Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-Benchmarked Analysis of Department-Specific ICU Exit Exam Performance: A Cross-Sectional Study

Ruijie Lin
The First Affiliated Hospital of Wenzhou Medical University
Yanche Lu
The Second Affiliated Hospital of Wenzhou Medical University
Qiuyu Zheng
The First Affiliated Hospital of Wenzhou Medical University
Qian Yu
Wenzhou Medical University
Lianmin Ye
ylmwork88@126.com

The First Affiliated Hospital of Wenzhou Medical University

Research Article

Keywords: artificial intelligence, large language model, intensive care unit, residency training, educational measurement

Posted Date: September 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10436751/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

In standardized residency programs, residents from most clinical departments undertake intensive care unit (ICU) rotations, yet post-rotation examination performance may differ by home department. We evaluated DeepSeek-V4-Pro on ICU exit examinations, examined department-based performance differences, and assessed an exploratory AI-benchmarked framework for identifying candidate learning needs.

Methods

We conducted a cross-sectional study of 412 resident physicians who completed one-month ICU rotations. Three ICU exit examination forms contained 300 items across four question types. DeepSeek-V4-Pro was tested on all items via a standardized API protocol. Exploratory person-level one-sample t-tests compared resident accuracy with the pooled AI benchmark. An additive model adjusted department estimates for examination version, and a prespecified dual-reference framework used leave-one-department-out peer means and the AI benchmark to identify candidate department-by-question-type learning needs.

Results

DeepSeek-V4-Pro achieved 85.7% overall accuracy versus a resident mean of 80.8% (95% CI 80.0–81.6%; $t(411) = -11.51$, $p < 0.001$, $d = -0.57$). The model outperformed residents on A1 and A2 questions, whereas residents had a small advantage on case-analysis questions (+ 2.7 percentage points; raw $p = 0.014$; Holm-adjusted $p = 0.028$; $d = 0.12$). The department effect remained significant in the additive model ($F(5, 404) = 6.85$, $p < 0.001$, $\eta^2 = 0.077$). Thirteen cells were below both reference values; eight also met the Bonferroni-adjusted threshold for comparison with the AI benchmark.

Conclusions

DeepSeek-V4-Pro performed better than residents on single-choice formats but not on case-analysis multiple-choice questions. Post-rotation examination performance differed by home department after statistical adjustment, although postgraduate year and examination version were inseparable by design. The exploratory dual-reference framework identifies candidate learning needs that require validation before use for curricular decisions.

Background

The rapid advancement of large language models (LLMs) has generated substantial interest in their potential applications in medical education, clinical practice, and medical research[1–12]. Recent studies have demonstrated that LLMs, including ChatGPT and DeepSeek, can achieve high accuracy on national medical licensing examinations and complex diagnostic challenges[13–20]. However, response patterns can vary across repeated testing, with different models showing different consistency profiles[21]. Although LLMs have been evaluated on national licensing and specialty examinations, their performance on institution-specific ICU rotation exit examinations and their use as a department-stratified reference remain underexplored.

In parallel, residency training programs face a persistent challenge: residents from most clinical departments undertake ICU rotations as part of standardized training, but post-rotation outcomes across diverse trainee backgrounds are poorly understood. Piquette et al. [22]conducted a multicentre assessment of resident competencies before and after ICU rotations but did not stratify residents by home specialty. The Japanese GM-ITE study subsequently found that residents pursuing different specialties differed in clinical proficiency[23]. Whether similar department-associated differences are present on institution-specific ICU exit examinations remains unclear.

The present study addresses these questions in an institution-specific setting. First, we evaluate DeepSeek-V4-Pro on ICU rotation exit examinations. Second, we examine whether resident physicians from different home departments show systematic differences in post-rotation examination performance after statistical adjustment for examination version. Third, we assess an exploratory dual-reference framework that compares department-by-question-type performance with both a peer reference and the model benchmark to identify candidate learning needs.

We selected DeepSeek-V4-Pro (API model identifier: deepseek-v4-pro; accessed June 20–21, 2026) because it was the current API-accessible DeepSeek model at the time of testing[24]. Prior studies of the DeepSeek model family on Chinese medical examinations supported its contextual relevance[13, 15], while its performance on institution-specific specialty examinations and complex multiple-selection items remained uncertain. API access also enabled a standardized item-level testing protocol.

Accordingly, the present study was designed to address the following three research questions:

RQ1

How does DeepSeek-V4-Pro's accuracy on ICU rotation exit examinations compare with that of resident physicians, and how does the AI–human gap vary across different question types?

RQ2

Do resident physicians from different clinical departments show systematic differences in ICU exit examination performance after one-month ICU rotations, after statistical adjustment for examination version?

RQ3

Does a prespecified dual-reference framework (leave-one-department-out peer mean plus AI benchmark) identify candidate department-specific, question-type-level learning needs for subsequent validation?

Methods

Study Design and Setting

This was a cross-sectional comparative study conducted at the First Affiliated Hospital of Wenzhou Medical University, a tertiary academic medical center in Zhejiang, China. The study involved two components: (1) evaluation of DeepSeek-V4-Pro's performance on ICU exit examinations, and (2) analysis of resident physician examination performance stratified by home department. This study is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement for cross-sectional studies[25]; the completed STROBE checklist is provided as Supplementary Material 1.

The study was reviewed and approved by the Ethics Committee in Clinical Research (ECCR) of the First Affiliated Hospital of Wenzhou Medical University (approval reference: KY2026-R288). The committee approved the secondary analysis of de-identified educational assessment data and a waiver of informed consent to participate because the study involved no patient data, clinical intervention, or direct participant contact. All resident data were de-identified before analysis; personally identifiable information was removed and unique numeric identifiers were used for all statistical procedures. All methods were performed in accordance with relevant institutional and national guidelines and regulations and with the Declaration of Helsinki.

Resident Participants

The study initially included all 437 examination records available for resident physicians who completed a one-month ICU rotation between November 2025 and April 2026. Twenty-five repeat records reflected sequential ICU rotations by residents with examinations in more than one postgraduate year (PGY-1 through PGY-3). For the primary analysis, only each resident's first (lowest PGY-level) examination record was retained to ensure independent observations, yielding an analytic census of 412 unique residents. A sensitivity analysis using all 437 records did not materially change the pattern of findings.

Residents were classified into six prespecified department groups: Internal Medicine (n = 156), Surgery (n = 98), Anesthesiology (n = 31), Orthopedics (n = 29), Critical Care (n = 20), and Other Clinical (n = 78). Other Clinical combined the remaining departments (general practice, rehabilitation medicine, infectious diseases, radiation oncology/radiotherapy, plastic surgery, neurology, psychiatry, otorhinolaryngology, and ultrasound medicine) plus one administratively coded record that could not be assigned more specifically. PGY level and examination version were perfectly aligned by design (Exam 1 for PGY-1,

Exam 2 for PGY-2, and Exam 3 for PGY-3). Their distributions differed across departments, so statistical adjustment could not separate examination-version effects from PGY effects.

Examination Materials

Three ICU rotation exit examination forms (Exam 1 for PGY-1, Exam 2 for PGY-2, and Exam 3 for PGY-3 residents) were used. Each form contained 100 items categorized as A1-type basic-knowledge single-choice questions (20–30 per form; 70 pooled), A2-type case-based single-choice questions (20–30 per form; 80 pooled), A3/A4-type case-series single-choice questions (29–41 per form; 101 pooled), and case-analysis multiple-choice questions requiring selection of all correct options (11–19 per form; 49 pooled). Internal consistency (Cronbach's α) ranged from 0.78 to 0.84 across the forms. Item difficulty indices ranged from 0.42 to 0.91, and item discrimination indices ranged from 0.21 to 0.54.

DeepSeek Testing Protocol

DeepSeek-V4-Pro (API model identifier: deepseek-v4-pro; accessed via https://api.deepseek.com; temperature: 1.0) was tested on all 300 items between June 20 and June 21, 2026. A standardized testing prompt instructed the model to output only answer letters without explanatory text. Each independent API session ensured zero context carryover between items. Model responses were programmatically parsed, cross-checked by a second researcher, and accuracy was computed by direct string comparison against the answer key. There were no refused answers (0/300 items).

Notably, each item was tested in a single pass. Recent evidence demonstrates that LLM response patterns exhibit substantial variability across repeated testing — Buhl et al. [21]found that different LLMs showed markedly different consistency profiles, with some models changing answers on up to 48% of items across 100 repeated trials. The single-pass design is a limitation discussed below.

Use of Generative AI

DeepSeek-V4-Pro was a research subject whose examination performance was evaluated. ChatGPT was used for language editing and data processing. The authors reviewed and verified all outputs and take responsibility for the manuscript.

Statistical Analysis

The AI–human analysis was prespecified as an exploratory pooled benchmark comparison. For each question type, the mean accuracy across 412 residents was compared with the model's pooled accuracy across the three forms using person-level one-sample t-tests, with Cohen's d as the effect size. Holm adjustment was applied across the four question-type comparisons; the overall comparison was reported separately. Because residents completed different forms according to PGY and item-level

resident responses were not available in the analytic dataset, the AI–human comparison was not form-matched and should be interpreted descriptively rather than as a controlled equivalence test.

For department-based comparisons, a two-way model initially included department, examination version, and their interaction. Because the interaction was non-significant and several department-by-version cells were sparse, the primary adjusted analysis used an additive ordinary least-squares model with department and examination version as fixed effects. Department-specific marginal means were standardized to the observed overall examination-version distribution and are reported with 95% confidence intervals. As a sensitivity analysis, accuracy was standardized within examination version and compared across departments using one-way ANOVA with Tukey's honestly significant difference procedure. Effect sizes were reported as eta-squared (η^2). ANOVA was retained because it is generally robust to moderate departures from normality in this setting[26]. PGY and examination version could not be disentangled because they were perfectly aligned.

For the dual-reference analysis, the peer reference for each department was the leave-one-department-out resident mean for the same question type. A cell was designated a candidate dual-reference learning need when the department mean was below both this peer reference and the pooled AI benchmark. One-sample t-tests then compared each department mean with the AI benchmark for all 24 department-by-question-type cells; Bonferroni correction set $\alpha = 0.05/24 = 0.0021$. Deficit severity was defined using the signed department-minus-AI difference: Critical (≤ -10 percentage points), High (> -10 to ≤ -5), and Moderate (> -5 to < 0). All analyses used Python 3.13 with two-sided $\alpha = 0.05$. The dataset contained no missing values.

Potential Sources of Bias

Selection bias may arise from the single-center setting. Test-administration bias was reduced through standardized procedures. Residual confounding remains because PGY was deterministically linked to examination version and several department-by-version cells were sparse. The pooled AI benchmark was not matched to each resident's examination form, and the administrative department label retained in Other Clinical could not be resolved further from the source dataset.

Results

Participant Characteristics

A total of 412 resident physicians were included. The Internal Medicine group was the largest (n = 156, 37.9%), followed by Surgery (n = 98, 23.8%) and Other Clinical (n = 78, 18.9%). The exam version distribution varied across departments (Table 1).

DeepSeek-V4-Pro Versus Resident Performance

DeepSeek-V4-Pro achieved 85.7% overall accuracy (257/300): 66/70 A1 items, 75/80 A2 items, 89/101 A3/A4 items, and 27/49 case-analysis items. The resident mean was 80.8% (95% CI 80.0–81.6%). The exploratory person-level comparison gave $t(411) = -11.51$, $p < 0.001$, $d = -0.57$. The model scored higher on A1 ($\Delta = -11.9$ percentage points, $d = -0.98$, $p < 0.001$) and A2 questions ($\Delta = -6.4$, $d = -0.72$, $p < 0.001$). Differences were small for A3/A4 ($\Delta = -1.0$, $d = -0.12$; raw $p = 0.015$, Holm-adjusted $p = 0.028$) and case analysis, on which residents had a 2.7-point advantage ($d = 0.12$; raw $p = 0.014$, Holm-adjusted $p = 0.028$) (Figure 1). These pooled comparisons were not matched by examination form.

Department-Based Performance Differences

The initial two-way model showed a department effect ($F(5, 394) = 6.81$, $p < 0.001$), a non-significant examination-version term ($F(2, 394) = 2.16$, $p = 0.117$), and no detected interaction ($F(10, 394) = 0.79$, $p = 0.638$). In the additive model, the department effect remained significant ($F(5, 404) = 6.85$, $p < 0.001$, $\eta^2 = 0.077$; model $R^2 = 0.106$). Examination-version-standardized marginal means (95% CI) were 82.7% (81.4–84.1) for Internal Medicine, 77.3% (75.6–78.9) for Surgery, 77.4% (74.5–80.3) for Anesthesiology, 82.6% (79.6–85.6) for Orthopedics, 79.6% (75.6–83.6) for Critical Care, and 82.4% (80.5–84.2) for Other Clinical. The within-version z-score sensitivity analysis also showed department differences ($F(5, 406) = 6.17$, $p < 0.001$). Because PGY and examination version were perfectly aligned, neither model isolates an independent examination-version or PGY effect.

Tukey's procedure on within-version z-scores indicated that Surgery scored lower than Internal Medicine ($p < 0.001$) and Other Clinical ($p = 0.002$); Anesthesiology scored lower than Internal Medicine ($p = 0.017$) and Other Clinical ($p = 0.032$). No other pairwise comparisons were significant (Table 1, Figure 2).

AI-Benchmarked Learning-Needs Analysis

Thirteen of the 24 department-by-question-type cells were below both the leave-one-department-out peer reference and the pooled AI benchmark; eight of these also met the Bonferroni-adjusted threshold for comparison with the AI benchmark (Supplementary Table S1). The largest supported deficits were Surgery × A1 (department − AI = −17.2 percentage points), Anesthesiology × A1 (−15.6), Other Clinical × A1 (−12.4), Orthopedics × A1 (−12.2), and Critical Care × A2 (−10.3); all had $p < 0.001$. Five additional candidate cells did not meet the adjusted threshold and are presented as exploratory findings (Figure 3).

Discussion

Principal Findings

This study evaluated DeepSeek-V4-Pro on ICU exit examinations and examined department-based differences using AI as a cognitive benchmark. Three findings emerged.

First (RQ1), DeepSeek-V4-Pro exceeded the pooled resident mean by 4.9 percentage points (d = − 0.57), consistent with evidence that LLMs can perform well on medical knowledge examinations[13–16]. The difference was largest for A1 and A2 items, while residents had a small 2.7-point advantage on case-analysis items (d = 0.12; Holm-adjusted p = 0.028). These written items assess the ‘knows’ and ‘knows how’ levels of Miller's framework, not observed clinical performance[27]. Moreover, the comparison pooled AI performance across forms and was not form-matched, so it should be interpreted as a descriptive benchmark rather than a controlled human–AI comparison.

Second (RQ2), department-associated differences remained after statistical adjustment for examination version ($F(5, 404) = 6.85$, $p < 0.001$, $\eta^2 = 0.077$). This finding is compatible with department-related differences in post-rotation examination performance and extends earlier work on specialty-associated variation[22, 23]. It does not establish a training effect: PGY and examination version were inseparable by design, department-by-version cells were unbalanced, and the cross-sectional post-only design cannot distinguish baseline differences from rotation-related learning.

Third (RQ3), 13 cells met the prespecified descriptive rule of falling below both the leave-one-department-out peer reference and the pooled AI benchmark; eight also met the Bonferroni threshold for comparison with the AI benchmark. These are candidate learning-needs signals rather than validated curricular priorities. The remaining five cells are exploratory and require confirmation.

LLM Response Variability

Buhl et al. [21]demonstrated that LLMs show markedly different consistency profiles across repeated testing — some changed answers on up to 48% of items. Our single-pass design means benchmark values may not represent stable performance. However, DeepSeek-V4-Pro's consistency across three examinations (85.0%, 86.0%, 86.0%) and the marked question-type asymmetry suggest the overall pattern is robust. Future studies should conduct 10–30 repeated tests to establish confidence intervals[21, 28].

Implications for Training

If validated in independent pre-post cohorts, the framework could support specialty-sensitive needs assessment before ICU rotations. The large Surgery × A1 deficit suggests foundational critical care topics for further assessment, while the Critical Care × A2 deficit suggests examining case-based reasoning content and item functioning. These signals should guide confirmatory evaluation rather than immediate curricular allocation.

Limitations

Key limitations include: (1) a single-center setting; (2) a cross-sectional post-only design that precludes causal inference about rotation effects; (3) perfect alignment of PGY with examination version and sparse, unbalanced department-by-version cells, preventing separation of their effects; (4) an exploratory pooled AI benchmark that was not matched by examination form; (5) single-pass LLM testing without repeated trials; (6) psychometric evidence from one institution and no independent demonstration of

form equivalence; (7) cognitive assessment limited to Miller's ‘knows/knows how’ levels; (8) one unresolved administrative department label retained in Other Clinical; and (9) a dual-reference framework that lacks external validation, with five of 13 candidate cells not meeting the multiple-comparison threshold.

Conclusions

DeepSeek-V4-Pro scored higher than the pooled resident mean on ICU exit examinations overall, whereas residents had a small advantage on case-analysis items. Department-associated post-rotation performance differences remained after statistical adjustment, but the design cannot separate PGY from examination version or attribute differences to the ICU rotation. Eight dual-reference candidate cells met the Bonferroni threshold and may inform confirmatory needs assessment. Multicentre pre-post studies with form-matched comparisons and external validation are needed before curricular use.

Abbreviations

AI, artificial intelligence; ICU, intensive care unit; LLM, large language model; PGY, postgraduate year; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate

This study was reviewed and approved by the Ethics Committee in Clinical Research (ECCR) of the First Affiliated Hospital of Wenzhou Medical University (approval reference: KY2026-R288). The requirement for informed consent to participate was waived by the committee because the study involved a secondary analysis of de-identified educational assessment data, with no patient data, clinical intervention, or direct participant contact. All methods were performed in accordance with relevant institutional and national guidelines and regulations and with the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

De-identified resident performance data and model response data are available from the corresponding author upon reasonable request. Examination item content is not publicly available to protect assessment integrity.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study. No funder had a role in study design, data analysis, interpretation, or manuscript preparation.

Authors' contributions

RL contributed to conceptualization, data curation, investigation, methodology, and drafting the manuscript. YL contributed to data curation, formal analysis, investigation, and validation. QZ contributed to data curation, investigation, and resources. QY contributed to formal analysis, visualization, and manuscript review and editing. LY supervised the study, administered the project, and reviewed and edited the manuscript. All authors critically revised and approved the final manuscript.

Acknowledgements

None.

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. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47(1):33.
3. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg. 2024;110(6):3701-6.
4. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ. 2023;9:e46885.
5. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589-96.
6. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in Medical Research: Current Status and Future Directions. J Multidiscip Healthc. 2023;16:1513-20.
7. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6).
8. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 2024;17(5):926-31.
9. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ. 2023;9:e48291.

10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
11. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-8.
12. Agarwal P, Agarwal R, Iezhitsa I. AI for Assessment in Medical Education in Post LLM Era: A Scoping Review. Med Sci Educ. 2026;36(2):1027-39.
13. Wang W, Zhou Y, Fu J, Hu K. Evaluating the Performance of DeepSeek-R1 and DeepSeek-V3 Versus OpenAI Models in the Chinese National Medical Licensing Examination: Cross-Sectional Comparative Study. JMIR Med Educ. 2025;11:e73469.
14. Hong DR, Huang CY. The performance of AI in medical examinations: an exploration of ChatGPT in ultrasound medical education. Front Med (Lausanne). 2024;11:1472006.
15. Wang X, Long Z, Zhu B, Cao Y, Tang H, He K, et al. Evaluation of DeepSeek-R1 and ChatGPT-4o on the Chinese national medical licensing examination: a multi-year comparative study. Sci Rep. 2026;16(1):2237.
16. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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. 2023;9:e45312.
17. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
18. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
19. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems2023 March 01, 2023:[arXiv:2303.13375 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2023arXiv230313375N.
20. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. Jama. 2023;330(1):78-80.
21. Buhl LK. The answer may vary: large language model response patterns challenge their use in test item analysis. Med Teach. 2025;47(11):1761-6.
22. Piquette D, Goffi A, Lee C, Brydges R, Walsh CM, Mema B, et al. Resident competencies before and after short intensive care unit rotations: a multicentre pilot observational study. Can J Anaesth. 2021;68(2):235-44.
23. Watari T, Nishizaki Y, Houchens N, Kataoka K, Sakaguchi K, Shiraishi Y, et al. Medical resident's pursuing specialty and differences in clinical proficiency among medical residents in Japan: a nationwide cross-sectional study. BMC Med Educ. 2023;23(1):464.
24. DeepSeek-AI, Xu A, Lin B, Xue B, Wang B, Xu B, et al. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence2026 April 01, 2026:[arXiv:2606.19348 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2026arXiv260619348D.

25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453-7.
26. Blanca MJ, Alarcón R, Arnau J, Bono R, Bendayan R. Non-normal data: Is ANOVA still a valid option? Psicothema. 2017;29(4):552-7.
27. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63-7.
28. Zhang H, Qu L, Bai H, Chen Y, Ji R, Cheng Z, et al. Beyond accuracy: evaluating the reliability of large language models for medical assessment. Front Artif Intell. 2026;9:1832829.

Tables

Table 1. Resident characteristics and department-specific examination performance.

Department	n	PGY-1 n (%)	PGY-2 n (%)	PGY-3 n (%)	A1 (%)	A2 (%)	A3/A4 (%)	Case (%)	Overall (%)
IM	156	1 (0.6)	54 (34.6)	101 (64.7)	86.4	89.0	88.5	61.3	83.0
Surg	98	2 (2.0)	77 (78.6)	19 (19.4)	77.1	84.6	84.8	51.2	76.9
Anesth	31	1 (3.2)	20 (64.5)	10 (32.3)	78.7	87.2	84.6	48.0	77.2
Ortho	29	2 (6.9)	21 (72.4)	6 (20.7)	82.1	88.4	87.7	62.7	82.3
CC	20	11 (55.0)	6 (30.0)	3 (15.0)	85.5	83.4	86.4	51.7	80.6
Other	78	10 (12.8)	49 (62.8)	19 (24.4)	81.8	88.0	88.4	62.7	82.3
Overall	412	27 (6.6)	227 (55.1)	158 (38.3)	82.4	87.3	87.1	57.8	80.8
AI benchmark	—	—	—	—	94.3	93.8	88.1	55.1	85.7

IM = Internal Medicine; Surg = Surgery; Anesth = Anesthesiology; Ortho = Orthopedics; CC = Critical Care. Department rows show unadjusted means. Other Clinical contains the remaining clinical departments and one unresolved administrative category; adjusted overall means are shown in Figure 2.

Table 2. DeepSeek-V4-Pro versus resident performance by question type (person-level analysis).

Question Type	AI Acc.	Resident (95% CI)	Δ (pp)	t(411)	p	d
A1 (Basic)	94.3%	82.4 (81.2–83.6)	−11.9	−19.79	<0.001	−0.98
A2 (Case-based)	93.8%	87.3 (86.4–88.2)	−6.4	−14.61	<0.001	−0.72
A3/A4 (Series)	88.1%	87.1 (86.4–87.9)	−1.0	−2.44	0.015 (0.028)	−0.12
Case Analysis	55.1%	57.8 (55.7–60.0)	+2.7	2.47	0.014 (0.028)	0.12
Overall	85.7%	80.8 (80.0–81.6)	−4.9	−11.51	<0.001	−0.57

Exploratory person-level one-sample t-tests. Cohen's d is reported. For A1, A2, A3/A4, and case analysis, cells show raw p values followed by Holm-adjusted p values in parentheses where raw p ≥ 0.001; the overall comparison was prespecified and reported separately. Comparisons were not matched by examination form.

Figures

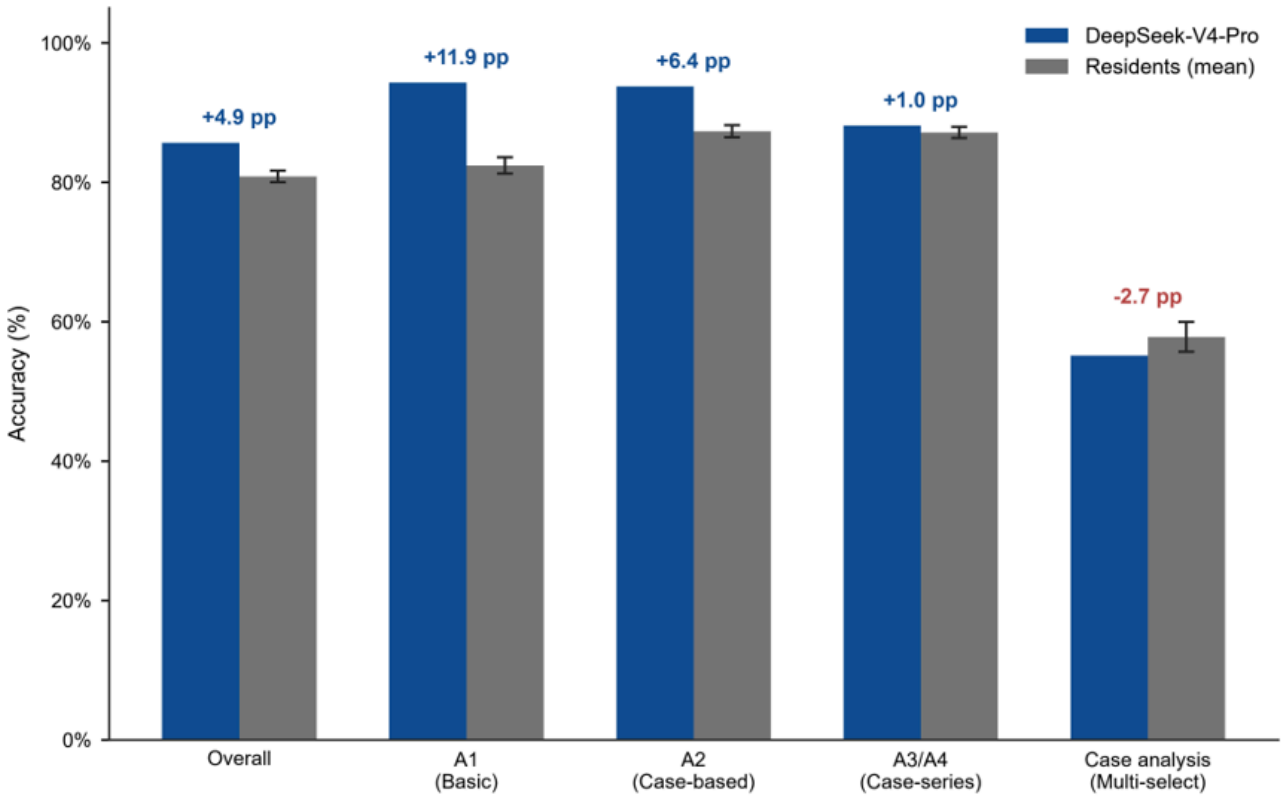


Figure 1

DeepSeek-V4-Pro versus resident physician performance by question type. Blue bars show single-pass model accuracy; grey bars show mean accuracy among 412 residents, with 95% confidence intervals. Numeric labels are AI minus resident and were calculated from the exact pooled AI proportions

(257/300 overall; 66/70 A1; 75/80 A2; 89/101 A3/A4; and 27/49 case-analysis items). Table 2 reports resident minus AI. Comparisons were exploratory and were not matched by examination form.

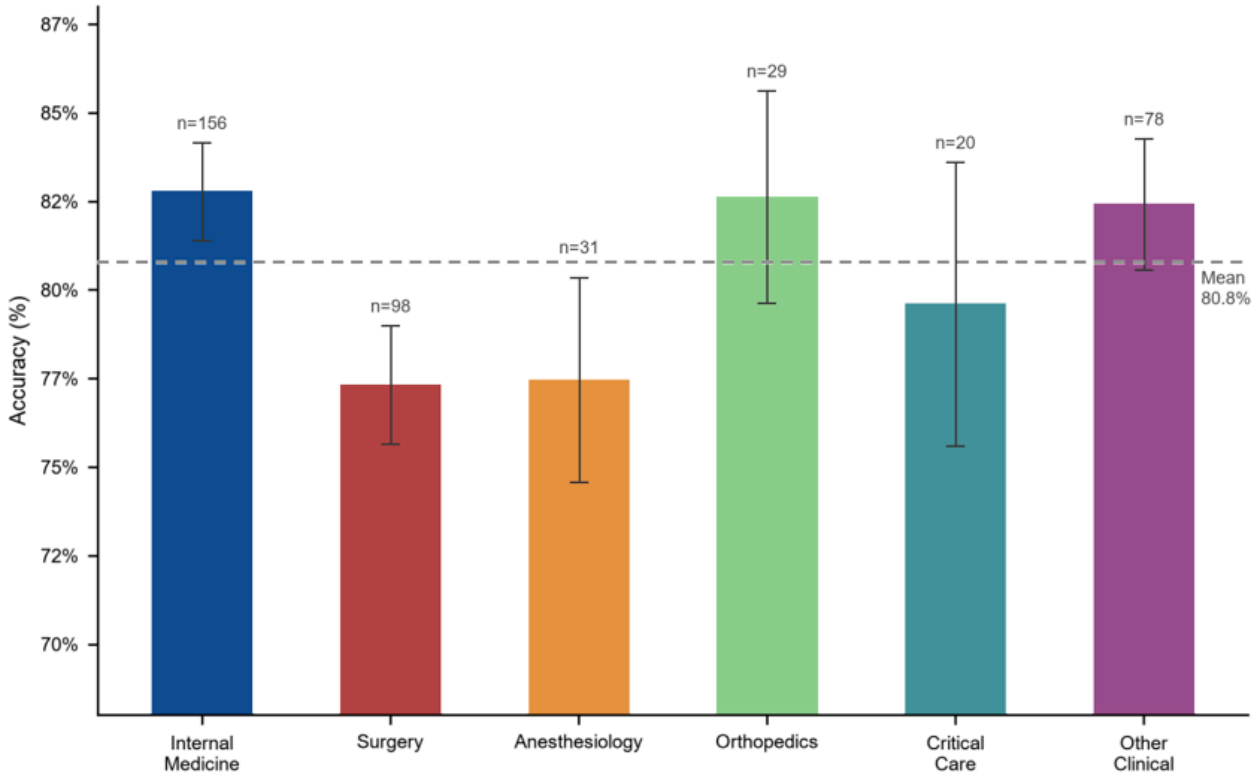


Figure 2

Department-specific examination performance adjusted for examination version. Bars show model-standardized marginal mean accuracy at the observed overall examination-version distribution; error bars show 95% confidence intervals. The dashed line indicates the overall resident mean (80.8%). Values above bars are the verified group sample sizes.

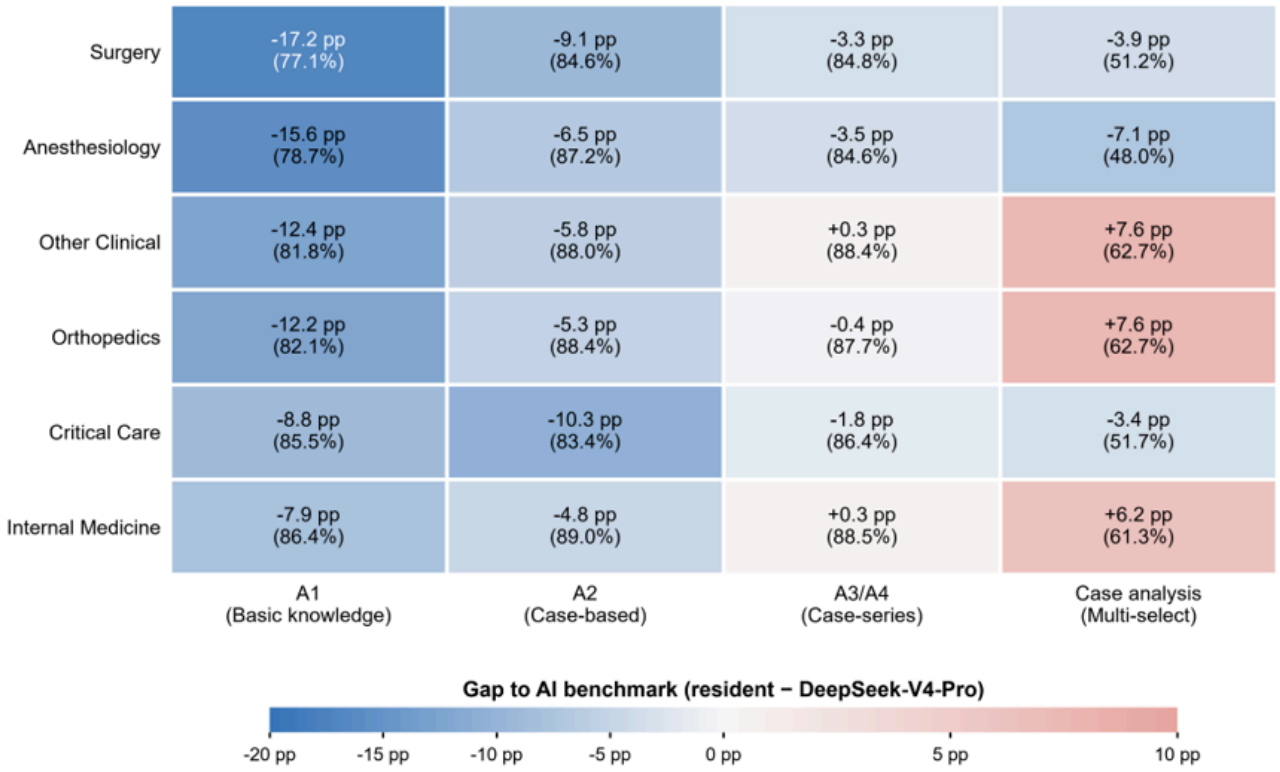


Figure 3

AI-benchmarked performance-gap heatmap (six departments × four question types). Cell values are department mean minus the exact pooled DeepSeek-V4-Pro benchmark, in percentage points; absolute department accuracy is shown in parentheses. Department sample sizes in displayed order are 98, 31, 78, 29, 20, and 156. Blue indicates performance below the AI benchmark and red indicates performance above it. Statistical confirmation is reported separately in Supplementary Table S1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- STROBEChecklistCrossSectionalBMC.docx
- SupplementaryMaterial2SensitivityAnalyses.docx
- SupplementaryMaterial3CompleteDualReferenceAnalysis.docx
- KY2026R288.pdf
- Supplementarymaterials.docx