Revista Española
de Educación Médica

Frontier Large Language Models on the 2026 Spanish MIR Examination: A Multimodal Cross-Sectional Evaluation.

Manuel Carpio Salmerón [1]*, Carlos Carazo-Casas [2], Pau Benito [3], Clemente García [4], Jesús Alonso-Carrillo [5], Beatriz Carratalá [6], Georgios Kyriakos [1], Pablo González-Castro [7].

[1], Department of Endocrinology and Nutrition, Santa Lucía University General Hospital, Cartagena, Spain. [2], Department of Otolaryngology, Ramón y Cajal Hospital, Madrid, Spain. Email: Carloscarazocasas@gmail.com ORCID: 0000-0001-7568-7140, [3], Department of Preventive Medicine and Epidemiology, Clinical Institute of Medicine and Dermatology (ICMiD), Hospital Clínic de Barcelona, Barcelona, Spain. Email: paubb13@gmail.com ORCID: 0000-0002-2480-9133, [4], Department of Radiology, Hospital Morales Meseguer, Murcia, Spain. Email: clemente292@gmail.com ORCID: 0009-0001-6672-2714, [5], Department of Internal Medicine, Hospital 12 de Octubre, Madrid, Spain. Email: j.alonsoc@faculty.amireducacion.com, [6], Innovation and Digital Projects Academic Department, Healthcademia, Madrid, Spain. Email: b.carratala@healthcademia.com, [7], Department of Plastic and Reconstructive Surgery, Virgen del Rocío University Hospital, Sevilla, Spain. Email: Pabloglez13@gmail.com ORCID: https://orcid.org/0009-0003-0077-126X

* Correspondence: manuelcarpios99@gmail.com
Received: 14/5/26; Accepted: 11/6/26; Published: 15/6/26

Abstract.

Introduction: Large language model-based AI systems are increasingly used in medical education, but their educational value in multilingual, high-stakes examination settings remains insufficiently defined. **Objective**: To assess the accuracy of latest-generation AI systems on the official 2026 Spanish Médico Interno Residente (MIR) examination and compare their performance with the Healthcademia chatbot, AMIR faculty experts, and average candidates. **Materials and Methods:** A cross-sectional quantitative evaluation analyzed all 200 valid questions from the official MIR 2026 Version 0 booklet. Performance was assessed using automated pipelines for text-only items and manual multimodal input for the 24 image-associated questions. Two instruction conditions, neutral and expert-role, were tested, with additional effort-level stratification for GPT-5.2. Accuracy was measured against the final official answer key. Human comparison groups included AMIR faculty experts and the average MIR candidate. **Results**: Several latest-generation AI systems showed very high accuracy on the MIR 2026 examination. The highest-performing configuration, GPT-5.2 high effort without expert-role instruction, achieved 199/200 correct answers (99.5%). Gemini 3 Flash with expert-role instruction achieved 198/200 (99.0%). AMIR faculty experts achieved 194/200 (97.0%), whereas the average candidate achieved 131/200 (65.7%). On image-associated questions, several multimodal configurations achieved 24/24 correct answers when the corresponding image was provided. **Conclusions**: Under the conditions evaluated in this study, several AI configurations achieved near-perfect accuracy on a complete, high-stakes national medical licensing examination. These findings support licensing-style examinations as benchmarks for educational AI and suggest potential use in supervised feedback and self-assessment. Performance on multiple-choice items should not be interpreted as evidence of autonomous clinical reasoning.

Keywords: Artificial intelligence, Academic research, Medical education, Medical residents, MIR examination.

Resumen.

Introducción: Las herramientas de inteligencia artificial basadas en modelos de lenguaje se utilizan cada vez más en educación médica, pero su valor educativo en exámenes multilingües y de alto impacto sigue estando insuficientemente definido. **Objetivo:** Evaluar la precisión de herramientas de

IA de última generación en el examen oficial Médico Interno Residente (MIR) español de 2026 y comparar su rendimiento con el chatbot educativo de Healthcademia, profesores de academia MIR y el candidato medio. **Materiales y métodos:** Una evaluación cuantitativa transversal analizó las 200 preguntas válidas del cuadernillo oficial Versión 0 del MIR 2026. El rendimiento se evaluó mediante procesos automatizados para los ítems de solo texto y entrada multimodal manual para las 24 preguntas asociadas a imagen. Se probaron dos condiciones de instrucciones a la IA, neutra y de rol experto, con estratificación adicional por nivel de esfuerzo para GPT-5.2. La precisión se midió frente a la plantilla oficial final de respuestas. Los grupos de comparación humanos incluyeron profesores de academia MIR y el candidato medio. **Resultados:** Varias herramientas de IA de última generación mostraron una precisión muy elevada en el examen MIR 2026. La configuración con mejor rendimiento, GPT-5.2 alto esfuerzo sin instrucción de rol experto, obtuvo 199/200 respuestas correctas (99,5 %). Gemini 3 Flash con instrucción de rol experto obtuvo 198/200 (99,0 %). Los profesores de academia MIR obtuvieron 194/200 (97,0 %), mientras que el candidato medio obtuvo 131/200 (65,7 %). En las preguntas asociadas a imagen, varias configuraciones multimodales obtuvieron 24/24 respuestas correctas cuando se proporcionó la imagen correspondiente. **Conclusiones:** En las condiciones evaluadas, varias herramientas de IA de última generación alcanzaron una precisión casi perfecta en un examen nacional de acceso a la residencia médica. Estos hallazgos apoyan el valor de los exámenes habilitantes como referentes para evaluar herramientas de inteligencia artificial aplicadas a la educación médica y sugieren utilidad para retroalimentación y autoevaluación supervisadas. El rendimiento en preguntas de opción múltiple no debe interpretarse como prueba de razonamiento clínico autónomo.

Palabras clave: Inteligencia artificial, Investigación académica, Educación médica, Residentes médicos, Examen MIR.

1. Introduction

Generative artificial intelligence has rapidly become part of medical education, particularly in question explanation, literature summarization, practice-item generation, and self-directed study. These uses create opportunities for faster feedback and broader access to educational support, but they also raise concerns about inaccurate explanations, overconfident responses, and variable performance across languages, specialties, and question formats. In high-stakes examination settings, the educational value of these tools depends not only on answer accuracy, but also on how reliably they can support supervised feedback and learning. Previous Spanish MIR studies have shown that general-purpose language models can pass or perform strongly on this examination, including early evaluations of ChatGPT and GPT-4 in the 2022 and 2023 calls (1-2). Subsequent work confirmed substantial differences across model generations and reported high accuracy for GPT-4o and OpenAI o1 on the 2024 MIR examination (3–5). These findings suggest rapid improvement, but evidence remains limited for latest-generation AI systems, multimodal item handling, and evaluation settings in which the exact examination content was released only after administration.

The Spanish competitive medical specialty access examination, commonly referred to as the "MIR exam", offers a particularly informative setting for such evaluation. The MIR exam is the mandatory gateway to publicly funded residency training in Spain and is taken annually by tens of thousands of candidates. The exam consists of a multiple-choice test with four response options per item, typically including 200 scored questions plus 10 reserve items that may replace annulled questions. In the 2026 examination, a total of 12,366 training positions were offered across the health professions, including 9,276 positions in medicine (6). The combination of high stakes, broad clinical coverage, and standardized scoring makes the MIR exam a useful benchmark for exam-oriented medical knowledge, although not a comprehensive measure of clinical competence.

From a benchmarking perspective, the MIR exam also has a practical advantage that is uncommon in medical datasets. Examination questions are not publicly available before the examination date, and the official exam booklets and provisional answer keys are released only after the test is administered (7). This policy substantially reduces the likelihood that candidates or third parties can curate targeted training material in advance, thereby supporting a more realistic evaluation of model performance on previously unseen content. In 2026, the benchmark value of the dataset was further strengthened by the release of bibliographic justifications accompanying the provisional answer key, enabling direct traceability between each question and its supporting evidence.

The MIR examination has also evolved toward increasingly contextualized items. Recent calls include clinical vignettes supported by laboratory data, imaging, electrocardiograms, clinical photographs, pathology, microbiology, and other diagnostic materials. This structure is relevant for AI evaluation because some items require integration of textual and visual information, whereas others may remain answerable from the vignette and response options alone. Therefore, performance on MIR items may reflect a combination of medical knowledge, recognition of recurring examination patterns, visual interpretation, and use of multiple-choice cues.

The present study evaluates the performance of contemporary frontier large language models (GPT-5.2, Claude 4 family, and Gemini 3 Flash) and the internal Healthcademia chatbot on the MIR 2026 examination (8–10). These systems differ in training paradigms, tool access, reasoning controls, and intended use cases, creating an opportunity to compare performance across different frontier general-purpose models and an internal educational chatbot comparator.

The primary objective was to quantify and compare accuracy on MIR 2026 multiple-choice questions across models under standardized conditions. Secondary objectives were to describe performance according to image availability, prompting strategy, and self-reported certainty during verification prompting. By leveraging a high-stakes Spanish-language examination with post hoc official release, this study aims to provide a current and methodologically transparent benchmark with potential relevance for medical learners, educators, and assessment researchers.

2. Methods

2.1 Study Design. A cross-sectional quantitative evaluation study was conducted to compare the performance of multiple generative artificial intelligence systems on the 2026 Spanish medical residency entrance examination, the Médico Interno Residente (MIR). The evaluation estimated item-level accuracy against the official final answer key and compared performance under two administration conditions: a neutral, no-instruction condition and an expert-role prompting condition. A standardized, item-independent workflow was used to minimize contextual carryover and improve reproducibility.

2.2 Data Sources and Dataset Construction. The dataset was constructed exclusively from the officially released Version 0 MIR 2026 exam booklet and the official final answer key (11). The MIR 2026 materials were publicly released on January 25, 2026 at 01:00 (Spain local time). Model runs were executed on January 27, 2026 at 01:10 (Spain local time). The MIR dataset included 210 questions, of which 25 were image-associated. Each item was transformed into a structured, question-level record including the question stem, four answer options labeled 1 through 4, an indicator for image association, and the final correct option. No image assets were missing in the source materials used for this study, and all image-associated questions were evaluated with their corresponding images when the evaluated system supported image input.

2.3 Inclusion and Exclusion Criteria. All questions in the official Version 0 booklet with a corresponding final official key were eligible. The annulled questions in the final answer key, 1 image-based question and 5 non-image questions, were considered invalid and were excluded from the accuracy denominator. These were replaced by reserve questions, all of which were non-image-based. A total of 10 reserve questions were available, of which 6 were ultimately used to replace the challenged items.

2.4 AI Systems Evaluated. The evaluated systems included GPT-5.2 accessed via the OpenAI API in two models: Fast and Thinking. For clarity, GPT-5.2 Fast is designated in this manuscript as low effort. GPT-5.2 Thinking was assessed in two configurations: without extended internal reasoning (medium effort) and with extended internal reasoning enabled (high effort). Additional evaluated systems included the Claude family (Haiku, Sonnet, Opus), Gemini 3 Flash, and the internal Healthcademia chatbot comparator. The Healthcademia chatbot is a proprietary institutional system, and details regarding its model architecture and training data are not publicly disclosed. It was included as a pragmatic educational comparator representing the performance of an internal chatbot available in the Healthcademia environment, rather than as a fully reproducible frontier-model benchmark. For each system, access dates, publicly available version identifiers when available, interface conditions, and relevant runtime metadata were recorded. When model interfaces allowed parameter control, available sampling parameters were set to their most deterministic configuration to reduce stochastic variability; otherwise, default platform settings were used and documented. No evaluated system was given access to the official answer key, bibliographic justifications, or previous item-level responses during testing.

2.5 AMIR faculty experts. The benchmark consisted of 51 AMIR faculty experts, all of whom were physicians involved in MIR preparation and had clinical experience in their respective specialties. This group was selected because these instructors are familiar with the structure, timing, and reasoning patterns of the MIR examination. Participants answered independently, without artificial intelligence tools, books, online searches, or bibliographic references. Aggregate and, when available, individual expert scores were summarized descriptively.

*2.6 Automated Processing Pipeline for Text-Only Questions***.** To enable standardized, high-throughput processing of MIR items without images, a Python-based pipeline was developed to connect an Excel workbook containing the structured question dataset to each model's application interface. The pipeline used the OpenAI SDK for GPT-5.2 and the corresponding official or platform-supported software development kits and interfaces for the other evaluated systems. This approach provided an automated, reliable, and rapid method for submitting questions, collecting responses, and storing outputs in a consistent tabular format. For each question, the pipeline initiated an item-independent interaction and captured structured outputs for downstream scoring and analysis.

2.7 Manual Processing for Image-Associated Questions. Because image-associated questions required multimodal input and consistent control over the interaction environment, these items were processed manually, one question at a time. For each image-associated question, a new conversation was started to preserve item independence. Internet access and model memory were disabled during these runs to reduce external information retrieval and cross-item contamination. The question text was provided together with the corresponding image, and outputs were recorded using the same structured fields as in the automated pipeline.

2.8 Prompting Conditions and Response Capture. Two administration conditions were used for each question, and all prompts were delivered in Spanish. The English translations below are provided for readability. In the neutral condition, only the question and answer options were provided, without additional instruction. In the expert-role condition, each question was preceded by the following

Spanish instruction: "Eres un especialista con experiencia en preparar el examen MIR. Antes de responder, razona paso a paso internamente, pero no muestres ese razonamiento." An English translation is: "You are a specialist physician with experience preparing candidates for the MIR exam. Before answering, reason step by step internally, but do not show that reasoning." Within each question session, a standardized multi-turn sequence was used to capture the selected option, a confirmation prompt, and a clinical rationale, all in Spanish. First, the model selected an answer option, recorded as the primary response. Immediately afterward, the model was asked "¿Cuánta seguridad tienes del 1-100?" which translates to "How confident are you, on a scale from 1 to 100?" and the response was recorded as a confidence-related signal. Third, the model was asked to justify the choice using clinical reasoning: "Describe la explicación de por qué esta respuesta es la correcta con razonamiento clínico," translated as "Describe why this answer is correct using clinical reasoning." The selected option was used for scoring, while confirmation and rationale text were retained for secondary analyses. Additionally, for image-based questions, the models were also evaluated by presenting the question without the image.

2.9 Outcomes. The primary outcome was accuracy, defined as agreement between the model's selected option and the final official answer key, calculated over valid (non-annulled) questions. Secondary outcomes included comparisons between neutral and expert-role conditions, stratified performance on image-associated versus non-image items where applicable, the confirmation response elicited by "How confident are you, on a scale from 1 to 100?" obtained through verification prompting in text-only conditions (non-image questions plus image questions without visual input). Additionally, accuracy and confidence were also compared for image-based questions when the image was provided versus when it was omitted.

2.10 Statistical Analysis. Primary analyses were descriptive. Accuracy was summarized as counts and percentages overall and by prespecified subgroups. Mean self-reported certainty and standard deviations were calculated separately for correct and incorrect answers in the verification-prompting analysis. Because the main purpose of the study was benchmark description rather than formal hypothesis testing, and because several comparisons involved small subgroup sizes (particularly image-associated items), no prespecified inferential statistical tests were used. Results should therefore be interpreted as descriptive performance estimates rather than as definitive tests of superiority between closely performing systems. Statistical analysis was performed using R version 4.4.1.

2.11 Data Management and Quality Control. All prompts, question text, images when applicable, raw outputs, extracted answer selections, and metadata (model name, variant or mode, and run date) were stored in a structured dataset. Standardized parsing rules were applied to extract the selected option from outputs. Quality-control checks were implemented to identify transcription errors in stems or options, option-mapping inconsistencies, and malformed outputs that did not contain a valid 1-to-4 response.

2.12 Ethics. This study involved responses provided by AMIR faculty experts and the use of aggregated, de-identified benchmark data from MIR candidates. The study used publicly released examination materials and did not involve patient data or identifiable personal information. A formal internal determination was issued by Healthcademia stating that the study qualified for an exception to informed consent requirements.

3. Results

A total of 200 valid MIR 2026 questions were analyzed after exclusion of annulled items. The final dataset included 176 text-only questions and 24 image-associated questions. Table 1 summarizes accuracy across frontier large language models, the Healthcademia chatbot, AMIR faculty experts,

and the average student benchmark, stratified by question type, image provision, and prompting condition.

3.1 Overall Accuracy. Frontier large language models showed high accuracy on the MIR 2026 examination (figure 1). The highest accuracy was achieved by GPT-5.2 "high effort" without prompting, with 199/200 correct answers (99.5%), followed by Gemini 3 Flash with prompting at 198/200 (99.0%) and GPT-5.2 "high effort" with prompting at 197/200 (98.5%). AMIR faculty experts achieved 194/200 correct answers (97.0%), the Healthcademia chatbot achieved 171/200 (85.5%), and the average student benchmark achieved 131/200 (65.7%). Full model-level results are shown in table 1.

3.2 Performance According to Question Type and Image Provision. On the 176 non-image questions, most frontier models exceeded 96% accuracy, whereas Claude Haiku variants ranged from 90.3% to 91.5%. For the 24 image-associated questions, several configurations achieved 24/24 correct answers when images were provided, including GPT-5.2 high-effort variants, Claude Opus 4.5 without prompting, and AMIR faculty experts (figure 2). When images were omitted, performance generally declined, particularly in smaller models. Detailed values are provided in table 1.

3.3 Effect of Prompting Strategy and Model Variants. The expert-role prompting strategy produced mixed effects across models (figure 3). Gemini 3 Flash improved from 98.0% to 99.0%, whereas several Claude configurations performed similarly or worse with prompting. These findings indicate that the effect of a single expert-role prompt was model-dependent.

3.4 Comparison with Human Benchmarks. The study incorporated two key human reference groups to contextualize LLM performance: a panel of AMIR faculty experts and the average MIR examinee.

3.5 AMIR faculty Benchmark. AMIR faculty experts achieved an overall accuracy of 194/200 correct answers (97.0%). This result was broken down as follows: 170/176 (96.6%) on text-only questions; 24/24 (100.0%) on image-associated questions when the diagnostic images were provided; and 24/24 (100.0%) on image-associated questions when the diagnostic images were not provided.

3.6 Average Student Benchmark. The performance of the average student was calculated from aggregated data from thousands of real MIR 2026 candidates who voluntarily submitted their answer sheets through the EstimAMIR application. They achieved 131/200 correct answers (65.7%), including 116/176 (66.1%) on questions without images and 15/24 (62.5%) on image-linked items.

3.7 Direct Comparison: LLMs vs. Humans. All frontier LLM configurations outperformed the average student benchmark. The best-performing GPT-5.2 and Gemini 3 Flash configurations also exceeded the AMIR faculty benchmark by 2.5 and 2.0 percentage points, respectively. In image-associated questions with images provided, several LLMs matched the 100% accuracy observed in AMIR faculty experts.

3.8 Analysis of Response Certainty. In the verification-prompting analysis, mean self-reported certainty was higher for correct than for incorrect answers across model families (Table 2). The separation was largest in GPT-5.2 configurations and smaller in Claude Haiku variants, while Gemini 3 Flash showed the highest absolute certainty values. Figure 4 shows certainty values for correct and incorrect answers by model and LLM family.

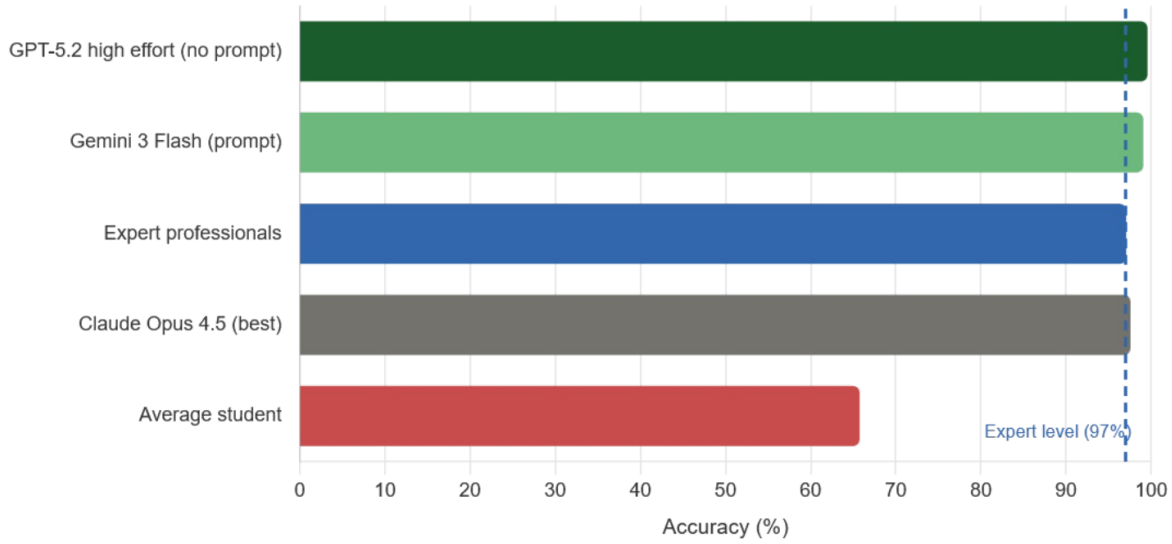

Figure 1. Global accuracy of frontier large language models, AMIR faculty experts, and average student. Several frontier configurations matched or exceeded the AMIR Faculty Benchmark on this benchmark, and all frontier models clearly outperformed the average-student benchmark.

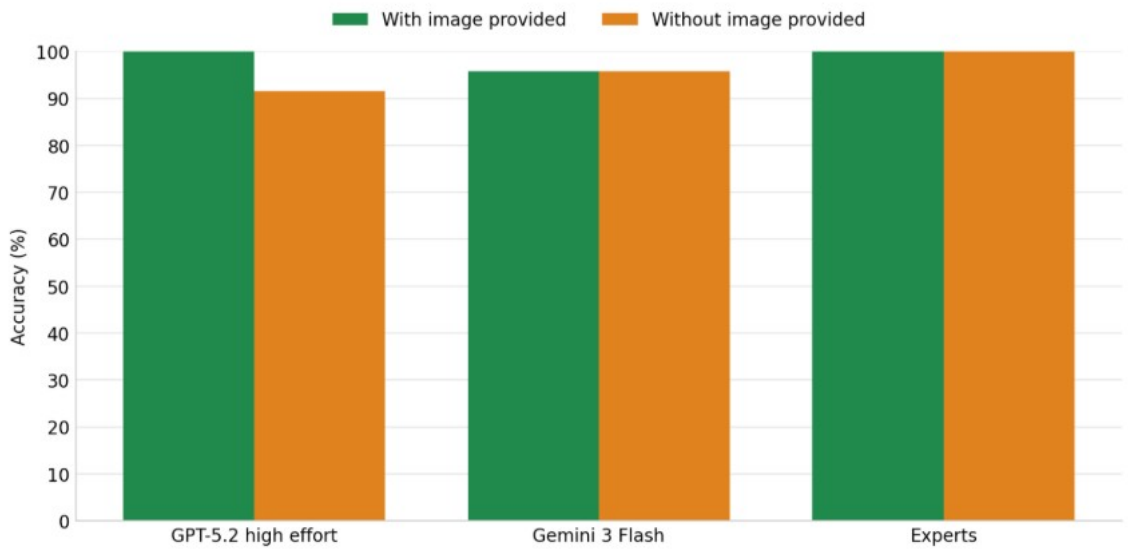

Figure 2. Accuracy on image-associated questions according to whether the diagnostic image was provided to multimodal models.

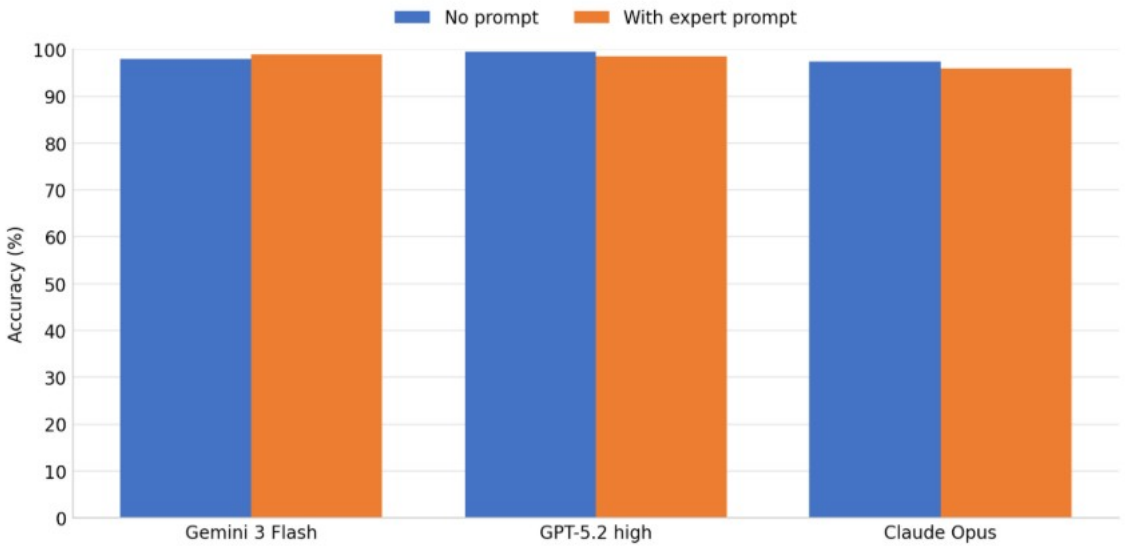

Figure 3. Impact of expert-role prompting on overall accuracy across LLM families.

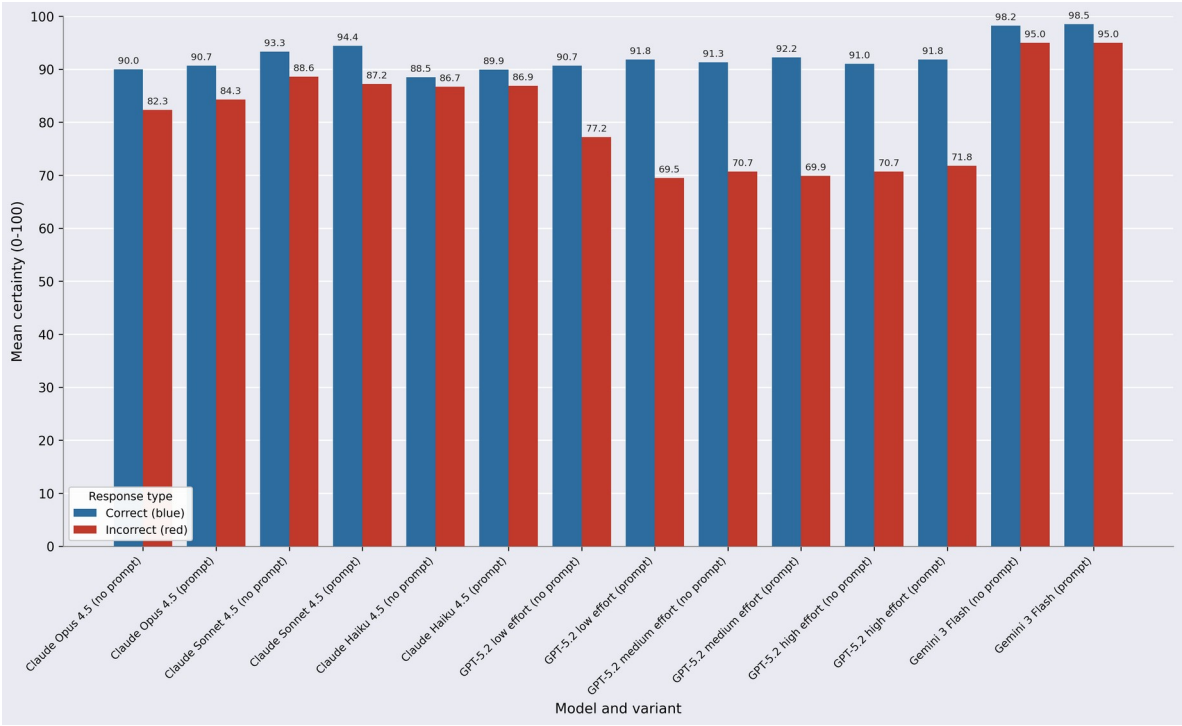

Figure 4. Self-reported certainty levels for correct versus incorrect answers by model and LLM family. Blue bars represent correct answers and red bars represent incorrect answers. Arrows highlight the particularly pronounced gap observed in GPT-5.2 across all effort levels.

4. Discussion

This study provides a contemporary benchmark of frontier LLMs on the MIR 2026 examination, a high-stakes Spanish national residency access test released only after examination day. Several frontier models achieved near-ceiling descriptive accuracy, with the best-performing configuration answering 199 of 200 valid questions correctly. These findings confirm very strong exam-oriented performance under standardized conditions, but they should be interpreted as performance on a multiple-choice benchmark rather than as evidence of autonomous clinical competence. The result is consistent with prior work showing rapid improvement across model generations and strong performance of newer systems on medical licensing examinations, including Spanish MIR-based evaluations (3–5).

The main contribution of the present study is therefore not merely to show that LLMs can "pass" a medical examination, which has already been reported for earlier systems, but to show how far performance has progressed in a particularly informative evaluation setting (3–5). The MIR exam is relevant as a benchmark because it is high stakes, spans multiple specialties, is written in Spanish, and was not publicly available before administration. These characteristics strengthen its value as a realistic assessment set compared with highly reused or openly circulating question banks, although they do not eliminate all possible sources of benchmark familiarity. In particular, strong performance may still reflect a mixture of broad medical knowledge, exposure to prior MIR-style questions, and skill in exploiting multiple-choice structure and contextual clues.

Our results also extend previous Spanish-language evidence (3, 5). Earlier work on ChatGPT in the MIR context suggested that advanced general-purpose LLMs could already achieve passing or strong performance in this examination environment, and a more recent evaluation on the MIR 2024 exam reported high accuracy for GPT-4o and OpenAI o1 (3, 5). Against that background, the present findings indicate a further upward shift in benchmark performance, with several 2026-era configurations clustering close to ceiling. This is also coherent with broader syntheses showing that GPT-4 class systems substantially outperform GPT-3.5 class systems in medical licensing examinations, while results remain sensitive to language, prompt design, and evaluation methodology (4).

At the same time, near-ceiling performance on MCQ examinations should not be conflated with reliable clinical reasoning in real care environments (12-13). Recent work has emphasized that high scores on medical QA benchmarks can coexist with limitations in metacognition, uncertainty estimation, and robustness when models are tested on more clinically grounded or less cue-rich tasks. For that reason, our findings are best interpreted as evidence of very strong exam-oriented knowledge retrieval and question-solving performance, not as proof that these systems are ready for unsupervised clinical use. This distinction is especially important in medicine, where the informational structure of multiple-choice examinations differs substantially from real diagnostic workflows (12–14).

Differences across model families were also informative. GPT-5.2 and Gemini 3 Flash occupied the top tier in our dataset, whereas Claude Opus and Sonnet remained strong but slightly lower, and Claude Haiku trailed the larger frontier systems. Because our analyses were descriptive and no formal superiority testing was prespecified, these differences should not be overinterpreted when models performed very closely. Even so, the ranking pattern is consistent with the broader literature in which performance varies materially across model generations, sizes, and optimization strategies, rather than all frontier systems behaving interchangeably (4-5, 15). From a practical perspective, our results suggest that "frontier LLM" is not a homogeneous category: performance claims should remain model-specific and configuration-specific (4, 15).

The findings on reasoning effort require similarly careful interpretation. In our study, the highest descriptive accuracy was obtained by the GPT-5.2 high-effort configuration, and the lower-effort variants performed slightly worse. However, our design does not permit causal attribution of this advantage to internal reasoning itself. The design did not observe latent reasoning traces, independently manipulate other system properties, or assess repeated-run stability. Accordingly, the appropriate interpretation is associative: within the tested GPT-5.2 configurations, higher-effort mode coincided with slightly better exam accuracy.

The prompting analysis likewise supports a restrained conclusion. Expert-role prompting produced mixed effects, with modest gains in some systems and neutral or adverse effects in others. This is directionally compatible with prior reviews suggesting that prompting can improve medical exam performance, but that the effect is neither uniform nor guaranteed across versions and tasks (4). In practical terms, our data argue against treating a single prompt template as universally beneficial. For benchmark studies, prompt specification should therefore be considered part of the intervention rather than a trivial implementation detail (4, 15).

The multimodal results are among the most interesting aspects of the study, but they also require nuanced interpretation. When images were provided, several top-performing systems achieved perfect accuracy on the 24 image-associated questions (Claude Opus 4.5 no prompt, GPT 5.2 "medium effort" no prompt, GPT 5.2 "high effort" with and without prompt). This suggests that contemporary multimodal systems can perform very strongly on small sets of clinically oriented examination items that include visual information (16). However, the companion finding that performance often remained high even when the image was withheld indicates that many image-linked MIR questions may still be answerable from the textual vignette and response options alone. Thus, our data do not allow us to quantify how much of the observed performance reflects genuine visual understanding as opposed to successful inference from text, exam conventions, or answer-option structure. This limitation is especially relevant given the small number of image-associated items and the growing literature emphasizing both the promise and the methodological complexity of medical multimodal LLM evaluation (16).

This observation has implications for the interpretation of image-based examination items more generally. In educational benchmarking, "image-associated" should not automatically be assumed to mean "image-dependent." Some questions may use images as confirmatory or supporting material rather than as the indispensable source of discriminative information. Future studies would benefit from item-level adjudication of visual necessity by human experts, ideally classifying questions as image-essential, image-supportive, or image-redundant before comparing multimodal and text-only performance.

The certainty analysis generated a secondary finding of potential educational interest. Across all model families, mean self-reported certainty was higher for correct than for incorrect responses, and the separation appeared especially marked in GPT-5.2. This suggests that elicited confidence may contain some discriminative signal about answer correctness in exam-style interactions (17–19). However, that signal should not be equated with good calibration (13, 17-18). Recent studies show that LLM confidence expression is often imperfectly aligned with true correctness probabilities, and that models may remain overconfident despite high accuracy (13, 19). Our study did not assess calibration formally, did not test operational thresholds, and included very few errors for the best-performing systems, which makes the incorrect-answer estimates relatively unstable. Accordingly, self-reported certainty should be viewed as an exploratory auxiliary feature rather than as a reliable safeguard or decision rule (17–19).

The comparison with human benchmarks deserves emphasis. All frontier models clearly outperformed the average-student benchmark, and some also exceeded the AMIR faculty benchmark used in this study. This finding is educationally relevant because it suggests that top-tier LLMs may support exam-oriented activities such as generating explanations, checking answers, or surfacing differential reasoning around MCQs (4, 14). Yet even this comparison must be interpreted carefully. Human examinees and AMIR faculty experts are not operating under the same cognitive conditions as an LLM; they bring time constraints, fatigue, and different incentives, and they solve the exam for a different purpose. Moreover, AMIR faculty experts are evaluated here only on the same narrow endpoint of item correctness, not on broader dimensions such as explanation quality, error recognition, or adaptability to incomplete information. The result is therefore best read as a benchmark comparison, not as a comprehensive statement that LLMs have surpassed medical experts in general medical reasoning.

Comparison with Prior Work. The present findings are consistent with continued improvement in LLM performance on medical examinations. In the most directly comparable MIR study, Benito et al. reported 90.9% accuracy for GPT-4o and 93.2% for OpenAI o1 on the 2024 examination (5). Earlier MIR evaluations reported lower performance for GPT-4 and other systems (20-21). In this context, the near-ceiling performance observed in the present study suggests further progress in examination-oriented accuracy. However, direct numerical comparisons across studies remain limited by differences in examination year, item composition, model access conditions, prompting strategy, and scoring workflow. The current results should therefore be understood as a contemporary benchmark under the specific conditions tested here, rather than as definitive evidence of superiority over all previously evaluated systems.

Strengths and Implications. This study has several strengths. First, it evaluated a complete, high-stakes national medical examination rather than a selected subset of items. Second, the assessment was performed shortly after official release of the exam materials, reducing the likelihood of prior direct exposure to the exact item set. Third, the study used an item-independent workflow and included both text-only and image-associated questions. Fourth, performance was contextualized using both an AMIR Faculty Benchmark and a student benchmark, which provides a more interpretable frame of reference than model-only comparisons. The findings may have practical relevance for medical education when interpreted as support for supervised formative use rather than as a substitute for teachers or clinical assessment. In MIR preparation, high-performing LLMs could be integrated into learning platforms to generate immediate explanations, identify recurrent error patterns, suggest targeted remediation, and create follow-up questions linked to the same curricular domain. In undergraduate medical education, these tools may be more useful when students are asked to critique AI-generated rationales, compare alternative explanations, and identify missing clinical information, rather than simply accept a selected answer. This framing preserves the role of educators while using LLMs to expand feedback, self-assessment, and deliberate practice. At the same time, the results highlight a limitation of multiple-choice benchmarks: strong performance on standardized items should not be equated with broad clinical competence, professional judgment, communication skills, uncertainty management, or safe real-world decision-making.

Limitations. Several limitations should be acknowledged. First, although the MIR examination offers a valuable real-world benchmark, some degree of conceptual overlap with previous years is likely, which may advantage systems trained on broad medical and educational corpora. Second, the AMIR Faculty Benchmark represents a highly prepared reference group with specific experience in MIR-style questions, and its performance should not be interpreted as representative of all practicing physicians. Third, the student benchmark was derived from aggregated data from candidates who voluntarily submitted responses through the EstimAMIR application and may therefore not perfectly represent the full distribution of all MIR examinees. Fourth, although efforts were made to

standardize item presentation and preserve independence across questions, full equivalence between platforms and interfaces cannot be guaranteed. Differences in model access, interface behavior, hidden defaults, and multimodal implementation may have influenced observed performance. In addition, the Healthcademia chatbot could not be fully characterized because its architecture and training data are proprietary; therefore, comparisons involving this comparator should be interpreted as pragmatic and institution-specific rather than fully reproducible model-to-model comparisons. Fifth, the image-associated subset was relatively small, limiting precision in subgroup interpretation. Sixth, the study focused primarily on accuracy and did not include a formal qualitative evaluation of explanation quality, hallucination patterns, calibration, or clinically unsafe reasoning. Finally, the field evolves rapidly, and the present results reflect model behavior during a narrow time window in January 2026; performance may differ in later versions or under different deployment conditions.

5. Conclusions

- Frontier large language models showed very high accuracy on the MIR 2026 examination, with several configurations achieving near-perfect performance under the tested conditions.
- The strongest models matched or exceeded the AMIR faculty benchmark included in this study and substantially outperformed the average-student benchmark.
- These findings support the use of complete licensing-style examinations as benchmarks for educational applications of LLMs, particularly in supervised feedback, self-assessment, and exam-oriented tutoring.
- However, performance on multiple-choice items should not be interpreted as equivalent to real-world clinical competence, professional judgment, or readiness for unsupervised clinical use.

Funding: The expenses associated with the preparation and publication of this study were funded by Healthcademia.

Declaration of conflict of interest: This study received funding from Healthcademia. Beatriz Carratalá is affiliated with Healthcademia, which developed the internal Healthcademia chatbot comparator evaluated in this study. Five authors of the current manuscript (PB, MCS, BC, CCC and PGC) are also coauthors of the comparator study by Benito et al. on the MIR 2024 examination, which is cited as a prior MIR-based LLM evaluation. The remaining authors declare no relevant financial or non-financial interests related to the content of this article.

Author contributions: MCS and CCC contributed equally to this work and share first authorship. MCS and CCC were responsible for the conception and design of the study, data collection, data interpretation, and drafting of the manuscript. PB contributed to the methodological design and statistical analysis. CG contributed to data acquisition and interpretation of radiological findings. JAC and PGC contributed to the clinical interpretation of the data and critically revised the manuscript for important intellectual content. BC and GK contributed to study development and manuscript revision. All authors read and approved the final version of the manuscript.

Acknowledgments: The authors would like to thank the students who entered their MIR examination answer sheets into the EstimAMIR application. The authors also wish to acknowledge the interest and collaboration of the AMIR faculty experts in this project, particularly the instructors, whose responses to the MIR 2026 examination questions made it possible to establish the AMIR faculty benchmark used in this study.

Data availability: The data from this study are available upon reasonable request.

6. References

1. Carrasco JP, García E, Sánchez DA, Porter E, De La Puente L, Navarro J, Cerame A. ¿Es Capaz "ChatGPT" de Aprobar El Examen MIR de 2022? Implicaciones de La Inteligencia Artificial En La Educación Médica En España. *Rev Esp Edu Med* **2023**, 4, https://doi.org/10.6018/edumed.556511
2. Cerame A, Juaneda J, Estrella-Porter P, De La Puente L, Navarro J, García E, Sánchez DA, Carrasco JP. ¿Es Capaz GPT-4 de Aprobar El MIR 2023? Comparativa Entre GPT-4 y ChatGPT-3 En Los Exámenes MIR 2022 y 2023. *Rev Esp Edu Med* **2024**, 5, https://doi.org/10.6018/edumed.604091.

3. Leis A, Mayer M-A, Mayer A. Bridging AI and Medical Expertise: ChatGPT's Success on the Medical Specialization Residency Admission Exam in Spain. In *Studies in Health Technology and Informatics*; Andrikopoulou E, Gallos P, Arvanitis TN, Austin R, Benis A, Cornet R, Chatzistergos P, Dejaco A, Dusseljee-Peute L, Mohasseb A, Natsiavas P, Nakkas H, Scott P, Eds.; IOS Press, **2025**. ISBN 978-1-64368-596-0.
4. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, Kiuchi T. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res* **2024**, 26, e60807, https://doi.org/10.2196/60807.
5. Benito P, Isla-Jover M, González-Castro P, Fernández Esparcia PJ, Carpio M, Blay-Simón I, Gutiérrez-Bedia P, Lapastora MJ, Carratalá B, Carazo-Casas C. GPT-4o and OpenAI O1 Performance on the 2024 Spanish Competitive Medical Specialty Access Examination: Cross-Sectional Quantitative Evaluation Study. *JMIR Med Educ* **2026**, 12, e75452–e75452, https://doi.org/10.2196/75452.
6. Ministry of Health of Spain. Specialized Healthcare Training. Madrid: Ministry of Health of Spain, **2026**. Available online: https://fse.sanidad.gob.es/fseweb/#/principal/escritorio (accessed on 31 May 2026).
7. Ministry of Health of Spain. Order SND/928/2025, of 14 August, Approving the Offer of Places and the Call for 2025 Selective Tests for Access in 2026 to Specialized Health Training Places for University Degree/Bachelor's/Diploma Programmes in Medicine, Pharmacy, Nursing and in the Fields of Psychology, Chemistry, Biology and Physics. **2025**. Available online: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2025-17059 (accessed on 31 May 2026).
8. Google AI for Developers. Gemini 3 Developer Guide. Google AI for Developers, **2026**.
9. Anthropic. Models Overview — Claude API Docs. **2026**. Available online: https://platform.claude.com/docs/en/about-claude/models/overview (accessed on 31 May 2026).
10. OpenAI. Introducing GPT-5.2. **2025**. Available online: https://openai.com/es-ES/index/introducing-gpt-5-2/ (accessed on 31 May 2026).
11. Ministry of Health of Spain. Exam Booklets — Previous Calls. Specialized Healthcare Training. **2026**. Available online: https://fse.sanidad.gob.es/fseweb/#/principal/datosAnteriores/cuadernosExamen (accessed on 31 May 2026).
12. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of Large Language Models in Clinical Problem-Solving Arising from Inflexible Reasoning. *Sci Rep* **2025**, 15, 39426, https://doi.org/10.1038/s41598-025-22940-0.
13. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large Language Models Lack Essential Metacognition for Reliable Medical Reasoning. *Nat Commun* **2025**, 16, 642, https://doi.org/10.1038/s41467-024-55628-6.
14. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, Hou L, Clark K, Pfohl SR, Cole-Lewis H, et al. Toward Expert-Level Medical Question Answering with Large Language Models. *Nat Med* **2025**, 31, 943–950, https://doi.org/10.1038/s41591-024-03423-7.
15. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating Large Language Models and Agents in Healthcare: Key Challenges in Clinical Applications. *Intelligent Medicine* **2025**, 5, 151–163, https://doi.org/10.1016/j.imed.2025.03.002.
16. Nam Y, Kim DY, Kyung S, Seo J, Song JM, Kwon J, Kim J, Jo W, Park H, Sung J, et al. Multimodal Large Language Models in Medical Imaging: Current State and Future Directions. *Korean J Radiol* **2025**, 26, 900, https://doi.org/10.3348/kjr.2025.0599.
17. Omar M, Agbareia R, Glicksberg BS, Nadkarni GN, Klang E. Benchmarking the Confidence of Large Language Models in Answering Clinical Questions: Cross-Sectional Evaluation Study. *JMIR Med Inform* **2025**, 13, e66917–e66917, https://doi.org/10.2196/66917.
18. Savage T, Wang J, Gallo R, Boukil A, Patel V, Safavi-Naini SAA, Soroush A, Chen JH. Large Language Model Uncertainty Proxies: Discrimination and Calibration for Medical Diagnosis and Treatment. *Journal of the American Medical Informatics Association* **2025**, 32, 139–149, https://doi.org/10.1093/jamia/ocae254.
19. Bentegeac R, Le Guellec B, Kuchcinski G, Amouyel P, Hamroun A. Token Probabilities to Mitigate Large Language Models Overconfidence in Answering Medical Questions: Quantitative Study. *J Med Internet Res* **2025**, 27, e64348–e64348, https://doi.org/10.2196/64348.
20. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, Montejo R, Aguinaga-Ontoso E, Barach P, Aguinaga-Ontoso I. Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. *Clinics and Practice* **2023**, 13, 1460–1487, https://doi.org/10.3390/clinpract13060130.

21. Vera CL, Picon IF, Nunez MT del V, Gandia JAG, Ancillo A de L, Arroyo VR, Figueredo CM. Evaluating Large Language Models on the Spanish Medical Intern Resident (MIR) Examination 2024/2025: A Comparative Analysis of Clinical Reasoning and Knowledge Application. *ArXiv Preprint* **2025**, arXiv:2503.00025. https://arxiv.org/abs/2503.00025

Table 1. Accuracy (number and percentage of correct answers) of evaluated large language models, chatbot, AMIR faculty experts, and average student on the MIR 2026 examination, stratified by image linkage and prompting condition (n=200 total questions).

Study arm / Type of LLM	No image (n=176) N (%)	Image provided (n=24) N (%)	All questions (n=200) N (%)
Claude LLMs			
Claude Opus 4.5 (no prompt)	171 (97.2)	24 (100.0)	195 (97.5)
Claude Opus 4.5 (prompt)	170 (96.6)	20 (83.3)	192 (96.0)
Claude Opus 4.6 (no prompt)	170 (96.6)	22 (91.7)	192 (96.0)
Claude Opus 4.6 (prompt)	171 (97.2)	22 (91.7)	193 (96.5)
Claude Sonnet 4.5 (no prompt)	172 (97.7)	22 (91.7)	194 (97.0)
Claude Sonnet 4.5 (prompt)	169 (96.0)	21 (87.5)	190 (95.0)
Claude Haiku 4.5 (no prompt)	161 (91.5)	17 (70.8)	178 (89.0)
Claude Haiku 4.5 (prompt)	159 (90.3)	16 (66.7)	175 (87.5)
GPT LLMs			
GPT-5.2 “high effort” (no prompt)	175 (99.4)	24 (100.0)	199 (99.5)
GPT-5.2 “high effort” (prompt)	173 (98.3)	24 (100.0)	197 (98.5)
GPT-5.2 “medium effort” (no prompt)	172 (97.7)	24 (100.0)	196 (98.0)
GPT-5.2 “medium effort” (prompt)	171 (97.2)	23 (95.8)	194 (97.0)
GPT-5.2 “low effort” (no prompt)	173 (98.3)	20 (83.3)	193 (96.5)
GPT-5.2 “low effort” (prompt)	174 (98.9)	22 (91.7)	196 (98.0)
Gemini LLMs			
Gemini 3 Flash (prompt)	175 (99.4)	23 (95.8)	198 (99.0)
Gemini 3 Flash (no prompt)	173 (98.3)	23 (95.8)	196 (98.0)
Other			
Chatbot Healthcademia	151 (85.8)	20 (83.3)	171 (85.5)
AMIR faculty experts	170 (96.6)	24 (100.0)	194 (97.0)
Average student	116 (66.1)	15 (62.5)	131 (65.7)

Table 2. Accuracy and mean certainty (SD) by model in verification prompting (text-only conditions: non-image questions plus image-based questions presented without the corresponding image).

Study arm / Type of LLM	Correct answers (N)	Correct answers (%)	Mean certainty correct (SD)	Mean certainty incorrect (SD)
Claude Opus 4.5 (no prompt)	193	96.5	90.0 (4.7)	82.3 (10.3)
Claude Opus 4.5 (prompt)	193	96.5	90.7 (4.7)	84.3 (7.3)
Claude Sonnet 4.5 (no prompt)	193	96.5	93.3 (4.8)	88.6 (4.8)
Claude Sonnet 4.5 (prompt)	191	95.5	94.4 (4.5)	87.2 (8.3)
Claude Haiku 4.5 (no prompt)	180	90.0	88.5 (6.1)	86.7 (6.1)
Claude Haiku 4.5 (prompt)	179	89.5	89.9 (5.5)	86.9 (6.5)
GPT-5.2 “low effort” (no prompt)	195	97.5	90.7 (7.9)	77.2 (11.2)
GPT-5.2 “low effort” (prompt)	196	98.0	91.8 (5.7)	69.5 (14.4)
GPT-5.2 “medium effort” (no prompt)	194	97.0	91.3 (6.5)	70.7 (12.4)
GPT-5.2 “medium effort” (prompt)	192	96.0	92.2 (5.3)	69.9 (10.9)
GPT-5.2 “high effort” (no prompt)	197	98.5	91.0 (7.0)	70.7 (16.0)
GPT-5.2 “high effort” (prompt)	195	97.5	91.8 (7.3)	71.8 (10.3)
Gemini 3 Flash (no prompt)	196	98.0	98.2 (2.5)	95.0 (0.0)
Gemini 3 Flash (prompt)	198	99.0	98.5 (2.3)	95.0 (0.0)