

GPT-5 versus Senior Anesthesiology-Intensive Care Residents on Sequential Clinical Cases: A Pilot Study of Documented Clinical Reasoning

Elmahdi Ezzikouri¹

mehdiforsocial@hotmail.com

Centre Hospitalier Universitaire Ibn Rochd, Morocco

Sabah Benhamza¹

Centre Hospitalier Universitaire Ibn Rochd, Morocco

Mohammed Bennani Othmani¹

Centre Hospitalier Universitaire Ibn Rochd, Morocco

Mohamed Lazraq¹

Centre Hospitalier Universitaire Ibn Rochd, Morocco

Research Article

Keywords: generative artificial intelligence, clinical reasoning, medical education, anesthesiology, R-IDEA, large language model

Posted Date: August 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10667066/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 have performed well on medical examinations, but examination accuracy does not show how explicitly a response documents clinical reasoning. We compared GPT-5 outputs with those of senior anesthesiology-intensive care residents using the Revised-IDEA (R-IDEA) clinical reasoning documentation score.

Methods

In this single-center cross-sectional comparative pilot study, 14 fourth-year residents each completed four of 20 sequential, fully anonymized clinical cases in French. GPT-5 completed all 20 cases once. Two faculty members developed case-specific reference rubrics by consensus, and one source-masked evaluator scored every response. The primary analysis paired each resident's mean score with GPT-5's mean score on the same four cases.

Results

All 14 eligible residents participated and completed 56 responses; GPT-5 generated 20. Resident-level mean R-IDEA scores were 7.0 (SD 2.4), compared with 9.5 (SD 0.9) for matched GPT-5 outputs (mean paired difference 2.5 points, 95% CI 1.4–3.6; $t(13) = 5.21$; $p < 0.001$; paired $d_z = 1.39$). GPT-5 scores were higher in each R-IDEA domain in exploratory analyses. Diagnostic accuracy was 68% for residents and 75% for GPT-5 and was treated descriptively.

Conclusions

Under a prompt aligned with R-IDEA components, GPT-5 outputs received higher scores for documented clinical reasoning than resident responses. The finding does not establish superior clinical competence or an educational benefit. It supports testing faculty-supervised GPT-5 worked examples as a reasoning-documentation scaffold in prospective educational trials.

Introduction

Large language models (LLMs) have rapidly entered medical education. A scoping review in Medical Science Educator found that the early literature was dominated by studies of performance on examinations and that comparatively few studies used rigorous empirical designs [1]. ChatGPT has achieved passing-level performance on the United States Medical Licensing Examination [2], and medical LLMs encode substantial clinical knowledge [3]. Free-response and board-style comparisons have also reported strong model performance relative to medical students or resident physicians [4, 5].

These results are relevant to knowledge assessment, but a correct final answer alone does not demonstrate how well a learner identifies and prioritizes the information that supports a diagnosis.

Clinical reasoning is context dependent and includes problem representation, generation and prioritization of diagnostic hypotheses, and interpretation of evidence [6, 7]. The Revised-IDEA (R-IDEA) instrument provides a structured assessment of clinical reasoning documentation across four domains: interpretive summary, differential diagnosis, explanation of the lead diagnosis, and explanation of alternative diagnoses [8]. It does not measure unobserved cognition or bedside performance; rather, it measures the quality of reasoning made visible in written work. This distinction is particularly important for LLMs, which are optimized to produce fluent, structured text.

Making reasoning visible is also an educational objective. A faculty member can give more specific feedback when a learner states which findings are discriminating, why one diagnosis is prioritized, and what evidence weighs against plausible alternatives. In anesthesiology and intensive care, that documentation must often integrate neurologic, respiratory, cardiovascular, and metabolic data within the same problem representation. However, explicitness has a measurement cost: a fluent or verbose response may satisfy a documentation rubric without demonstrating that its author could gather the same data, recognize deterioration, or act safely at the bedside. Comparisons between LLMs and trainees therefore require an outcome that is clearly named and a deliberately narrow interpretation.

Cabral et al. used sequential clinical cases and R-IDEA scoring to compare a generative AI model with physicians and reported higher documentation scores for the model [9]. Conceptual replication in other languages, specialties, and training environments is needed because model behavior is prompt and version dependent, and clinical reasoning is sensitive to context. Evidence from French-language postgraduate training in North Africa is particularly limited. Senior anesthesiology-intensive care residents provide a relevant comparison group because their work requires rapid synthesis of multisystem information and explicit prioritization of competing diagnoses.

The primary objective of this pilot study was to compare the quality of documented clinical reasoning in GPT-5 and fourth-year anesthesiology-intensive care resident responses using the total R-IDEA score. Secondary objectives were to describe differences across R-IDEA domains, score variability, and diagnostic accuracy. We interpreted the comparison as an assessment of written outputs under a defined task, not as a test of whether an LLM can replace a physician.

Materials and Methods

Study design and setting

We conducted a single-center cross-sectional comparative pilot study in the Department of Anesthesia and Intensive Care, Hospital 20 August 1953, CHU Ibn Rochd, Casablanca, Morocco. Resident data were collected from September 17 to October 8, 2025; GPT-5 responses were generated on October 10, 2025. The study compared written responses to sequential clinical cases. It was not an educational

intervention and did not evaluate change in resident learning. Reporting was informed by relevant STROBE principles for observational studies.

The protocol specified total R-IDEA as the primary outcome before response scoring. Domain scores, variability, and diagnostic accuracy were secondary or exploratory outcomes. The model was treated as a response generator under one defined configuration, not as a participant with demographic or training characteristics. Human and model responses were retained in the same text-based format for application of the case-specific rubrics.

Participants and sample

All fourth-year anesthesiology-intensive care residents in the department were eligible if they were available during data collection and provided written informed consent. Residents in years 1–3 and residents who did not consent or were unavailable were excluded. We used an exhaustive census of the fixed eligible cohort rather than probability sampling: all 14 eligible residents participated (14/14, 100%). Because this was a pilot study of a fixed training cohort, no formal a priori sample-size calculation was used. The resident, not the individual case response, was the primary unit of analysis.

Clinical cases and allocation

Twenty real clinical cases were selected from departmental records dated 2020–2024. All direct and indirect identifiers were removed before use, and the narrative format was standardized. Two faculty members reviewed the cases and agreed on the intended diagnosis and a case-specific R-IDEA reference rubric. Cases were classified by their dominant clinical category as neurology ($n = 7$), pulmonology ($n = 4$), cardiology ($n = 3$), endocrinology ($n = 3$), or other ($n = 3$).

Each case was presented in two successive parts. Part 1 contained the history, medical background, and physical examination. Participants wrote a 100–300 word analytic problem representation and a prioritized list of investigations with brief justifications. Part 2 then revealed laboratory and imaging results. After advancing, participants could not return to Part 1. They listed at least three differential diagnoses in descending order of probability and supplied evidence for and against each diagnosis. A pilot administration with one non-study trainee assessed clarity and technical function; the observed completion time was 18 minutes per case on average (SD 7). No enforced time limit was used in the study.

A constrained randomization procedure assigned four distinct cases to each resident while balancing coverage. Sixteen cases received three resident responses and four cases received two, yielding 56 resident responses. GPT-5 received all 20 cases once. This incomplete-block design allowed every resident to be matched with the model on the same four cases while limiting participant burden.

GPT-5 configuration and prompt

GPT-5 was accessed through the OpenAI application programming interface using the model alias gpt-5 [10], rather than through the consumer ChatGPT interface. The study log recorded reasoning effort as medium, a maximum output of 2,000 tokens, temperature 0.3, and top_p 0.9. The dated snapshot behind the alias was not retained; the findings therefore apply to the alias as accessed on October 10, 2025. All clinical inputs were fully anonymized before API transfer, and each case was submitted once. The task, prompt, resident responses, and model outputs were in French.

The standardized prompt assigned the role of a clinician specialized in anesthesiology-intensive care and mirrored the two-stage resident task. It asked the model to begin each section with a three- to seven-item conceptual checklist; after Part 1, to provide a concise problem representation and prioritized, justified investigations; and after Part 2, to provide at least three prioritized differential diagnoses with explicit supporting and opposing evidence. The prompt therefore directly elicited components scored by R-IDEA. This alignment was intentional for task standardization but was considered when interpreting ceiling effects.

The prompt did not provide the reference diagnosis or the scoring rubric. GPT-5 received the same clinical information sequence as residents: Part 1 was supplied before the model's initial response, and Part 2 was supplied before its differential diagnosis response. The case-specific faculty rubrics were used only after all responses had been produced. Thus, the model was instructed about the required response structure but was not given the answer key.

Outcome assessment

The primary outcome was the total R-IDEA score (0–10) [8]. The instrument comprises interpretive summary (I; 0–4), differential diagnosis (D; 0–2), explanation of the lead diagnosis (E; 0–2), and explanation of alternative diagnoses (A; 0–2). The case-specific reference rubrics identified expected elements, acceptable formulations, optional elements, and critical errors. The two faculty developers first reviewed each case independently and then resolved differences by consensus before scoring began.

One evaluator scored all resident and GPT-5 responses against the case-specific rubrics with the response source concealed. Because writing style could still have suggested source, we describe this as source masking rather than guaranteed blinding. A second rater was not used, so inter-rater reliability could not be estimated. Diagnostic accuracy was recorded when the first-ranked diagnosis agreed with the faculty reference diagnosis.

Statistical analysis

The primary analysis used 14 resident-level matched pairs. For each resident, we calculated the mean R-IDEA score across the four assigned cases. We then calculated GPT-5's mean score on those same four cases. The distribution of within-pair differences was compatible with normality (Shapiro-Wilk $W = 0.943$, $p = 0.467$), and the means were compared with a two-sided paired Student t test at $\alpha = 0.05$. We report the mean paired difference with its 95% confidence interval and the paired standardized effect size d_z , calculated as $t/\text{square root of } n$.

Domain-level paired comparisons used the same resident-level aggregation and were exploratory; their p values were not adjusted for multiple testing. Coefficients of variation and distributional summaries were descriptive. Because diagnostic accuracy involved repeated resident observations, unequal case replication, and an incomplete-block allocation, percentages were reported descriptively and no unclustered Fisher exact or chi-square test was performed. There were no missing resident responses.

Ethical considerations

The protocol was reviewed by the Comité d'Éthique de la Recherche, CHU Ibn Rochd, Casablanca, Morocco, and was determined to be exempt from formal ethics approval. The study used an educational comparison and fully anonymized clinical cases. All resident participants provided written informed consent; participation was voluntary, withdrawal was permitted without prejudice, and individual performance had no academic consequences. Resident data were pseudonymized, and no identifying patient information was retained.

Use of generative AI in manuscript preparation

After completion of data collection and the original analyses, OpenAI's ChatGPT was used to assist with English-language editing, reference checking, and journal-specific organization of the manuscript. The authors reviewed and verified all text, calculations, and references and take full responsibility for the final manuscript.

Results

Participants and response set

All 14 eligible residents enrolled and completed all four assigned cases, producing 56 resident responses. GPT-5 produced one response for each of the 20 cases. The primary analysis therefore comprised 14 resident-level matched pairs, each based on a resident's four assigned cases and GPT-5's responses to those same cases.

Primary outcome

The mean resident-level R-IDEA score was 7.0 (SD 2.4; median 7.0; range 3.0–10.0), compared with 9.5 (SD 0.9; median 9.75; range 8.0–10.0) for the matched GPT-5 outputs. The mean paired difference was 2.5 points in favor of GPT-5 (95% CI 1.4–3.6; standard error 0.48; $t(13) = 5.21$; $p < 0.001$). The paired standardized effect size was $d_z = 1.39$. The coefficient of variation was 34% for resident means and 9% for matched GPT-5 means.

The interquartile range was 5.5–8.5 for resident means and 9.0–10.0 for matched GPT-5 means. These summaries show that several residents produced high-scoring work while the model scores were concentrated near the top of the scale; they do not imply that individual residents can be ranked against a deterministic model benchmark.

R-IDEA domains and diagnostic accuracy

GPT-5 outputs received higher mean scores in all four R-IDEA domains (Table 1). The largest absolute difference was in interpretive summary (1.0 point on a four-point domain), followed by explanation of alternative diagnoses (0.6 point on a two-point domain). Domain-level p values were exploratory and unadjusted. Scores for GPT-5 approached the maximum in each domain, indicating a potential ceiling effect.

Table 1
Resident-level paired R-IDEA outcomes

Outcome (score range)	Residents, mean (SD)	GPT-5, mean (SD)	Paired difference, mean (95% CI)	p value
Interpretive summary (0–4)	2.8 (1.0)	3.8 (0.4)	1.0 (0.5–1.5)	< 0.001
Differential diagnosis (0–2)	1.5 (0.6)	1.9 (0.2)	0.4 (0.1–0.7)	0.008
Explanation of lead diagnosis (0–2)	1.4 (0.7)	1.9 (0.2)	0.5 (0.2–0.8)	0.004
Explanation of alternatives (0–2)	1.3 (0.8)	1.9 (0.3)	0.6 (0.2–1.0)	0.002
Total R-IDEA (0–10)	7.0 (2.4)	9.5 (0.9)	2.5 (1.4–3.6)	< 0.001
Note. $n = 14$ resident-level matched pairs. Differences are GPT-5 minus resident. The total score was the prespecified primary outcome. Domain comparisons were exploratory and p values were not adjusted for multiple testing. CI, confidence interval; R-IDEA, Revised-IDEA; SD, standard deviation.				

GPT-5 identified the faculty reference diagnosis in 15 of 20 cases (75%). The resident mean diagnostic accuracy across assigned cases was 68%. These percentages were not subjected to an inferential comparison because the underlying observations were not independent and case replication differed. The accuracy outcome therefore does not establish equivalence or a difference between groups.

Discussion

In this single-center pilot study, GPT-5 outputs received higher total R-IDEA scores than the written responses of senior anesthesiology-intensive care residents when both were evaluated on matched sequential cases. The 2.5-point paired difference was accompanied by a narrower distribution of GPT-5 scores. Differences were present across all four R-IDEA domains, with the largest absolute difference in interpretive summary. Diagnostic accuracy was 75% for GPT-5 and 68% for residents, but the study design supported descriptive, not inferential, comparison of those percentages. The main finding is therefore about the explicit structure of written reasoning, not diagnostic superiority.

The result is directionally consistent with prior free-response comparisons [4, 5, 9]. In particular, Cabral et al. reported higher R-IDEA scores for a generative model than for physicians completing sequential clinical cases [9]. Our study extends that approach to a French-language task, a Moroccan academic center, GPT-5, and an advanced anesthesiology-intensive care cohort. It should nevertheless be viewed as a conceptual replication rather than proof of general performance: the cases, prompt, participant population, evaluator, and model version all differed.

The R-IDEA result is plausible because the prompt explicitly requested a concise problem representation, a prioritized differential diagnosis, and evidence for and against competing diagnoses. These are the same visible behaviors rewarded by the rubric. The model could produce complete and consistently organized text without the time, fatigue, and working-environment constraints that affect clinicians. Residents may also use efficient pattern recognition and omit reasoning that they consider obvious. Consequently, a higher R-IDEA score should not be interpreted as access to a better internal reasoning process. R-IDEA was developed to assess documentation and support feedback [8], and the current task adapted it beyond its original use in admission notes.

The descriptive accuracy findings reinforce the need to separate documentation quality from diagnostic correctness. A response can be well organized yet prioritize the wrong diagnosis, while a clinician may recognize the correct pattern but document only part of the supporting logic. The current study was not designed to determine whether these two outcomes diverge systematically: residents answered different subsets of cases, cases were replicated unequally, and each model case had only one run. For education, both outcomes matter. A worked example should be factually correct and should expose a defensible path from findings to hypotheses; strength on either dimension cannot compensate automatically for weakness on the other.

The educational relevance lies in what the outputs make visible. A faculty member could use a carefully checked GPT-5 answer as one worked example, ask residents to compare it with their own problem representation, and discuss omitted discriminating evidence or overinclusive alternatives. This use aligns with the need to move medical education research beyond examination passing toward theory-informed, measurable learning activities [1]. It also keeps the educator in control of the case standard, local epidemiology, and feedback. The present study did not test such an intervention, so any learning benefit remains a hypothesis. A next study should randomize residents to faculty feedback with or

without a model-generated worked example and assess change in independently rated R-IDEA scores over time.

A practical teaching sequence could begin with an unaided resident response, followed by comparison with a faculty-vetted model example and a short self-explanation of the most important differences. Faculty could then ask the resident to revise the problem representation, remove nondiscriminating details, and justify why an alternative moved up or down the differential. R-IDEA could provide a common language for feedback without making the model output the answer key. Such a sequence would be inexpensive to pilot, but its value should be judged by transfer to a new case and retention over time, not by learner satisfaction alone. It should also include an explicit step in which learners identify a model error or unsupported assertion, preserving critical appraisal rather than passive acceptance.

Recent work in Medical Science Educator illustrates why reproducible prompts, explicit rubrics, and reliability analysis matter when generative AI is used with open-ended responses [11]. Those principles apply equally when the model is the respondent. A high score can depend on prompt-rubric alignment, and small rubric changes may alter apparent performance. Future replications should preregister the prompt and scoring plan, preserve a dated model snapshot, repeat model runs to quantify stochastic variability, and use at least two independent raters with inter-rater reliability and adjudication. They should also consider word-count- or time-matched conditions and mixed-effects models that account for both resident and case.

The study setting adds contextual value but also raises an implementation question. A French-language model example may improve access to structured reasoning material where faculty time and locally authored cases are limited, yet imported model knowledge may not reflect Moroccan epidemiology, test availability, or local care pathways. Educational use should therefore begin with locally curated cases and reference standards. Multi-center replication across Moroccan training programs could test whether the observed pattern persists across institutions while allowing investigators to examine language, specialty, and case-complexity effects.

Faculty oversight is essential. Fluent model output can contain incorrect or weakly calibrated clinical claims, and formal completeness can create unwarranted confidence [12, 13]. The model missed the reference diagnosis in five cases in this study. Before use in teaching, every generated example should therefore be checked for factual accuracy, uncertainty, local relevance, and alignment with the intended learning objective. The appropriate educational question is not whether the model should replace a resident, but whether a supervised activity using the model improves a learner outcome without adding automation bias or weakening independent reasoning.

Limitations

Several limitations constrain interpretation. First, this was a single-center study of 14 residents and 20 cases; the fixed cohort improved participation but limited precision and generalizability. Second, one evaluator scored all responses. Source masking reduced but could not eliminate recognition of model

writing style, and the absence of a second rater prevented reliability estimation. Third, the prompt was deliberately aligned with R-IDEA components, which may have favored explicit completeness and contributed to ceiling effects. Fourth, R-IDEA evaluates documentation, not bedside data gathering, treatment planning, communication, professional judgment, or patient outcomes. Fifth, each GPT-5 case was run once and the exact dated snapshot behind the API alias was not retained, limiting reproducibility. Sixth, the resident-level paired t test accounted for each resident's matched cases but did not fully model crossed clustering by case in the incomplete-block design. Domain analyses were exploratory, and no multiplicity adjustment was applied. Finally, diagnostic accuracy was summarized using different response structures and was not suitable for a simple unclustered test. These limitations make the findings hypothesis generating.

Conclusions

GPT-5 outputs received higher R-IDEA scores than senior anesthesiology-intensive care resident responses on matched sequential clinical cases, particularly for explicit problem representation and explanation of alternatives. This finding concerns written clinical reasoning documentation under a rubric-aligned prompt; it does not establish superior clinical competence, diagnostic equivalence, or educational effectiveness. The most defensible next use is as a faculty-supervised worked-example scaffold evaluated in prospective, multi-rater educational trials.

Declarations

Ethics approval and consent to participate

The protocol was reviewed by the Comité d'Éthique de la Recherche, CHU Ibn Rochd, Casablanca, Morocco, and was determined to be exempt from formal ethics approval. All resident participants provided written informed consent. The clinical cases were fully anonymized before use, resident data were pseudonymized, and individual performance had no academic consequences.

Consent for publication

Not applicable. The manuscript contains no identifiable individual participant or patient information.

Data availability

The de-identified response-level dataset, case-specific scoring rubrics, and original French prompt are available from the corresponding author on reasonable request, subject to institutional and ethical restrictions.

Funding

The authors received no specific funding for this work.

Competing interests

The authors declare no relevant financial or non-financial interests.

Preprint

A previous version of this manuscript was posted on Research Square: <https://doi.org/10.21203/rs.3.rs-9540519/v1>.

Author Contribution

E.E. conceived the study, developed the methodology, conducted the investigation, curated the data, performed the formal analysis, drafted the original manuscript, and coordinated the project. S.B. supervised the study, performed validation, contributed to the development of the case-specific scoring rubrics, and critically reviewed and edited the manuscript. M.B.O. supervised the study, contributed to the methodology and development of the scoring rubrics, and critically reviewed and edited the manuscript. M.L. selected the clinical cases, contributed to the investigation, performed response scoring, and critically reviewed and edited the manuscript. All authors read and approved the final manuscript and agree to be accountable for the work.

References

1. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - scoping literature review. *Med Sci Educ*. 2025;35:555-567. <https://doi.org/10.1007/s40670-024-02206-6>
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:e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
3. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180. <https://doi.org/10.1038/s41586-023-06291-2>
4. Strong E, DiGiammarino A, Weng Y, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Intern Med*. 2023;183:1028-1030. <https://doi.org/10.1001/jamainternmed.2023.2909>
5. Katz U, Cohen E, Shachar E, et al. GPT versus resident physicians - a benchmark based on official board scores. *NEJM AI*. 2024;1(5). <https://doi.org/10.1056/Aldbp2300192>

6. Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ*. 2005;39:98-106. <https://doi.org/10.1111/j.1365-2929.2004.01972.x>
7. Norman GR, Eva KW. Diagnostic error and clinical reasoning. *Med Educ*. 2010;44:94-100. <https://doi.org/10.1111/j.1365-2923.2009.03507.x>
8. Schaye V, Miller L, Kudlowitz D, et al. Development of a clinical reasoning documentation assessment tool for resident and fellow admission notes: a shared mental model for feedback. *J Gen Intern Med*. 2022;37:507-512. <https://doi.org/10.1007/s11606-021-06805-6>
9. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern Med*. 2024;184:581-583. <https://doi.org/10.1001/jamainternmed.2024.0295>
10. OpenAI. GPT-5 model. OpenAI API documentation. <https://developers.openai.com/api/docs/models/gpt-5>. Accessed 9 August 2026.
11. Olvet DM, Kruidering M, Fulton TB, et al. Can generative artificial intelligence reliably score open-ended question assessments in undergraduate medical education? *Med Sci Educ*. 2026;36:1539-1552. <https://doi.org/10.1007/s40670-026-02638-2>
12. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388:1233-1239. <https://doi.org/10.1056/NEJMSr2214184>
13. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30:2613-2622. <https://doi.org/10.1038/s41591-024-03097-1>