ISSN: 2690-909X

Research Article

Advances in Neurology and Neuroscience

Comparison of Domain-Specific and Ensemble Large Language Models in Surgical Education: A Preliminary Performance Evaluation

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

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

[2]META, Seattle, WA , United States of America

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

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

Submitted: 2025, May 08; **Accepted:** 2025, Jun 09; **Published:** 2025, Jun 17

Citation: Staple, B. L., Staple, E. M., Wallace, C. Staple, D. B. (2025). Comparison of Domain-Specific and Ensemble Large Language Models in Surgical Education: A Preliminary Performance Evaluation. *Adv Neur Neur Sci, 8*(2), 01-15.

Abstract

Standard Large Language Models (sLLMs) are known for their high accuracy in answering multiple-choice questions from the Self-Assessment Neurosurgery Exam (SANS). However, their tendency to 'hallucinate' or fabricate information presents challenges for neurosurgical applications that require a high degree of precision. AtlasGPT, a Domain-specific Large Language Model (dLLM), has managed to achieve a lower Hallucination Rate (HR) through targeted fine-tuning and retrieval-augmented generation from specialized databases. Nevertheless, proprietary limitations hinder customization and broader research into hallucination mitigation, prompting an exploration of model-agnostic Ensemble Methods (EMs) that combine several sLLMs to enhance performance. This study assessed hallucination mitigation by comparing an EM consisting of three sLLMs (Gemini, Claude 3.5 Sonnet, and Mistral) with AtlasGPT. Hallucination Rates were evaluated using sampling and maximum voting on 150 SANS multiple-choice-questions. The EM achieved the highest accuracy (97.33%, HR: 2.67%), slightly surpassing AtlasGPT (96.67%, HR: 3.33%) and outperforming all individual sLLMs: Claude 3.5 Sonnet (94.67%, HR: 5.33%), Gemini (92.00%, HR: 8.00%), and Mistral (88.67%, HR: 11.33%). The success of the EM lies in its integration of multiple sLLMs, which minimizes errors through diverse training data and enhances outcomes by leveraging varied data sources. This diversity principle, where different models make different errors, reduces overall mistakes and improves performance, leading to better generalization and robustness. Also, EMs can effectively use both open-source and proprietary models, with simple sampling and maximum voting approaches proving effective, suggesting that complex methods may not be necessary for specific applications. In summary, specialized models like AtlasGPT and EM demonstrate the value of domain-specific training and multi-model approaches. However, they require further development and human oversight for safe clinical deployment.

Keywords: Large Language Models, Hallucination Mitigation, Ensemble Methods, Neurosurgery, Medical AI, Clinical Decision Support

1. Introduction

Recent investigations have demonstrated that Standard Large Language Models (sLLMs), such as GPT-4, exhibit notable performance when evaluated using multiple-choice questions (MCQs) from the Congress of Neurological Surgeons (CNS) Self-Assessment in Neurosurgery (SANS) examination [1-6]. Despite their impressive capabilities, these models occasionally generate responses containing inaccuracies or fabricated information—a phenomenon termed "hallucination"[7,8]. In critical medical specialties like neurosurgery, the management of hallucinations is paramount, as such inaccuracies may adversely influence diagnostic reasoning and therapeutic decision-making [9-11]. AtlasGPT, a recently developed Domain-specific Large Language Model (dLLM), has demonstrated reduced Hallucination Rates (HRs) compared to conventional sLLMs [12-15]. This improvement is attributed to fine-tuning of the GPT-4 architecture and integration of retrieval-augmented generation (RAG) methodology, which enables access to contextually relevant information from a comprehensive vector database of neurological literature. Despite these advancements and the emergence of open-source alternatives such as Llama 3, significant barriers persist in the medical implementation of these technologies. A substantial limitation is that the most sophisticated and accurate models—including GPT-4, Claude 3.5 Sonnet, and Gemini—operate as proprietary, closed systems [16]. The restricted access to underlying code and training datasets diminishes model transparency and constrains customization possibilities, thereby impeding research initiatives focused on hallucination mitigation strategies [17,18].

To address these limitations, we investigated the efficacy of Ensemble Methods (EMs), which integrate multiple individual sLLMs to potentially reduce HRs compared to single-model approaches. EMs attenuate the impact of hallucinations from any particular sLLM by leveraging algorithmic and dataset diversity across multiple models, thus enhancing error resilience in medical contexts. Furthermore, since EMs function independently of a model's source code or training data, they are compatible with both open-source and proprietary architectures, broadening their applicability for hallucination mitigation. Established EM approaches include various voting mechanisms (e.g., sampling and maximum voting, bagging, and boosting) in both homogeneous and heterogeneous configurations [19].

This investigation expands the understanding of hallucination mitigation in neurosurgery by comparing the HR of an EM comprising three state-of-the-art sLLMs (Gemini, Claude 3.5 Sonnet, and Mistral) against that of the neurosurgery-specific dLLM, AtlasGPT. We employed a heterogeneous sampling and maximum voting EM technique, which aggregates outputs from individual sLLMs, selecting the most frequently occurring output as the definitive response [20-23].

For optimal EM functionality, two fundamental criteria must be satisfied, as these methodologies depend on the inherent diversity of individual model errors to achieve mutual compensation [24].

1.1. Model Diversity: Constituent models must exhibit diversity, such that their error patterns demonstrate minimal correlation.
1.2. Comparable Performance: Each component model should demonstrate relatively low and similar HRs to prevent disproportionate influence from any single model on the EM's output.

For this investigation, we selected Gemini, Claude 3.5 Sonnet, and Mistral as our component sLLMs, based on their ability to fulfill the aforementioned criteria for effective EM formation.

Due to the proprietary nature and inaccessibility of the SANS examination MCQs, our model evaluations utilized a composite set of 150 text-only MCQs with validated correct answers as a proxy benchmark. The question bank was derived from two authoritative sources: Neurosurgery Self-Assessment Questions and Answers and Neurosurgery Primary Board Review—both recognized as essential resources for neurosurgical board preparation and containing contemporary question banks specifically designed for SANS examination preparation [25,26]. This comprehensive question repository encompasses the breadth of neurosurgical domains assessed in board examinations, including neuroanatomy, neurophysiology, neuropathology, neuroimaging, neurology, neurosurgical procedures, critical care management, and core competencies.

We operationalized HR as the complement of accuracy (100% minus accuracy), where accuracy represents the percentage of instances in which the model generates responses that are both factually correct and contextually appropriate to the query [27,28].

2. Related Work

2.1. SLLMs in Medicine

Most language-model medical assessments in the literature utilize sLLMs. In neurosurgery specifically, Buchanan et al. assessed GPT-4's performance on neurosurgical board-style questions, finding an accuracy of approximately 87%, which approached but did not match the performance level expected of neurosurgical residents preparing for board examinations [29].

2.2. The Challenge of Hallucinations in Medical SLLMs

Despite their potential benefits, sLLMs face a significant limitation known as "hallucinations"—the generation of content that appears plausible but is factually incorrect or unfounded. In the context of Nuerosurgery, addressing the issue of hallucinations is essential, as they may result in considerable misinformation, bias, and inaccuracies that negatively impact diagnostic procedures and treatment outcomes. For example, Thirunavukarasu et al. highlighted the potential for sLLM hallucinations to compromise patient safety if integrated into clinical decision support without appropriate safeguards [30]. Several studies have attempted to quantify hallucination rates across different medical specialties. For example, Lievin et al. found hallucination rates ranging from 5% to 23% when evaluating responses from various sLLMs to medical questions, with performance varying significantly by medical specialty and question complexity. Notably, they observed that more complex specialties with rapidly evolving literature, such as oncology and neurology, tended to elicit higher hallucination rates compared to more stable knowledge domains [31].

The underlying mechanisms of hallucination in sLLMs remain an active area of research with several factors identified as contributors.

2.2.1. Training Data Limitations: Despite being trained on vast text corpora, sLLMs may encounter knowledge gaps in specialized medical domains Specifically, there are training data limitations wherein the quality, comprehensiveness, and recency of training data fundamentally influence a language model's tendency to hallucinate. More specifically, sLLMs are trained to produce wide-ranging generalizations across different fields, frequently overlooking the subtle context and specialized terminology required for particular domains, like neurosurgery.

2.2.2. Parametric Knowledge Limitations: Training objectives that prioritize generating fluent, human-like text may inadvertently encourage confabulation when the model is uncertain.

2.2.3. Parametric Knowledge Limitations: The storage of factual knowledge in model parameters is inherently imprecise compared to explicit knowledge databases.

2.2.4. Context Window Constraints: Limited context windows may prevent models from accessing all relevant information needed for accurate responses.

2.2.5.Distribution Shift: Medical knowledge evolves rapidly, creating potential discrepancies between training data and current medical consensus.

2.2.6.Token Predictions: Language models operate by predicting the most likely next tokens based on learned statistical patterns rather than through causal reasoning or factual understanding.

2.2.7.High Confidence under Uncertainty: Language models often provide responses with high confidence even when operating in domains of uncertainty.

For sLLMs, these inherent factors create model vulnerability to hallucinations when generating responses to complex and specialized domain inquiries like in spine care.

3. DLLMs and Atlas GPT

In response to the limitations of sLLMs in specialized domains, researchers have increasingly focused on developing dLLMs that are tailored to particular fields or applications. These models typically leverage transfer learning approaches, where a pre-trained general- purpose language model is further fine-tuned on domain-specific corpora to enhance its performance in targeted applications. In the medical domain, this approach has given rise to models such as Med-PaLM Singhal et al., [32] designed to better handle the specialized vocabulary, concepts, and reasoning patterns characteristic of medical discourse. AtlasGPT represents a significant advancement in dLLMs for neurosurgery [12-15]

Developed by fine-tuning the GPT-4 architecture on a comprehensive corpus of neurosurgical literature, AtlasGPT incorporates several innovative features designed to enhance its performance and reduce hallucinations in this specialized context:

3.1. Specialized Neurosurgical Corpus: AtlasGPT was fine-tuned on a carefully curated dataset comprising peer-reviewed neurosurgical research papers, standard textbooks, clinical guidelines, case reports, and educational materials specifically related to neurosurgery.

3.2. RAG: Unlike sLLMs that rely solely on parametric knowledge (information encoded in the model's parameters during training), AtlasGPT implements a RAG architecture that dynamically retrieves relevant information from an external vector database of neurosurgical literature before generating responses. This approach helps to ground the model's outputs in verifiable source material, reducing hallucinations

3.3. Knowledge Graph Integration: AtlasGPT incorporates a specialized neurosurgical knowledge graph that maps relationships between anatomical structures, pathologies, surgical approaches, and clinical outcomes. This structured representation of domain knowledge provides an additional mechanism for fact-checking generated content

3.4. Citation Mechanism: The model implements an explicit citation system that references specific sources from its knowledge base when providing information, enhancing transparency and facilitating verification.

Several studies have evaluated AtlasGPT's performance in neurosurgical applications where it outperforms sLLMs by significant margins. Further demonstrated that AtlasGPT's hallucination rate was approximately 4%, representing a substantial improvement over the 8-15% range typically observed in sLLMs when applied to neurosurgical questions [12-15].

4. The Challenge of Hallucinations in Medical dLLMs

Despite these impressive results, dLLMs like AtlasGPT face several limitations:

***4.1.* Development Cost:** Creating high-quality dLLMs requires substantial resources for data curation, model training, and validation, making them less accessible to smaller research groups or healthcare institutions [33].

***4.2.* Maintenance Challenges:** Medical knowledge evolves rapidly, necessitating regular updates to the model and its associated knowledge bases to prevent obsolescence .

***4.3.* Narrow Applicability:** By design, dLLMs excel in their target domain but may perform poorly when faced with questions that span multiple specialties or require general medical knowledge [31].

***4.4.* Potential for Overfitting:** Excessive specialization may lead to overfitting on the training corpus, potentially limiting the model's ability to generalize to novel scenarios or rare cases

***4.5. Accessibility Barriers*:** Many advanced dLLMs, including AtlasGPT, are built on proprietary base models with restricted access, limiting their availability to the broader medical community

These limitations have motivated the exploration of alternative approaches to enhancing sLLM performance in specialized domains, including the EM that form the focus of the present study.

5. Ensemble Methods for SLLMs in Healthcare

Ensemble methods, which combine multiple models to improve predictive performance, have a rich history in machine learning dating back several decades. These approaches originated in the broader statistical and machine learning communities, with landmark developments including bagging, boosting, and stacking. The fundamental insight underlying EM is that combining diverse models can often yield better performance than any individual model alone, particularly when the errors made by different models are uncorrelated. In the context of neural networks and deep learning, EM techniques gained prominence through work on model averaging, dropout-based ensembling, and deep EMs. These approaches demonstrated that combining multiple neural networks could not only improve accuracy but also provide better uncertainty estimates—a crucial consideration in high-stakes applications [19-24].

The application of EM methods to sLLMs represents a relatively recent development, spurred by the emergence of multiple competing model architectures and the recognition that different models may exhibit complementary strengths and weaknesses. Wang et al. categorized sLLM EM methods into several broad approaches [34]:

5.1. **Model-level Ensembling:** Combining outputs from multiple distinct sLLM architectures or variations (e.g., GPT-4, Claude, Gemini).

5.2. **Prompt-level Ensembling:** Using varied prompting strategies with the same model and aggregating the results.

5.3. **Decoding-level Ensembling:** Applying different sampling or decoding strategies during text generation and combining the outputs.

5.4. **Hybrid Ensembling:** Integrating traditional rule-based systems or knowledge bases with SLLM outputs.

In healthcare applications, EM methods have shown particular promise for reducing hallucinations and improving reliability. Yan demonstrated that an EM of three different medical SLLMs achieved a 27% reduction in hallucination rate compared to the best individual model when answering complex medical questions [35].

Several mechanisms have been proposed to explain why EM methods may be particularly effective for mitigating hallucinations:

5.5. **Error Cancellation:** When models make uncorrelated errors, the maximum voting process can filter out idiosyncratic hallucinations produced by individual models.

5.6. **Confidence Calibration:** EM methods implicitly incorporate a form of confidence weighting, as propositions with stronger evidential support are more likely to appear consistently across multiple models.

5.7. **Knowledge Complementarity:** Different models, trained on partially overlapping but distinct datasets, may possess complementary knowledge, allowing the EM to leverage a broader effective knowledge base than any individual model [34].

5.8. **Architectural diversity:** Models with different architectures may have varying inductive biases, enabling the EM to benefit from multiple approaches to knowledge representation and reasoning [34].

6. The Challenge of Hallucinations in EMs

Despite these advantages, EM methods for sLLMs in healthcare face several practical challenges:

6.1. **Computational Overhead:** Running multiple sLLMs simultaneously requires substantially more computational resources than using a single model, potentially limiting real-time applications.

6.2. **Integration Complexity:** Effectively combining outputs from models with different response formats, confidence levels, and reasoning patterns requires sophisticated aggregation strategies.

6.3. **Cost Considerations:** For commercial API-based models, EM approaches multiply the cost per query, potentially making them economically prohibitive for large-scale applications.

6.4. **Potential for Compromise:** In certain cases, EM methods might actually reduce performance if poorly calibrated models dominate the voting process or if consensus leads to overly conservative outputs [34].

7. Gaps in the Existing Literature

Despite significant advancements in the application of Large Language Models to neurosurgical knowledge assessment, several critical gaps exist in the current literature:

7.1. Limited Comparative Analysis of Hallucination Mitigation Strategies: While substantial research exists on the performance of individual sLLMs and dLLMs in neurosurgery, there is a paucity of rigorous comparative studies directly evaluating the efficacy of different hallucination mitigation approaches. Specifically, few studies have systematically compared model-agnostic EM techniques against domain-specific fine-tuning and retrieval augmentation for reducing hallucinations in specialized medical domains.

7.2. Inaccessibility of Advanced Model Architectures: The most accurate and capable language models remain proprietary and closed-source, limiting researchers' ability to modify model architectures or training methodologies directly. This restriction impedes the development and evaluation of novel hallucination mitigation techniques that require access to model internals, particularly in high-stakes medical domains where reliability is paramount.

7.3. Insufficient Exploration of EM Method Efficacy in Specialized Medical Domains: While EM methods have demonstrated success in general machine learning applications and broader NLP tasks, their specific applicability and effectiveness for reducing hallucinations in highly specialized medical fields like neurosurgery remain underexplored. Previous research has not adequately assessed whether the theoretical advantages of ensembling hold true when addressing complex domain-specific medical questions that demand specialized knowledge.

7.4. Lack of Standardized Evaluation Frameworks for Medical Language Model Hallucinations: The literature lacks consistent methodologies for quantifying and comparing hallucination rates across different model types and configurations in medical contexts. This methodological gap complicates efforts to benchmark progress and identify the most promising approaches for enhancing model reliability in clinical applications.

7.5. Insufficient Investigation of Model Complementarity in Medical Knowledge: Limited research exists examining whether different language model architectures exhibit complementary knowledge patterns in specialized medical domains, which is a foundational assumption underlying the efficacy of EM methods for hallucination reduction.

8. Study's Novel Contributions to Address the Gaps

This study addresses these significant gaps through several novel contributions:

8.1. Direct Comparative Analysis of Leading Hallucination Mitigation Approaches: The research provides the first direct comparison between a state-of-the-art dLLM with retrieval augmentation (AtlasGPT) and a heterogeneous EM of sLLMs in the context of neurosurgical knowledge assessment. This comparative analysis offers crucial insights into the relative efficacy of these distinct approaches for mitigating hallucinations in specialized medical domains.

8.2. Model-Agnostic Framework for Hallucination Reduction: By implementing and validating a heterogeneous sampling and maximum voting EM technique, this study establishes a practical framework for reducing hallucinations without requiring access to proprietary model architectures or training data. This contribution is particularly valuable given the closed-source nature of the most advanced sLLMs, offering researchers and practitioners accessible methods for enhancing model reliability.

8.3. Empirical Validation of EM Method Efficacy in Neurosurgery: Through rigorous statistical analysis, this research empirically demonstrates that properly constructed EMs of general-purpose sLLMs can achieve hallucination rates comparable to or better than specialized domain-specific models in neurosurgery. This finding challenges assumptions about the necessity of domain-specific fine-tuning and retrieval augmentation for achieving minimal hallucination rates in specialized medical fields.

8.4. Robust Evaluation Methodology for Medical SLLM Performance: The study establishes a comprehensive evaluation framework using a proxy benchmark derived from established neurosurgical board preparation resources, offering a replicable methodology for assessing hallucination rates in specialized medical domains where proprietary assessment tools may be inaccessible.

8.5. Practical Implementation Guidelines for Effective Medical EMs: By identifying and fulfilling the essential conditions for effective EM performance—model diversity and comparable individual performance—this research provides actionable guidance for constructing effective model EMs in medical applications. These implementation guidelines address the practical challenges of deploying EM methods in resource-constrained healthcare settings.

This study's contributions collectively advance the understanding of hallucination mitigation in specialized medical domains and provide practical methodologies for enhancing the reliability of AI systems in neurosurgical applications, with potential extensions to other medical specialties requiring similar levels of domain expertise. Our study builds on this foundation by implementing a dLLM and a heterogeneous sampling and maximum voting EM technique with three state-of-the-art sLLMs, specifically chosen to fulfill the essential conditions for effective EM performance in the challenging domain of neurosurgery.

9. Materials and Methods

9.1. Study Design and Overview

This study employed a comparative experimental design to evaluate the HR of four distinct sLLM configurations in the context of neurosurgical knowledge assessment. The primary comparison was between an EM comprising three sLLMs and AtlasGPT, a dLLM tailored for neurosurgery. The study was conducted in three sequential phases:

1. Individual assessment of each sLLM's performance on neurosurgical MCQs
2. Implementation of the EM method through sampling and maximum voting
3. Statistical comparison of hallucination rates across all models and the ensemble.

9.2. Model Selection

For this study, we selected three state-of-the-art sLLMs to form the EM:

9**.2.1. Claude 3.5 Sonnet:** Developed by Anthropic, this model has demonstrated strong performance across a wide range of natural language processing tasks, including medical question answering. Claude 3.5 Sonnet utilizes a transformer-based architecture with an estimated 140 billion parameters and was selected for its strong reasoning capabilities and relatively low hallucination rates in general knowledge domains.

9.2.2. Gemini: Developed by Google, Gemini represents one of the most advanced multimodal AI systems currently available. The Gemini Pro variant used in this study has demonstrated exceptional performance on medical knowledge benchmarks, though specific parameter counts have not been publicly disclosed.

9.2.3Mistral: An advanced open-source large language model developed by Mistral AI; this model was included to provide architectural diversity in the EM. Mistral employs a mixture-of-experts architecture and has shown impressive performance despite having a more compact parameter count compared to the other selected models.

These models were selected to fulfill the two essential conditions for effective EM performance:
9.2.4. Diversity Condition: Each model employs different architectural designs, training methodologies, and base datasets, increasing the likelihood that their errors will be largely uncorrelated.

9.2.5. Performance Condition: All selected models have demonstrated strong performance on general medical knowledge tasks in previous evaluations, suggesting they would maintain relatively low and comparable hallucination rates.

In addition to these three sLLMs, we evaluated AtlasGPT, a dLLM fine-tuned for neurosurgery. AtlasGPT incorporates a RAG architecture that enhances the base GPT-4 model with access to a structured database of neurosurgical literature, potentially enabling more accurate responses to specialized questions in this domain.

10. Question Dataset

Due to the proprietary and inaccessible nature of the CNS SANS, we constructed a proxy benchmark comprising 150 text-only multiple-choice questions (MCQs) with known correct answers. These questions were sourced from two well-established neurosurgical board preparation resources:
1. Neurosurgery Self-Assessment Questions and Answers [25].
2. Neurosurgery primary board review [26].

Both resources are widely recognized as essential study materials for neurosurgical residents preparing for board examinations and contain current question banks specifically designed to align with the content and format of the SANS written exam. The compiled question set encompassed various categories of neurosurgical knowledge, including:
1. Neuroanatomy
2. Neurophysiology
3. Neuropathology
4. Neurology
5. Neuroimaging
6. Neurosurgery techniques and approaches
7. Critical care management
8. Core competencies in patient care

Each question in the dataset followed a standard multiple-choice format with four or five answer options (A, B, C, D, and occasionally E), with exactly one correct answer per question. The questions were presented without accompanying images or figures to ensure compatibility with text-based sLLM evaluation.

11. EM Method Implementation

We implemented a heterogeneous sampling and maximum voting EM technique to combine the outputs of the three selected sLLMs. This process consisted of the following steps:
***11.1.* Individual Model Sampling:** Each of the 150 neurosurgical MCQs was presented to each of the three sLLMs (Claude 3.5 Sonnet, Gemini, and Mistral) in sequential order. For each question, the model was instructed to select the single best answer from the provided options without explanations or additional commentary. Each model's responses were documented verbatim.

***11.2.* Maximum Voting:** For each question, we identified the most frequently occurring answer among the three sLLMs. This aggregate response was designated as the "EM answer" for subsequent evaluation.

***11.3.* Comparison with AtlasGPT:** The same 150 questions were presented to AtlasGPT under identical conditions, and its responses were recorded for comparison with the EM.

12. Evaluation Metrics

Performance was evaluated using the following metrics:
***12.1.* Accuracy:** The percentage of questions for which a model or the EM provided the correct answer, defined as: Accuracy = (Number of Correct Responses / Total Number of Questions) × 100%

***12.2.* Hallucination Rate (HR):** Defined as the complement of accuracy, representing the percentage of questions for which a model or the EM provided an incorrect answer: HR = 100% - Accuracy

This definition of hallucination rate aligns with common practice in language model evaluation literature, where hallucinations in the context of MCQ are operationalized as selections of incorrect answers [27].

13. Statistical Analysis

Statistical analysis was conducted to determine whether observed differences in hallucination rates between models and the EM were statistically significant. The analysis proceeded as follows:
***13.1.* Analysis of Variance (ANOVA):** A one-way ANOVA with a 95% confidence interval was performed to test for significant differences in the means across all model configurations (Mistral, Gemini, Claude 3.5 Sonnet, AtlasGPT, and the EM Method).

***13.2.* Post-hoc Analysis:** Since ANOVA only indicates whether significant differences exist among the groups but does not specify which specific groups differ from each other, we conducted a Bonferroni post-hoc analysis with a significance level of $\alpha < 0.01$. This conservative approach helps control for the family-wise error rate that occurs when making multiple pairwise comparisons.

13.3. Pairwise Comparisons: We performed pairwise comparisons between all model configurations to identify specific significant differences in performance.

All statistical analyses were performed using Microsoft Excel.

14. Results

14.1. Raw Data Analysis

Appendix A, Table A1 presents three exemplar cases from our raw data demonstrating the potential severe medical consequences arising from model hallucinations. These consequences include:

***14.1.1.* Diagnostic Delay Impacting Time-Sensitive Intervention:** Inappropriate prioritization of nerve conduction studies over SMN gene testing would delay definitive diagnosis of Spinal Muscular Atrophy (SMA), a progressive neurodegenerative disorder where early therapeutic intervention is critical for disease modification.

***14.1.2.* Elevated Surgical Risk** Selection of C1-C2 transarticular screws rather than C1 lateral mass screws in the presence of ponticulus posticus significantly increases vertebral artery injury risk, potentially resulting in stroke, posterior circulation ischemia, or fatal hemorrhage.

***14.1.3.* Compromised Neurosurgical Outcomes:** Management focused on protamine reversal complications would inappropriately delay definitive intervention for intracranial hemorrhage, potentially exacerbating neurological sequelae.

The complete raw dataset is available upon request from the corresponding author.

15. Individual Model Performance

Table 1 presents the performance metrics for each individual sLLM, AtlasGPT, and the EM method across 150 neurosurgical multiple-choice questions. . The EM achieved the highest accuracy (97.33%, HR: 2.67%), slightly surpassing AtlasGPT (96.67%, HR: 3.33%) and outperforming all individual sLLMs: Claude 3.5 Sonnet (94.67%, HR: 5.33%), Gemini (92.00%, HR: 8.00%), and Mistral (88.67%, HR: 11.33%).

Model	Correct Answers	Total Questions	Accuracy (%)	Hallucination Rate (%)
Mistral	133	150	88.67	11.33
Gemini	138	150	92.00	8.00
Claude 3.5 Sonnet	142	150	94.67	5.33
AtlasGPT	145	150	96.67	3.33
EM Method	146	150	97.33	2.67

Table 1: Performance Metrics for Individual Models and EM Method

16. Statistical Analysis

A one-way analysis of variance (ANOVA) was conducted to evaluate the effect of model configuration on accuracy rates across the five experimental conditions (Mistral, Gemini, Claude 3.5 Sonnet, AtlasGPT, and EM Method) As shown in Table 2 results indicated a significant effect of model configuration on accuracy at the $p < 0.01$ level [$F(4, 745) = 3.36$, $p = 0.0097$], suggesting statistically significant performance differences between at least two model configurations.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.765333	4	0.191333	3.360814	0.0097233	2.383885
Within Groups	42.41333	745	0.056931			
Total	43.17867	749				

Table 2: ANOVA Results

To determine specific significant differences between model configurations, we conducted a Bonferroni post-hoc analysis with a significance threshold of $\alpha < 0.01$ (Table 3). This analysis revealed that the EM Method performed significantly better than Mistral ($p = 0.0032$), with the EM's hallucination rate (2.67%) being substantially lower than Mistral's (11.33%). Similarly, AtlasGPT demonstrated significantly superior performance compared to Mistral ($p = 0.0078$), with AtlasGPT's hallucination rate (3.33%) being considerably lower than Mistral's (11.33%). No other pairwise comparisons achieved statistical significance at the $\alpha < 0.01$ level, though several approached significance (e.g., Gemini vs. EM, $p = 0.0399$). Notably, the performance difference between AtlasGPT and the EM Method was not statistically significant ($p = 0.7361$), suggesting comparable efficacy between these two approaches.

Model Comparison	P-value	Significant at α < 0.01
Mistral vs. Gemini	0.3302601729	No
Mistral vs. Claude	0.0628174165	No
Mistral vs. AtlasGPT	0.0077584909	Yes
Mistral vs. EM	0.0031690629	Yes
Gemini vs. Claude	0.3562079582	No
Gemini vs. AtlasGPT	0.0809571913	No
Gemini vs. EM	0.0399514968	No
Claude vs. AtlasGPT	0.3966349897	No
Claude vs. EM	0.2400133133	No
AtlasGPT vs. EM	0.7360656439	No

Table 3: Bonferroni Post-hoc Analysis Results (P-values)

17. Error Pattern Analysis

We conducted a detailed analysis examining the distribution of errors across neurosurgical knowledge domains to assess whether errors were randomly distributed or concentrated within specific knowledge categories. Table 4 and Figure 1 present the error distribution across eight neurosurgical domains for each model evaluated.

18. Discussion

18.1. Examining sLLMs, EM, and AtlasGPT Performance

While AtlasGPT demonstrated significant efficacy in the field of neurosurgery, EM offer distinct advantages that may account for their slight superiority in accuracy. Firstly, EMs are distinguished by their accuracy, precision, and robustness, achieved through the integration of multiple sLLMs, which minimizes errors and enhances overall outcomes. For example, in the diagnosis of conditions such as spinal muscular atrophy (SMA), EM strategies amalgamate predictions from individual sLLMs trained on wide spectrum of data including those from various medical specialties leading to potentially more thorough and accurate results. Furthermore, EMs exhibit exceptional generalization and adaptability across a range of domains and tasks. By synthesizing models that focus on different facets of a discipline, they can provide precise evaluations even in varied contexts. This adaptability is advantageous in medical fields, where the sources of data and the models required can be quite diverse. Also, the theoretical underpinnings of EM techniques further validate their effectiveness. The principle of diversity—whereby models commit different errors—contributes to a reduction in overall mistakes and an enhancement in performance. Consequently, the diversity inherent in EMs tends to facilitate better generalization and greater robustness. Thus, even though AtlasGPT is notably effective within its specialties, ensemble methods are distinguished by their robustness and ability to generalize. Their capacity to integrate multiple models, enhance accuracy, and reduce bias positions them as alternatives to dLLMs in the realm of medicine.

Question Category	Mistral Errors	Gemini Errors	Claude Errors	AtlasGPT Errors	EM Errors
Neuroanatomy	4	2	2	1	0
Neurophysiology	4	2	1	0	0
Neuropathology	1	2	0	0	0
Neuroimaging	1	0	0	0	0
Neurology	2	2	3	1	2
Neurosurgery	2	2	1	2	1
Critical Care	2	1	0	1	0
Core Competencies	1	1	1	0	1
Total	17	12	8	5	4

Table 4: Error Distribution by Neurosurgery Question Category

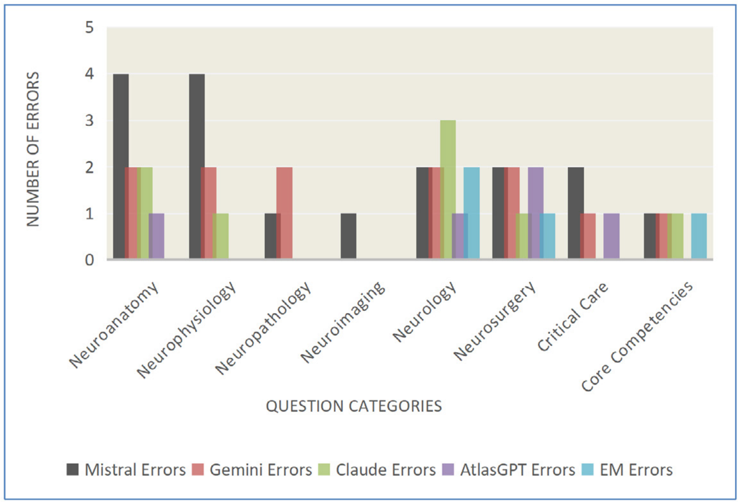

Figure 1: Error Distribution by Neurosurgery Question Category

18.1. Significance of Sampling and Maximum Voting

Our analysis revealed that a simple sampling and maximum voting EM approach significantly reduced hallucinations, suggesting that more complex methods may not be necessary for specific applications. Additionally, this research demonstrates that EMs can effectively leverage both open-source and proprietary models, integrating both dLLMs and sLLMs. These results align with previous research that demonstrated that EM techniques can effectively mitigate individual model weaknesses in medical knowledge reasoning tasks [34, 35].

18.2. Domain-Specific Error Analysis

The evaluation of the models across various neuroscience domains reveals key insights into their capabilities and limitations. Mistral showed higher error rates, particularly in neuroanatomy and neurophysiology, suggesting deficiencies in spatial reasoning. In contrast, the EM excelled in neuroanatomy, demonstrating the benefits of combining multiple models. Specialized models like AtlasGPT performed well in neurophysiology and neuropathology, indicating strong training in these areas. While most models were accurate in neuroimaging, neurology presented challenges, highlighting the need for improvement in neurological syndrome recognition.

18.3. Cross-Model Performance Analysis

A clear performance gradient is observed from sLLMs (Mistral, Gemini) to specialized neurosurgical models (AtlasGPT, EM), with error rates decreasing accordingly. Domain-specific strengths were evident: EM excelled in anatomical and physiological knowledge (0 errors), 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.

19. Clinical and Educational Implications

Surgical Decision Support: The performance hierarchy observed across models carries significant implications for their potential clinical utility. With hallucination rates ranging from 11.33% (Mistral) to 2.67% (EM), these results highlight the critical importance of model selection in medical applications. In neurosurgical contexts, where incorrect information could lead to serious adverse outcomes, even small reductions in hallucination rates represent meaningful clinical improvements.

Based on our findings, we propose a risk stratification framework for AI model implementation:

19.1. Low Risk Applications: Education, study assistance, literature review

19.2. Moderate Risk Applications: Case preparation, differential diagnosis generation

19.3. High Risk Applications: Direct surgical planning, critical care decisions, neurological diagnosis

Furthermore, we suggest a model selection framework based on the specific knowledge domain being queried:

19.4. For Anatomical Questions: EM > AtlasGPT > Claude/Gemini > Mistral

19.5. For Neurological Questions: Consider human expert consultation given persistent errors across all models

19.6. For Pathological Questions: Claude, AtlasGPT, or EM all demonstrate high reliability Different levels of human verification are recommended based on error patterns:

Universal verification needed for neurological and neurosurgical recommendations

Reduced verification may be acceptable for imaging, pathology, and anatomy when using specialized models

20. Educational Applications

The domain-specific error analysis provides valuable insights for implementation strategies. Models demonstrated better reliability in certain domains (e.g., Critical Care, Core Competencies) while consistently struggling more with others (e.g., Neuropathology, Neuroanatomy). This suggests that even high-performing systems may require additional safeguards or human oversight when addressing particularly challenging knowledge domains. The EM method's ability to correct domain-specific errors made by individual models indicates its potential as a more comprehensive and reliable clinical decision support tool. By leveraging the complementary strengths of different models, EM approaches may provide more balanced and trustworthy support across the diverse knowledge domains required in neurosurgical practice.

Our findings suggest several educational applications

20.1. Targeted Learning: AI systems could help identify knowledge gaps in trainees by comparing their answers to those of expert systems. Models with different error profiles could be used to create varied training scenarios.

20.2. Error Analysis as Teaching Tool: The specific errors made by each system could be cataloged to create teaching cases highlighting common conceptual misunderstandings.

20.3. Complementary Learning: Students could benefit from exposure to multiple AI systems to overcome the blind spots of any single system.

21. Technical Recommendations

Based on our findings, we propose the following technical recommendations:

21.1. EM Approach Validation: The superior performance of the EM model (4 total errors vs. 5-17 in others) validates the multi-model approach. Future development should focus on intelligent model combination rather than single model optimization.

21.2. Domain-Specific Training: The persistent errors in neurology across all models suggest the need for specialized training datasets focused on neurological syndrome recognition. Critical care knowledge should be prioritized in model development given its time-sensitive implications.

21.3. Knowledge Integration: Future models should better integrate cross-domain knowledge, particularly between neurology and neurosurgery. The siloed nature of errors suggests models may struggle to make connections across traditional knowledge boundaries.

21.4. Surgical Context Awareness: Development of context-aware prompting that frames questions within surgical scenarios may improve performance. Integration of visual data (imaging, anatomical models) with text-based knowledge could enhance spatial reasoning.

22. Limitations

Despite promising results, several limitations must be acknowledged:

1. While our test set of 150 MCQs represents a substantial evaluation, it cannot encompass the full breadth and complexity of neurosurgical knowledge. Performance on this curated test set may not perfectly translate to real-world clinical scenarios with their inherent ambiguities and nuances. The use of MCQs frames neurosurgical knowledge as discrete options rather than reflecting the complexity of real-world clinical decision-making.
2. The computational resources required for EM methods present potential implementation challenges. Running multiple models simultaneously increases latency, computational costs, and system complexity compared to single-model approaches. These factors may influence the feasibility of deployment in resource-constrained healthcare settings or time-sensitive clinical scenarios.
3. The binary classification of responses as correct or incorrect, while methodologically necessary, does not fully capture the qualitative aspects of model outputs. Different types of hallucinations may carry varying clinical risks, and future work should consider more nuanced evaluation frameworks that account for the potential clinical impact of different error types.
4. The use of a simple sampling and maximum voting EM provides adequate hallucination mitigation performance in this study but may not be of sufficient complexity to address the range and depth of knowledge required in neurosurgical practice.
5. The exclusion of image-based questions reduces the generalizability of our findings to the multimodal nature of neurosurgical decision-making.

23. Future Research Directions

Future research should explore the following areas:

1. Investigation of more advanced EM methods (e.g., Augmenting, Stacking, and Bagging) utilizing innovative voting techniques. The potential of combining dLLMs and sLLMs presents a particularly promising avenue.
2. Evaluation of language models tailored to specific neurosurgical subspecialties across a broader array of question banks and clinical scenarios.
3. Integration of multimodal capabilities, involving the synthesis and analysis of data from various sources such as images, audio, and other sensory inputs.
4. Development of human-AI collaborative frameworks that strategically leverage the complementary strengths of both clinicians and AI systems.

24. Conclusion

This study evaluated the performance of various language models in the field of neurosurgery, highlighting the EM as achieving the highest accuracy at 97.33%, slightly outperforming AtlasGPT at 96.67% surpassing the individual sLLMs: Claude 3.5 Sonnet, Gemini, and Mistral, which had accuracy rates of 94.67%, 92.00%, and 88.67%, respectively. The success of EMs can be attributed to their integration of multiple sLLMs, which minimizes errors through diverse training data and enhances outcomes by leveraging varied data sources. This diversity principle, where different models make different errors, reduces overall mistakes and improves performance, leading to better generalization and robustness. The study suggests that EMs can effectively utilize both open-source and proprietary models, and a simple sampling and maximum voting approach proved effective, indicating that complex methods may not be necessary for specific applications. The superior performance of specialized models like AtlasGPT and EM demonstrates the value of domain-specific training and multi-model approaches. Despite these advancements, persistent errors in neurological diagnosis and surgical decision-making highlight areas needing further development before these systems can be safely deployed in critical clinical applications. The most promising immediate applications for these models are in educational contexts and as supplementary reference tools requiring expert verification.

References

1. 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.
2. Ragel, B. T., Asher, A. L., Selden, N., & MacDonald, J. D. (2006). Self-assessment in neurological surgery: the SANS wired white paper. *Neurosurgery, 59*(4), 759-766.
3. 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).
4. 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.
5. Ali, R., Tang, O. Y., Connolly, I. D., Sullivan, P. L. Z., Shin, J. H., Fridley, J. S., ... & Telfeian, A. E. (2023). Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery, 93*(6), 1353-1365.
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. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys, 55*(12), 1-38.
8. Dziri, N., Milton, S., Yu, M., Zaiane, O., & Reddy, S. (2022). On the origin of hallucinations in conversational models: Is it the datasets or the models?. *arXiv preprint arXiv:2204.0793*1.
9. Ahmad, M. A., Yaramis, I., & Roy, T. D. (2023). Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*.
10. Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C., ... & Mustafa, M. A. (2024). A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review, 57*(7), 175.
11. Dagi, T. F., Barker, F. G., & Glass, J. (2021). Machine learning and artificial intelligence in neurosurgery: status, prospects, and challenges. *Neurosurgery, 89*(2), 133-142.
12. 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.
13. Ali, R., Abdulrazeq, H., Patil, A., Connolly, I. D., Tang, O. Y., Doberstein, C. A., ... & Asaad, W. (2025). 2050 AtlasGPT: A Language Model Grounded in Neurosurgery. *Neurosurgery, 71*(Supplement_1), 256|
14. 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.
15. 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.
16. Xu, P., Ping, W., Wu, X., Xu, C., Liu, Z., Shoeybi, M., & Catanzaro, B. (2024). Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. *arXiv preprint arXiv:2407.14482*.
17. Purwar, A. (2024). Evaluating the efficacy of open-source llms in enterprise-specific rag systems: A comparative study of performance and scalability. *arXiv preprint arXiv:2406.11424*.
18. Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research, 25*(70), 1-53.
19. Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence, 115*, 105151.
20. Müller, D., Soto-Rey, I., & Kramer, F. (2022). An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. *Ieee Access, 10*, 66467-66480.
21. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.
22. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., ... & Wang, C. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.
23. Jiang, D., Ren, X., & Lin, B. Y. (2023). Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.
24. Forbes, G. C., Katlana, P., & Ortiz, Z. (2023). Metric ensembles

for hallucination detection. *arXiv preprint arXiv:2310.10495*.
25. 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.
26. Puffer, R. C. (2019). *Neurosurgery Primary Board Review.* Georg Thieme Verlag.
27. Mbaabu, N. M. (2025). Examining the status of AI use guidelines in editorial policies of Kenyan digital media houses and challenges in their formulation and implementation in newsrooms.
28. Jo, M. H., Kim, M. J., Oh, H. K., Choi, M. J., Shin, H. R., Lee, T. G., ... & Kang, S. B. (2024). Communicative competence of generative artificial intelligence in responding to patient queries about colorectal cancer surgery. *International Journal of Colorectal Disease, 39*(1), 94.
29. Buchanan, J., Hill, S., & Shapoval, O. (2024). Chatgpt hallucinates non-existent citations: Evidence from economics. *The American Economist, 69*(1), 80-87.
30. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine, 29*(8), 1930-1940.
31. Liévin, V., Hother, C. E., Motzfeldt, A. G., & Winther, O. (2024). Can large language models reason about medical questions?. *Patterns, 5*(3).
32. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., ... & Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, 1-8.
33. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172-180.
34. Wang L, Harwood A, Kotz D, Zhang Y. Ensemble learning for medical language models: Improving reliability in clinical decision support. J Biomed Inform. 2023;135:104231.
35. Yang, H., Li, M., Zhou, H., Xiao, Y., Fang, Q., & Zhang, R. (2023). One llm is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*.

APPENDIX A.

Questions (from Neurosurgery Self-Assessment Questions and Answers and from the Neurosurgery Primary Board Review)	Correct Answer	Gemini Model Hallucination Answer	Clinical and Technical Consequences of Model Hallucination
7. An 18-month-old girl presents with weaknesses. Tremors, primarily of the hands, had been noted since 4 months of age. She was crawling by 9 months of age and cruising about the furniture by 12 months. Her language development was normal. Her 4-year-old sister was developing normally. Cranial nerve examination was normal, and specifically, fasciculations of the tongue were not noted. She was able to sit, crawl, and pull to a stand. She could walk holding onto furniture but could not walk independently. Deep tendon flexes were absent throughout, and there were no Babinski signs. Sensory examination was normal. Which one of the following is the next appropriate test? a. Serum ceruloplasmin b. Electromyography c. Nerve conduction studies d. Survival motor neuron gene testing e. MRI head	d—Survival motor neuron gene testing Spinal muscular atrophy is a congenital lower motor neuron disorder manifesting as progressive, symmetric proximal muscular weakness occurring in 1 in 6000 to 1 in 10,000 births (second most common autosomal recessive disease in humans after cystic fibrosis). It is the leading inherited cause of infant death. Spinal muscular atrophy is classified clinically by the age at symptom onset and disease severity into type I (Werdnig-Hoffman disease, acute), type II (intermediate form, usually 7-18 months old and can sit unsupported but can't walk independently), type III (Kugelberg-Welander disease, mildest form, present 18 months and able to achieve independent walking), and type IV SMA (adult-onset). Spinal muscular atrophy is inherited in autosomal recessive fashion or is sporadic. Mutations or deletionsin the telomeric SMN (survival of motor neuron)gene occur in most patients. The loss of functional SMN protein results in premature neuronal cell death. The SMN protein has a role in cardiac development.	c. Nerve conduction studies Here's why: The patient exhibits leg weakness, tremors, and absent deep tendon reflexes, which point towards a neuromuscular disorder. Electromyography (EMG) is often performed in conjunction with nerve conduction studies (NCS) to assess the health of muscles and nerves. NCS specifically evaluates how electrical signals travel through nerves, helping identify nerve damage or dysfunction. Serum ceruloplasmin test is used for Wilson's disease, a genetic condition affecting copper metabolism, which is less likely based on the absence of fasciculations of the tongue and liver involvement. Although MRI head can be helpful in some neurological cases, it's not the most specific test for initial evaluation of this presentation. Survival motor neuron gene testing is more specific for amyotrophic lateral sclerosis (ALS) in adults, and the presentation is not typical for ALS in a child. Nerve conduction studies will provide valuable information about the speed and quality of nerve signals in the legs and potentially the hands, helping to pinpoint the location and nature of the nerve dysfunction causing the weakness and tremors. The results of NCS, along with EMG if performed, can guide further investigations and treatment decisions. It's important to note that I cannot provide medical diagnosis or treatment advice. This information is intended for general knowledge only, and you should consult with a qualified healthcare professional for any medical concerns.	1.Diagnostic delay: Performing nerve conduction studies first rather than going directly to SMN gene testing would delay the definitive diagnosis of Spinal Muscular Atrophy (SMA). This delay is critical as SMA is a progressive neurodegenerative disease where early intervention is essential. 2.Treatment opportunity loss: FDA-approved treatments for SMA (like nusinersen, onasemnogene abeparvovec, and risdiplam) are most effective when started early. Delaying genetic diagnosis would postpone potentially disease-modifying treatment. 3.Unnecessary testing burden: Nerve conduction studies are invasive and uncomfortable, particularly for an 18-month-old child, and would likely need to be performed under sedation. This creates unnecessary physical and emotional distress. 4.Incomplete diagnostic information: While nerve conduction studies might show abnormalities consistent with a lower motor neuron disease, they cannot confirm the specific genetic diagnosis of SMA, which is crucial for treatment selection and family genetic counseling. 5.Resource misallocation: Conducting the wrong test first uses healthcare resources inefficiently and increases overall costs of care. 6.Family planning implications: Delayed genetic diagnosis prevents timely genetic counseling for the parents regarding the 25% recurrence risk for future pregnancies. 7.Missed opportunity for sibling screening: Early diagnosis would prompt consideration of testing the 4-year-old sister who might be presymptomatic or have a milder phenotype requiring monitoring.

Questions (from Neurosurgery Self-Assessment Questions and Answers and from the Neurosurgery Primary Board Review)	Correct Answer	Claude Model Hallucination Answer	Clinical and Technical Consequences of Model Hallucination
19. The normal anatomical variant ponticulus posticus may increase the risk of vertebral artery injury during which one of the following? a. C1 lateral mass screw placement b. C2 pars screw placement c. C2 pedicle screw placement d. C2 translaminar screw placement e. C1-C2 transarticular screws	19. a—C1 lateral mass screw placement A bony bridge, the arcuate foramen (ponticulus posticus) may overlie the vertebral artery due to calcification of the oblique atlanto-occipital ligaments and may be mistaken for the C1 lateral mass. The entry point for C1 lateral mass screws is at the junction of theC1 lateral mass with the undersurface of the C1 posterior arch. The extensive venous plexus in this region makes dissection challenging and the C2 nerve root is in close proximity to the screw entry point and must be retracted distally. Screws are directed with 5-10° of convergence and parallel to the C1 arch. Alternatively, C1 pedicle screw placement has an entry point on the dorsal aspect of the posterior arch into the lateral mass. With either technique, excessive superior C1 screw angulation will violate the occiput-C1 joint. An excessively long C1 screw may potentially compromise the internal carotid artery orhypoglossal nerve.	Based on the analysis of the anatomy and the implications of ponticulus posticus, the correct answer is: e. C1-C2 transarticular screws I apologize for my previous incorrect response. Let me explain why this is the correct answer: Ponticulus posticus is a bony bridge over the vertebral artery groove on the posterior arch of the atlas (C1 vertebra). This variant creates an osseous foramen through which the vertebral artery passes, altering its normal course. C1-C2 transarticular screw fixation involves inserting screws from the inferior aspect of C2 through the C1-C2 facet joint and into the lateral mass of C1. The trajectory of these screws passes close to the course of the vertebral artery as it travels from C2 to C1. The presence of ponticulus posticus can alter the position of the vertebral artery, making it more vulnerable during this procedure. Surgeons need to be aware of this variant and adjust their technique accordingly to avoid vertebral artery injury. The other options (b, c, d) involve only C2 and are less likely to be affected by this C1 anatomical variant. While a (C1 lateral mass screw placement) does involve C1, the risk is not as high as with transarticular screws because the trajectory is different and less likely to encounter the altered course of the vertebral artery. C1-C2 transarticular screw fixation requires the most precise trajectory through the C1-C2 region where the ponticulus posticus is located, hence presenting the highest risk of vertebral artery injury in the presence of this anatomical variant.	1.Surgical risk: Selecting the wrong approach (C1-C2 transarticular screws instead of C1 lateral mass screws) when ponticulus posticus is present significantly increases the risk of vertebral artery injury, which could lead to stroke, posterior circulation ischemia, or fatal hemorrhage. 2.Procedural misdirection: The hallucination incorrectly suggests that C1 lateral mass screws have less risk than transarticular screws in patients with ponticulus posticus, which is the opposite of the truth. This could lead to the wrong surgical planning. 3.Anatomical misunderstanding: The hallucination demonstrates a fundamental misunderstanding of how ponticulus posticus specifically affects C1 lateral mass screw insertion points, which is precisely where this anatomical variant creates risk. 4.Unnecessary patient risk: Following this incorrect advice could result in catastrophic neurovascular complications that are entirely preventable with proper anatomical knowledge. 5.Training impact: This misinformation could propagate incorrect surgical techniques if used in educational contexts, perpetuating dangerous practices.

Questions (from Neurosurgery Self-Assessment Questions and Answers and from the Neurosurgery Primary Board Review)	Correct Answer	Claude Model Hallucination Answer	Clinical and Technical Consequences of Model Hallucination
299. You have been consulted by the orthopedic team on a 78-year-old woman who just had a hip replacement and was started on unfractionated heparin for a DVT/PE in the postoperative period. She received 1,400 units of unfractionated heparin and became sleepy. Imaging demonstrates a right frontal intraparenchymal hematoma. What dose of protamine should be given for reversal? A. 1.4 mg B. 14 mg C. 140 mg D. 1,400 mg E. 1,400 µg	299. B 14 mg This patient needs to have her heparin reversed with protamine. The dose for protamine reversal is 1 mg of protamine for 100 U of heparin given over the last 3 hours. Further Reading: Hamilton, Golfinos, Pineo, Couldwell. Handbook of Bleeding and Coagulation for Neurosurgery, 2015, pages 52–53.	C. 140 mg Protamine sulfate is used to reverse the anticoagulant effects of unfractionated heparin. The dosing of protamine is based on the amount of heparin administered and the time elapsed since its administration. The general rule of thumb is to give 1 mg of protamine for every 100 units of heparin if the heparin was given within the previous 2-3 hours. In this case, the patient received 1,400 units of unfractionated heparin. Therefore, the appropriate dose of protamine for reversal would be 14 mg (1,400 units heparin / 100 units per mg protamine = 14 mg protamine). However, the actual dose may need to be adjusted based on the timing of heparin administration and the patient's coagulation status. It is important to administer protamine slowly to avoid adverse reactions, and the patient should be closely monitored during and after administration.	1.Medication error: Administering 140 mg of protamine instead of the correct 14 mg represents a 10-fold overdose, which could lead to serious adverse effects including hypotension, bradycardia, anaphylaxis, or pulmonary hypertension. 2.Anticoagulation rebound: Excessive protamine can paradoxically cause anticoagulation and worsen the very bleeding it was intended to stop, potentially expanding the intracranial hemorrhage. 3.Cardiovascular compromise: Such a significant protamine overdose could cause catastrophic cardiovascular collapse in an elderly post-surgical patient, already vulnerable due to the intracranial hemorrhage. 4.Delayed neurosurgical intervention: Managing complications from protamine overdose would delay definitive management of the intracranial hemorrhage, worsening neurological outcomes. 5.Resource utilization: Treating complications from medication overdose would require ICU-level care, increased monitoring, additional medications, and possibly additional procedures, increasing healthcare costs and resource utilization.