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.

Large Language Models in Israeli Medical Administration Board Exam: Performance, Calibration, and Implications for Health-System Governance

Alon Moore Galindo

alonmor289@gmail.com

Tel Aviv Sourasky Medical Center

Anat Loewenstein

Tel Aviv Sourasky Medical Center

David Schwartzman

Saint Joseph's University

Yuval Weiss

Israel Medical Association

Ofer Almog

Tel Aviv Sourasky Medical Center

Gil Fire

Tel Aviv Sourasky Medical Center

Research Article

Keywords: Large language models, artificial intelligence, medical administration, medical directorship, board examinations, calibration, ChatGPT, Gemini

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11250421/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:

Large language models (LLMs) are increasingly used in healthcare, but their reliability in medical administration remains poorly understood, particularly in areas involving local regulation, policy, and health-system governance.

Methods:

We evaluated six LLMs using 596 multiple-choice questions from Israeli Medical Administration Board examinations conducted between 2020 and 2025. Official resident benchmark data were available for 296 matched items. Outcomes included overall accuracy, comparison with resident performance, stratified performance by administrative domain, question modality, and examination year, and confidence-based discrimination and calibration metrics.

Results:

Overall accuracy varied across models, with Gemini-Pro-3 achieving the highest performance (77.9%; 95% CI, 74.3–81.1). In the matched subset, its accuracy was 78.0%, compared with a mean resident accuracy of 74.0%. More importantly, performance was significantly lower in domains requiring locally specific knowledge, including Israeli Ministry of Health Directives, Reports and Committees, and Legislation and Regulations, compared with Statistics and Epidemiology. Confidence reliability also varied across models: Gemini-Flash-3 showed the highest confidence-based discrimination, ChatGPT-5.4 had the lowest expected calibration error, and mean self-reported confidence exceeded observed accuracy in all models.

Conclusions

Contemporary LLMs showed variable performance across different areas of medical administration. Performance was weaker in locally specific regulatory and governance domains, and confidence did not consistently identify reliable responses. These findings support local validation, authoritative source verification, and human oversight when LLMs are used for health-system tasks.

Introduction

Artificial intelligence is becoming part of routine medical practice, with the American Medical Association reporting growing use of AI among physicians [1–7]. Large language models (LLMs) have also shown strong performance on medical licensing and specialty examinations, although results vary substantially across models, specialties, task types, and subspecialty domains [8–17].

Furthermore, the rapid expansion of LLM capabilities and the continuous release of new models have been accompanied by persistent concerns regarding accuracy, reliability, calibration, and safety [18–20]. Although prior benchmarks have primarily evaluated clinical knowledge and specialty-specific medical examinations, comparative evidence in medical administration remains significantly limited[21, 22]. This gap is particularly

important because medical administration requires the integration of clinical judgment with health policy, epidemiology, ethics, medico-legal risk, regulation, quality assurance, and organizational governance. A central unanswered question is whether contemporary LLMs can reliably navigate the locally specific regulations, governance structures, and health-system policies on which many administrative decisions depend. Medical Administration is a recognized physician subspecialty in Israel that prepares physicians for senior leadership roles across the health system. Its national board examinations assess the core knowledge required for these roles and therefore provide a rigorous, standardized benchmark for evaluating LLM performance in administrative medicine.

We therefore evaluated six contemporary LLMs using official Israeli Medical Administration Board examination questions. We examined overall and domain-specific accuracy, performance relative to residents, and confidence-based reliability. Particular attention was given to locally specific regulatory and governance domains. Together, these analyses were designed to clarify the potential utility of LLMs in medical administration and the limitations that currently constrain their use.

Methods

Study design and dataset

This comparative study evaluated the performance of six contemporary LLMs using multiple-choice questions from Israeli medical administration board examinations administered between 2020 and 2025. These official exams are designed to assess the essential knowledge required for a healthcare administrator. Medical administration is a formally recognized medical subspecialty in Israel[23]. It is intended for physicians who have already completed at least one clinical specialty and who seek to qualify for senior leadership and management positions within the Israeli healthcare system. The requirements for certification in this subspecialty include completion of a master's degree in Medical Administration, Public Health, or Business Administration (MHA, MPH, or MBA); two supervised rotations with a combined duration of two years in two of the following settings: the Israeli Ministry of Health, hospital management, or health-fund management; and successful completion of two board examinations, comprising a written examination and an oral examination. Before analysis, to ensure diagnostic clarity, Items with more than two correct answer options or questions with linked multi-part structure were excluded. Consequently, the analysis was performed on 596 unique questions.

Official resident (n = 51, 2023–2025) benchmark data were available for 296 matched items. No resident-identifiable information was available to the investigators, and resident examination scores or item-level responses could not be linked to individual examinees. Resident performance was analyzed as mean per-question accuracy on this subset to enable direct comparison with the same question set answered by each model. In addition to administrative domains, questions were stratified by modality (image-based vs. text-only) and annotated according to exam year (Fig. 1).

Two senior healthcare administrators classified each question into one of six predefined medical administration domains according to its primary content area: Legislation and Regulations (n = 170), Statistics and Epidemiology (n = 127), Israeli Ministry of Health Directives (n = 93), Health Economics (n = 87), International Health Systems (n = 83), and Reports and Committees (n = 36).

Data collection and model evaluation

Board official exams published by the Scientific Council of the Israeli Medical Association and official answer keys from 2020 to 2025 were imported into a custom application that segmented question stems, answer options, embedded images, and answer keys into a structured dataset. The extracted dataset was manually reviewed for fidelity against the source documents. Each model received the same standardized prompt and was instructed to provide a single final answer choice and a self-reported confidence score from 0 to 100 using an API (application programming interface). The evaluated systems were Gemini-Pro-3 (Google, USA), Gemini-Flash-3 (Google, USA), Claude Opus 4.6 (Anthropic, USA), ChatGPT-5.4 (OpenAI, USA), Grok-4-1-fast-reasoning (xAI, USA), and Perplexity-Sonar-Deep Research (Perplexity AI, USA), with a temperature 0. No model was manually provided with external documents, regulatory sources, or answer keys. Perplexity-Sonar-Deep Research was evaluated with its native retrieval functionality, whereas no external retrieval source was added to the other models.

Statistical analysis

Overall LLM accuracy was calculated across the full 596-question database. Human–AI comparisons were restricted to the matched 2023–2025 subset (n = 296), for which official resident performance data were available. For this matched item set, accuracies with 95% confidence intervals (CIs) were calculated for residents and each model. A total of 51 residents completed the board examination during the relevant study period. Mean resident accuracy was defined as the mean proportion of examinees answering each question correctly. Resident 95% CIs were calculated from the item-level standard error of the mean, whereas LLM 95% CIs were calculated using the Wilson score method. All the models answered all the questions.

To account for the non-independence of observations, Generalized Estimating Equations (GEE) logistic regression was used with an exchangeable working correlation structure and robust standard errors, clustered by question ID. A multivariable GEE logistic model was fitted to the same 3,576 model–question responses, clustered within the 596 question IDs, to evaluate the simultaneous impact of model, modality, topic, and examination year on the odds of a correct response. Gemini-Pro-3, Statistics and Epidemiology, Text-only, and examination year 2020 were used as the respective reference categories. To avoid Type I error across multiple comparisons, p-values were adjusted using the Holm-Bonferroni method. McNemar's test and Cochran's Q test were used for pairwise and multi-model matched comparisons, respectively.

Model confidence was evaluated in terms of discrimination, calibration, and confidence-threshold performance. The 0–100 values represented model-generated, self-reported confidence scores and were not assumed to be formally calibrated probabilities. Scores were rescaled to [0,1] for analysis. Discrimination between correct and incorrect responses was quantified separately for each model using the area under the receiver operating characteristic curve (AUC), with 95% confidence intervals calculated using DeLong's method. Calibration was assessed using the expected calibration error (ECE) across 10 equal-width confidence bins and the Brier score, calculated at the individual model–question response level. Two-sided 95% confidence intervals for ECE and Brier scores were obtained from 2,000 nonparametric bootstrap resamples of the 596 question-level confidence–outcome pairs within each model. Selective performance was additionally evaluated at self-reported confidence thresholds of ≥ 70%, ≥ 80%, and ≥ 90%; for each threshold, accuracy and coverage were reported.

Accuracy confidence intervals were calculated using the exact Clopper–Pearson method. Analyses were performed using R version 4.5.2 (R Foundation for Statistical Computing, Vienna, Austria).

A post hoc alternative-prompt analysis was performed to assess the robustness of model performance across task framing and to determine whether the primary findings were dependent on the initial prompt structure (Supplementary Table 1).

Comparison with residents

In the matched resident-comparison subset of 296 items, mean resident accuracy was 74.0% (95% CI, 71.2–76.7). Gemini-Pro-3 achieved the highest model performance, with an accuracy of 78.0% (95% CI, 72.9–82.6) and was the only LLM with a point estimate above the resident benchmark. The remaining models performed below residents, with ChatGPT-5.4, Perplexity-Sonar-Deep Research, and Claude Opus 4.6 forming an intermediate performance group, followed by Grok-4-1-fast-reasoning and Gemini-Flash-3. (Fig. 2)

Overall model performance

Overall accuracy differed substantially across the six evaluated systems. Gemini-Pro-3 was the leading model, correctly answering 464 of 596 questions (77.9%; 95% CI, 74.3–81.1), followed by ChatGPT-5.4, Claude Opus 4.6, Grok-4-1-fast-reasoning, Perplexity-Sonar-Deep Research, and lastly Gemini-Flash-3, showed lower and closely clustered performance (Fig. 3). Global matched comparison confirmed significant heterogeneity across models (Cochran's $Q = 96.51$, $df = 5$; $P < 0.001$). In pairwise McNemar testing with Holm–Bonferroni correction, Gemini-Pro-3 significantly outperformed every other model, whereas no statistically significant differences remained among the other five systems after correction. Consistently, in multivariable GEE analysis adjusted for model, topic, modality, and examination year, all non-reference models had significantly lower odds of a correct response compared with Gemini-Pro-3, with adjusted odds ratios ranging from 0.42 to 0.57 (all Holm-adjusted $P < 0.001$) (Table 1,2).

Table 1
Adjusted odds of correct LLM responses in multivariable GEE analysis

Variable	Level	Correct/Total (%)	Adjusted OR (95% CI)	Holm-adjusted P
Model	ChatGPT-5.4	400/596 (67.1%)	0.57 (0.46–0.69)	< 0.001
	Claude Opus 4.6	391/596 (65.6%)	0.53 (0.43–0.65)	< 0.001
	Gemini-Flash-3	362/596 (60.7%)	0.42 (0.34–0.53)	< 0.001
	Gemini-Pro-3	464/596 (77.9%)	1.00 (Reference)	
	Grok-4-1-fast-reasoning	368/596 (61.7%)	0.44 (0.36–0.54)	< 0.001
	Perplexity-Sonar-Deep Research	365/596 (61.2%)	0.43 (0.35–0.54)	< 0.001
Topic	Israeli Ministry of Health Directives	285/558 (51.1%)	0.33 (0.22–0.51)	< 0.001
	Reports and Committee	113/216 (52.3%)	0.33 (0.19–0.59)	0.002
	Health Economics	370/522 (70.9%)	0.81 (0.52–1.27)	1.000
	International Health Systems	368/498 (73.9%)	0.94 (0.60–1.50)	1.000
	Legislation and Regulations	642/1020 (62.9%)	0.54 (0.36–0.79)	0.013
	Statistics and Epidemiology	572/762 (75.1%)	1.00 (Reference)	
Question format	Image-based	153/216 (70.8%)	0.84 (0.50–1.43)	1.000
	Text-only	2197/3360 (65.4%)	1.00 (Reference)	
Exam year	2020	395/600 (65.8%)	1.00 (Reference)	
	2021	373/600 (62.2%)	0.85 (0.57–1.25)	1.000
	2022	370/600 (61.7%)	0.85 (0.57–1.28)	1.000
	2023	373/588 (63.4%)	0.90 (0.59–1.38)	1.000
	2024	378/594 (63.6%)	0.96 (0.62–1.48)	1.000
	2025	461/594 (77.6%)	2.00 (1.32–3.04)	0.010

Legend: Binomial GEE model with logit link, exchangeable working correlation, robust standard errors, and clustering by question ID. The model included LLM, administrative domain, question format, and examination year. Reference categories were Gemini-Pro-3, Statistics and Epidemiology, text-only questions, and examination year 2020. P values were adjusted using the Holm–Bonferroni method.

Table 2
Pairwise comparisons of LLM accuracy

Model A	Model B	McNemar chi-square	Holm-adjusted P
Gemini-Pro-3	Gemini-Flash-3	56.05	< 0.001
Gemini-Pro-3	Claude Opus 4.6	33.02	< 0.001
Gemini-Pro-3	ChatGPT-5.4	30.07	< 0.001
Gemini-Pro-3	Grok-4-1-fast-reasoning	55.03	< 0.001
Gemini-Pro-3	Perplexity-Sonar-Deep Research	52.48	< 0.001
Gemini-Flash-3	Claude Opus 4.6	5.41	0.140
Gemini-Flash-3	ChatGPT-5.4	7.78	0.053
Gemini-Flash-3	Grok-4-1-fast-reasoning	0.17	1.000
Gemini-Flash-3	Perplexity-Sonar-Deep Research	0.02	1.000
Claude Opus 4.6	ChatGPT-5.4	0.44	1.000
Claude Opus 4.6	Grok-4-1-fast-reasoning	3.81	0.305
Claude Opus 4.6	Perplexity-Sonar-Deep Research	3.81	0.305
ChatGPT-5.4	Grok-4-1-fast-reasoning	6.96	0.075
ChatGPT-5.4	Perplexity-Sonar-Deep Research	6.92	0.075
Grok-4-1-fast-reasoning	Perplexity-Sonar-Deep Research	0.03	1.000

Legend: Pairwise matched comparisons were performed across all 596 examination items using continuity-corrected McNemar tests. P values were adjusted for multiple comparisons using the Holm–Bonferroni method.

Topic-specific performance patterns

Performance was significantly lower in domains requiring locally specific regulatory and governance knowledge (Fig. 4). Compared with Statistics and Epidemiology, which had the highest aggregate accuracy (572/762; 75.1%), the adjusted odds of a correct response were substantially lower for Israeli Ministry of Health Directives (285/558; 51.1%; adjusted OR, 0.33; 95% CI, 0.22–0.51; Holm-adjusted P < 0.001), Reports and Committees (113/216; 52.3%; adjusted OR, 0.33; 95% CI, 0.19–0.59; Holm-adjusted P = 0.002), and Legislation and Regulations (642/1020; 62.9%; adjusted OR, 0.54; 95% CI, 0.36–0.79; Holm-adjusted P = 0.013), after adjustment for model, modality, and examination year. Health Economics and International Health Systems did not differ significantly from Statistics and Epidemiology.

Descriptively, models varied on internationally oriented domains. In Statistics and Epidemiology, Gemini-Pro-3 achieved the highest accuracy (115/127; 90.6%; 95% CI, 84.1–95.0), followed by Grok-4-1-fast-reasoning

(102/127; 80.3%; 95% CI, 72.3–86.8). In International Health Systems, ChatGPT-5.4 achieved the highest accuracy (72/83; 86.7%; 95% CI, 77.5–93.2), followed by Gemini-Pro-3 (69/83; 83.1%; 95% CI, 73.3–90.5). Thus, the leading model differed by examination domain, indicating domain-level variation in model ranking.

Modality and Exam Year

As shown in Table 1, image-based questions were not significantly associated with higher adjusted odds of a correct response (adjusted OR, 0.84; 95% CI, 0.50–1.43; adjusted P = 1.000), using text-only questions as the reference category. Descriptively, modality effects varied by model: Gemini-Pro-3, Gemini-Flash-3, and Grok-4-1-fast-reasoning performed numerically better on image-based questions, whereas ChatGPT-5.4 and Perplexity-Sonar-Deep Research performed numerically better on text-only questions (supplementary 1). Given the small number of image-based items, these findings should be interpreted cautiously (36 image-based questions).

The examination year showed limited heterogeneity overall. Compared with 2020, only 2025 was associated with significantly higher adjusted odds of a correct response (adjusted OR, 2.00; 95% CI, 1.32–3.04; adjusted P = 0.010). The remaining examination years did not differ significantly from 2020 after adjustment. (Fig. 5)

Confidence-based discrimination and calibration

Confidence-based reliability differed meaningfully across models (Table 3). Gemini-Flash-3 had the highest AUC for confidence-based discrimination (0.720; 95% CI, 0.680–0.760), followed closely by ChatGPT-5.4 (0.717; 95% CI, 0.676–0.757), Claude Opus 4.6, Grok-4-1-fast-reasoning, Gemini-Pro-3, and lastly Perplexity-Sonar-Deep Research (0.648; 95% CI, 0.603–0.693).

Table 3
Confidence discrimination and calibration of LLM responses

Model	N	Mean confidence (%)	Accuracy (%)	AUC (95% CI)	ECE (95% CI)	Brier (95% CI)
Gemini-Pro-3	596	93.7	77.9	0.683 (0.640–0.726)	0.161 (0.127–0.195)	0.191 (0.162–0.219)
Gemini-Flash-3	596	64.8	60.7	0.720 (0.680–0.760)	0.273 (0.239–0.309)	0.287 (0.254–0.319)
Claude Opus 4.6	596	76.4	65.6	0.706 (0.667–0.745)	0.124 (0.094–0.161)	0.216 (0.195–0.235)
ChatGPT-5.4	596	71.9	67.1	0.717 (0.676–0.757)	0.067 (0.041–0.106)	0.196 (0.179–0.212)
Grok-4-1-fast-reasoning	596	88.6	61.7	0.694 (0.654–0.735)	0.268 (0.230–0.309)	0.291 (0.260–0.322)
Perplexity-Sonar-Deep Research	596	80.6	61.2	0.648 (0.603–0.693)	0.226 (0.193–0.266)	0.273 (0.243–0.301)

Legend: AUC measures how well self-reported confidence distinguished correct from incorrect responses; 95% CIs were calculated using DeLong's method. ECE quantifies the average difference between stated confidence and observed accuracy, while the Brier score reflects overall probabilistic accuracy. Lower ECE and Brier scores indicate better calibration.

Mean self-reported confidence exceeded observed accuracy for every model, demonstrating systematic overconfidence across all six systems. The confidence–accuracy gap was largest for Grok-4-1-fast-reasoning (88.6% confidence versus 61.7% accuracy), followed by Perplexity-Sonar-Deep Research (80.6% versus 61.2%) and Gemini-Pro-3 (93.7% versus 77.9%). More moderate overconfidence was observed for Claude Opus 4.6 (76.4% versus 65.6%). The closest alignment between confidence and observed accuracy was seen for Gemini-Flash-3 (64.8% versus 60.7%) and ChatGPT-5.4 (71.9% versus 67.1%).

Applying confidence thresholds revealed that self-reported certainty did not function as a consistent reliability signal across models. For ChatGPT-5.4, restricting analysis to higher-confidence responses substantially enriched performance, with accuracy increasing from 67.1% overall to 79.7% at confidence ≥ 70%, and to 92.0% and 92.3% at the ≥ 80% and ≥ 90% thresholds, respectively. Claude Opus 4.6 showed a similar pattern, reaching 86.7% accuracy at confidence ≥ 80% and 88.2% at confidence ≥ 90%. In contrast, Gemini-Pro-3 assigned high confidence to nearly all responses, limiting the discriminatory value of confidence thresholds despite its superior overall accuracy. Grok-4-1-fast-reasoning, Perplexity-Sonar-Deep Research, and Gemini-Flash-3 demonstrated only modest or inconsistent enrichment at higher confidence thresholds. These findings indicate that self-reported confidence is not a uniform reliability marker across LLMs; rather, its clinical and administrative utility depends on model-specific calibration behavior and should be interpreted alongside coverage, accuracy, and task context.

Calibration metrics showed a different hierarchy. ChatGPT-5.4 had the lowest ECE (0.067; 95% CI, 0.041–0.106), indicating the closest average alignment between stated confidence and observed accuracy. Whereas Gemini-Pro-3 had an ECE of 0.161 (95% CI, 0.127–0.195). The least favorable ECE values were observed for Perplexity-Sonar-Deep Research (0.226; 95% CI, 0.193–0.266), Grok-4-1-fast-reasoning (0.268; 95% CI, 0.230–0.309), and Gemini-Flash-3 (0.273; 95% CI, 0.239–0.309). Brier scores were lowest for Gemini-Pro-3 (0.191; 95% CI, 0.162–0.219) and ChatGPT-5.4 (0.196; 95% CI, 0.179–0.212), followed by Claude Opus 4.6 (0.216; 95% CI, 0.195–0.235), Perplexity-Sonar-Deep Research (0.273; 95% CI, 0.243–0.301), Gemini-Flash-3 (0.287; 95% CI, 0.254–0.319), and Grok-4-1-fast-reasoning (0.291; 95% CI, 0.260–0.322).

Post hoc alternative prompt

The alternative-prompt analysis produced findings consistent with the primary analysis. Gemini-Pro-3 remained the highest-performing model, increasing from 77.9% to 79.4%, while most models changed by only 0.5–1.5 percentage points. Perplexity-Sonar-Deep Research showed the largest increase, from 61.2% to 66.8%. Overall, the main performance pattern was preserved, suggesting that the principal findings were not driven by the wording of a single prompt.

Discussion

Overall performance

In this comparative evaluation of six contemporary LLMs on medical administration board examination questions, performance varied across models and across areas of administrative knowledge. The most important finding was the significantly weaker performance in locally specific domains, including Israeli Ministry of Health Directives, Reports and Committees, and Legislation and Regulations, suggesting a particular limitation in regulatory and governance knowledge. Model-reported confidence also did not consistently reflect correctness, with overconfidence observed across all systems. Gemini-Pro-3 achieved the highest overall accuracy and was the only model whose point estimate exceeded the resident benchmark in the matched subset. Together, these findings show that strong overall performance does not necessarily translate into reliable performance across the full range of knowledge required in medical administration.

The available health-administration literature remains insufficient to support conclusions about LLM competency in medical administration. Liu et al.[21] evaluated earlier-generation systems, including ChatGPT-4o, on Master of Health Administration educational tasks and reported substantial model-level heterogeneity, including weaker performance than human students in applied case analysis. Avramovic and Avramovic assessed ChatGPT-3 in course-based health administration and policy examinations and reported variable performance across assessment levels, with a highest score of 83.3/100 on a PhD comprehensive examination[22]. Although useful as educational benchmarks, these studies do not establish board-level medical administration competency: they relied on academic course assessments rather than official board-examination material, evaluated older or narrower model sets, lacked an official resident comparator, included limited question cohorts that precluded robust domain-specific analysis, and did not assess confidence-based discrimination or calibration. The present study, therefore, addresses a distinct evidentiary gap by evaluating six contemporary LLMs on official medical administration board-style questions, incorporating resident benchmarking, domain-level performance analysis, and calibration-aware measures of reliability.

Topic-specific

The domain-specific findings may be particularly relevant to local health-system implementation. Models performed better in standardized or internationally shared knowledge domains and worse in areas that depend on jurisdiction-specific legislation, regulatory directives, and governance structures. Importantly, this pattern persisted in the alternative-prompt analysis even when the relevant national context was explicitly stated, suggesting that the weakness was not simply due to failure to recognize the setting. Rather, it may reflect difficulty retrieving or applying locally specific administrative knowledge. Because this benchmark cannot fully distinguish retrieval from reasoning failures, this interpretation should remain cautious. These findings support using current authoritative sources, explicit citations, and expert review when applying LLMs to locally regulated administrative tasks.

Examination-year effects and modality

Compared with the 2020 reference examination, the 2025 question set was associated with higher adjusted odds of a correct response across models. This finding should be interpreted with caution, as annual examinations may differ in content composition, item difficulty, and domain distribution. Therefore, it cannot be used as a causal measure of model improvement over time. One possible explanation is greater accessibility of contemporary health-system information to newer models and retrieval-augmented systems; however, this remains hypothesis-generating and cannot be established from the present analysis.

Question modality did not independently affect performance after adjustment. This finding differs from many clinical AI benchmarks, in which image-based questions often reduce performance[24–26]. A plausible explanation is that image-based questions in medical administration may rely primarily on graphs, tables, diagrams, or structured administrative figures rather than complex clinical imaging such as radiographs, OCT, CT, MRI, or pathology slides. Such visual materials are more standardized and may require quantitative interpretation rather than clinical pattern recognition. Nevertheless, the image-based subset was relatively small and should not be interpreted as evidence of robust multimodal reliability.

Confidence as a core dimension of safe deployment

Accuracy and calibration captured different aspects of model performance. The highest-accuracy model was not the best calibrated by every metric, underscoring the need to evaluate confidence reliability separately from accuracy. Mean self-reported confidence exceeded observed accuracy across all models. ChatGPT-5.4 showed the lowest ECE, indicating the closest average agreement between confidence and accuracy, whereas Gemini-Pro-3 had the lowest Brier score, reflecting the best overall probabilistic accuracy by that metric.

Confidence thresholds also identified higher-accuracy subsets for some models, particularly ChatGPT-5.4 and Claude Opus 4.6, but at the cost of substantially reduced coverage and with inconsistent benefit across systems. In medical administration, this distinction has practical safety implications: high confidence should not be interpreted as a uniform marker of reliability. These findings support evaluating calibration alongside accuracy and reinforce the need for authoritative source verification and expert review when LLMs are used for high-stakes administrative tasks.

Limitations

First, performance was evaluated using board-examination questions rather than real-time administrative workflows. Although board questions provide a standardized benchmark, they do not fully capture the complexity of medical administration. Second, resident comparison was available only for a matched subset of 296 items from 2023–2025, limiting direct human–AI comparisons across the full dataset. Third, the image-based subset was relatively small, reducing power to detect modality-specific effects and limiting conclusions regarding multimodal performance. Fourth, the benchmark was based on medical administration Israeli board exams, and some findings may therefore be Israeli-specific; however, two of the six domains: Statistics and Epidemiology, and International Health Systems, reflect internationally generalizable content. Fifth, commercial LLMs are trained on large and non-transparent corpora; we could not exclude potential prior exposure to publicly available examination-related content. In addition, performance may vary with prompt template, model version, temperature settings, and future model updates. Finally, the LLM landscape evolves rapidly; therefore, these findings are inherently dependent on the specific model versions.

Conclusion

Contemporary LLMs show meaningful but uneven promise for medical administration and remain insufficiently reliable for unsupervised use. Gemini-Pro-3 achieved the strongest overall accuracy and was the only model whose point estimate exceeded the resident benchmark on matched items; however, all models demonstrated systematic overconfidence. Performance was strongest in broadly standardized knowledge domains and significantly weaker in areas dependent on Israeli legislation, Ministry of Health directives, and local governance structures. Model confidence also did not consistently provide a reliable indication of correctness. These findings suggest that LLMs may be useful as supervised educational and preparatory tools for medical administration, including knowledge review, policy summarization, document drafting, and structured decision support. However, high-stakes regulatory, medico-legal, institutional, and patient-safety decisions require source-grounded, locally validated, and confidence-aware outputs reviewed by expert humans, with final accountability remaining with physician medical administrators rather than the model.

Tables

Table 4
Accuracy by self-reported confidence threshold

Model	Overall accuracy, % (95% CI)	Accuracy >= 70% confidence, % (95% CI)	Coverage >= 70%, n/N (%)	Accuracy >= 80% confidence, % (95% CI)	Coverage >= 80%, n/N (%)	Accuracy >= 90% confidence, % (95% CI)	Coverage >= 90%, n/N (%)
Gemini-Pro-3	77.9 (74.3–81.1)	77.9 (74.4–81.2)	594/596 (99.7)	77.9 (74.4–81.2)	594/596 (99.7)	78.2 (74.6–81.6)	570/596 (95.6)
ChatGPT-5.4	67.1 (63.2–70.9)	79.7 (74.9–83.9)	325/596 (54.5)	92.0 (87.0-95.6)	176/596 (29.5)	92.3 (86.3–96.2)	130/596 (21.8)
Claude Opus 4.6	65.6 (61.6–69.4)	69.6 (65.2–73.6)	483/596 (81.0)	86.7 (81.5–90.9)	218/596 (36.6)	88.2 (82.7–92.5)	187/596 (31.4)
Grok-4-1-fast-reasoning	61.7 (57.7–65.7)	62.9 (58.8–66.9)	574/596 (96.3)	64.9 (60.6–69.0)	518/596 (86.9)	67.5 (62.9–71.9)	440/596 (73.8)
Perplexity-Sonar-Deep Research	61.2 (57.2–65.2)	65.9 (61.5–70.1)	490/596 (82.2)	68.0 (63.1–72.7)	388/596 (65.1)	68.7 (63.7–73.4)	374/596 (62.8)
Gemini-Flash-3	60.7 (56.7–64.7)	70.0 (65.2–74.4)	406/596 (68.1)	71.5 (66.7–76.0)	386/596 (64.8)	74.1 (69.1–78.6)	343/596 (57.6)

Legend: Overall accuracy is shown for all 596 items per model and for responses meeting self-reported confidence thresholds of ≥ 70%, ≥ 80%, and ≥ 90%. Coverage indicates the number and proportion of responses meeting each threshold. Confidence was reported on a 0–100 scale.

Figures

Overview of dataset construction, question classification, standardized LLM evaluation, resident benchmark comparison, and statistical analysis

Accuracy of each large language model was compared with mean resident performance in the matched subset of 296 Israeli medical administration board-examination questions from 2023–2025 for which resident benchmark data were available. Resident accuracy represents the mean proportion of examinees answering each item correctly. Error bars denote 95% confidence intervals.

Overall accuracy across the full dataset of 596 medical administration board-examination questions administered between 2020 and 2025.

Accuracy of each large language model across the six medical administration domains. Bars represent model-specific accuracy within each topic, and error bars denote 95% confidence intervals. Adjusted odds ratios for each domain were derived from multivariable generalized estimating equation models, with Statistics and Epidemiology as the reference domain.

Figure 5 | Accuracy trends across examination years by large language model.

Model accuracy stratified by examination year from 2020 to 2025. Points represent the annual accuracy of each LLM, and error bars denote 95% confidence intervals. Accuracy varied across years and models, with the 2025 examination year showing higher adjusted odds of a correct response compared with 2020 in multivariable analysis.

Declarations

Competing interests

The authors declare no competing interests.

Consent for publication

Not applicable.

Funding

This study did not receive specific funding from any public, commercial, or not-for-profit funding agency.

Author Contribution

A.M.G., G.F., O.A., and A.L. contributed to the conception and design of the study and development of the study methodology. D.S. implemented the standardized prompting workflow and contributed to model-output acquisition and data collection. A.M.G. performed the statistical analyses and contributed to data interpretation and manuscript preparation. Y.W. contributed expertise regarding the examination materials and critically reviewed the manuscript. All authors contributed to the interpretation of the findings, critically revised the manuscript for important intellectual content, approved the final version, and agreed to be accountable for the work.

Acknowledgements

The authors thank the Israel Medical Association for providing the Medical Administration board-examination materials and the corresponding official answer keys. The authors also gratefully acknowledge the provision of official resident-performance statistics, which enabled the matched human benchmark analysis.

Data Availability

Data and materials supporting the findings of this study are available from the corresponding author upon reasonable request, subject to applicable institutional and data-use restrictions.

References

1. Roumeliotis KI, Tselikas ND. Chatgpt and open-ai models: A preliminary review. Future Internet; 2023.
2. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in artificial intelligence 2023.
3. Pal A, Sankarasubbu M. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In: *Proceedings of the 6th Clinical Natural Language Processing Workshop*Anonymous, pp.21–46.
4. Li J, Zhou Z, Lyu H et al. Large language models-powered clinical decision support: enhancing or replacing human expertise? Intell Med 2025.
5. Webster P. Six ways large language models are changing healthcare. Nat Med. 2023. 10.1038/s41591-023-02700-1.
6. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA 2023. 10.1001/jama.2023.14217
7. American MA. 2026 Physician Survey on Augmented Intelligence.
8. Yaneva V, Baldwin P, Jurich DP, et al. Examining ChatGPT performance on USMLE sample items and implications for assessment. Academic Medicine; 2024.
9. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS digital health; 2023.
10. Shvartz E, Zur O, Attal L et al. GPT-4o vs. residency exams in ophthalmology—a performance analysis. J Med Artif Intell 2025.
11. Li SW, Kemp MW, Logan SJ, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Obstet Gynecol; 2023.
12. Riedel M, Kaefinger K, Stuehrenberg A et al. ChatGPT's performance in German OB/GYN exams–paving the way for AI-enhanced medical education and clinical practice. Frontiers in medicine 2023.
13. Chen R, Fettel KD, Nguyen DH, et al. Evaluating the performance of ChatGPT on dermatology board-style exams: A meta-analysis of text-based and image-based question accuracy. J Am Acad Dermatol. 2025. 10.1016/j.jaad.2025.04.004.
14. Le M, Davis M. ChatGPT yields a passing score on a pediatric board preparatory exam but raises red flags. Global Pediatric Health; 2024.
15. Rydzewski NR, Dinakaran D, Zhao SG et al. Comparative evaluation of LLMs in clinical oncology. Nejm ai 2024.
16. Patil A, Serrato P, Chisvo N et al. Large language models in neurosurgery: a systematic review and meta-analysis. Acta Neurochir 2024.
17. Galindo AM, Saadi R, Kerman T et al. A large comprehensive comparison of large language models on ophthalmology board exams. Graefe's Archive for Clinical and Experimental Ophthalmology 2026. 10.1007/s00417-026-07264-7
18. Khlaif ZN, Mousa A, Hattab MK, et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. JMIR medical education; 2023.
19. Sharma S. Benefits or concerns of AI: A multistakeholder responsibility. Futures; 2024.
20. Li F, Ruijs N, Lu Y. Ethics & AI: A systematic review on ethical concerns and related strategies for designing with AI in healthcare. Ai 2022.

21. Liu X, Gaynor M, Gillespie K et al. Evaluation and implications of generative AI in health administration education. J Health Adm Educ 2025.
22. Avramovic S, Avramovic I. Exploring the potential benefits and limitations of using an AI text-generation tool in education: An examination of ChatGPT's performance on assessments. J Health Adm Educ 2024.
23. Angel Y, Goldshtein H, Barel N, et al. Trends in Israel's Medical Administration subspecialty, 1987–2022. Isr J Health Policy Res. 2025. 10.1186/s13584-025-00666-8.
24. Payne DL, Purohit K, Borrero WM et al. Performance of GPT-4 on the American College of Radiology in-training examination: evaluating accuracy, model drift, and fine-tuning. Acad Radiol 2024.
25. Hayden N, Gilbert S, Poisson LM et al. Performance of GPT-4 with vision on text-and image-based ACR diagnostic radiology in-training examination questions. Radiology. 2024.
26. Jin Q, Chen F, Zhou Y, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. NPJ Digital Medicine; 2024.

Figures

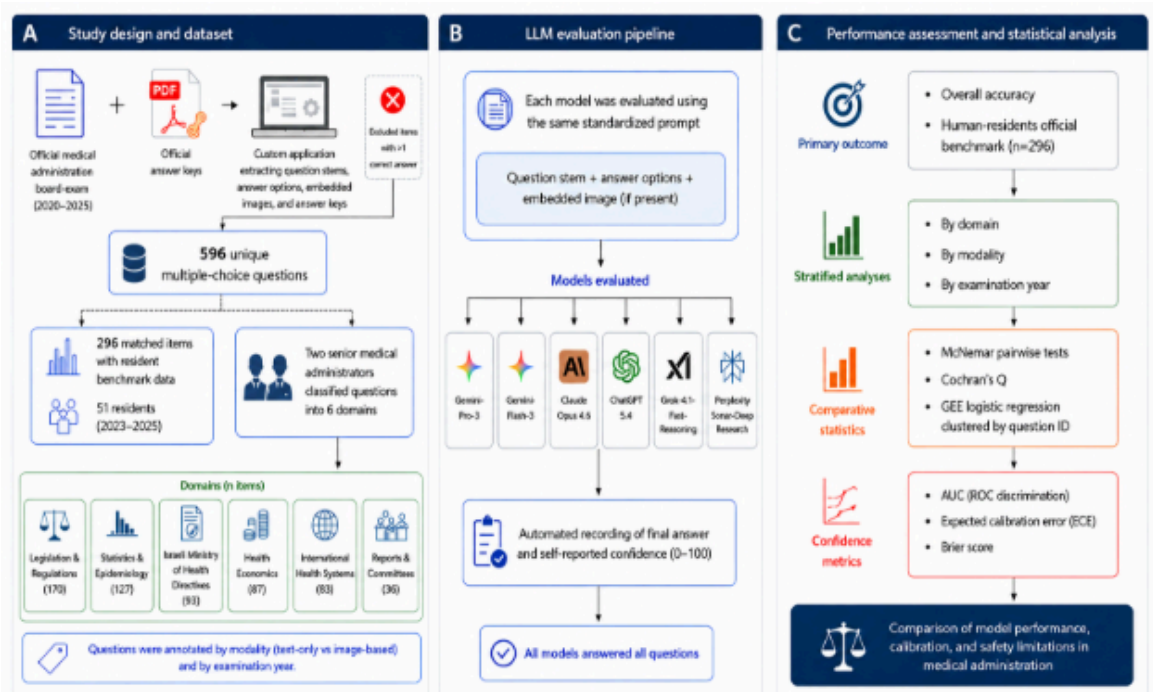


Figure 1

Study design and model-evaluation workflow.

Overview of dataset construction, question classification, standardized LLM evaluation, resident benchmark comparison, and statistical analysis

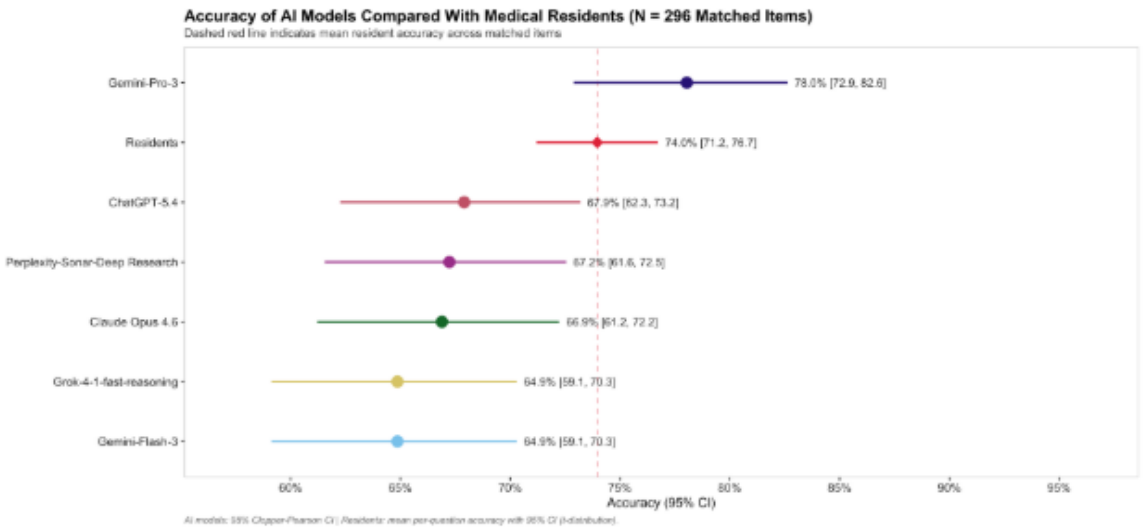


Figure 2

Performance of large language models compared with residents

Accuracy of each large language model was compared with mean resident performance in the matched subset of 296 Israeli medical administration board-examination questions from 2023–2025 for which resident benchmark data were available. Resident accuracy represents the mean proportion of examinees answering each item correctly. Error bars denote 95% confidence intervals.

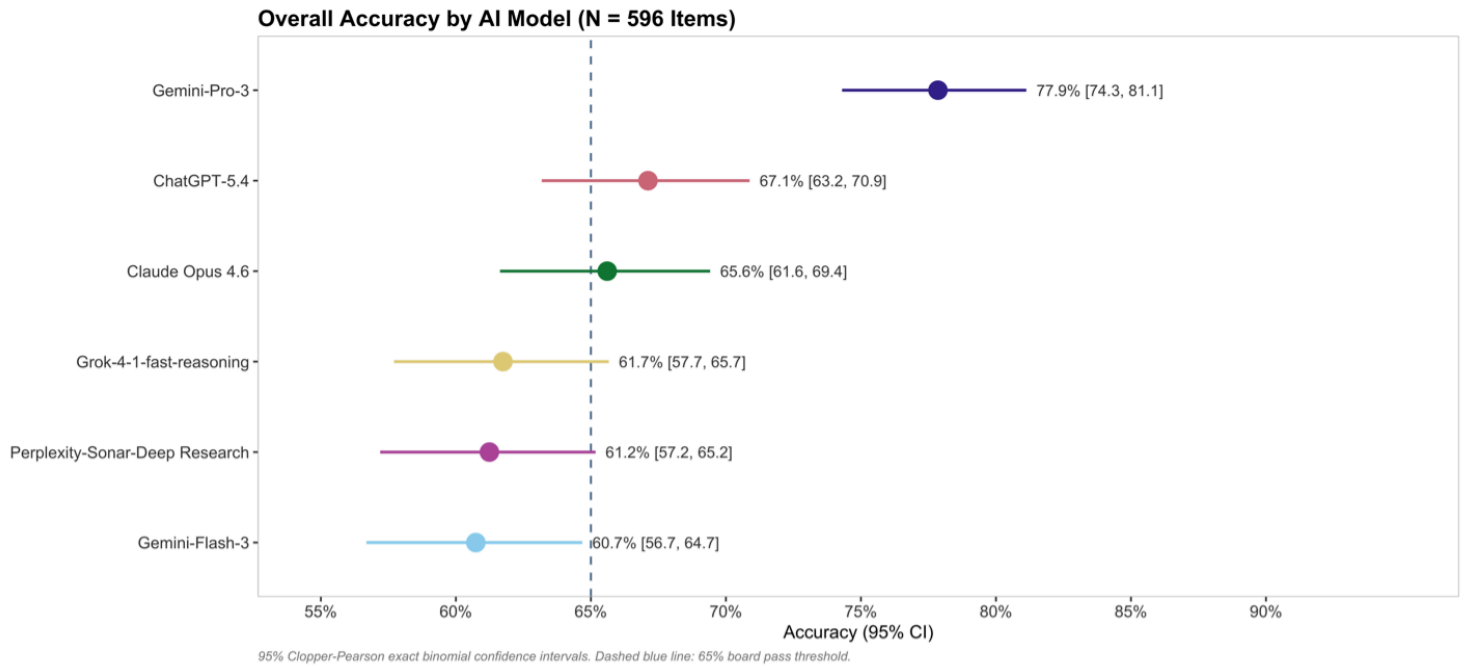


Figure 3

Overall accuracy of six contemporary large language models

Overall accuracy across the full dataset of 596 medical administration board-examination questions administered between 2020 and 2025.

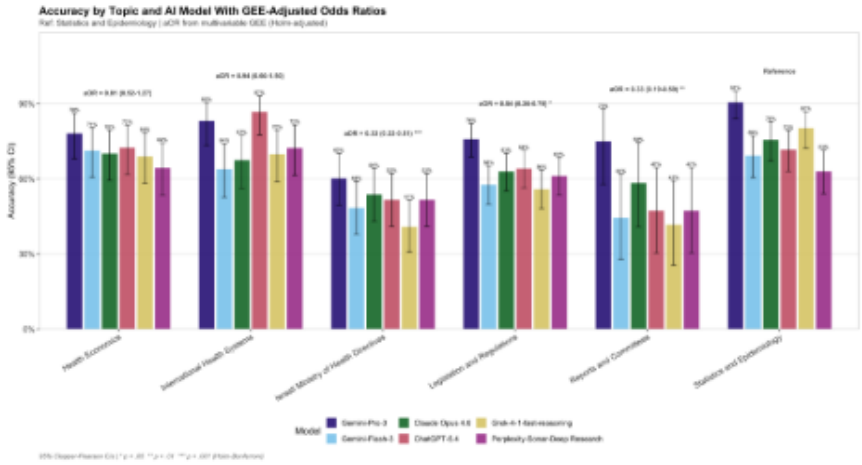


Figure 4

Topic-specific accuracy and adjusted odds of correct response by model.

Accuracy of each large language model across the six medical administration domains. Bars represent model-specific accuracy within each topic, and error bars denote 95% confidence intervals. Adjusted odds ratios for each domain were derived from multivariable generalized estimating equation models, with Statistics and Epidemiology as the reference domain.

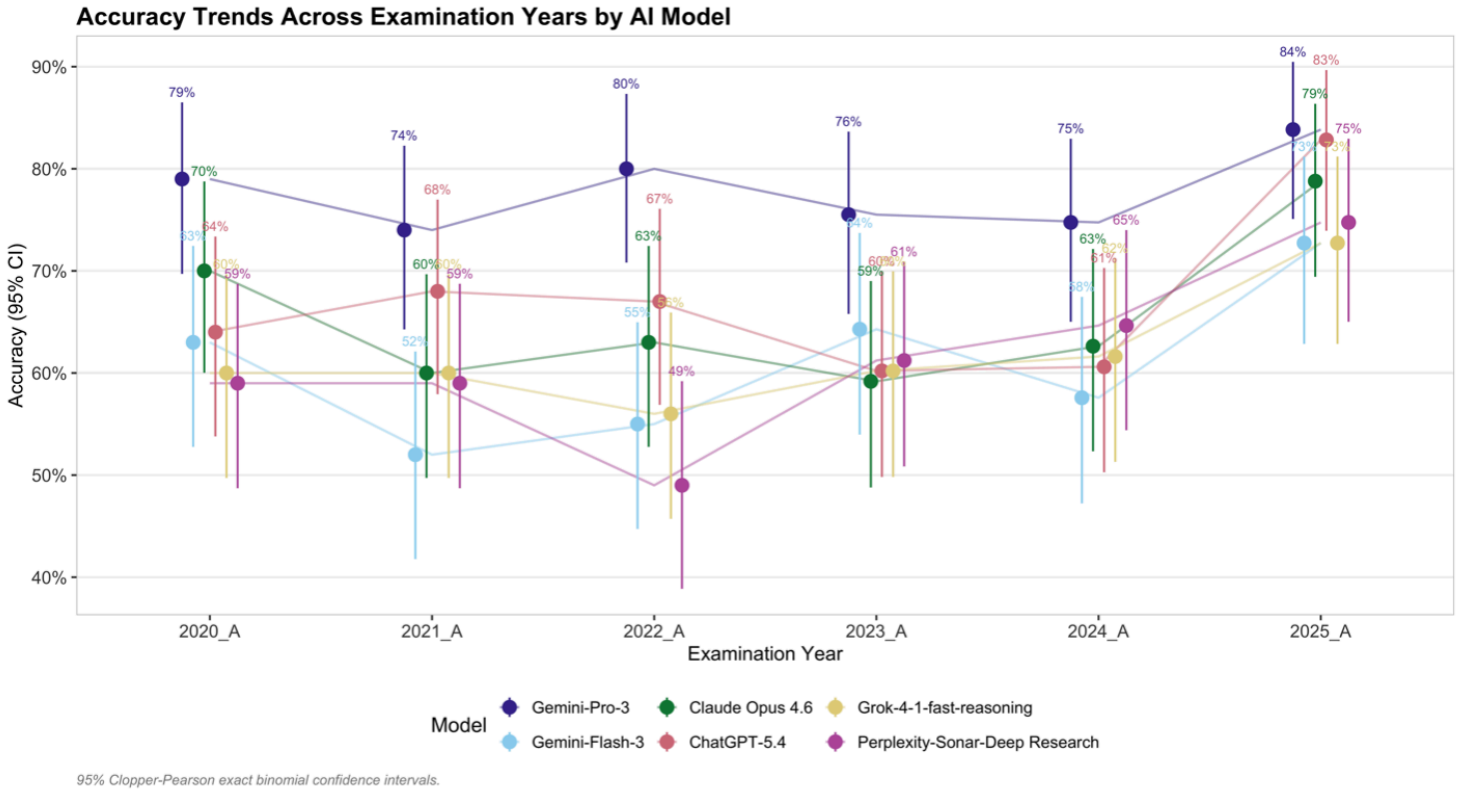


Figure 5

Accuracy trends across examination years by large language model.

Model accuracy stratified by examination year from 2020 to 2025. Points represent the annual accuracy of each LLM, and error bars denote 95% confidence intervals. Accuracy varied across years and models, with the 2025 examination year showing higher adjusted odds of a correct response compared with 2020 in multivariable analysis.