Revista Española
de Educación Médica

Evaluating the Performance of DeepSeek 3, Claude Sonnet 4, and Gemini 2.5 in the Chilean Medical Licensing Examination: Observational Study.

Evaluación del desempeño de DeepSeek 3, Claude Sonnet 4 y Gemini 2.5 en el examen de licencia médica chileno: estudio observacional.

Anaís Aracelly Lancellotti Guajardo[1], Oscar Jerez Yañez[2], Vicente Alberto Edgardo Jesus Silva Arroyo[1,3], Marcos Jeremías Giovanny Vera Cartes[1,4], Álvaro Andrés Herrera Alcaíno[1,5].

1, Faculty of Medicine, University of Chile, Santiago, Chile. ORCID: 0009-0003-2254-0470, anaislancellotti@ug.uchile.cl. 2, Department of Health Sciences Education, Faculty of Medicine, University of Chile, Santiago, Chile. ORCID: 0000-0003-0869-5938, ojerez@uchile.cl, 3, Faculty of Medicine, University of Chile, Santiago, Chile. ORCID: 0009-0001-4182-0115, vicente.silva.a@ug.uchile.cl 4, Faculty of Medicine, University of Chile, Santiago, Chile. ORCID: 0009-0009-9156-7419, marcosvera@ug.uchile.cl, 5, Faculty of Medicine, University of Chile, Santiago, Chile, and Faculty of Medicine, San Sebastián University, Santiago, Chile. ORCID: 0009-0007-4861-2144, levarito@uchile.cl

Correspondence to Anaís Lancellotti, anaislancellotti@ug.uchile.cl

Recibido: 12/9/25; Aceptado: 7/10/25; Publicado: 9/10/25

Abstract.
Introduction: Artificial intelligences and their continuous improvement have revolutionized medical education, but their performance in specific evaluative contexts still requires further exploration. **Methods:** This study qualitatively evaluated and compared the performance of three state-of-the-art language models — Claude Sonnet 4, Gemini 2.5, and DeepSeek 3 — in simulations of the National Medical Knowledge Examination (EUNACOM) in Chile. Three mock exams with 180 questions each were used, covering various medical areas and question types, including those based on clinical cases. **Results**: The results show that all AI models consistently passed the exams, with Claude Sonnet 4 achieving the highest overall performance (89% accuracy) and the greatest consistency across attempts. Clinical case-based questions were answered more accurately than theoretical knowledge questions, highlighting the models' strength in contextual clinical reasoning. Claude excelled in Internal Medicine and Psychiatry, DeepSeek in Surgery, and Gemini demonstrated balanced performance. However, specific gaps were identified in areas such as Public Health and clinical follow-up, suggesting the need for model-specific adjustments. **Conclusion:** The findings support the educational potential of these tools but also emphasize the importance of their ethical, supervised, and complementary use alongside traditional medical training. This study contributes to understanding the emerging role of artificial intelligence in professional assessments, as well as its limitations and opportunities within the Chilean medical context.

Keywords: artificial intelligence, medical education, EUNACOM, clinical reasoning, language models, medical assessment.

Resumen.
Introducción: La inteligencias artificial y su mejora continua han revolucionado la educación médica, pero su desempeño en contextos evaluativos específicos aún requiere mayor exploración. **Métodos:** Este estudio evaluó y comparó cualitativamente el desempeño de tres modelos de lenguaje de última generación —Claude Sonnet 4, Gemini 2.5 y DeepSeek 3— en simulaciones del Examen Nacional de Conocimientos Médicos (EUNACOM) en Chile. Se utilizaron tres exámenes simulados con 180 preguntas cada uno, que abarcaban diversas áreas médicas y tipos de preguntas, incluidas las basadas en casos clínicos. **Resultados:** Los resultados muestran que todos los modelos de IA aprobaron los exámenes de forma consistente, y Claude Sonnet 4 logró el mayor desempeño general (89% de precisión) y la mayor consistencia en todos los intentos. Las preguntas basadas en casos clínicos se respondieron con mayor precisión que las preguntas de conocimiento teórico, lo que destaca la fortaleza de los modelos en el razonamiento clínico contextual. Claude sobresalió en Medicina Interna y Psiquiatría, DeepSeek en Cirugía y Gemini demostró un desempeño equilibrado. Sin embargo, se identificaron deficiencias específicas en áreas como la salud pública y el seguimiento clínico, lo que sugiere la necesidad de realizar ajustes específicos a cada modelo. **Conclusión:** Los hallazgos respaldan el potencial educativo de estas herramientas, pero también enfatizan la importancia de su uso ético, supervisado y complementario a la formación médica tradicional. Este estudio contribuye a comprender el papel emergente de la inteligencia artificial en las evaluaciones profesionales, así como sus limitaciones y oportunidades en el contexto médico chileno.

Palabras clave: inteligencia artificial, educación médica, EUNACOM, razonamiento clínico, modelos de lenguaje, evaluación médica.

1. Introducción

The widespread use of generative artificial intelligence (AI) has experienced exponential growth since the launch of ChatGPT-3.5 by OpenAI in November 2022, transforming areas such as medical education (1-2). This advancement was further solidified with the arrival of more sophisticated models: GPT-4V (September 2023), Claude Sonnet 4 by Anthropic (May 2025), Gemini 2.5 by Google (June 2025), and DeepSeek 3 by DeepSeek AI (March 2025), each offering different technical and interactional approaches (3-6).

In medicine, the use of AI-powered chatbots has gained relevance due to the growing trust in these technologies. This trust is particularly strong among individuals under 35 and over 64 years old, who appreciate their benefits for diagnosis and patient follow-up (7). In this context, the medical performance of various AIs has been widely evaluated through internationally recognized standardized tests, such as the United States Medical Licensing Examination (USMLE) or the Spanish MIR exam, with promising results reported (8-9). However, there is still limited evidence regarding the medical competence of these systems in specific linguistic and cultural contexts, particularly in Latin America.

In Chile, the National Medical Knowledge Examination (EUNACOM) is a critical requirement for medical practice, assessing comprehensive knowledge in areas such as internal medicine, surgery, pediatrics, obstetrics and gynecology, psychiatry, specialties, and public health. Passing this exam is mandatory for foreign doctors who wish to practice in Chile, as well as for Chilean medical students completing their studies (10-11). A recent Chilean study evaluated the performance of ChatGPT on the EUNACOM, showing that the GPT-3.5, GPT-4, and GPT-4V versions were able to pass the exam successfully, with GPT-4 standing out with a 70.67% score (12). However, comparative studies that include other generative AIs developed by different companies—with distinct engines and approaches—are still lacking. This highlights the need to expand this line of research to better understand the medical competence of these systems in this specific context.

Therefore, this study aims to comparatively evaluate the performance of different versions of generative artificial intelligence (Claude, Gemini, and DeepSeek) on practical EUNACOM tests. The results will provide empirical evidence on the current capabilities of these tools in specific Chilean clinical contexts, guide future technological and educational improvements, and contribute meaningfully to the global debate on the potential, limitations, and educational role AI should play in contemporary medical training.

2. Métodos

We evaluated the performance of different artificial intelligences in the field of medical knowledge and clinical skills. To do so, we conducted a quantitative, cross-sectional, and descriptive study using a series of questions based on the EUNACOM, classified according to their content. The AI models used were: Gemini 2.5, Deepseek 3, and Claude Sonnet 4.

Undergraduate Question Dataset

For professional, ethical, and academic reasons, it is not possible to access exact original copies of the EUNACOM. Therefore, we relied on two official reconstructions (13,14) to build exams as close as possible to the real EUNACOM. We compiled three mock exams, each with 180 multiple-choice questions (five options, only one correct). The number of questions per medical category followed the EUNACOM standards (15), with the overall distribution across the three exams as follows: 6.9% surgery, 5.6% specialties, 42.8% internal medicine, 18.3% obstetrics and gynecology, 14.3% pediatrics, 7.4% psychiatry, and 4.8% public health.

Question Classification

To classify the questions, we used three sub-items to later deepen the analysis of our results. The categorization followed the same approach as Carrasco et al. (16) in 2023 regarding the Spanish Internal Medical Resident Examination:

- Medical area: Internal medicine, pediatrics, obstetrics and gynecology, surgery, psychiatry, specialties (including otorhinolaryngology, ophthalmology, and dermatology), and public health.
- Question type: Clinical case or medical knowledge, depending on whether the question required clinical skills to analyze a case and respond based on the context, or explicit factual knowledge, respectively.
- Focus: For clinical case-type questions, we classified them according to whether the question targeted diagnosis, treatment, or follow-up processes.

Prompting and Application of the AIs

To administer the questions to the AI models, we evaluated the strengths, weaknesses, and requirements of each system to obtain optimal performance. The same prompt was used for all three AIs. It was designed to be clear and direct, encouraging reasoning and critical thinking, and formatted for easy review. To facilitate question delivery, they were sent in sections of no more than 45 questions per message. Each exam was conducted in separate chats and different accounts to avoid exceeding free usage limits and prevent memory effects between responses. We avoided using images to reduce interference with text comprehension by the models. Additionally, we did not disclose the true nature of the exam, to avoid refusal to respond due to internal AI policies against participating in official assessments.

We used Gemini 2.5, DeepSeek 3, and Claude Sonnet 4 to answer the three EUNACOM mock exams in June 2025. Each exam was completed three times by each AI, using the following prompt:

> "Read each question carefully and analyze the options.
> Reason step by step and explain why each option is correct or incorrect.
> At the end, give your final answer by writing:
> Final answer: (Letter)"

Data Analysis

Data analysis was performed using Julius AI (Pro Plan). We calculated the percentage of correct answers for each mock exam and set the passing score at >51%, in line with the EUNACOM standard (10). We also calculated consistency across attempts and the error rate by question type and EUNACOM category. Statistical analysis was performed using Chi-square tests.

Rationale for Excluding OpenAI Models

We opted not to include ChatGPT (GPT-4 and GPT-4V), despite its status as a widely studied benchmark model, due to methodological and practical limitations. Specifically, access to the advanced versions requires a subscription that was unavailable for this study, and the free version imposes usage restrictions that prevent the full and repeated execution of the evaluations. Furthermore, the model's internal policies may block responses to evaluative content. Consequently, the focus was placed on less explored emerging models (Gemini 2.5, DeepSeek 3, and Claude Sonnet 4) to contribute new comparative evidence. We acknowledge this limits comparability with prior research and recommend incorporating OpenAI models in future investigations.

Methodological Limitations

A principal limitation of this study is the reliance on reconstructed exams rather than the official EUNACOM, which may not adequately represent the complexity, phrasing, or distribution of question types in the authentic examination, thereby constraining the validity and generalizability of our findings. Furthermore, the absence of clinical-image items impedes the assessment of critical visual diagnostic skills. Future research should endeavor to incorporate officially authorized exams and high-fidelity image-based items, ideally through collaboration with examination boards, to more rigorously evaluate AI performance across the full spectrum of clinical reasoning.

Ethical Considerations

The Research Ethics Committee on Human Subjects of the Faculty of Medicine, University of Chile, determined that this study did not raise ethical concerns requiring institutional review board oversight. We used authorized EUNACOM practice exams from the University of Chile's School of Medicine and freely available reconstructions (13-14), as access to the official exam is restricted. The authors declare no conflicts of interest.

3. Results

The three AIs evaluated successfully passed all three EUNACOM mock exams across their 27 total attempts, as shown in table 1. Test 3 was where all models achieved their best performance, while Test 1 posed the greatest challenge. Claude stood out as the model with the best overall performance, achieving an average accuracy of 89%, outperforming Gemini and DeepSeek. Furthermore, all AIs showed high consistency across different attempts, with Claude being the most consistent, averaging 96.36%. A chi-square test revealed a statistically significant difference between the models (p = 0.025), indicating that Claude's superior performance was not due to chance.

Table 1. Correct answers of Deepseek 3, Claude Sonnet 4 Y Gemini 2.5 on each of the EUNACOM a drills (each with 180 multiple-choice questions) per attempt.

EUNACOM drill and attempt	Correct answers provided by each version of IA, n (%)			
	Test	Deep Seek 3	Claude sonnet 4	Gemini 2.5
Drill 1	1	133 (73.89)	133 (73.89)	149 (82.78)
	2	165 (91.67)	167 (92.78)	167 (92.78)
	3	164 (91.11)	173 (96.11)	173 (96.11)
Drill 2	1	129 (71.67)	150 (83.33)	140 (77.78)
	2	169 (93.89)	169 (93.89)	166 (92.22)
	3	132 (90.0)	170 (94.44)	170 (94.44)
Drill 3	1	135 (75.0)	147 (81.67)	138 (76.67)
	2	169 (93.89)	165 (91.67)	145 (80.56)
	3	167 (92.78)	170 (94.44)	171 (95.0)

This finding is further illuminated by analyzing the 95% confidence intervals for each individual test. The intervals frequently overlapped, suggesting that in many head-to-head attempts the performance gaps were not statistically significant, particularly between Claude and Gemini. This means Claude's average performance was higher, but the statistically significant differences varied from test to test.

In total, the three exams included 357 clinical case-based questions and 183 medical knowledge questions. Interestingly, all three AIs performed better on clinical questions (average error rate: 10.52%) compared to general medical knowledge questions (average error rate: 15.67%).

From a percentage perspective, Claude had the lowest overall error rate (7–16%) and stood out particularly in Internal Medicine and Psychiatry. DeepSeek showed a more heterogeneous performance: it achieved excellent results in Surgery, but showed significant weaknesses in Public Health and Obstetrics-Gynecology. Gemini placed in the middle: it did not exhibit the marked weaknesses DeepSeek showed in Public Health but still had relatively high error rates in Surgery and Obstetrics-Gynecology. These findings may suggest different potentials for each AI depending on the subject area. However, given the limited number of questions per sub-specialty (e.g., only ~5% in Public Health), such trends should be interpreted with caution. Drawing firm conclusions about model superiority in specific areas (e.g., DeepSeek in Surgery, Claude in Psychiatry) may be premature. Therefore, we frame these observations as preliminary trends rather than definitive outcomes.

In clinical case questions, Claude achieved the best accuracy rates in diagnosis (0.92), follow-up (0.83), and treatment (0.90) subtypes. In contrast, DeepSeek had the largest performance gap in follow-up questions, followed by Gemini 2.5.

Overall, Claude proved to be the most consistent and reliable model for tackling the EUNACOM, while DeepSeek and Gemini would require specific adjustments by medical area, particularly in Public Health and surgical specialties.

Detailed data on average performance by medical area and question type are presented in tables 2 through 7.

4. Discusión

This study shows that DeepSeek 3, Claude Sonnet 4, and Gemini 2.5 successfully passed the EUNACOM, with Claude Sonnet 4 standing out for its superior performance, consistency, and accuracy. Claude's statistical advantage suggests significant differences in medical reasoning capabilities among the models.

Interestingly, all three AIs performed better on clinical case-based questions than on general medical knowledge questions. These differences are likely due to the outdated nature of the mock exams used compared to the current knowledge held by the AIs, the complexity level of the questions, and the type of content emphasized in their training. AI models tend to be more exposed to clinical practice than to basic theory, and the narrative structure of clinical questions aligns better with how these models process language (17).

All versions showed competent performance across various medical specialties, with Claude Sonnet 4 and Gemini 2.5 excelling in Internal Medicine and Psychiatry, and DeepSeek 3 performing best in Surgery. However, variations in accuracy by specialty may be attributed to each field's inherent complexity, use of specific terminology, or the format of the questions—factors that may be more or less represented in the training data of each model.

In clinical case questions, Claude continued to lead across the three clinical subtypes, especially in diagnosis, while DeepSeek and Gemini showed the largest gaps in follow-up questions. These differences may be due to Claude's stronger ability to manage the logical sequence of clinical decision-making, whereas DeepSeek and Gemini may have had less exposure or training in continuity-of-care scenarios, where integrating prior information and anticipating clinical behavior is essential.

Our results partially align with those reported by Rojas et al. (12), who also observed that more recent versions of AI models—namely ChatGPT-4 and 4V—outperformed earlier versions in the EUNACOM, particularly in clinical questions. As in our study, they also found better performance in clinical scenarios than in general knowledge questions, suggesting a shared trend among language models: stronger performance in contextual clinical reasoning compared to memorization of theoretical content. However, unlike Rojas et al., who highlighted the newer models' performance in Surgery, our results showed better outcomes for Claude Sonnet 4 and Gemini 2.5 in Internal Medicine and Psychiatry, and for DeepSeek 3 in Surgery.

Among the strengths of this study is the use of three recently released large language models (Gemini 2.5, Claude Sonnet 4, and DeepSeek 3), each representing distinct design approaches and reasoning mechanisms. We also employed a broad and diverse question bank validated academically and distributed according to the official EUNACOM area proportions, enhancing the relevance, thematic representativeness, and validity of our findings.

As for limitations, the question bank used did not include image-based questions, which are an important component of the real EUNACOM exam. This limits the direct extrapolation of results to the official test setting. Moreover, due to ethical and access constraints, we could not use exact copies of the real exam, relying instead on authorized reconstructions with varying difficulty levels, which may have affected the consistency and comparability of results across tests and models. Consistent with external validity considerations, our findings should therefore be interpreted cautiously for image-rich domains (e.g., radiology, dermatology) and under authentic test conditions. Despite these constraints, internal reliability within our dataset was high, with stable cross-attempt consistency observed across models (see table 2), reinforcing the robustness of comparative patterns observed.

Given that all models surpassed the passing threshold and performed best in contextual clinical reasoning, we propose a phased, curriculum-aligned introduction focused on formative, not high-stakes, use: (1) Preclinical years: AI-assisted case-based learning (CBL) and problem-based learning (PBL) "co-tutors" to prompt differential diagnosis reasoning and justifyability of answers; (2) Clinical rotations: structured "AI-augmented ward rounds" using de-identified vignettes aligned with national guidelines (GES/AUGE), where students must critique the model's rationale and reconcile it with local protocols; (3) EUNACOM preparation: faculty-curated item banks where AIs generate rationales and counter-explanations, with automatic tagging of cognitive level and competency mapping to EUNACOM domains; (4) OSCE preparation: simulation of history-taking and clinical reasoning stations (no image items) with rubric-based feedback; (5) Faculty development: short courses on AI literacy, prompt engineering for education, and assessment design to ensure alignment with WFME standards; and (6) Equity measures: institutional access to approved AI tools to avoid widening gaps between students with and without paid access. These actions align with recent guidance that positions AI as a tutor/assessment aid while emphasizing competency frameworks, educator upskilling, and explicit acknowledgment of AI use (18).

In conclusion, the evidence shows Claude as the most robust and reliable model for the comprehensive demands of the EUNACOM exam overall, but with different profiles of efficacy depending on the medical area and type of questioning. These patterns, while informative, should be considered exploratory and subject to further validation.

Ethical and regulatory considerations.

Institutional policies should anchor AI use in health professions education to international health ethics guidance for large multi-modal models (LMMs), including transparency, accountability, data governance, bias mitigation, and human oversight, especially when generating clinical advice or feedback (19). For Chile, two developments are salient: (i) the updated National AI Policy and risk-based AI bill announced in 2024, which frames acceptable and high-risk uses (20); and (ii) the new Personal Data Protection Law (Law 21.719), published on December 13, 2024 and entering fully into force in December 2026, creating a Data Protection Authority and stronger obligations (privacy-by-design, incident reporting, lawful bases, sensitive-data safeguards) (21). Medical schools should therefore implement: model disclosure to learners; audit trails for AI-assisted tasks; strict prohibition of AI in summative high-stakes assessments (unless validated and proctored); dataset localization to Chilean clinical practice; and privacy impact assessments and DPA-aligned governance for any processing of student/clinical data.

Theoretically, our findings reinforce that current AIs tend to perform better on case-based contextual questions than on isolated theoretical content, likely due to the nature of their training and architecture. This pattern supports using AIs to scaffold diagnostic reasoning and metacognition, while preserving human-led instruction for foundational biomedical knowledge and image-dependent tasks, until multimodal capabilities are locally validated in Spanish/Chile-specific settings. Recent policy and education guidance similarly recommend a human-centred, age-/stage-appropriate deployment that prioritizes safety, equity, and methodological rigor.

Future studies should include question banks with clinical images to better reflect the complexity of the real EUNACOM and assess the visual reasoning abilities of different models. It is also suggested to evaluate, when possible, the performance of AIs on official exams, and compare it with that of medical students and practicing physicians, to establish a human benchmark. Additionally, experimental work should test proctoring solutions, disclosure norms, and item-generation pipelines that meet Chile's evolving regulatory landscape, and report psychometrics (item difficulty/discrimination, DIF by specialty, reliability) for AI-augmented assessments. Comparative trials of prompting strategies and system configurations in weaker areas (Public Health, follow-up) should be paired with faculty-development interventions to determine whether performance gains translate to learner outcomes (22).

Additionally, it is important to investigate how different prompting strategies and model configurations affect clinical reasoning performance—especially in areas where significant weaknesses were observed (such as Public Health or follow-up scenarios). Finally, exploring the integration of these tools into formal educational programs would be valuable, assessing their potential to enhance learning and preparation of future physicians—without encouraging technological dependence or misuse in official evaluations.

5. Conclusions

- This study demonstrates that Claude Sonnet 4, Gemini 2.5, and DeepSeek 3 are capable of consistently passing the EUNACOM, exceeding the required passing threshold and showing particularly strong performance in clinical reasoning. Claude Sonnet 4 proved to be the most accurate and consistent model, while DeepSeek 3 stood out in Surgery, and Gemini 2.5 maintained a balanced performance.
- While these AIs show great potential as educational support tools, they still present limitations in specific areas and in theoretical knowledge-based questions. Their implementation in educational or certification contexts should be regulated to maximize benefits and minimize risks. In practical terms for Chile, we recommend limiting AI to supervised formative uses (CBL/PBL co-tutor, OSCE practice, rationale generation for item review), ensuring equity of access, and explicitly prohibiting unsupervised AI use in summative high-stakes examinations until locally validated under Chile's risk-based AI policy and new data protection regime. Institutions should adopt clear disclosure rules, audit trails, and privacy-by-design processes consistent with WHO/UNESCO guidance and the forthcoming enforcement of Law 21.719.
- Finally, although external validity is constrained by the absence of image-based items and the use of reconstructed exams, the high internal reliability observed supports the robustness of comparative conclusions. As multimodal evaluation with clinical images becomes feasible and governance matures, these systems may evolve from adjuncts to validated components of medical education, always as complements to, not replacements for, human expertise.

Funding: No funding was required.

Conflicts of interest: No conflicts of interest are declared.

Contributions of authors: Anaís Aracelly Lancellotti Guajardo estuvo a cargo de la conceptualización del estudio, mientras que ella junto a Vicente Alberto Edgardo Jesús Silva Arroyo y Marcos Jeremías Giovanny Vera Cartes desarrollaron la metodología, recolectaron los datos y redactaron, revisaron y editaron el borrador original. El análisis formal fue realizado por Óscar Jerez Yañez, Anaís Lancellotti, Vicente Silva, Marcos Vera y Álvaro Andrés Herrera Alcaíno. Finalmente, la supervisión general del proyecto estuvo a cargo de Óscar Jerez y Álvaro Herrera, quien además facilitó el acceso a documentos oficiales necesarios para la elaboración del estudio.

Referencias

1. Heng JJY, Teo DB, Tan LF. The impact of Chat Generative Pre-trained Transformer (ChatGPT) on medical education. *Postgrad Med J* **2023,** 99(1176),1125–1127. https://doi.org/10.1093/postmj/qgad058
2. 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. https://doi.org/10.2196/46885
3. OpenAI. GPT-4V(ision) system card. In: OpenAI Research. OpenAI **2023**. https://openai.com/research/gpt-4v-system-card. Accessed July 20, 2025.
4. Anthropic. Claude Opus 4. In: Claude Models. Anthropic **2023**. https://www.anthropic.com/claude/opus. Accessed July 20, 2025.
5. Google Cloud. Gemini 2.5 Flash. In: Generative Models Documentation. Google Cloud **2025**. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash. Accessed July 20, 2025.
6. DeepSeek. DeepSeek-V3-0324 Release. In: DeepSeek API Docs. DeepSeek **2025**. https://api-docs.deepseek.com/news/news250325. Accessed July 20, 2025.
7. Institute of Knowledge Engineering. Trust and interest in AI applications in the health sector. In: Health with AI. Institute of Knowledge Engineering n.d. https://www.iic.uam.es/lasalud/confianza-e-interes-en-la-aplicacion-de-la-ia-en-el-sector-salud/. Accessed July 20, 2025.
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* **2023,** 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

9. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): Promising horizons for AI in clinical medicine. *Clin Pract* **2023,** 13, 1460–1487. https://doi.org/10.3390/clinpract13060130
10. Eunacom. Official Regulations. In: National Medical Knowledge Exam. **2023.** https://www.eunacom.cl/reglamentacion/NormativaOficial.pdf. Accessed July 21, 2025.
11. Chile. Law No. 20.261: Creates a national unified medical knowledge exam, incorporates specified posts into the Senior Public Management System, and amends Law No. 19,664. *Diario Oficial de la República de Chile.* **2008** Apr 19. https://www.bcn.cl/leychile/navegar?idNorma=270584.
12. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploration of the performance of versions 3.5, 4, and 4 with vision of ChatGPT in the Chilean National Medical Exam: Observational study. *JMIR Med Educ* **2024,** 10, e55048. https://doi.org/10.2196/55048
13. Guevara DR. 180 EUNACOM-style questions. In: Study material for the medical exam. DR Guevara **2024**. https://www.drguevara.cl/material-y-pruebas-gratis/180-preguntas-tipo-eunacom/. Accessed July 21, 2025.
14. Faculty of Medicine. Official EUNACOM mock exam. In: Academic Portal, University of Chile. University of Chile **2024**. https://medicina.uchile.cl/. Accessed July 21, 2025.
15. EUNACOM. Sample official questions. In: Official website of the National Medical Knowledge Exam. National Health Service **2023**. https://www.eunacom.cl/contenidos/muestra.html. Accessed July 21, 2025.
16. Carrasco JP, García E, Sánchez DA, Porter E, De La Puente L, Navarro J, Cerame A. Is "ChatGPT" capable of passing the 2022 MIR exam? Implications of artificial intelligence in medical education in Spain. *Revista Española de Educación Médica,* **2024,** 4(1). https://doi.org/10.6018/edumed.556511
17. Gaspar Casal Foundation. Clinical decisions and artificial intelligence. In: Publications on health innovation. Gaspar Casal Foundation **2020**. https://fundaciongasparcasal.org/wp-content/uploads/2020/12/Decisiones-clinicas-e-inteligencia-artificial.pdf. Accessed July 21, 2025.
18. Masters K, MacNeil H, Benjamin J, Carver T, Nemethy K, Valanci-Aroesty S, et al. Artificial intelligence in health professions education assessment: AMEE Guide No. 178. *Med Teach.* **2025,** 47(9), 1410-1424. doi:10.1080/0142159X.2024.2445037.
19. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: Large Multi-Modal Models. WHO Guidance. World Health Organization, 18 Jan. **2024**, www.who.int/publications/i/item/9789240084759 . Accessed October 6, 2025.
20. Chile. Law No. 21.719: Regulates the protection and processing of personal data and creates the Data Protection Agency. Official Gazette of the Republic of Chile. **2024** Dec 13. Available from: https://www.bcn.cl/leychile/navegar?idNorma=1209272. Accessed October 6, 2025.
21. Chamber of Deputies of Chile. Bill regulating artificial intelligence systems [Docket No. 16.821-19]. Valparaíso; **2024** May 7. Available from: https://www.camara.cl/legislacion/ProyectosDeLey/tramitacion.aspx?prmBOLETIN=16821&prmID=17429. Accessed October 6, 2025.
22. Miao F, Holmes W. Guidance for generative AI in education and research. Paris, France: UNESCO; **2023**. https://unesdoc.unesco.org/ark:/48223/pf0000386693. Accessed October 6, 2025.

Table 2. Consistency analysis (%).

IAs	Average consistency (%)	Standard deviation	Minimal consistency	Maximum consistency
Deep seek 3	95.8	11.79	33.33	100.0
Claude S4	96.36	10.8	33.33	100.0
Gemini 2.5	93.77	13.78	33.33	100.0

Table 3. Error rate by question type (%).

IAs	Clinical case (%)	Medical knowledge (%)
Claude S4	8.9	14.21
Deep seek	12.27	16.94
Gemini	10.39	15.85

Table 4. Error rate by EUNACOM area (%).

IAs	Surgery	Specialties	Internal Medicine	Obstetrics Gynecology	Pediatrics	Psychiatry	Public health
Deep seek 3	6.3	12.2	12.0	18.5	16.5	10.0	23.1
Claude S4	12.6	13.3	8.7	15.8	10.4	8.3	9.0
Gemini 2.5	15.3	15.6	10.4	17.2	10.8	8.3	11.5

Table 5. Overall accurate in each test.

IAs	Correct answers	Test 1(%)	Test 2(%)	Test 3(%)	Global (%)
Deep seek 3	1393	73.52	93.15	91.3	85.99
Claude S4	1444	79.63	92.78	95.0	89.14
Gemini 2.5	1419	79.07	88.52	95.19	87.59

Table 6. Accuracy in case clinic questions by subtype.

IAs	Deep seek 3	Claude S4	Gemini 2.5
Diagnosis	0.89	0.92	0.9
Monitoring	0.62	0.83	0.71
Tratment	0.87	0.9	0.89

Table 7. Confidence intervals for AIs performance.

EUNACOM drill & attempt	Test	Deep Seek 3	Claude Sonnet 4	Gemini 2.5
Drill 1	1	121.4 – 144.6	121.4 – 144.6	139.1 – 158.9
	2	157.7 – 172.3	160.1 – 173.9	160.1 – 173.9
	3	156.4 – 171.6	167.8 – 178.2	167.8 – 178.2
Drill 2	1	117.2 – 140.8	140.2 – 159.8	129.0 – 151.0
	2	162.6 – 175.4	162.6 – 175.4	158.9 – 173.1
	3	120.3 – 143.6*	163.9 – 176.1	163.9 – 176.1
Drill 3	1	123.6 – 146.4	136.8 – 157.2	126.8 – 149.2
	2	162.6 – 175.4	157.7 – 172.3	134.5 – 155.5
	3	160.1 – 173.9	163.9 – 176.1	165.3 – 176.7