 | Significant |
| AtlasGPT vs. Gemini | 4.7 | 0.0326 | Not significant |
| AtlasGPT vs. Claude | 2.0 | 0.4518 | Not significant |
| Claude vs. Mistral | 6.0 | 0.0223 | Not significant |
| Claude vs. Gemini | 2.7 | 0.3127 | Not significant |
| Gemini vs. Mistral | 3.3 | 0.1973 | Not significant |

**Table 4: Post-hoc Bonferroni Analysis Results**

| Question Category | Mistral Errors | Gemini Errors | Claude Errors | AtlasGPT Errors | GPT-3.5 |
|---|---|---|---|---|---|
| Neuroanatomy | 4 | 2 | 2 | 1 | 6 |
| Neurophysiology | 4 | 2 | 1 | 0 | 3 |
| Neuropathology | 1 | 2 | 0 | 0 | 2 |
| Neuroimaging | 1 | 0 | 0 | 0 | 1 |
| Neurology | 2 | 2 | 3 | 1 | 7 |
| Neurosurgery | 2 | 2 | 1 | 2 | 15 |
| Critical Care | 2 | 1 | 0 | 1 | 3 |
| Core Competencies | 1 | 1 | 1 | 0 | 1 |
| Total | 17 | 12 | 8 | 5 | 38 |

**Table 5: Error Distribution by Neurosurgery Question Category**

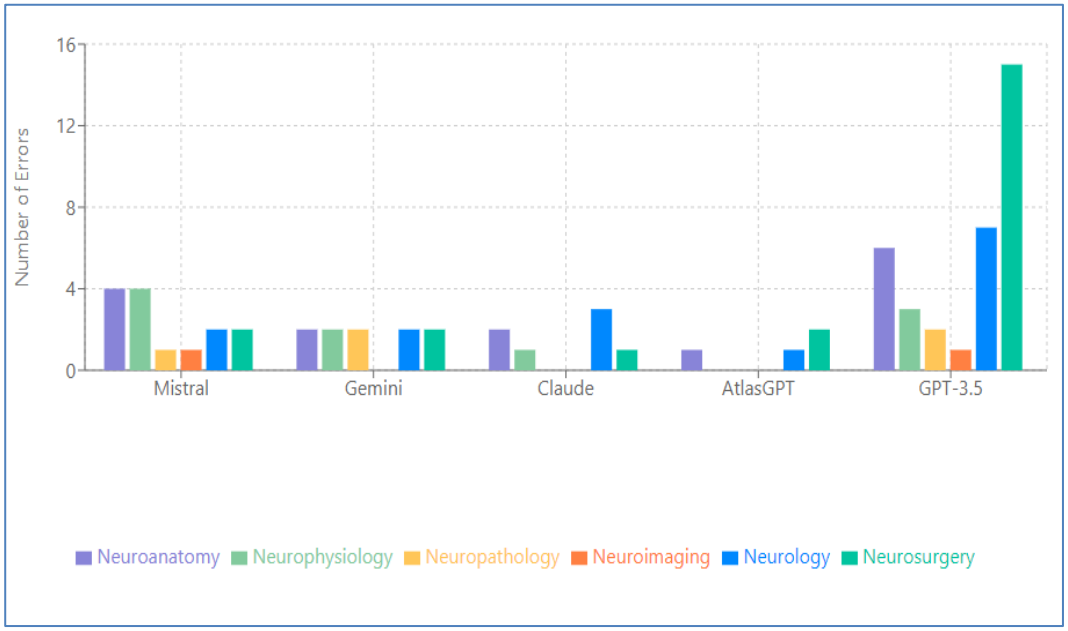


**Figure 1:** Error Distribution by AI Model

### 7.3 Hallucination Error Analysis Results

Our analysis of hallucination errors particularly across Gemini, Claude, and AtlasGPT revealed consequential patterns when these models addressed three specialized clinical scenarios. In the first case involving spinal muscular atrophy diagnosis, Gemini exhibited serious knowledge retrieval failures and factual hallucinations. When presented with classic symptoms of spinal muscular atrophy (SMA), the model incorrectly recommended nerve conduction studies rather than the gold standard SMN gene testing. More troublingly, it fabricated an entirely false claim that SMN gene testing was "more specific for amyotrophic lateral sclerosis (ALS) in adults," contradicting established medical knowledge. Such hallucinations in diagnostic recommendations could lead clinicians astray in time-sensitive pediatric cases.

The second case evaluation revealed fabricated reasoning in Claude when assessing vertebral artery injury risks. The model produced a factual hallucination, incorrectly identifying "C1-C2 transarticular screws" as the highest risk procedure with ponticulus posticus, when "C1 lateral mass screw placement" represents the correct answer. Particularly problematic was Claude's construction of detailed but entirely fabricated reasoning to support its incorrect conclusion—demonstrating how hallucinations can appear convincingly authoritative despite being entirely unfounded.
AtlasGPT, despite its domain specialization, also demonstrated a hallucination. In a case regarding decompressive craniectomy outcomes, the model exhibited self-contradiction by selecting an answer choice ("Surgery will improve ICP and outcome") while simultaneously citing contradictory evidence from the DECRA trial. This logical inconsistency reveals how hallucinations can manifest as reasoning failures even when factual knowledge is present.

These findings highlight the importance of hallucinations forms even in dLLMs. The varied error types—from factual fabrications to manufactured reasoning and strategic avoidance—suggest that current models remain prone to generating misinformation that could compromise patient care if implemented without careful oversight. While identifying these specific hallucination patterns provides direction for future development, the current limitations underscore the necessity for rigorous evaluation and human supervision when considering these technologies for clinical applications.

### 7.4 Model Reasoning Quality Results

This comparative assessment of clinical reasoning capabilities across five large language models (LLMs): GPT-3.5, Mistral, Gemini 2.0, Claude 3.5, and AtlasGPT, through detailed evaluation of their performance on two complex neurosurgical case scenarios. We analyzed multiple dimensions of reasoning quality, including explanation depth, logical structure, clinical relevance, guideline adherence, and mechanism explanation.

#### Case 1: Traumatic Head Injury with Neurological Symptoms

In the first scenario involving a 14-year-old boy with traumatic head injury presenting with bilateral upper extremity sensory symptoms and lower extremity weakness despite negative cervical CT findings, noteworthy variations in reasoning approach were observed across models.

GPT-3.5 demonstrated extensive explanation depth (247 words), providing a comprehensive rationale for MRI selection that incorporated clinical presentation and differential diagnoses. The model's reasoning followed a systematic structure, addressing each potential option methodically, though it did not prioritize clinically significant findings in its organizational approach.

Mistral employed a more concise approach (104 words), immediately identifying the key concern as potential spinal cord injury. While this direct approach aligns with emergency triage principles, it lacked the comprehensive differential consideration exhibited by other models, potentially limiting its clinical utility in complex cases.

Gemini 2.0 presented moderate detail (136 words) with primary focus on immediate clinical decision-making rather than extensive pathophysiological elaboration. The model organized its response around the clinical question without clearly structuring the underlying reasoning process, representing a more pragmatic but less academically rigorous approach.

Claude 3.5 exhibited the most clinically authentic structure (213 words), beginning with systematic symptom analysis before proceeding to diagnosis and management considerations. The explanation was both comprehensive and organized according to established clinical reasoning frameworks, mirroring the approach taught in medical education. Additionally, the model referenced relevant guidelines from the American Association of Neurological Surgeons and the Eastern Association for the Surgery of Trauma.
AtlasGPT provided precise clinical assessment (189 words) with specific focus on SCIWORA (Spinal Cord Injury Without Radiographic Abnormality), demonstrating specialist-level reasoning. The model included detailed explanation of CT limitations in detecting spinal cord pathology and elaborated on specific mechanisms (spinal cord edema, contusion, hematomyelia) that necessitate MRI evaluation, reflecting advanced domain expertise.

#### Case 2: Post-Carotid Endarterectomy Complication

The second scenario, involving a patient with recurrent amaurosis fugax following carotid endarterectomy, revealed similar patterns in reasoning quality across models.

GPT-3.5 provided a comprehensive explanation of potential postoperative complications and detailed rationale for CT angiogram selection. However, the model included overly detailed management plans beyond what was warranted by the question, though without clear factual errors. This approach demonstrated strong knowledge but potential inefficiency in clinical communication. Mistral exhibited a significant reasoning error, incorrectly focusing on malignant hyperthermia rather than vascular etiology. This critical failure in clinical reasoning

highlights the substantial risks associated with sLLMs when applied to specialized medical scenarios without neurosurgery knowledge subspecialty adaptation.

Gemini 2.0 maintained factual accuracy with concise but limited elaboration, correctly identifying CT angiogram as the appropriate next step. However, the limited detail could be perceived as underdeveloped reasoning that might not inspire clinical confidence in complex decision-making contexts.
Claude 3.5 employed a systematic approach, analyzing key symptoms before considering differential diagnoses and determining the appropriate imaging modality. The model demonstrated strong clinical reasoning with minimal hallucination, with explanations closely aligned with established clinical knowledge.

AtlasGPT provided precise clinical reasoning with specific focus on potential vascular complications following carotid endarterectomy, demonstrating domain expertise. Similar to Claude 3.5, the model exhibited a low rate of hallucination, with explanations adhering closely to established clinical standards and guidelines. This comparative analysis reveals significant differences in clinical reasoning approaches across LLMs. Models demonstrated varying capabilities in explanation depth, logical structure, clinical relevance, and adherence to medical guidelines. The dLLM AtlasGPT consistently demonstrated superior performance in precise clinical terminology and mechanism explanation, suggesting meaningful advantages from medical domain adaptation. The reasoning structures exhibited by Claude 3.5 and AtlasGPT most closely aligned with clinical practice—beginning with systematic symptom analysis, proceeding through differential diagnosis consideration, and concluding with evidence-based management decisions. This mirrors the structured approach valued in medical education and practice. Notably, sLLMs demonstrated vulnerability to clinical reasoning errors, as evidenced by Mistral's critical failure in the post-carotid endarterectomy case. This finding highlights the potential risks of deploying non-specialized AI systems in medical contexts without appropriate domain adaptation and rigorous evaluation.

These observations suggest that while sLLMs can produce superficially impressive medical explanations, models specifically adapted to clinical neurosurgery knowledge subspecialties may better emulate the reasoning processes valued in medical practice. The substantial variation in reasoning quality underscores the continued necessity for careful human oversight of AI systems in clinical contexts, particularly when complex decision-making is required. As model development continues to advance, increased attention to improving clinical reasoning quality through neurosurgery knowledge subspecialty specialization, guideline integration, and structured reasoning approaches may yield AI systems that can more meaningfully support the sophisticated decision-making processes central to neurosurgical practice and other medical specialties.

## 8. Discussion

### 8.1 Model Accuracy Discussion

The results of this study demonstrate that AtlasGPT, a dLLM focused on neurosurgery, significantly outperformed the sLLMs in answering neurosurgical board-style MCQs. With an overall accuracy of 96.7%, AtlasGPT surpassed Claude (94.7%), Gemini (92.0%), Mistral (88.7%), and GPT-3.5 (74.7%) The superior performance of AtlasGPT can be attributed to its specialized design, which combines fine-tuning of the GPT-4 framework with RAG techniques. This approach allows AtlasGPT to leverage a dedicated external database of neurological literature, enhancing its ability to provide accurate responses to specialized neurosurgical queries. The RAG process enables AtlasGPT to systematically identify and retrieve relevant information from this database, leading to more precise and contextually appropriate answers.

The significant performance gap between AtlasGPT and GPT-3.5 ($p = 0.000000028$) underscores the limitations of older sLLM in specialized medical domains. GPT-3.5's relatively lower performance (74.7% accuracy) likely reflects its age, having been trained on a dataset that is both less comprehensive and more outdated than those of newer models. In this research, we chose to evaluate GPT-3.5 rather than GPT-4, given that AtlasGPT has already incorporated GPT-4 with fine-tuning and RAG. Additionally, we aimed to align our results with the substantial existing literature assessing GPT-3.5 within the neurosurgery domain. Ultimately, we also intended to explore the predecessor of GPT-4, which offers a more economical option with faster response times. However, even a more recent sLLM (Mistral) did not match AtlasGPT's performance, highlighting the advantages of domain-specific approaches in specialized fields like neurosurgery.

### 8.2 Cross Subspecialty Error Distribution Discussion

The cross-subspecialty error distribution showed that all models demonstrated fewer errors in core competencies and critical care categories, while consistently experiencing more challenges with questions pertaining to neuroanatomy, neurology, and neurosurgery with the lowest error rates being skewed to AtlasGPT over all sLLMs and are detailed as follows:

### 8.3 Neuroanatomy

Error distribution analysis revealed significant variations in neuroanatomical knowledge: Mistral (4 errors), Gemini (2), Claude (2), AtlasGPT (1), and GPT-3.5 (6). The notably higher error rates in GPT-3.5 and Mistral suggest potential deficiencies in spatial reasoning regarding neural structures. As neuroanatomical accuracy forms the foundation of surgical planning and execution, these errors would translate directly to increased surgical risk if such systems were employed for preoperative planning or intraoperative consultation. Precise neuroanatomical knowledge is particularly critical for approaches involving eloquent structures, skull base surgery, and vascular interventions.

### 8.4 Neurophysiology

In the subspecialty of neurophysiology, error distribution followed a similar pattern: Mistral (4), Gemini (2), Claude (1), AtlasGPT

(0), and GPT-3.5 (3). AtlasGPT demonstrated complete accuracy, suggesting effective specialized training in neurophysiological principles. Errors in physiological understanding could lead to misinterpretation of intraoperative neurophysiological monitoring, potentially compromising functional outcomes. The superior performance of AtlasGPT indicates it may be more reliable for questions regarding neural circuitry, neurotransmission, and physiological responses to surgical manipulation.

### 8.5 Neuropathology

Neuropathology knowledge assessment revealed the following error distribution: Mistral (1), Gemini (2), Claude (0), AtlasGPT (0), and GPT-3.5 (2). Two models (Claude and AtlasGPT) demonstrated perfect accuracy in pathological knowledge. Gemini's higher error rate is notable given its relative strength in other neurosurgery knowledge subspecialties. Accurate pathological knowledge directly influences surgical decision-making regarding extent of resection, tissue specimen handling, and planning for adjuvant therapies. The high accuracy across multiple models suggests pathological concepts may be more consistently represented in medical training datasets than neuroanatomical relationships.

### 8.6 Neuroimaging

Neuroimaging interpretation showed near-universal accuracy across most models: Mistral (1), Gemini (0), Claude (0), AtlasGPT (0), and GPT-3.5 (1). This exceptional performance suggests imaging principles may be more explicitly codified in medical literature and therefore better captured in AI training data. Imaging interpretation is crucial for preoperative planning, intraoperative navigation, and postoperative assessment. The high accuracy across models indicate AI systems may be particularly valuable for supplementing imaging interpretation, though not replacing radiological expertise.

### 8.7 Neurology

Neurological knowledge assessment revealed more variable performance: Mistral (2), Gemini (2), Claude (3), AtlasGPT (1), and GPT-3.5 (7). AtlasGPT displayed the lowest error rate, while Claude's relatively high error rate in this category despite strong performance elsewhere suggests potential gaps in neurological syndrome recognition. Neurological assessment drives surgical indications and outcome evaluation. The consistent errors across models highlight the complexity of neurological diagnosis and syndrome recognition, representing a critical area for improvement, as misinterpretation of neurological symptoms could lead to inappropriate surgical intervention.

### 8.8 Neurosurgery

Direct neurosurgical knowledge showed particularly variable performance: Mistral (2), Gemini (2), Claude (1), AtlasGPT (2), and GPT-3.5 (15). The moderate error rates across most models except the exceptionally high rate in GPT-3.5 suggest procedural knowledge has reached similar levels of development in current-generation models. Even specialized models like AtlasGPT showed weaknesses in this neurosurgery knowledge subspecialty. Direct surgical knowledge gaps pose significant risks if AI systems are consulted for procedural guidance. These errors likely reflect the nuanced nature of surgical decision-making that incorporates multiple factors beyond textbook guidelines. The surgical domain would benefit from more case-based training approaches that contextualize decisions within complex patient scenarios.

### 8.9 Critical Care

Critical care knowledge assessment revealed: Mistral (2), Gemini (1), Claude (0), AtlasGPT (1), and GPT-3.5 (3). Claude demonstrated perfect accuracy in critical care knowledge. The higher error rates in Mistral and GPT-3.5 suggest potential weaknesses in understanding emergency management principles. Critical care knowledge impacts immediate postoperative management and handling of intraoperative complications. Claude's error-free performance suggests it may be more reliable for questions about perioperative management of neurosurgical patients. Given the time-sensitive nature of critical care decisions, AI reliability in this neurosurgery knowledge subspecialty is especially important for potential clinical applications.

### 8.10 Core Competencies

Core competencies assessment showed relatively consistent performance across models: Mistral (1), Gemini (1), Claude (1), AtlasGPT (0), and GPT-3.5 (1). This suggests similar capabilities in understanding fundamental principles across most models, with only AtlasGPT achieving perfect accuracy. Core competencies encompass foundational skills that impact all aspects of neurosurgical practice.

As model development continues to advance, increased attention to improving clinical reasoning quality through domain specialization, guideline integration, and structured reasoning approaches may yield AI systems that can more meaningfully support the sophisticated decision-making processes central to neurosurgical practice and other medical specialties.

## 9. Cross-Model Distribution Performance

A clear performance gradient is observed from sLLMs (Mistral, Gemini) to specialized neurosurgical models (AtlasGPT), with error rates decreasing accordingly. Neurosurgery knowledge subspecialty-specific strengths were evident: Claude demonstrated particular strength in neuropathology and critical care (0 errors), and AtlasGPT showed expertise in neurophysiology, neuropathology, and core competencies. Persistent challenge areas included neurology, which remained difficult across all models, and procedural neurosurgical knowledge, which showed moderate error rates even in specialized models.

## 10. Error Analysis of Hallucination Cases Discussion

Our analysis of Gemini, Claude, and AtlasGPT models in clinical neurosurgical scenarios reveals valuable insights into current AI capabilities and opportunities for advancement. First, while AI systems demonstrate impressive medical knowledge breadth, specific cases like Gemini's error recommendation for SMA diagnostics highlight opportunities to enhance precision in specialized neurosurgery knowledge subspecialties. Also,

AtlasGPT's incorrect handling of the decompressive craniectomy case shows how models can accurately retrieve clinical trial data while revealing promising areas to strengthen reasoning connections between evidence and conclusions. Finally, Claude's detailed but incorrect reasoning about vertebral artery injury reflects sophisticated medical language capabilities that, with improved factual grounding, could provide valuable clinical decision support. The sophisticated medical reasoning capabilities demonstrated across different models showcase how far AI technology has advanced, even as specific error patterns indicate areas for continued refinement. With thoughtful implementation and ongoing vigilance, AI systems have tremendous potential to enhance medical practice while supporting—never replacing—clinical judgment. These findings don't diminish AI's promise in healthcare but rather illuminate the path toward more reliable, helpful clinical AI tools that can meaningfully support delivering optimal care.

## 11. Model Reasoning Quality Comparisons Discussion

The comparative assessment of clinical reasoning capabilities across five models: GPT-3.5, Mistral, Gemini 2.0, Claude 3.5, and AtlasGPT was performed through analysis of their performance on two neurosurgical case scenarios, over multiple dimensions of reasoning quality, including explanation depth, adherence to clinical guidelines, logical coherence, and hallucination rates. Our investigation revealed significant variations in reasoning approaches and quality across the examined models. The dLLM, AtlasGPT, demonstrated superior performance in structured clinical reasoning compared to sLLM alternatives. This suggests meaningful advantages derived from medical domain adaptation and specialization.

A notable finding was the presence of a depth versus precision trade-off in model responses. While some models like GPT-3.5 provided extensive explanations, these often lacked the clinical focus and precision exhibited by more specialized models. Conversely, dLLMs generally demonstrated greater efficiency in delivering clinically relevant information without superfluous content.

Regarding reasoning structure, Claude 3.5 and AtlasGPT exhibited patterns most closely aligned with clinical practice standards. Their approaches typically began with systematic symptom analysis, proceeded through differential diagnosis consideration, and concluded with evidence-based management decisions—mirroring the methodical reasoning process valued in medical practice.

The dLLM AtlasGPT consistently excelled in clinical terminology precision and mechanism explanation compared to the sLLMs alternatives. This observation highlights the tangible benefits of domain adaptation in medical AI applications, where specialized knowledge and reasoning patterns are critical for clinical utility.

Our analysis also revealed concerning vulnerabilities to hallucinations across several models. Particularly, Mistral demonstrated susceptibility to significant clinical reasoning errors. This observation underscores the potential risks associated with deploying non-specialized AI systems in medical contexts where reasoning failures could have serious consequences. These findings illuminate meaningful differences in how various models approach clinical reasoning tasks. While sLLMs can produce superficially impressive explanations, those specifically adapted to clinical domains (i.e., dLLMs) appear better equipped to emulate the reasoning processes valued in medical practice. This study emphasizes the importance of evaluating not merely the factual accuracy of model outputs but also the quality of their underlying reasoning processes when considering potential healthcare applications.

The substantial variation in reasoning quality—particularly the potential for significant reasoning failures even when reaching correct conclusions—underscores the continued necessity for careful human oversight of AI systems in clinical contexts. As model development advances, increased focus on improving clinical reasoning quality through domain specialization, guideline integration, and structured reasoning approaches may yield AI systems that can more meaningfully support the complex decision-making processes central to medical practice.

## 12. Clinical Implications

The specialized AtlasGPT (96.7% accuracy) demonstrated superior performance in neurosurgical board-style MCQs compared to sLLMs like Claude (94.7%), Gemini (92.0%), Mistral (88.7%), and GPT-3.5 (74.7%). This hierarchy suggests that dLLMs may provide more reliable clinical decision support in specialized fields. Additionally, since all models showed stronger performance in core competencies and critical care, with more challenges in neuroanatomy, neurology, and neurosurgical procedures, these variations highlight the importance of selecting appropriate AI tools for specific clinical tasks. Moreover, the hallucination error analysis revealed potential risks, including incorrect diagnostic recommendations, factual errors, and fabricated reasoning even in high-performing models demonstrated errors in procedural neurosurgical knowledge, suggesting human oversight remains essential for patient safety.

### 12.1 Educational Implications

High-performing models show potential as educational resources for medical trainees, particularly for case-based learning. Different models' strengths across Neurosurgery knowledge subspecialties suggest opportunities for targeted educational applications (e.g., using Claude for critical care concepts). Analysis of reasoning quality reveals models with clinical reasoning structures most aligned with medical practice (Claude 3.5 and AtlasGPT). These patterns can inform the development of educational tools that mirror proper clinical reasoning processes. With thoughtful implementation and ongoing vigilance, these AI systems have significant potential to enhance both clinical practice and medical education while supporting—never replacing—clinical expertise and judgment.

### 12.2 Implications for AI Development:

The success of the RAG approach in AtlasGPT highlights the importance of combining general language capabilities with specialized knowledge sources. This hybrid approach represents a promising direction for AI development, particularly in neurosurgery knowledge subspecialties requiring deep expertise and access to specialized information.

## 13. Comparison with Previous Research

The results of this study align with and extend previous research on the application of large language models in medical education and neurosurgery specifically. Prior studies by Ali et al. [4], Hopkins et al. [5], and Guerra et al. [6] demonstrated promising performance of sLLMs like GPT- 4 on neurosurgical board examinations. Our findings confirm the capabilities of advanced sLLMs while also highlighting the enhanced performance achievable through domain-specific approaches (AtlasGPT).

The accuracy rates observed for sLLMs in our study are generally consistent with those reported in previous evaluations, though direct comparisons are challenging due to variations in question sets and evaluation methodologies. The performance advantage of AtlasGPT over LLMs supports the preliminary findings reported by Hopkins et al. [10] , Ali et al. [11], and Basaran et al. [13] regarding the potential benefits of neurosurgery-specific language models. Our results also align with the broader literature on RAG, such as the work by Ram et al.[8] and Zakka et al. [9], which demonstrated improvements in model performance through the integration of external knowledge sources. The success of AtlasGPT provides further evidence for the effectiveness of this approach in specialized domains.

## 14. Limitations

Several limitations must be acknowledged in our current analysis:

***14.1 Limited Case Diversity:*** While our selected cases represent complex clinical scenarios, a broader range of clinical presentations across multiple specialties would provide more generalizable insights.

***14.2 Evaluation Subjectivity:*** Despite using standardized rubrics, assessment of reasoning quality contains inherent subjectivity that could be addressed through larger evaluation panels.

***14.3 Static Evaluation:*** Our analysis evaluates single-turn responses rather than interactive clinical reasoning, which may not fully capture the iterative nature of clinical decision-making.

***14.4 Given that AtlasGPT is built on GPT-4:*** a direct comparison with GPT-4 would better isolate the impact of domain-specific fine-tuning and RAG. Moreover, comparison with more state-of-the-art models would be good as well.

***14.5 Text-Only Questions:*** The evaluation was limited to text-only MCQs, excluding image-based questions that constitute an important component of neurosurgical assessment and practice. The ability to interpret and reason from medical images represents a distinct skill set that was not evaluated in this study.

***14.5 MCQ Format Limitations:*** Multiple-choice questions, while standardized and quantifiable, may not fully capture the complexity of medical reasoning and decision-making. The binary correct/incorrect scoring system does not account for nuances in understanding or reasoning processes.

***14.6 Absence of Real-World Clinical Scenarios:*** The board-style questions used in this evaluation, while designed to assess neurosurgical knowledge, differ from the complexity and ambiguity of real-world clinical scenarios. Performance on these questions may not directly translate to performance in clinical practice.

***14.7 Temporal Limitations:*** The evaluation was conducted at a specific point in time with specific versions of each model. Given the rapid pace of development in AI, the relative performance of these models may change as they are updated and improved.

***14.8 Limited Information about Model Specifics:*** Detailed information about the training data, architectural specifications, and fine-tuning processes of some models, particularly the proprietary ones, was not fully available. This limits our ability to analyze the specific factors contributing to performance differences.

***14.9 Single Domain Focus:*** The study focused exclusively on neurosurgery, and findings may not generalize to other medical specialties.

## 15. Future Research Directions

Based on the findings and limitations of this study, several promising directions for future research emerge:

***15.1 Multimodal Evaluation:*** Future studies should incorporate image-based questions and other multimodal content to provide a more comprehensive assessment of model capabilities relevant to medical practice.

***15.2 Real-World Clinical Applications:*** Evaluating model performance on real-world clinical scenarios, including complex cases with ambiguity and incomplete information, would provide insights into their potential utility in clinical practice.

***15.3 Longitudinal Assessment:*** Tracking model performance over time as new versions are released would help understand the trajectory of improvement and identify areas of persistent challenge.

***15.4 User Experience Studies:*** Investigating how medical students, residents, and practicing clinicians interact with and perceive these models would provide valuable insights into their practical utility and integration into medical education and practice.

***15.5 Comparative Analysis of Domain-Specific Approaches:*** Comparing different approaches to domain specialization (fine-tuning, RAG, prompt engineering, etc.) across multiple medical specialties would help identify the most effective strategies for developing specialized medical AI tools.

***15.6 Error Analysis:*** Detailed analysis of the types of questions and topics where models make errors would provide insights into specific knowledge gaps and areas for improvement.

***15.7 Reasoning Quality Studies:*** Expanding to diverse clinical scenarios across multiple specialties, developing more standardized metrics for reasoning quality assessment, implementing interactive evaluation protocols to assess iterative clinical reasoning, and conducting prospective studies on the impact of LLM reasoning assistance on clinical decision quality

***15.8 Ethical and Implementation Considerations:*** Exploring the ethical implications of AI integration in medical education and practice, including issues related to accountability, transparency,

and appropriate reliance on AI tools.

## 16. Conclusions

This comprehensive evaluation of language models in neurosurgery demonstrates the significant potential of domain-specific approaches like AtlasGPT. The performance hierarchy observed—with the domain-specific model outperforming even advanced sLLMs—provides compelling evidence for the value of specialized knowledge integration in AI systems for healthcare applications. Our multidimensional analysis, which examined not only accuracy but also error distribution, hallucination patterns, and reasoning quality with the lowest occurrences being skewed to AtlasGPT over all sLLMs and revealing both the promise and persistent limitations of current AI technologies in neurosurgery. The findings suggest several important implications for clinical practice and medical education. While high-performing models show potential as educational tools and decision support systems, the identified hallucination patterns and reasoning limitations underscore the continued necessity for human oversight. The RAG approach implemented in AtlasGPT demonstrates a particularly promising direction for AI development in specialized domains, combining the strengths of general language capabilities with access to domain-specific knowledge. Despite limitations including the text-only format and static evaluation approach, this study advances our understanding of how AI systems can be optimized for specialized medical applications. Future research should expand to multimodal content, real-world clinical scenarios, and longitudinal assessment to further refine these technologies. As AI continues to evolve, thoughtful implementation focusing on domain specialization, structured reasoning approaches, and appropriate integration into clinical workflows will be essential to realize the potential of these technologies while prioritizing patient safety and care quality.

## References


1. Channa, R., Wolf, R. M., Abràmoff, M. D., & Lehmann, H. P. (2023). Effectiveness of artificial intelligence screening in preventing vision loss from diabetes: a policy model. *NPJ digital medicine, 6*(1), 53.
2. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health, 2*(2), e0000198.
3. Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375.*
4. Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Sullivan, P. L. Z., ... & Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery, 93*(5), 1090-1098.
5. Hopkins, B. S., Nguyen, V. N., Dallas, J., Texakalidis, P., Yang, M., Renn, A., ... & Mack, W. J. (2023). ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board–style questions. *Journal of Neurosurgery, 139*(3), 904-911.
6. Guerra, G. A., Hofmann, H., Sobhani, S., Hofmann, G., Gomez, D., Soroudi, D., ... & Zada, G. (2023). GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. *World neurosurgery, 179*, e160-e165.
7. Bečulić, H., Begagić, E., Skomorac, R., Mašović, A., Selimović, E., & Pojskić, M. (2024). ChatGPT's contributions to the evolution of neurosurgical practice and education: a systematic review of benefits, concerns and limitations. *Medicinski Glasnik, 21*(1).
8. Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics, 11,* 1316-1331
9. Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., ... & Hiesinger, W. (2024). Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai, 1*(2), AIoa2300068.
10. Hopkins, B. S., Carter, B., Lord, J., Rutka, J. T., & Cohen-Gadol, A. A. (2024). AtlasGPT: dawn of a new era in neurosurgery for intelligent care augmentation, operative planning, and performance. *Journal of Neurosurgery, 140*(5), 1211-1214.
11. Ali, R., Abdulrazeq, H. F., Patil, A., Cheatham, M., Connolly, I. D., Tang, O. Y., ... & Asaad, W. F. (2025). AtlasGPT: a language model grounded in neurosurgery with domain-specific data and document retrieval. *Journal of Neurosurgery, 1*(aop), 1-8.
12. O'Malley, G. R., Sarwar, S., Kazarian, E., Chan, A., Kilgallon, J., Ali, M., ... & Patel, N. V. (2025). 1133 Evaluating the Accuracy, Variation, And Sources of Large Language Model Responses to Patient Questions in Neurosurgery: An Assessment of Popular Large Language Models and Emerging Platforms. *Neurosurgery, 71*(Supplement_1), 173.
13. Basaran, A. E., Güresir, A., Knoch, H., Vychopen, M., Güresir, E., & Wach, J. (2025). Beyond traditional prognostics: integrating RAG-enhanced AtlasGPT and ChatGPT 4.0 into aneurysmal subarachnoid hemorrhage outcome prediction. *Neurosurgical Review, 48*(1), 1-10.
14. Mbakwe, A. B., Lourentzou, I., Celi, L. A., Mechanic, O. J., & Dagan, A. (2023). ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS digital health, 2*(2), e0000205.
15. Ejaz, H., McGrath, H., Wong, B. L., Guise, A., Vercauteren, T., & Shapey, J. (2022). Artificial intelligence and medical education: A global mixed-methods study of medical students' perspectives. *Digital Health, 8*, 20552076221089099.
16. Shah, R. S., Cadoux-Hudson, T. A., Van Gompel, J. J., & Pereira, E. (2016). *Neurosurgery Self-Assessment E-Book: Questions and Answers*. Elsevier Health Sciences.
17. Puffer, R. C. (2019). *Neurosurgery Primary Board Review*. Georg Thieme Verlag.



**Copyright:** *©2025 Brandon L. Staple, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*