

Original Paper

Performance and Hallucination Analysis of Large Language Models on European Anesthesiology Examinations: Cross-Sectional Comparative Study

Stefan Andrei^{1,2}, MD, PhD; Thibault Giet¹, BS; Alexis Belouard¹, BS; Mihai Stefan^{3,4}, MD; Mihai Popescu^{3,5}, MD, PhD; Sébastien Tanaka^{1,6}, MD, PhD; Philippe Montravers^{1,7}, MD, PhD; Aurélie Gouel^{1,8}, MD, PhD

¹Department of Anesthesiology and Intensive Care, CHU Bichat Claude Bernard, Assistance Publique, Hôpitaux de Paris, Paris, Île-de-France, France

²Group of Data Modeling, Computational Biology, and Predictive Medicine, Applied Mathematics, Institut de Biologie de l'École Normale Supérieure, Paris, Île-de-France, France

³Discipline of Anaesthesiology and Intensive Care, Carol Davila University of Medicine and Pharmacy, Bucharest, București, Romania

⁴2nd of Department of Anaesthesiology and Intensive Care, Institutul de Urgență pentru Boli Cardiovasculare "Prof.Dr. C.C. Iliescu", Bucharest, București, Romania

⁵Department of Anaesthesiology and Intensive Care, Bucharest Emergency University Hospital, Bucharest, București, Romania

⁶Diabète athérothrombose Thérapies Réunion Océan Indien (DéTROi), INSERM UMR 1188, University of Reunion Island, Saint-Pierre de La Réunion, Réunion, Réunion

⁷Physiopathology and Epidemiology of Respiratory Diseases, INSERM UMR1152, Université Paris Cité, Paris, Île-de-France, France

⁸Antibodies in Therapy and Pathology, Institut Pasteur, INSERM UMR1222, Université Paris Cité, Paris, Île-de-France, France

Corresponding Author:

Mihai Stefan, MD

Discipline of Anaesthesiology and Intensive Care, Carol Davila University of Medicine and Pharmacy

Sos Fundeni, 258

Bucharest, București 022322

Romania

Phone: 40 742185533

Email: mihai.stefan@umfcd.ro

Abstract

Background: Large language models (LLMs) have shown promising performance on medical examinations across specialties. However, comparative evaluations of current-generation LLMs across multiple European anesthesiology examinations, alongside structured assessment of hallucinations vs question-related confusion, remain lacking.

Objective: This study aimed to compare the performance of 4 state-of-the-art LLMs on anesthesiology and intensive medicine examination questions and assess their hallucination rates.

Methods: This computational comparative study analyzed 437 multiple-choice questions (1748 queries) from 3 sources: nurse anesthetist school examinations (infirmier anesthésiste diplômé d'État [registered nurse anesthetist]; n=100, 22.9%), European Diploma in Anaesthesiology and Intensive Care (EDAIC; n=219, 50.1%), and EDAIC On-Line Assessment (n=118, 27.0%). Each question was submitted to 4 LLMs (Claude Sonnet 4.5, Gemini 2.5 Pro, GPT-5, and Grok 4) using standardized prompts via default web interface settings. Responses were evaluated through structured consensus review by 2 examiners for accuracy, hallucinations, and question-related confusion. Statistical analysis included Friedman and Wilcoxon signed-rank tests with Holm-Bonferroni correction, the Cochran Q test, and generalized estimating equations.

Results: Average success rates ranged from 86% (SD 18%) to 94% (SD 10%) across LLMs and examination types, exceeding the EDAIC part I passing threshold, representing substantial improvement over previously reported GPT-3.5 performance. For the EDAIC, overall intermodel differences were significant (Friedman $\chi^2_3=13.9$; $P=.003$; $W=0.02$), with Gemini outperforming GPT-5 as the only pairwise difference. Hallucination rates ranged from 11% (11/100) to 20.1% (44/219) without significant intermodel differences. All models exceeded the EDAIC passing threshold.

Conclusions: Current-generation LLMs demonstrated consistently high performance across multiple European anesthesiology examinations but continue to produce clinically relevant hallucinations, supporting their role as supervised educational tools

rather than autonomous learning resources. These findings underscore the need for structured integration frameworks and systematic verification when deploying LLMs as learning tools in medical education.

JMIR Form Res 2026;10:e95859; doi: [10.2196/95859](https://doi.org/10.2196/95859)

Keywords: artificial intelligence; AI; large language models; medical education; anesthesiology; intensive care medicine; hallucinations; examination; multiple-choice questions

Introduction

Large language models (LLMs) have gained increasing attention in health care and medical education due to their demonstrated capacity to generate clinically relevant text, answer medical questions, and support reasoning tasks. Benchmark analyses using different standardized assessments have shown that LLMs can achieve substantial accuracy, sometimes approaching or surpassing human performance [1-4]. However, empirical evidence also reveals wide variability across tasks and domains, with persistent limitations in multimodal reasoning, complex clinical scenarios, and domain-dependent weaknesses, raising concerns about the consistency, precision, and clinical applicability of these systems [5].

LLMs may support the preparation of future health care professionals for clinical environments where AI-assisted reasoning will become increasingly prevalent. These competencies may become more relevant for modern medical practice [6]. While no universally accepted definition exists, the French National Commission on Informatics and Liberty describes LLMs as statistical models of linguistic unit distribution that predict subsequent words in a sequence [7]. This text completion mechanism, trained on vast datasets including Wikipedia, internet sources, and books, enables coherent text generation but also introduces the potential for error propagation from training data [7,8].

Despite enthusiasm regarding their potential, emerging evidence indicates that LLMs exhibit systematic vulnerabilities, including hallucinations, overconfidence, and errors in complex reasoning tasks. Operating on probabilistic text completion rather than genuine reasoning or comprehension [9], LLMs generate plausible-appearing text with perfect syntax and grammar that may diverge from contextual accuracy, user instructions, or factual knowledge [7]. These errors, collectively termed “hallucinations,” are defined by Huang et al [10] as phenomena where generated content appears nonsensical or unfaithful to the source material. Furthermore, analyses of LLM performance across cognitive domains suggest that, while these models perform adequately in factual recall and substantial clinical knowledge, their accuracy declines in tasks requiring higher-order reasoning and clinical judgment [11,12].

In a discipline requiring high-level expertise and practical knowledge, evaluating whether LLMs meet sufficient standards for pedagogical use is imperative. This study aimed to assess LLM performance and their assessment potential in anesthesiology education using standard examination questions. To our knowledge, no study has simultaneously compared multiple current-generation LLMs across different

European anesthesiology examinations while systematically distinguishing hallucinations from question-related confusion.

Methods

Study Design

This computational comparative study evaluated 4 LLMs: Claude Sonnet 4.5 (Anthropic), Gemini 2.5 Pro (Google), GPT-5 (OpenAI), and Grok 4 (xAI), accessed via their respective publicly available web interfaces in October 2025. Models were accessed using default parameters (temperature and other generation settings were not modified). Exact internal model checkpoints were not disclosed by model providers.

Objectives and End Points

The primary objective was to comparatively evaluate LLM performance on anesthesiology examination questions. The primary end point was the percentage of correct responses across the 4 LLMs. Secondary objectives included assessment of hallucination rates and evaluation of question clarity through LLM-identified confusion. Secondary end points comprised the proportion of serious errors (hallucinations) and the proportion of questions flagged as “confusing” by the LLMs.

Questions and Selection

Questions were selected using a multidimensional approach drawing from (1) nurse anesthetist school examinations from Assistance Publique - Hôpitaux de Paris, designated as infirmier anesthésiste diplômé d'état (IADE; in English state-registered nurse anesthetist) questions; (2) European Diploma in Anaesthesiology and Intensive Care (EDAIC) examinations; and (3) EDAIC On-Line Assessment (OLA) practice examinations. All questions were available in online public repositories from prior examinations. None of the study authors were involved in the creation or selection of IADE examination questions.

The EDAIC, offered by the European Society of Anaesthesiology and Intensive Care, comprises written and oral components. The written examination includes 60 basic science multiple-choice questions (MCQs) followed by 60 clinical MCQs covering internal and emergency medicine; general and regional anesthesia; and specialized anesthesia, including pain management, resuscitation, and intensive care [13]. The OLA follows identical formatting for examination preparation. All questions were exclusively multiple choice to facilitate objective evaluation as MCQ responses are less subject to examiner subjectivity [14]. The IADE examination is the qualifying examination for French state-registered

nurse anesthetists. The selected questions consisted of MCQs covering 5 domains (physiology, pharmacology, anesthetic techniques and implementation, procedure-related considerations, and patient-specific factors) and were administered in French, whereas EDAIC and OLA questions were administered in English. Questions were analyzed in their original language without translation.

Study Protocol

Questions were submitted identically to the 4 LLMs following a standardized protocol. All models were queried using default web interface settings. Each question comprised a stem with 4 to 6 possible responses, with correct answers ranging from none to all options. A new conversation thread was initiated for each query to prevent cross-contamination of responses. Each individual answer option was scored as correct or incorrect, yielding a percentage success score for each question (ie, the proportion of correctly classified options out of all available options per question). This item-level scoring approach was chosen to capture partial knowledge and provide a more granular assessment of model performance than binary whole-question scoring. This scoring approach mirrors the official EDAIC examination marking scheme, in which each individual option is independently scored as correct or incorrect and partial credit is awarded accordingly.

The standardized prompt template was as follows: “[Question stem verbatim] Options: A) [option A] B) [option B] C) [option C] D) [option D] [E) and F) if applicable]. Please identify all correct answers and explain your reasoning.” This format was maintained consistently across all queries to ensure comparability.

All responses were evaluated through a structured consensus process by 2 examiners (TG and AB), who jointly reviewed each response, discussed discrepancies, and reached agreement on classification. When an LLM achieved less than 100% accuracy, joint review was conducted to distinguish between question-related confusion and true hallucination [7, 10].

A hallucination was operationally defined as a response containing factually incorrect, fabricated, or unsupported medical information despite apparent comprehension of the question stem. In contrast, question-related confusion was defined as an incorrect response associated with misinterpretation of ambiguous or complex wording despite otherwise medically coherent reasoning. Representative examples illustrating these classification criteria are provided in [Multimedia Appendix 1](#).

Hallucination and confusion analysis was performed for IADE and EDAIC questions only. The OLA was included as an external validation set for the primary end point (comparative success rates). Structured dual-reviewer classification of hallucinations and confusion for the OLA set would have required evaluation of 472 additional responses (118 × 4); the IADE and EDAIC provided 319 questions (1276 evaluated responses) across 2 distinct cognitive levels, which was considered sufficient to characterize error patterns. Given

the fact that the OLA is not conceptually different from the EDAIC (it is an EDAIC simulation), it was retained as an external validation dataset for performance assessment rather than for detailed error pattern characterization.

Statistical Analysis

Statistical analyses were performed using SPSS (version 29.0; IBM Corp) and Python (version 3.11; Python Software Foundation) with the *statsmodels* package (version 0.14), with the significance threshold fixed at a *P* value below .05 for all analyses. Quantitative variables are presented as means and SDs with 95% CIs where appropriate; qualitative variables are presented as counts and percentages. Normal distribution was assessed visually using histograms and Shapiro-Wilk testing. Results were analyzed by examination group (IADE, EDAIC, and OLA). Paired tests included Friedman and Wilcoxon signed-rank tests for quantitative variables and the Cochran *Q* test for qualitative variables. Pairwise post hoc comparisons following significant Friedman tests were performed using Wilcoxon signed-rank tests with exclusion of zero differences, and effect sizes were reported as $r = Z/\sqrt{n}$. To control the familywise error rate across the 6 pairwise comparisons per examination group while avoiding the excessive conservativeness of the Bonferroni method with correlated comparisons, the Holm-Bonferroni sequential procedure was applied. Holm-adjusted *P* values (denoted as “adjusted *P*”) are reported for all pairwise post hoc and domain-level comparisons and were compared against the fixed .05 significance threshold; unadjusted *P* values are reported only for omnibus tests, to which no multiplicity correction applies. The Kendall *W* was reported as the effect size for the Friedman and Cochran *Q* tests. To account for the clustered structure of the binary outcomes (each question was answered by the 4 models), pairwise intermodel comparisons of hallucination and confusion rates were performed using generalized estimating equations (GEE) with a logit link, a binomial distribution, an exchangeable working correlation structure, and robust SEs, with questions specified as clusters. All 6 pairwise model contrasts were estimated for each outcome, and the resulting *P* values were Holm adjusted.

Ethical Considerations

This computational study did not involve human participants, patient data, biological material, or identifiable personal information. This study exclusively analyzed publicly available examination questions and AI-generated responses. According to French regulations governing biomedical research (Loi n° 2012-300 du 5 mars 2012, relative aux recherches impliquant la personne humaine, “Loi Jardé,” modified by Ordonnance n° 2016-800 du 16 juin 2016 [15]), research not involving human participants, identifiable personal data, or biological samples falls outside the scope of mandatory ethics committee review. The Commission Nationale de l’Informatique et des Libertés (CNIL) similarly does not require declaration for research that does not process personal data. Ethics committee approval was therefore not required for the present study. No patient data were collected, stored, or analyzed at any stage of the study.

Results

2 separate examinations, and 118 (27.0%) OLA questions. Complete results are shown in [Tables 1-4](#) and [Figure 1](#).

Overview

A total of 437 questions (1748 queries) were analyzed: 100 (22.9%) IADE questions, 219 (50.1%) EDAIC questions from

Table 1. Success rates and hallucination and confusion rates by examination type.^a

Examination and variables	Claude Sonnet 4.5	Gemini 2.5 Pro	GPT-5	Grok 4	P value
IADE ^b (n=100 questions)					
Success rate (%), mean (SD)	93 (12)	91 (13)	91 (13)	92 (14)	.16
Hallucinations, n (%)	11 (11)	16 (16)	13 (13)	16 (16)	.34
Confusion, n (%)	8 (8)	13 (13)	14 (14)	8 (8)	.049
EDAIC ^c (n=219 questions)					
Success rate (%), mean (SD)	89 (16)	89 (17)	86 (18)	87 (17)	.003
Hallucinations, n (%)	36 (16.4)	31 (14.2)	40 (18.3)	44 (20.1)	.10
Confusion, n (%)	51 (23.3)	53 (24.2)	67 (30.6)	59 (26.9)	.10
OLA ^d (n=118 questions)					
Success rate (%), mean (SD)	94 (10)	93 (12)	90 (14)	90 (14)	.008

^aSuccess rates were compared using the Friedman test, and hallucination and confusion rates were compared using the Cochran *Q* test; test statistics are reported in the Results section.

^bIADE: Infirmier Anesthésiste Diplômé d'État.

^cEDAIC: European Diploma in Anaesthesiology and Intensive Care.

^dOLA: On-Line Assessment.

Table 2. Infirmier Anesthésiste Diplômé d'État examination success rates by subject domain. *P* values are Holm adjusted across the 5 domain-level Friedman tests.

Subjects	Claude Sonnet 4.5, mean (SD)	Gemini 2.5 Pro, mean (SD)	GPT-5, mean (SD)	Grok 4, mean (SD)	Adjusted <i>P</i> value
Procedure-related considerations (%; n=21)	87 (17)	90 (16)	89 (15)	86 (19)	>.99
Pharmacology (%; n=20)	92 (13)	92 (14)	90 (15)	89 (16)	>.99
Physiology (%; n=21)	99 (4)	93 (12)	93 (11)	95 (8)	.06
Techniques (%; n=14)	93 (9)	86 (12)	84 (11)	93 (9)	.16
Patient-specific factors (%; n=24)	93 (11)	93 (12)	94 (10)	95 (8)	>.99

Table 3. Pairwise comparisons of hallucination and confusion rates between large language models for Infirmier Anesthésiste Diplômé d'État questions: generalized estimating equation analysis.^a

Comparisons	OR ^b (95% CI)	Adjusted <i>P</i> value
Hallucinations		
Gemini 2.5 Pro vs Claude Sonnet 4.5	1.54 (0.84-2.84)	.82
GPT-5 vs Claude Sonnet 4.5	1.21 (0.63-2.30)	>.99
Grok 4 vs Claude Sonnet 4.5	1.54 (0.88-2.70)	.78
GPT-5 vs Gemini 2.5 Pro	0.78 (0.49-1.26)	>.99
Grok 4 vs Gemini 2.5 Pro	1.00 (0.63-1.59)	>.99
Grok 4 vs GPT-5	1.27 (0.79-2.05)	>.99
Confusion		
Gemini 2.5 Pro vs Claude Sonnet 4.5	1.72 (0.91-3.24)	.34
GPT-5 vs Claude Sonnet 4.5	1.87 (0.98-3.57)	.34
Grok 4 vs Claude Sonnet 4.5	1.00 (0.59-1.70)	>.99
GPT-5 vs Gemini 2.5 Pro	1.09 (0.70-1.70)	>.99
Grok 4 vs Gemini 2.5 Pro	0.58 (0.33-1.02)	.34
Grok 4 vs GPT-5	0.53 (0.28-1.02)	.34

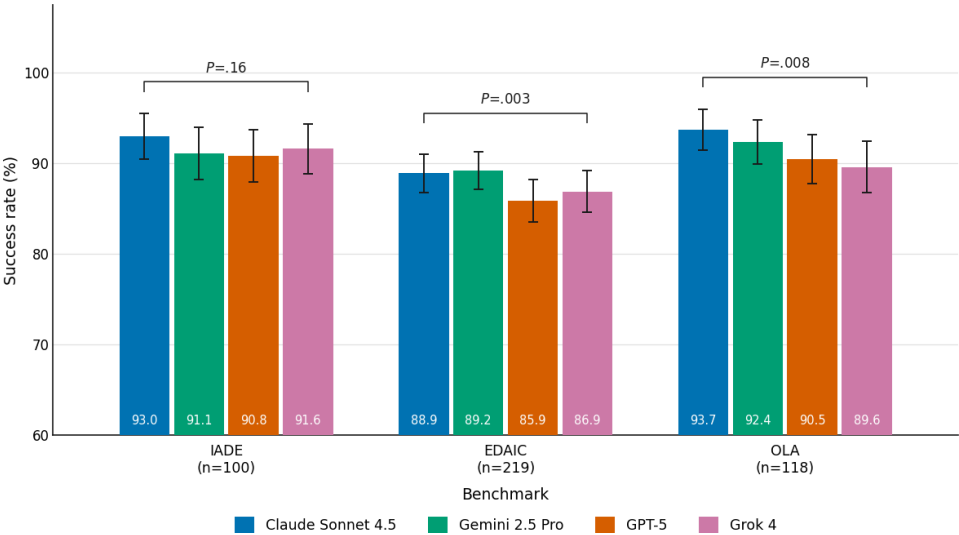
^aOdds ratios were estimated using generalized estimating equations with a logit link, an exchangeable working correlation structure, and robust SEs, with questions specified as clusters; *P* values are Holm adjusted across the 6 pairwise comparisons within each outcome.
^bOR: odds ratio.

Table 4. Pairwise comparisons of hallucination and confusion rates between large language models for European Diploma in Anaesthesiology and Intensive Care questions: generalized estimating equation analysis.^a

Comparisons	OR ^b (95% CI)	Adjusted <i>P</i> value
Hallucinations		
Gemini 2.5 Pro vs Claude Sonnet 4.5	0.84 (0.59-1.20)	>.99
GPT-5 vs Claude Sonnet 4.5	1.14 (0.79-1.63)	>.99
Grok 4 vs Claude Sonnet 4.5	1.28 (0.91-1.79)	.63
GPT-5 vs Gemini 2.5 Pro	1.36 (0.99-1.86)	.30
Grok 4 vs Gemini 2.5 Pro	1.52 (1.08-2.14)	.09
Grok 4 vs GPT-5	1.13 (0.80-1.58)	>.99
Confusion		
Gemini 2.5 Pro vs Claude Sonnet 4.5	1.05 (0.76-1.46)	>.99
GPT-5 vs Claude Sonnet 4.5	1.45 (1.03-2.05)	.21
Grok 4 vs Claude Sonnet 4.5	1.21 (0.87-1.70)	>.99
GPT-5 vs Gemini 2.5 Pro	1.38 (1.01-1.88)	.21
Grok 4 vs Gemini 2.5 Pro	1.15 (0.86-1.55)	>.99
Grok 4 vs GPT-5	0.84 (0.60-1.16)	>.99

^aOdds ratios were estimated using generalized estimating equations with a logit link, an exchangeable working correlation structure, and robust SEs, with questions specified as clusters; *P* values are Holm adjusted across the 6 pairwise comparisons within each outcome.
^bOR: odds ratio.

Figure 1. Success rates of 4 large language models across 3 anesthesiology examination types (Infirmier Anesthésiste Diplômé d’État [IADE]: n=100 questions; European Diploma in Anaesthesiology and Intensive Care [EDAIC]: n=219 questions; On-Line Assessment [OLA]: n=118 questions). Bars represent mean success rates; error bars denote 95% CIs of the mean. Brackets above each examination group indicate *P* values from the Friedman test for overall intermodel comparison.



IADE Examination Performance

IADE performance results are summarized in Tables 1 and 2. Overall performance was high across all models, with no statistically significant intermodel differences observed (Friedman $\chi^2_3=5.2$; *P*=.16; *W*=0.02).

Questions were categorized into 5 domains: physiology (21/100, 21%), pharmacology (20/100, 20%), anesthetic techniques and implementation (14/100, 14%), procedure-related considerations (21/100, 21%), and patient-specific factors (24/100, 24%). After Holm adjustment across the

5 domain-level Friedman tests, no significant intermodel differences were observed (physiology: $\chi^2_3=11.1$, adjusted *P*=.06, and *W*=0.18; anesthetic techniques: $\chi^2_3=8.3$, adjusted *P*=.16, and *W*=0.20; pharmacology, procedure-related considerations, and patient-specific factors: adjusted *P*>.99). Given the small per-domain sample sizes, these analyses should be interpreted as exploratory.

EDAIC Examination Performance

EDAIC performance results are summarized in Table 1. Overall performance remained high across all models, with

significant overall intermodel differences observed (Friedman $\chi^2_3=13.9$; $P=.003$; $W=0.02$). Claude Sonnet 4.5 and Gemini 2.5 Pro achieved the highest overall performances, whereas GPT-5 and Grok 4 showed lower mean scores.

In pairwise Wilcoxon signed-rank comparisons, Gemini 2.5 Pro significantly outperformed GPT-5 ($Z=2.75$; $r=0.32$; adjusted $P=.04$). No other pairwise comparisons were significant (adjusted $P\geq.09$ in all cases).

OLA Examination Performance

OLA performance results are summarized in Table 1. Overall performance remained high across all models, with statistically significant overall intermodel differences observed (Friedman $\chi^2_3=12.0$; $P=.008$; $W=0.03$). Claude Sonnet 4.5 achieved the highest performance, whereas GPT-5 and Grok 4 showed lower mean scores. In pairwise Wilcoxon signed-rank comparisons, Claude Sonnet 4.5 significantly outperformed Grok 4 ($Z=3.18$; $r=0.53$; adjusted $P=.009$). No other pairwise comparisons were significant (adjusted $P\geq.08$ in all cases).

Hallucination and Confusion Analysis

Hallucination and confusion rates for each model are shown in Table 1. For IADE questions, hallucination rates did not differ significantly across models (Cochran $Q_3=3.38$; $P=.34$; $W=0.01$). The omnibus test for confusion rates reached nominal significance (Cochran $Q_3=7.85$; $P=.049$; $W=0.03$); however, the effect size was negligible, and no pairwise contrast was significant after Holm adjustment in the GEE analysis (Table 3; adjusted $P\geq.34$ in all cases).

For EDAIC questions, neither hallucination rates (Cochran $Q_3=6.22$; $P=.10$; $W=0.01$) nor confusion rates (Cochran $Q_3=6.28$; $P=.10$; $W=0.01$) differed significantly across models. In the pairwise GEE analyses (Tables 3 and 4), no contrast was significant after Holm adjustment for either outcome; the largest contrast was observed for EDAIC hallucinations between Grok 4 and Gemini 2.5 Pro (odds ratio 1.52, 95% CI 1.08–2.14; adjusted $P=.09$).

Discussion

Principal Findings

In this study, we compared the performance of 4 current-generation LLMs across multiple European anesthesiology examination datasets and evaluated hallucination and question-related confusion rates. All evaluated models demonstrated high overall performance, with success rates approaching or exceeding typical human passing thresholds for the EDAIC examination. Although overall intermodel differences were detected, a single pairwise difference per examination remained significant after correction, indicating that no model was consistently superior. Importantly, hallucination rates remained nonnegligible across all models, highlighting an important limitation for unsupervised educational use. Notably, LLM performance surpassed typical first-attempt pass scores for the EDAIC, which range between

60% and 75% (2025 passing scores: 67% for part A and 73.5% for part B), indicating that LLMs may outperform average examinees on standardized theoretical examinations [1,16].

The higher performance observed on IADE compared with EDAIC questions may reflect differences in cognitive complexity, as questions requiring lower-order cognitive skills (recall and comprehension) may be more amenable to current LLM capabilities than those requiring higher-order reasoning (analysis, synthesis, and evaluation) [5, 17]. However, because IADE questions were administered exclusively in French whereas EDAIC and OLA questions were administered exclusively in English, the effects of examination type and language cannot be disentangled. Consequently, the observed performance differences may reflect differences in cognitive complexity, language, or examination format or a combination of these factors.

After Holm-Bonferroni correction, a single robust pairwise difference remained per examination: Gemini outperformed GPT-5 on the EDAIC, and Claude Sonnet 4.5 outperformed Grok 4 on the OLA. However, absolute performance differences between models were modest (2–4 percentage points), suggesting that, while statistically detectable, these disparities may have limited practical significance. The clinically relevant finding is that all models performed comparably well, and hallucination rates represent a more meaningful differentiator for educational deployment.

Comparison With Prior Work

These findings represent substantial improvement from previous reports. We evaluated GPT-3.5 in 2023 for the EDAIC part 1, finding a 70.5% success rate [1]. Other authors have documented similar performance with earlier model versions [18,19]. Our results demonstrate the rapid evolution of these technologies, consistent with other work showing progressive improvement across ChatGPT versions on medical licensing examinations worldwide [20]. In the United States, a recent exploratory study showed that ChatGPT can perform similarly to humans in standardized oral examinations as graded by American Board of Anesthesiology examiners [21], and other data show that current LLMs can surpass the minimum competency required for American Board of Anesthesiology certification [22]. Remarkable data have also emerged from various medical fields in recent years suggesting that LLMs can achieve high levels of performance in knowledge-based tasks while exhibiting important limitations in complex clinical reasoning and reliability [23, 24].

Regarding hallucinations, our data do not support declaring any LLM superior. Confusion rates showed no significant intermodel differences except for a marginal effect on the IADE set ($P=.049$; $W=0.03$), which was of negligible magnitude and not corroborated by the pairwise GEE analysis. Training data may naturally contain errors that LLMs perpetuate, and users lack direct control over consulted sources, making verification challenging. Research on GPT-4o using Chilean national examination questions demonstrated that, despite strong factual recall and

comprehension, the model exhibited limitations in complex reasoning and diagnostic judgment, producing unfounded medical assertions and imprecise conclusions [11]. Importantly, the hallucination rates of 11% (11/100) to 20.1% (44/219) in our study carry significant practical implications for medical education. A student relying on an LLM for self-directed learning may encounter erroneous information in approximately 1 out of every 5 to 10 queries, potentially reinforcing incorrect medical knowledge if not critically evaluated. Unlike factual errors that may be easily identified, LLM hallucinations are typically presented with high confidence and syntactic fluency, making them particularly insidious in an educational context. This underscores the necessity for supervised use and systematic cross-referencing with authoritative sources when LLMs are used as learning aids.

Implications for Education and Practice

From an educational perspective, these findings suggest that LLMs should not be used as primary knowledge sources but rather as guided learning assistants. A pragmatic framework for safe integration may include (1) mandatory cross-verification with authoritative resources, (2) supervised use in structured teaching environments, and (3) explicit training of students to recognize and critically appraise AI-generated content. Without such safeguards, the observed hallucination rates could lead to reinforcement of incorrect medical knowledge during self-directed learning.

Our standardized prompt approach, while ensuring reproducibility, may not reflect optimal performance achievable through advanced prompting techniques. Chain-of-thought prompting, few-shot learning, or retrieval-augmented generation could improve accuracy and reduce hallucinations. Future studies should systematically evaluate the impact of prompt engineering on medical examination performance.

The high acceptance of AI tools among medical students further supports the relevance of these findings [25]. Surveys indicate widespread enthusiasm for AI integration into medical curricula, with students anticipating improvements in learning conditions and educational efficiency [26]. LLMs enable personalized educational approaches through tutorial dialogue simulation, complex concept summarization, and enhanced knowledge retention, positioning them as valuable complementary tools in medical education [20].

Beyond student support, these tools may reduce educator administrative burden by automating repetitive tasks, allowing more time for personalized student guidance and improving pedagogical quality [27,28]. The integration of LLMs into medical education raises important regulatory and policy considerations. While tools for self-directed learning may require less oversight than clinical decision support systems, educational institutions should develop clear policies regarding appropriate use, required supervision, and assessment of AI-generated content accuracy. Professional societies and accreditation bodies may need to establish guidelines for LLM use in medical education curricula.

Future studies examining open-ended questions or clinical cases would be valuable for testing LLM clinical reasoning capabilities. Given the exponential evolution of these technologies, our results will likely become obsolete within months, warranting repeated evaluation of subsequent model versions. Additionally, intervention studies evaluating the efficacy of LLM-assisted learning compared to traditional methods would help establish evidence-based recommendations for educational integration.

Limitations

Several limitations warrant consideration. The consensus-based evaluation approach, while ensuring classification consistency between examiners, did not involve independent parallel scoring and, therefore, did not allow for formal quantification of interrater agreement through metrics such as the Cohen κ . Future studies should incorporate independent blinded evaluations to enable formal reliability assessment. Question difficulty was not formally graded. The evaluation relied on MCQs, which primarily assess knowledge recall and pattern recognition rather than complex clinical reasoning.

Questions originated from official institutions evaluating future professionals (EDAIC and IADE), and dual review reduced evaluation bias. While our substantial question sample merits expansion for enhanced statistical reliability, the exclusive use of MCQs (rather than open-ended questions) attenuated evaluation bias. Prompt formulation was standardized, with each LLM receiving identical questions in new conversation threads to prevent cross-contamination, although prior model exposure to these question banks cannot be excluded.

An important limitation is the potential presence of some examination questions or closely related content within LLM training datasets ("training data contamination"). Consequently, high performance may partially reflect memorization or prior exposure rather than genuine reasoning ability. However, several factors mitigate this concern: IADE questions were administered in French from an institutional repository with limited online dissemination, EDAIC questions are regularly updated by the examination committee, and the use of 3 distinct sources with different formats and languages reduces the likelihood of systematic contamination across all question sets. This concern has been increasingly acknowledged in the LLM benchmarking literature: prior work has demonstrated that training data contamination can artificially inflate performance on widely used medical examination benchmarks such as MedQA and US Medical Licensing Examination-style datasets, and dedicated detection methodologies have been proposed to estimate the extent of such contamination [29,30].

Additional methodological limitations include the fact that (1) questions were administered in their original language (French for the IADE and English for the EDAIC and OLA) and potential translation or language processing effects were not explored; (2) exact model version checkpoints were not documented beyond the October 2025 time frame, limiting precise reproducibility; (3) no test-retest reliability

assessment was performed to evaluate response consistency—this single-query design was chosen to reflect typical real-world use patterns, where students or clinicians query an LLM once per question rather than repeatedly. However, future studies should assess response reproducibility across multiple runs.

Because the IADE questions were administered exclusively in French and the EDAIC and OLA questions were administered exclusively in English, language and examination type are inherently confounded in this design; consequently, the observed performance differences cannot be attributed to examination type alone and should be interpreted with this limitation in mind. Finally, there is a temporal limitation given the rapid LLM evolution as these results reflect October 2025 model versions.

Conclusions

Current-generation LLMs demonstrated high performance on anesthesiology examination questions. All models

substantially exceeded human passing thresholds, consistent with previously reported progressive improvements in successive LLM iterations. However, hallucination rates remained persistent across the 4 models, emphasizing the importance of user awareness regarding this phenomenon. AI-generated outputs continue to require critical verification and supervised use in educational settings, but these models represent tools with substantial pedagogical and assessment potential. These findings highlight both the educational potential and the limitations of LLMs as supplementary learning tools for medical trainees preparing for standardized examinations. Educational deployment frameworks must therefore prioritize systematic verification mechanisms over reliance on model accuracy alone.

Acknowledgments

This manuscript was prepared with language editing assistance from ChatGPT (OpenAI) and Claude (Anthropic). These AI tools had no role in the study design, data acquisition, or interpretation of results. The final content was verified and approved by the authors.

Funding

The authors declared no financial support was received for this work.

Data Availability

The question corpus cannot be disclosed due to institutional confidentiality requirements. Statistical outputs are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SA, MS, AG

Data curation: SA, TG, AB

Formal analysis: SA

Methodology: SA, MS, AG

Validation: TG, AB, MS, AG

Writing—original draft: SA, TG, AB, MS

Writing—review and editing: MP, ST, PM, AG

All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Representative examples of hallucination and question-related confusion classifications.

[DOCX File (Microsoft Word File), 21 KB-Multimedia Appendix 1]

References

1. Andrei S, Longrois D, Stefan MG, Stefan G. Chat-based generative pretrained transformers sits the European Diploma in Anaesthesiology and Intensive Care part I examination: a computational study. *Eur J Anaesthesiol*. Apr 1, 2024;41(4):323-325. [doi: [10.1097/EJA.0000000000001969](https://doi.org/10.1097/EJA.0000000000001969)] [Medline: [38426256](https://pubmed.ncbi.nlm.nih.gov/38426256/)]
2. Benito P, Isla-Jover M, González-Castro P, et al. GPT-4o and OpenAI o1 performance on the 2024 Spanish competitive medical specialty access examination: cross-sectional quantitative evaluation study. *JMIR Med Educ*. Jan 12, 2026;12:e75452. [doi: [10.2196/75452](https://doi.org/10.2196/75452)] [Medline: [41525685](https://pubmed.ncbi.nlm.nih.gov/41525685/)]
3. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. Feb 8, 2023;9:e45312. [doi: [10.2196/45312](https://doi.org/10.2196/45312)] [Medline: [36753318](https://pubmed.ncbi.nlm.nih.gov/36753318/)]

4. Siam MK, Varela A, Faruk MJ, et al. Benchmarking large language models on the United States Medical Licensing Examination for clinical reasoning and medical licensing scenarios. *Sci Rep*. Dec 3, 2025;16(1):1387. [doi: [10.1038/s41598-025-31010-4](https://doi.org/10.1038/s41598-025-31010-4)] [Medline: [41339739](https://pubmed.ncbi.nlm.nih.gov/41339739/)]
5. Ferretti R, Rizzi JS, Requena LS, Bicudo AM, Hamamoto Filho PT. Factors associated with lower performance of artificial intelligence on answering undergraduate medical education multiple-choice questions. *Med Sci Educ*. 2025;35(4):2145-2152. [doi: [10.1007/s40670-025-02426-4](https://doi.org/10.1007/s40670-025-02426-4)] [Medline: [41112890](https://pubmed.ncbi.nlm.nih.gov/41112890/)]
6. Ravi A, Neinstein A, Murray SG. Large language models and medical education: preparing for a rapid transformation in how trainees will learn to be doctors. *ATS Sch*. 2023;4(3):282-292. [doi: [10.34197/ats-scholar.2023-0036PS](https://doi.org/10.34197/ats-scholar.2023-0036PS)] [Medline: [37795112](https://pubmed.ncbi.nlm.nih.gov/37795112/)]
7. Anh-Hoang D, Tran V, Nguyen LM. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Front Artif Intell*. 2025;8:1622292. [doi: [10.3389/frai.2025.1622292](https://doi.org/10.3389/frai.2025.1622292)] [Medline: [41098969](https://pubmed.ncbi.nlm.nih.gov/41098969/)]
8. Metheniti E, Bhar S, Asher N. Une revue sur les hallucinations des LLM. In: Bechet F, Chifu AG, Pinel-Sauvagnat K, Favre B, Maes E, Nurbakova D, editors. *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN). ATALA & ARIA*; 2025:756-779. URL: <https://aclanthology.org/2025.jeptalnrecital-taln.46/> [Accessed 2026-08-13]
9. Han Z, Battaglia F, Mansuria K, Heyman Y, Terlecky SR. Beyond text generation: assessing large language models' ability to reason logically and follow strict rules. *AI*. 2025;6(1):12. [doi: [10.3390/ai6010012](https://doi.org/10.3390/ai6010012)]
10. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst*. 2025;43(2):1-55. [doi: [10.1145/3703155](https://doi.org/10.1145/3703155)]
11. Altermatt FR, Neyem A, Sumonte NI, et al. Evaluating GPT-4o in high-stakes medical assessments: performance and error analysis on a Chilean anesthesiology exam. *BMC Med Educ*. Oct 27, 2025;25(1):1499. [doi: [10.1186/s12909-025-08084-9](https://doi.org/10.1186/s12909-025-08084-9)] [Medline: [41146119](https://pubmed.ncbi.nlm.nih.gov/41146119/)]
12. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Aug 2023;620(7972):172-180. [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
13. Ateleanu B, Goldik Z, Varvinskiy A, Moisin T, Hasan N. An overview of the European Diploma of Anaesthesia and Intensive Care and of other important initiatives of the European Society of Anaesthesiology. *Jurnalul Român de Anestezie Terapie Intensivă*. 2013;20(2):137-144. URL: https://www.researchgate.net/publication/288566101_An_overview_of_the_European_Diploma_of_Anaesthesia_and_Intensive_Care_and_of_other_important_initiatives_of_the_European_Society_of_Anaesthesiology [Accessed 2026-08-18]
14. Brogly N, Varvinskiy A, Varosyan A, et al. On-Line Assessment (OLA) as a preparation for the European Diploma in Anaesthesiology and Intensive Care (EDAIC), a retrospective observational study on the results and the potential impact on the part-I examination. *Rev Esp Anesthesiol Reanim (Engl Ed)*. Oct 2022;69(8):454-462. [doi: [10.1016/j.redare.2022.08.006](https://doi.org/10.1016/j.redare.2022.08.006)] [Medline: [36089526](https://pubmed.ncbi.nlm.nih.gov/36089526/)]
15. LOI n° 2012-300 du 5 mars 2012 relative aux recherches impliquant la personne humaine (1). République Française. URL: <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000025441587/> [Accessed 2026-08-18]
16. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. Mar 30, 2023;388(13):1233-1239. [doi: [10.1056/NEJMSr2214184](https://doi.org/10.1056/NEJMSr2214184)] [Medline: [36988602](https://pubmed.ncbi.nlm.nih.gov/36988602/)]
17. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. Nov 2024;58(11):1276-1285. [doi: [10.1111/medu.15402](https://doi.org/10.1111/medu.15402)] [Medline: [38639098](https://pubmed.ncbi.nlm.nih.gov/38639098/)]
18. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. Feb 2023;2(2):e0000198. [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
19. Yoon SH, Oh SK, Lim BG, Lee HJ. Performance of ChatGPT in the in-training examination for anesthesiology and pain medicine residents in South Korea: observational study. *JMIR Med Educ*. Sep 16, 2024;10:e56859. [doi: [10.2196/56859](https://doi.org/10.2196/56859)] [Medline: [39284182](https://pubmed.ncbi.nlm.nih.gov/39284182/)]
20. Khakpaki A. Advancements in artificial intelligence transforming medical education: a comprehensive overview. *Med Educ Online*. Dec 2025;30(1):2542807. [doi: [10.1080/10872981.2025.2542807](https://doi.org/10.1080/10872981.2025.2542807)] [Medline: [40798935](https://pubmed.ncbi.nlm.nih.gov/40798935/)]
21. Blacker SN, Chen F, Winecoff D, et al. An exploratory analysis of ChatGPT compared to human performance with the anesthesiology oral board examination: initial insights and implications. *Anesth Analg*. Jun 1, 2025;140(6):1253-1262. [doi: [10.1213/ANE.0000000000006875](https://doi.org/10.1213/ANE.0000000000006875)] [Medline: [39269908](https://pubmed.ncbi.nlm.nih.gov/39269908/)]
22. Patel S, Ngo V, Wilhelmi B. Evaluating large language models on American Board of Anesthesiology-style anesthesiology questions: accuracy, domain consistency, and clinical implications. *J Cardiothorac Vasc Anesth*. Sep 2025;39(9):2511-2515. [doi: [10.1053/j.jvca.2025.05.033](https://doi.org/10.1053/j.jvca.2025.05.033)] [Medline: [40518333](https://pubmed.ncbi.nlm.nih.gov/40518333/)]

23. Chan-Chia Lin C, Sun JS, Chang CH, Chang YH, Zwei-Chieng Chang J. Performance of artificial intelligence chatbots in national dental licensing examination. *J Dent Sci*. Oct 2025;20(4):2307-2314. [doi: [10.1016/j.jds.2025.05.012](https://doi.org/10.1016/j.jds.2025.05.012)] [Medline: [41040620](https://pubmed.ncbi.nlm.nih.gov/41040620/)]
24. Genovese A, Borna S, Gomez-Cabello CA, et al. The current landscape of artificial intelligence in plastic surgery education and training: a systematic review. *J Surg Educ*. Aug 2025;82(8):103519. [doi: [10.1016/j.jsurg.2025.103519](https://doi.org/10.1016/j.jsurg.2025.103519)] [Medline: [40378641](https://pubmed.ncbi.nlm.nih.gov/40378641/)]
25. Luo J, Peng Q, Cao H, Li W, Tan S. Medical undergraduate students' readiness and anxiety toward artificial intelligence: a systematic review and meta-analysis. *BMC Med Educ*. Dec 4, 2025;26(1):44. [doi: [10.1186/s12909-025-08388-w](https://doi.org/10.1186/s12909-025-08388-w)] [Medline: [41339885](https://pubmed.ncbi.nlm.nih.gov/41339885/)]
26. Chakri I, El Khayali O, Lahlou L. Knowledge and perceptions of AI among medical students in Morocco: cross-sectional study. *JMIR Form Res*. Sep 19, 2025;9:e66156. [doi: [10.2196/66156](https://doi.org/10.2196/66156)] [Medline: [40971792](https://pubmed.ncbi.nlm.nih.gov/40971792/)]
27. Li Z, Li F, Fu Q, et al. Large language models and medical education: a paradigm shift in educator roles. *Smart Learn Environ*. 2024;11:26. [doi: [10.1186/s40561-024-00313-w](https://doi.org/10.1186/s40561-024-00313-w)]
28. Roveta A, Castello LM, Massarino C, Francese A, Ugo F, Maconi A. Artificial intelligence in medical education: a narrative review on implementation, evaluation, and methodological challenges. *AI*. 2025;6(9):227. [doi: [10.3390/ai6090227](https://doi.org/10.3390/ai6090227)]
29. Jacovi A, Caciularu A, Goldman O, Goldberg Y. Stop uploading test data in plain text: practical strategies for mitigating data contamination by evaluation benchmarks. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2023:5075-5084. [doi: [10.18653/v1/2023.emnlp-main.308](https://doi.org/10.18653/v1/2023.emnlp-main.308)]
30. Sainz O, Campos JA, García-Ferrero I, Etxaniz J, de Lacalle OL, Agirre E. NLP evaluation in trouble: on the need to measure LLM data contamination for each benchmark. *arXiv*. Preprint posted online on Oct 27, 2023. [doi: [10.48550/arXiv.2310.18018](https://doi.org/10.48550/arXiv.2310.18018)]

Abbreviations

EDAIC: European Diploma in Anaesthesiology and Intensive Care

GEE: generalized estimating equations

IADE: infirmier anesthésiste diplômé d'état

LLM: large language model

MCQ: multiple-choice question

OLA: On-Line Assessment

Edited by Luke MacNeill; peer-reviewed by Alousious Kasagga, Niyazi Cetin; submitted 22.Mar.2026; final revised version received 23.Jul.2026; accepted 30.Jul.2026; published 01.Sep.2026

Please cite as:

Andrei S, Giet T, Belouard A, Stefan M, Popescu M, Tanaka S, Montravers P, Gouel A

Performance and Hallucination Analysis of Large Language Models on European Anesthesiology Examinations: Cross-Sectional Comparative Study

JMIR Form Res 2026;10:e95859

URL: <https://formative.jmir.org/2026/1/e95859>

doi: [10.2196/95859](https://doi.org/10.2196/95859)

© Stefan Andrei, Thibault Giet, Alexis Belouard, Mihai Stefan, Mihai Popescu, Sébastien Tanaka, Philippe Montravers, Aurélie Gouel. Originally published in *JMIR Formative Research* (<https://formative.jmir.org>), 01.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Formative Research*, is properly cited. The complete bibliographic information, a link to the original publication on <https://formative.jmir.org>, as well as this copyright and license information must be included.