



Large Language Model Accuracy on Subspecialty Anaesthesiology Examination Questions

Büyük Dil Modellerinin Yan Dal Anesteziyoloji Sınav Sorularındaki Doğruluğu

Muhammed Nezh KOÇ¹, Sait Ramazan GÜLBAY²

¹Department of Anaesthesiology and Reanimation, Konya City Hospital, Konya, Türkiye
²Department of Medical Microbiology, Faculty of Medicine, Necmettin Erbakan University, Konya, Türkiye

Abstract

Aim: In large language model (LLM) examination studies, models answer the same items repeatedly, so observations are clustered and conventional intervals overstate precision. We aimed to measure the accuracy of seven LLMs on a subspecialty-level anaesthesiology question bank and to determine which between-model and subgroup differences remain distinguishable under a cluster-respecting analysis.

Material and Method: A 120-item, five-option multiple-choice bank in the style of the Anaesthesiology and Reanimation Subspecialty Examination, written by the authors and reviewed against current guidelines and textbooks, was administered to seven LLMs in five runs at temperature 0, with option order re-randomised per run (4,200 responses from 120 unique items). Accuracy is reported with cluster bootstrap 95% confidence intervals (CIs) resampling items (4,000 replicates) and naive Wilson intervals for comparison; subgroup analyses are exploratory.

Results: Pooled accuracy was 80.1% (cluster-robust 95% CI 75.8-83.9; naive Wilson 78.8-81.3). Between-model differences were large and robust, spanning 60.8% (54.0-67.5) to 94.0% (90.3-97.0) with non-overlapping extremes. Cluster-robust intervals were wider in 25 of 26 estimates (median factor 2.1, range 0.8-4.0). Most subgroup comparisons were not distinguishable, intervals overlapping substantially: vignette (75.4%, 62.5-86.5) versus non-vignette (80.9%, 76.6-84.9) and guideline-dependent (73.0%, 64.6-80.7) versus other items (80.9%, 76.6-84.9). Only contrasts between weakest and strongest domains persisted. Between-run standard deviation was 1.26-2.80 percentage points.

Conclusion: Between-model differences and run-to-run instability are robust; commonly emphasised domain and item-characteristic differences are mostly not distinguishable once clustering is respected. Benchmarks of 120 items can separate models whose accuracies differ widely but cannot reliably localise their weaknesses.

Keywords: Artificial intelligence, anesthesiology, educational measurement, confidence intervals, natural language processing

Öz

Amaç: Büyük dil modeli (BDM) sınav çalışmalarında aynı maddeler modellerce tekrar tekrar yanıtlandığından gözlemler kümelenmiştir ve alışılmış aralıklar kesinliği olduğundan yüksek gösterir. Yedi BDM'nin yan dal düzeyinde bir anesteziyoloji soru bankasındaki doğruluğunu ölçmeyi ve modeller arası farklar ile alt grup farklarından hangilerinin kümelenmeyi gözetin bir analizde ayırt edilebilir kaldığını belirlemeyi amaçladık.

Gereç ve Yöntem: Anesteziyoloji ve Reanimasyon Yan Dal Sınavı tarzında, yazarlar tarafından yazılan ve güncel kılavuzlar ile ders kitapları üzerinden denetlenen 120 maddelik beş seçenekli bir soru bankası, yedi BDM'ye sıcaklık 0'da ve her denemede seçenek sırası yeniden rastgeleleştirilerek beş denemede soruldu (120 benzersiz maddeden 4.200 yanıt). Doğruluk, maddeleri yeniden örnekleyen küme önyüklemesi %95 güven aralıkları (GA; 4.000 yinleme) ve karşılaştırma için naif Wilson aralıklarıyla sunuldu; alt grup analizleri keşifeldir.

Bulgular: Birleşik doğruluk %80,1 idi (küme-dayanıklı %95 GA 75,8-83,9; naif Wilson 78,8-81,3). Modeller arası farklar büyük ve dayanıklıydı: %60,8 (54,0-67,5) ile %94,0 (90,3-97,0); uçların aralıkları örtüşmedi. Küme-dayanıklı aralıklar 26 tahminin 25'inde daha genişti (ortanca 2,1 kat; aralık 0,8-4,0). Alt grup karşılaştırmalarının çoğu ayırt edilebilir değildi, aralıklar belirgin örtüştü: vinyetli (%75,4; 62,5-86,5) ile vinyetsiz (%80,9; 76,6-84,9) ve kılavuz bağımlı (%73,0; 64,6-80,7) ile diğer maddeler (%80,9; 76,6-84,9). Yalnızca en zayıf ve en güçlü alanlar arasındaki karşıtlıklar korundu. Denemeler arası standart sapma 1,26-2,80 puandı.

Sonuç: Modeller arası farklar ve denemeler arası kararsızlık dayanıklıdır; sıkça vurgulanan alan ve madde niteliği farkları, kümelenme gözetildiğinde çoğunlukla ayırt edilebilir değildir. 120 maddelik ölçütler doğrulukları belirgin ayrılan modelleri ayırt edebilir ancak zayıflıklarını güvenilir biçimde yerelleştiremez.

Anahtar Kelimeler: Yapay zekâ, anesteziyoloji, eğitimsel ölçme, güven aralıkları, doğal dil işleme



INTRODUCTION

Large language models (LLMs) have moved rapidly from general-purpose text systems to tools under examination for clinical roles.^[1] The first widely cited demonstrations used licensing examinations: ChatGPT performed at or near the passing threshold of the United States Medical Licensing Examination,^[2,3] and instruction-tuned models subsequently exceeded it.^[4] Benchmarking has since extended to specialty content, including anaesthesiology, where recent models have matched or exceeded human examinees on national board examinations.^[5-8]

Anaesthesiology is an informative test case, and one in which artificial intelligence has been discussed for some years.^[9] Its knowledge base combines basic science, device physics, procedural anatomy and guideline-defined algorithms, so a single aggregate figure may conceal very different levels of competence. This reasoning motivates the subgroup analyses now standard in the field: accuracy reported by content domain, cognitive level and question format.

Those analyses rest on a statistical assumption that is rarely examined. A study reporting "4,200 observations" has not observed 4,200 independent facts; it has observed far fewer examination items, each answered repeatedly. Responses to the same item are correlated because difficulty is a property of the item. Intervals computed as though every response were independent are therefore too narrow, and the error is often greatest where it matters most: in subgroups defined by a handful of items.

Two further gaps persist. Most reports give one accuracy figure per model from one run; because outputs are not fully deterministic even at temperature 0, and accuracy is sensitive to answer-option order,^[10-12] a single run provides no estimate of measurement uncertainty. Scores on public examinations may also be inflated by exposure to the items during training;^[13,14] items written afresh for a study close this route, at the price of losing the psychometric record of an administered examination.

We therefore evaluated seven openly accessible LLMs on a bank of 120 multiple-choice items written for this study in the style of the Anaesthesiology and Reanimation Subspecialty Examination, administered five times each, and analysed the results with confidence intervals that respect the clustered structure of the data. We asked how accurate the models are, and which of the conclusions that a study of this kind would conventionally draw actually survive an appropriate analysis.

MATERIAL AND METHOD

Ethics

The study used examination-style items created by the authors and the outputs of the evaluated large language models; no human participants or patient data were involved, and institutional review board approval was not required.

Question Development

The question bank comprised 120 single-best-answer items, each with five options, written by the authors in the style of the Anaesthesiology and Reanimation Subspecialty Examination. Item scope followed the Turkish national core curriculum, version 2.5,^[15] and the examination's official application guide;^[16] topic balance was checked against the American Board of Anesthesiology content outline^[17] and the European Diploma in Anaesthesiology and Intensive Care (EDAIC) syllabus;^[18] and item writing followed the National Board of Medical Examiners (NBME) item-writing guide^[19] and the principles of Haladyna et al.^[20] with vignette items using a fixed template and the medical content of every item anchored to current international guidelines.^[21-25]

Because not all evaluated models accept image input, two figure-based items produced during writing were replaced with equivalent text-based items. The replacement items were administered to all seven models in five runs using the same protocol, and their responses replaced those of the original items in the final dataset. Thus, the final 120-item bank was text-only and every model received identical text for every analysed item. Both authors reviewed all items against standard reference textbooks;^[26-29] items were written with the aim of requiring application or analysis rather than recall alone,^[30] although cognitive level was not formally classified and this intention was not verified empirically. Because the items were not published before the model queries, none can have appeared in the training data of any evaluated model.

Item Classification

Each item was classified by a single anaesthesiology and reanimation specialist (one of the authors) into 12 mutually exclusive content domains, using the American Board of Anesthesiology content outline^[17] as the primary framework, aligned with the EDAIC subject division^[18] and cross-checked against the national core curriculum;^[15] items were assigned to the competence being tested rather than the setting of the stem. Three binary characteristics were defined operationally: guideline dependence (the key derives directly from a named guideline or consensus definition and could change with its revision), clinical vignette (a patient scenario with clinical data precedes the question) and negatively worded stem (the lead-in asks for the incorrect or exception option). The classification was not independently duplicated, no inter-rater agreement statistic is available, and all classification-dependent analyses are exploratory.

Models and Administration Protocol

Seven openly accessible LLMs were evaluated. "Openly accessible" here means reachable through a provider application programming interface without a paid subscription at the time of the study, and implies neither open weights nor an open-source licence. Complete model identifiers, query dates and inference settings, together with the verbatim administration prompt, are given in **Supplementary Table S1**. Each model answered every item in five independent runs at temperature 0,

with option order re-randomised per run using a deterministic seed derived from the item, model and run identifiers;^[10-12] responses were mapped back through the permutation and compared with the answer key, yielding 7 models × 5 runs × 120 items = 4,200 responses from 120 unique items.

Statistical Analysis

The outcome was the proportion of correct responses. Because the 4,200 responses arise from 120 unique items, each answered by 7 models in 5 runs, the primary analysis uses a cluster bootstrap in which the resampling unit is the item;^[31,32] in each of 4,000 replicates, 120 items were drawn with replacement with all their responses, and the 2.5th and 97.5th percentiles give the 95% confidence interval. Naive Wilson intervals,^[33,34] which assume independence, are shown for comparison only. Model-level intervals treat the seven models as fixed; for the pooled estimate an interval resampling both items and models is reported in the Discussion. Between-run variability is summarised as the standard deviation (SD) of the five run-level accuracies, in percentage points. Analyses are descriptive, with no formal significance testing: differences are described as distinguishable only when cluster-robust intervals do not overlap, and all subgroup analyses are exploratory, presented with the number of unique items on which they rest. Item difficulty was calculated as the proportion of the 35 model-run observations per item that were correct, and item discrimination as the corrected item-total correlation between the item and the total score across the remaining items; both are reported descriptively.

RESULTS

Across 4,200 responses to 120 unique items, pooled accuracy was 80.1%. The cluster-robust 95% confidence interval was 75.8-83.9, compared with 78.8-81.3 under the naive independence assumption: the appropriate interval is 3.2 times wider (Table 1). Model accuracy spanned 33.2 percentage points, from 60.8% to 94.0% (Table 1, Figure 1). These differences are robust: the intervals of the two highest-scoring models do not overlap those of any other model, and the interval of the weakest model sits below all others. Two models exceeded 90%, and the four models between 74.7% and 83.5% had mutually overlapping intervals, so their ordering among themselves

should not be interpreted. Agreement between models is summarised in Supplementary Table S2: all seven models were correct together in 40.0% of item-run decisions and all seven were wrong in 2.2%, and a majority vote across the seven models (92.2%) did not exceed the best single model (94.0%).

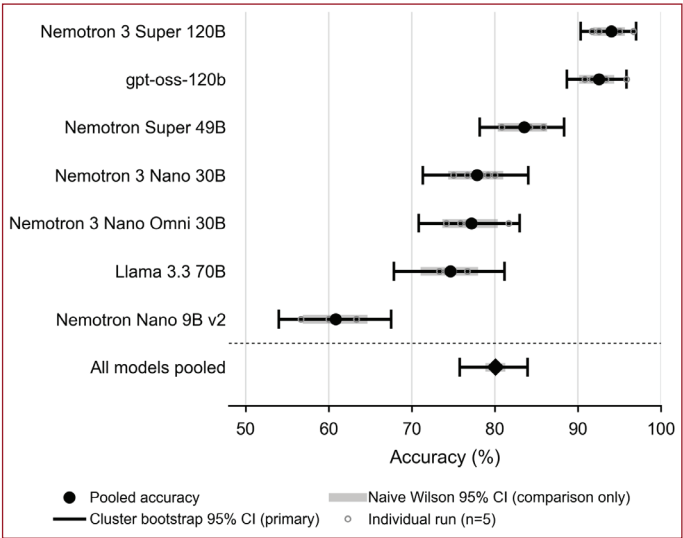


Figure 1. Accuracy of seven large language models on the 120-item anaesthesiology question bank. Filled circles show accuracy pooled across five independent administrations (600 responses per model from 120 unique items); open circles show individual run accuracies. Thick bars are cluster bootstrap 95% confidence intervals resampling items (4,000 replicates); thin grey inner bars are naive Wilson intervals, shown for comparison only and not used for inference. The bottom row shows all seven models combined. CI: confidence interval. Model names are abbreviated; complete model identifiers are given in Supplementary Table S1.

Supplementary Table S2. Agreement among the seven models across 600 item-run decisions		
Measure	n/600	%
Majority vote across seven models	553	92.2
Decisions with 7 of 7 models correct	240	40.0
Decisions with 6 of 7 models correct	151	25.2
Decisions with 5 of 7 models correct	81	13.5
Decisions with 4 of 7 models correct	60	10.0
Decisions with 3 of 7 models correct	30	5.0
Decisions with 2 of 7 models correct	17	2.8
Decisions with 1 of 7 models correct	8	1.3
Decisions with 0 of 7 models correct	13	2.2

A decision is one item in one run (120 items × 5 runs = 600 decisions per model set). Majority-vote accuracy was 1.8 percentage points below the best single model (94.0%), with 11 tied decisions. Pairwise agreement between models ranged from 58.7% (Llama 3.3 70B and Nemotron Nano 9B v2) to 90.7% (Nemotron 3 Super 120B and gpt-oss-120b).

Supplementary Table S1. Model identifiers, snapshots, query dates and inference settings					
Model (manuscript name)	Provider	Complete model identifier	Query dates	Temp.	n
Nemotron 3 Super 120B A12B	NVIDIA Build (NIM)	nvidia/nemotron-3-super-120b-a12b	31 Jul 2026	0	600
gpt-oss-120b	Groq	openai/gpt-oss-120b	30-31 Jul 2026	0	600
Llama-3.3-Nemotron Super 49B v1.5	NVIDIA Build (NIM)	nvidia/llama-3.3-nemotron-super-49b-v1.5	30-31 Jul 2026	0	600
Nemotron 3 Nano 30B A3B	NVIDIA Build (NIM)	nvidia/nemotron-3-nano-30b-a3b	31 Jul 2026	0	600
Nemotron 3 Nano Omni 30B (reasoning)	NVIDIA Build (NIM)	nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	31 Jul 2026	0	600
Llama 3.3 70B	Groq	llama-3.3-70b-versatile	30-31 Jul 2026	0	600
Nemotron Nano 9B v2	NVIDIA Build (NIM)	nvidia/nvidia-nemotron-nano-9b-v2	31 Jul 2026	0	600

Every model returned a single snapshot identifier across all 600 of its calls, confirming that no model was silently updated mid-study. All calls used temperature 0 and the provider's default maximum output length.

Table 1. Accuracy of seven large language models, with cluster-robust and naive confidence intervals (4,200 responses from 120 unique items)

Model	Accuracy (%)	Cluster-robust 95% CI*	Naive Wilson 95% CI†	Ratio‡	SD across runs§
Nemotron 3 Super 120B	94.0	90.3-97.0	91.8-95.6	1.8	1.99
gpt-oss-120b	92.5	88.7-95.8	90.1-94.3	1.7	2.12
Nemotron Super 49B	83.5	78.2-88.3	80.3-86.3	1.7	1.81
Nemotron 3 Nano 30B	77.8	71.3-84.0	74.3-81.0	1.9	2.01
Nemotron 3 Nano Omni 30B	77.2	70.8-83.0	73.6-80.3	1.8	2.80
Llama 3.3 70B	74.7	67.8-81.2	71.0-78.0	1.9	1.26
Nemotron Nano 9B v2	60.8	54.0-67.5	56.9-64.7	1.7	2.76
All models pooled	80.1	75.8-83.9	78.8-81.3	3.2	NA

* Primary analysis: bootstrap resampling examination items (4,000 replicates), which preserves the correlation between responses to the same item. Model-level intervals treat the seven models as fixed.

† Assumes all 4,200 responses are independent; shown for comparison only.

‡ Width of the cluster-robust interval divided by the width of the naive interval.

§ Standard deviation of the five run-level accuracies, in percentage points; a within-model quantity not affected by item clustering.

CI: confidence interval; NA: not applicable; SD: standard deviation.

Accuracy increased with parameter count within one lineage (from 60.8% at 9B to 94.0% at 120B; **Table 1**), but this is descriptive across different released models rather than a controlled manipulation of scale, and did not hold across lineages: a 49B model outperformed a 70B model from another family.

Two variants of a 30B model, one released with a reasoning mode, returned nearly identical pooled accuracy (77.8% and 77.2%). Because these are different released model variants rather than the same model with reasoning enabled and disabled, this observation does not establish that reasoning mode is without effect, and we draw no such conclusion.

Run-level accuracy varied within every model. The SD across the five runs reached 2.80 percentage points, and the widest within-model spread was 7.5 percentage points, despite deterministic decoding. Variability did not track accuracy: the lowest SD of the series belonged to one of the weaker models rather than the strongest (**Supplementary Table S3**). The weakest model returned the same result in all five runs for only 50.8% of items, against 89.2% for the strongest.

Content-domain accuracy ranged from 62.0% to 95.7% as point estimates, but most domains rest on very few unique items and their intervals are correspondingly wide (**Table 2, Figure 2**). The three-item subspecialty domain yielded an interval spanning 29 points, and the narrow intervals of some of the smallest domains reflect low between-item heterogeneity and the coarseness of a bootstrap resting on very few items rather than adequate item sampling. The contrasts that remained distinguishable were confined to the extremes of the range:

regional anaesthesia (62.0%, 16 items) was separated from physiology and pathophysiology, pharmacology, critical care medicine, anatomy and obstetric anaesthesia (**Table 2**).

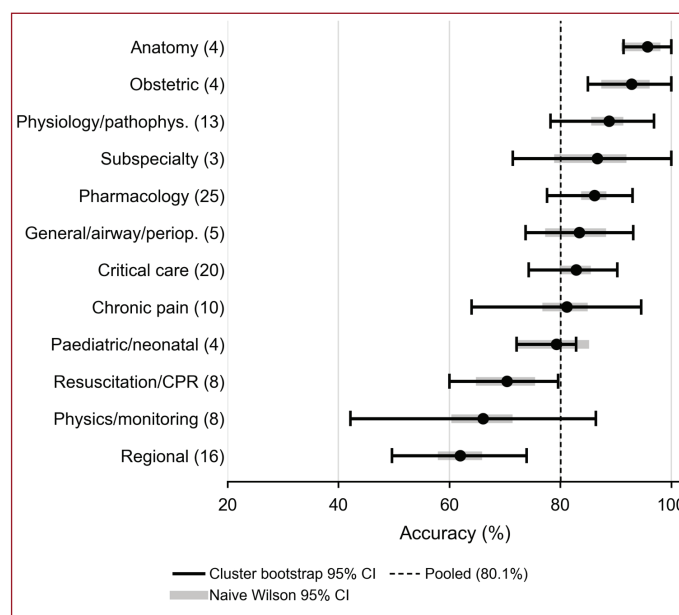


Figure 2. Accuracy by content domain, pooled across seven models and five runs. Thick bars are cluster bootstrap 95% confidence intervals resampling items (4,000 replicates); thin grey inner bars are naive Wilson intervals, shown for comparison only and not used for inference. The number of unique items contributing to each domain is given in parentheses; intervals generally widen sharply as this number falls, although with very few items the width also depends on between-item heterogeneity and the discreteness of the bootstrap distribution; the smallest domains cannot support inference. The dashed vertical line marks pooled accuracy across all 120 items. All domain analyses are exploratory. CI: confidence interval.

Supplementary Table S3. Run-level accuracy and within-model consistency across five administrations

Model	Run 1 %	Run 2 %	Run 3 %	Run 4 %	Run 5 %	Mean %	SD	Identical across runs, n/120 (%)
Nemotron 3 Super 120B	94.2	91.7	96.7	95.0	92.5	94.0	1.99	107/120 (89.2)
gpt-oss-120b	91.7	95.8	90.8	90.8	93.3	92.5	2.12	102/120 (85.0)
Nemotron Super 49B	83.3	84.2	83.3	80.8	85.8	83.5	1.81	85/120 (70.8)
Nemotron 3 Nano 30B	78.3	75.0	76.7	80.0	79.2	77.8	2.01	84/120 (70.0)
Nemotron 3 Nano Omni 30B	76.7	74.2	81.7	75.8	77.5	77.2	2.80	79/120 (65.8)
Llama 3.3 70B	74.2	76.7	75.0	74.2	73.3	74.7	1.26	89/120 (74.2)
Nemotron Nano 9B v2	60.0	56.7	63.3	63.3	60.8	60.8	2.76	61/120 (50.8)

SD: standard deviation of the five run-level accuracies, in percentage points. Identical: items scored the same way (correct or incorrect) in all five runs of that model, although option order was re-randomised per run. Model names are abbreviated; complete model identifiers are given in Supplementary Table S1.

Table 2. Exploratory analysis by content domain, pooled across seven models and five runs

Content domain	Unique items (n)	Accuracy (%)	Cluster-robust 95% CI*	Naive Wilson 95% CI†	Ratio‡
Anatomy	4	95.7	91.4-100.0	91.0-98.0	1.2
Obstetric anaesthesia	4	92.9	85.0-100.0	87.4-96.1	1.7
Physiology and pathophysiology	13	88.8	78.2-96.9	85.6-91.4	3.2
Subspecialty anaesthesia	3	86.7	71.4-100.0	78.9-91.9	2.2
Pharmacology	25	86.2	77.6-93.0	83.7-88.3	3.4
General anaesthesia, airway and perioperative care	5	83.4	73.7-93.1	77.2-88.2	1.8
Critical care medicine	20	82.9	74.3-90.3	79.9-85.5	2.9
Chronic pain medicine	10	81.1	64.0-94.6	76.7-84.9	3.7
Paediatric and neonatal anaesthesia	4	79.3	72.1-82.9	71.8-85.2	0.8
Resuscitation and cardiopulmonary resuscitation	8	70.4	60.0-79.6	64.8-75.4	1.8
Physics, monitoring and anaesthetic equipment	8	66.1	42.1-86.4	60.3-71.4	4.0
Regional anaesthesia	16	62.0	49.6-73.9	57.9-65.9	3.0

* Primary analysis, resampling items. † Shown for comparison only. ‡ Width of the cluster-robust interval divided by the width of the naive interval.

All comparisons in this table are exploratory. The widest intervals span more than 40 percentage points, and the narrow intervals of the smallest domains reflect low between-item heterogeneity and the coarseness of a bootstrap based on very few items rather than adequate item sampling; distinguishable contrasts are confined to those between the strongest and weakest domains. Domains are mutually exclusive and sum to 120 items.

CI: confidence interval.

No item-characteristic comparison survived (**Table 3**). Items containing a clinical vignette were answered correctly in 75.4% of responses against 80.9% for other items (18 versus 102 items), guideline-dependent items scored 73.0% against 80.9% (13 versus 107 items), and negatively worded stems scored 81.0% against 79.4% (50 versus 70 items); all three cluster-robust intervals overlapped substantially (**Table 3**). Under the naive analysis the vignette and guideline differences would have appeared clear.

Across the 26 estimates reported in this paper, the cluster-robust interval was wider than the naive interval in 25, by a median factor of 2.1 (range 0.8 to 4.0) (**Figure 3**). The single exception was a four-item domain in which the cluster-robust interval was marginally narrower than the naive one: with so few items, the width of the bootstrap interval is governed by between-item heterogeneity, which was low in this domain, and by the discreteness of a distribution resting on only four resampled items. The inflation was not a simple function of the number of unique items: model-level estimates, each resting on all 120 items, were inflated 1.7 to 1.9 times, whereas the pooled estimate across models was inflated 3.2 times and individual domains ranged from 0.8 to 4.0 times.

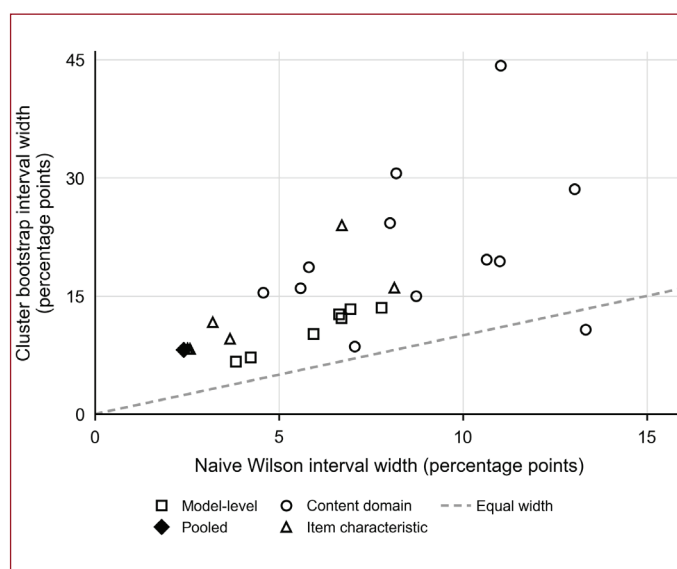


Figure 3. Width of the cluster bootstrap confidence interval plotted against the width of the naive Wilson interval for each of the 26 estimates reported in this paper. All but one point lies above the line of equality: the naive interval was narrower in 25 of the 26 analyses; median inflation was 2.1-fold and the maximum 4.0-fold. Marker shape distinguishes model-level estimates, the pooled estimate, content domains and item characteristics.

Table 3. Exploratory analysis by item characteristic

Item characteristic	Unique items (n)	Accuracy (%)	Cluster-robust 95% CI*	Naive Wilson 95% CI†	Distinguishable?‡
Clinical vignette present	18	75.4	62.5-86.5	71.9-78.6	No
Clinical vignette absent	102	80.9	76.6-84.9	79.6-82.2	
Guideline-dependent	13	73.0	64.6-80.7	68.7-76.8	No
Not guideline-dependent	107	80.9	76.6-84.9	79.6-82.2	
Negatively worded stem	50	81.0	75.9-85.4	79.1-82.8	No
Positively worded stem	70	79.4	73.3-84.9	77.7-80.9	

* Primary analysis, resampling items. † Shown for comparison only. ‡ Cluster-robust intervals for the two categories do not overlap.

Under the naive analysis the vignette and guideline contrasts would both have appeared clear; neither survives an analysis respecting item clustering. Definitions: guideline-dependent, the key derives directly from a named guideline or consensus definition; clinical vignette, the item presents a patient scenario preceding the question. The two comparator rows are largely overlapping subsets of the same 120 items and differ by less than 0.05 percentage points in accuracy, so their intervals coincide after rounding.

CI: confidence interval.

Item statistics derived from the model responses describe an item bank that was easy for the evaluated models overall but not uniformly so (Figure 4). Median item difficulty was 85.7% of the 35 model-run observations per item; 28 items were answered correctly in every observation and one in none. Among the remaining 91 items, whose scores varied across observations, the median corrected item-total correlation was 0.359; nine had a correlation at or below zero, indicating that they did not discriminate between stronger and weaker models. These items were retained because removing them post hoc would select on the outcome, but they are a further reason to treat the domain-level estimates as exploratory.

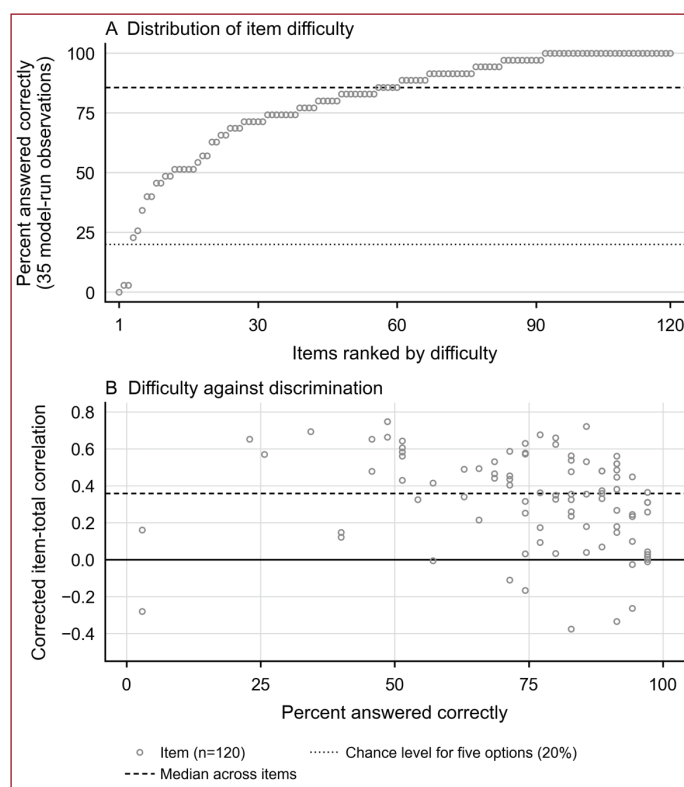


Figure 4. Psychometric properties of the 120-item bank derived from model responses. (A) Items ranked by difficulty, expressed as the percentage of the 35 model-run observations per item answered correctly; the dashed line marks the median across items and the dotted line the 20% expected by chance with five options. (B) Item difficulty plotted against corrected item-total correlation for the 91 items whose scores varied across observations (the 28 items answered correctly in every observation and the one answered correctly in none have no defined correlation), with the dashed line marking the median. These statistics describe the bank as the evaluated models encountered it and are not equivalent to psychometric statistics obtained from human examinees.

DISCUSSION

Seven openly accessible LLMs answered 120 subspecialty anaesthesiology examination questions with a pooled accuracy of 80.1%, and the best model reached 94.0%, within the range recent models have achieved on specialty board examinations.^[5-8] The contribution of this study is less this figure than the demonstration of which accompanying claims an appropriate analysis supports. Between-model differences

are robust: the 33-point separation between strongest and weakest survives with a wide margin, and the ordering into a high-performing pair, a middle group and a weak model is stable. A study of this design can rank models at this coarse level, which is a useful function.

Localising a weakness is a different matter. Our point estimates would support the familiar narrative that models are weak in regional anaesthesia and equipment physics and falter on vignettes and guideline-dependent content. Under cluster-robust analysis most of this is not distinguishable: an eight-item domain yields an interval more than 44 points wide, and the vignette and guideline contrasts overlap substantially once the 18 and 13 items on which they rest are treated as clusters. Only coarse contrasts between the weakest and strongest domains persist. This concern extends beyond the present dataset, because subgroup reporting of this kind is routine in LLM examination studies with similar item counts.^[5-8] An examination of about 120 items spread over a dozen domains provides only a handful of items in many of them (3 to 25 in this bank). Repeating it five times with seven models multiplies rows but adds no items, and it is items that carry the information about content. Studies wishing to make domain-level claims should either use much larger item banks or label such analyses exploratory.

Our repeated-administration findings are unaffected by item clustering, because between-run SD is computed within a model across runs of the identical item set. That SD reached 2.80 percentage points at temperature 0, with a widest within-model spread of 7.5 points. A single administration therefore carries this much unreported uncertainty, a concern raised early in this literature^[10] and reported again recently.^[5] Because option order was re-randomised per run, part of the spread reflects the known sensitivity of LLMs to option position;^[11,12] a model whose answer changes with option order provides a weak basis for the inference the score is taken to support.

The regional anaesthesia result, though its interval is wide, illustrates an instructive failure mode. The hardest item in that domain asked which fluid should be injected to confirm needle-tip position during nerve stimulation; models persistently selected normal saline, which is conductive and abolishes the stimulation response, rather than 5% dextrose, the non-conducting solution indicated for electrical nerve localisation.^[35,36] The hardest equipment item asked which volatile agent has the highest vapour pressure at room temperature; no model answered correctly in any run, the memorised association between desflurane and high vapour pressure becoming wrong once nitrous oxide is among the options.^[37] These are illustrative failure modes in this item bank, not established domain-level deficits.

Writing the items afresh for this study changes the trade-offs of the benchmark rather than removing them. Items that did not exist before the study cannot have leaked into any model's training data, which closes the contamination route that public

examinations leave open.^[13,14] The price is that the bank has no history: no candidate has sat it, no psychometric statistics from human examinees exist, and its difficulty cannot be equated with that of the official examination. Absolute accuracies should be read with this in mind; the findings on interval width and run variability do not depend on it. Generalisation beyond the evaluated models is not supported either. Five of the seven models belong to related lineages, and no leading proprietary model was included. When the pooled estimate is recomputed resampling both items and models, the interval widens to 70.9–88.5, which is the appropriate expression of how little seven related models constrain a statement about LLMs as a class. Our conclusions refer to the evaluated models, and the pooled figure should not be read as representative of contemporary LLM performance in general.

Limitations

First, the study rests on 120 unique items classified by a single assessor: repeated administration adds no information about content, the classification was not independently duplicated and was re-derived during revision (moving a small number of boundary items), and no inter-rater agreement statistic exists. All classification-dependent analyses are therefore exploratory, several rest on three to four items, and no recognition-versus-management dissociation can be claimed.

Second, the items were written by the authors rather than drawn from an administered examination: the bank has no psychometric record from human examinees, no human comparator, and its difficulty is not calibrated to any official examination; the item statistics in **Figure 4** derive from model responses alone.

Finally, evaluation was confined to models reachable without a paid subscription, and examination performance is not evidence of clinical competence; nothing in these data supports the use of any of these models in patient care.

Implications

Three recommendations follow. Studies should report multiple administrations with a measure of dispersion. Confidence intervals should respect the clustered structure of repeated administration, since naive intervals were too narrow in 25 of our 26 estimates, by a median factor of 2.1. And domain-level claims should be made only when the number of unique items supports them, or labelled exploratory.

CONCLUSION

Seven openly accessible large language models answered a 120-item subspecialty-level anaesthesiology question bank with a pooled accuracy of 80.1%, and the strongest reached 94.0%. Differences between models were large and robust to an analysis respecting the clustered design, as was run-to-run instability of up to 2.80 percentage points at temperature 0. In contrast, the domain and item-characteristic differences that such studies conventionally emphasise were, with few

exceptions, not distinguishable once the 120 unique items were treated as the clusters they are. An examination-style benchmark of this size can separate models whose accuracies differ widely; it cannot reliably tell us where their weaknesses lie.

ETHICAL DECLARATIONS

Ethics Committee Approval: The study used examination-style items created by the authors and the outputs of computational models. No human participants, patient data or identifiable personal information were involved, and institutional review board approval was not required.

Informed Consent: Informed consent was not required for this study.

Referee Evaluation Process: Externally peer-reviewed.

Conflict of Interest Statement: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Author Contributions: All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29(8):1930–40.
2. 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* 2023;2(2):e0000198.
3. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? *JMIR Med Educ* 2023;9:e45312.
4. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023;620(7972):172–80.
5. Vishwanath K, Alyakin A, Ghosh M, et al. Evaluating the Performance and Fragility of Large Language Models on the Self-Assessment for Neurological Surgeons. *Neurosurgery*. 2026;99(3):563–70.
6. Angel MC, Rinehart JB, Cannesson MP, Baldi P. Clinical knowledge and reasoning abilities of AI large language models in anesthesiology: a comparative study on the American Board of Anesthesiology examination. *Anesth Analg* 2024;139(2):349–56.
7. Altermatt FR, Neyem A, Sumonte NI, Villagrán I, Mendoza M, Lacassie HJ. Evaluating the performance of large language models on the CONACEM anesthesiology certification exam: a comparison with human participants. *Appl Sci* 2025;15(11):6245.
8. Ronen A, Fein S, Orbach-Zinger S, et al. Large language models versus human examinee performance on Israeli anesthesiology board examinations. *Sci Rep* 2026;16(1):14978.
9. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. *Anesthesiology* 2020;132(2):379–94.
10. Epstein RH, Dexter F. Variability in large language models' responses to medical licensing and certification examinations. *JMIR Med Educ* 2023;9:e48305.
11. Zheng C, Zhou H, Meng F, Zhou J, Huang M. Large language models are not robust multiple choice selectors. In: *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, Vienna, 2024.
12. Pezeshkpour P, Hruschka E. Large language models sensitivity to the order of options in multiple-choice questions. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City: Association for Computational Linguistics, 2024:2006–17.
13. Xu C, Guan S, Greene D, Kechadi MT. Benchmark data contamination of large language models: a survey. *arXiv* 2024;arXiv:2406.04244.

14. Acar A. Data contamination in AI evaluation. *JMIR Med Inform* 2025;13:e80987.
15. Republic of Türkiye Ministry of Health, Medical Specialty Board (TUK). Anaesthesiology and Reanimation specialty training core curriculum, v.2.5. Ankara: TUK [cited 29 Jul 2026]. Available from: <https://tuk.saglik.gov.tr/TR-87040/v-25-dokumanlari.html>
16. Measuring, Selection and Placement Centre (ÖSYM). 2026 Subspecialty Training Entrance Examination (YDUS) application guide. Ankara: ÖSYM; 2026. Available at: <https://www.osym.gov.tr> (accessed 29 July 2026).
17. The American Board of Anesthesiology. Initial certification in anesthesiology content outline. Raleigh: ABA; 2023 [cited 29 Jul 2026]. Available from: <https://www.theaba.org>
18. European Society of Anaesthesiology and Intensive Care. EDAIC syllabus. Brussels: ESAIC [cited 29 Jul 2026]. Available from: <https://www.esaic.org>
19. National Board of Medical Examiners. NBME item-writing guide: constructing written test questions for the health sciences. 6th ed. Philadelphia: NBME; 2020.
20. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ* 2002;15(3):309-34.
21. Apfelbaum JL, Hagberg CA, Connis RT, et al. 2022 American Society of Anesthesiologists practice guidelines for management of the difficult airway. *Anesthesiology* 2022;136(1):31-81.
22. Joshi GP, Abdelmalak BB, Weigel WA, et al. 2023 American Society of Anesthesiologists practice guidelines for preoperative fasting: carbohydrate-containing clear liquids with or without protein, chewing gum, and pediatric fasting duration. *Anesthesiology* 2023;138(2):132-51.
23. Wigginton JG, Agarwal S, Bartos JA, et al. Part 9: adult advanced life support: 2025 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation* 2025;152(Suppl 2):S538-77.
24. Soar J, Böttiger BW, Carli P, et al. European Resuscitation Council guidelines 2025: adult advanced life support. *Resuscitation* 2025;215(Suppl 1):110769.
25. Kopp SL, Vandermeulen E, McBane RD, Perlas A, Leffert L, Horlocker T. Regional anesthesia in the patient receiving antithrombotic or thrombolytic therapy: American Society of Regional Anesthesia and Pain Medicine evidence-based guidelines (fifth edition). *Reg Anesth Pain Med* 2025. doi:10.1136/rapm-2024-105766.
26. Gropper MA, Eriksson LI, Fleisher LA, Cohen NH, Leslie K, Johnson-Akeju O, editors. *Miller's anesthesia*. 10th ed. Philadelphia: Elsevier; 2025.
27. Butterworth JF, Mackey DC, Wasnick JD. Morgan & Mikhail's clinical anesthesiology. 7th ed. New York: McGraw Hill; 2022.
28. Cullen BF, Stock MC, Ortega R, et al. editors. Barash, Cullen, and Stoelting's clinical anesthesia. 9th ed. Philadelphia: Wolters Kluwer; 2023.
29. Flood P, Rathmell JP, Urman RD, editors. Stoelting's pharmacology and physiology in anesthetic practice. 6th ed. Philadelphia: Wolters Kluwer; 2022.
30. Anderson LW, Krathwohl DR, editors. *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. New York: Longman; 2001.
31. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman and Hall; 1993.
32. Cameron AC, Gelbach JB, Miller DL. Bootstrap-based improvements for inference with clustered errors. *Rev Econ Stat* 2008;90(3):414-27.
33. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 1927;22(158):209-12.
34. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci* 2001;16(2):101-33.
35. Tsui BCH, Kropelin B, Ganapathy S, Finucane B. Dextrose 5% in water: fluid medium maintaining electrical stimulation of peripheral nerve during stimulating catheter placement. *Acta Anaesthesiol Scand* 2005;49(10):1562-5.
36. Tsui BCH, Kropelin B. The electrophysiological effect of dextrose 5% in water on single-shot peripheral nerve stimulation. *Anesth Analg* 2005;100(6):1837-9.
37. Eger EI 2nd. New inhaled anesthetics. *Anesthesiology* 1994;80(4):906-22.