Universidad Católica de Santa María

Facultad de Medicina Humana

Escuela Profesional de Medicina Humana

Análisis comparativo de los puntajes obtenidos por Chat GPT-4o, Gemini Advanced y Copilot al aplicarlas en el examen del Concurso Nacional de Admisión al Residentado Médico, realizado por el Consejo Nacional de Residentado Medico (CONAREME), de los años 2017-2024. Perú.

Tesis presentada por los Bachilleres:

Calderon Paiva, Gabriel Alberto

ORCID: 0000-0002-5668-9295

Medina Suarez, Joshua Andree

ORCID: 0000-0003-1933-2454

para optar el Título Profesional de Médico Cirujano

Asesora:

Dra. Apaza Tososcahua de Palma, Sandra Leonor

ORCID: 0000-0002-1234-6098

Arequipa - Perú

2025

UCSM-ERP

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

MEDICINA HUMANA

TITULACIÓN CON TESIS

DICTAMEN APROBACIÓN DE BORRADOR

Arequipa, 21 de Enero del 2025

Dictamen: 014230-C-EPMH-2025

Visto el borrador del expediente 014230, presentado por:

2016800121 - MEDINA SUAREZ JOSHUA ANDREE
2018810231 - CALDERON PAIVA GABRIEL ALBERTO

Titulado:

ANÁLISIS COMPARATIVO DE LOS PUNTAJES OBTENIDOS POR CHAT GPT-4O, GEMINI ADVANCED Y COPILOT AL APLICARLAS EN EL EXAMEN DEL CONCURSO NACIONAL DE ADMISIÓN AL RESIDENTADO MÉDICO, REALIZADO POR EL CONSEJO NACIONAL DE RESIDENTADO MEDICO (CONAREME), DE LOS AÑOS 2017-2024. PERÚ.

Nuestro dictamen es:

APROBADO

Titulo Profesional/Titulo de Segunda Especialidad/Grado Académico a optar:

MEDICO CIRUJANO

30401320 - FARFAN DELGADO MIGUEL FERNANDO
DICTAMINADOR

29296240 - MONTANCHEZ CARAZAS EDGAR CUSTODIO GASPAR
DICTAMINADOR

40374914 - ALPACA CANO CESAR GUILLERMO
DICTAMINADOR

Análisis comparativo de los puntajes obtenidos por Chat GPT-4o, Gemini Advanced y Copilot al aplicarlas en el examen del Concurso Nacional de Admisión al Residentado Médico, realizado por el Consejo N

INFORME DE ORIGINALIDAD

16%	15%	7%	6%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	downtoearthnw.com Fuente de Internet	1%
2	Submitted to Universidad Católica de Santa María Trabajo del estudiante	1%
3	tesis.ucsm.edu.pe Fuente de Internet	1%
4	www.unprg.edu.pe Fuente de Internet	1%
5	repositorio.ucsm.edu.pe Fuente de Internet	<1%
6	1library.co Fuente de Internet	<1%
7	revistas.um.es Fuente de Internet	<1%
8	www.coursehero.com Fuente de Internet	<1%
9	core.ac.uk Fuente de Internet	<1%
10	Submitted to Universidad del Istmo de Panamá Trabajo del estudiante	<1%

Dedicatoria

Dedicamos este trabajo a Dios, fuente infinita de sabiduría y amor, quien nos ha guiado en este camino. A Él dedicamos este trabajo, fruto de su inspiración y bendición.

A nuestras familias, cuyo amor incondicional y constante apoyo han sido la base de nuestro crecimiento personal y profesional. Nuestra mayor fuente de inspiración.

A nuestros padres, por enseñarnos el valor del esfuerzo, la disciplina y la perseverancia.

A mi tío, quien ha sabido ser un apoyo incondicional durante mi carrera y para la elaboración de esta tesis.

También dedicamos esta tesis a todos los médicos que han sido parte de nuestra formación que, con su vocación y compromiso, inspiran a continuar investigando y que gracias a sus enseñanzas conforman parte de nuestra consciencia y ser médico. Su labor nos motiva a aportar nuestro granito de arena al desarrollo de la Medicina.

A nuestros amigos, quienes con su amistad sincera y compañía inigualable han hecho de esta etapa un viaje inolvidable. Gracias por ser un ancla en los momentos más difíciles de la carrera. Su amistad es de los mayores tesoros que nos dio la Medicina.

Finalmente, dedicamos este proyecto a nuestros compañeros de carrera en formación, con la esperanza de que el conocimiento aquí compartido sirva como una herramienta útil en su camino hacia la excelencia profesional.

Agradecimientos

A la Facultad de Medicina Humana de la Universidad Católica de Santa María, por brindarnos una formación integral que nos ha preparado no solo como profesionales competentes, sino también como ciudadanos comprometidos con el bienestar de la sociedad. Su enfoque en el desarrollo humano y académico nos ha inspirado a perseguir la excelencia en cada paso de nuestro camino.

A los doctores y docentes de la facultad, por su invaluable dedicación, enseñanzas y consejos. Su compromiso con nuestra educación, así como su interés en nuestro crecimiento profesional y personal, han sido fundamentales para la culminación de este proyecto.

A la Dra. Sandra Leonor Apaza Tososcahua de Palma, por su liderazgo, paciencia y apoyo constante como mentora y asesora de esta tesis. Su entusiasmo por la innovación tecnológica y su compromiso con el desarrollo de la medicina fueron clave para la realización de este trabajo.

Finalmente, agradecemos a nuestras familias y amigos, cuyo apoyo incondicional y palabras de aliento nos dieron la fuerza necesaria para completar este proyecto. Su compañía y comprensión han sido nuestro pilar a lo largo de este proceso.

Epígrafe

"La curiosidad ilumina el camino correcto a cualquier cosa en la vida. Si no tienes curiosidad, es cuando tu cerebro está empezando a morir"

- **V. S. Ramachandran**

RESUMEN

Introducción: La inteligencia artificial (IA) ha revolucionado la educación médica y la práctica clínica, proporcionando herramientas innovadoras para la resolución de preguntas complejas. En el contexto del Concurso Nacional de Admisión al Residentado Médico en Perú, es necesario evaluar y comparar el desempeño de distintos modelos de IA para determinar su aplicabilidad y utilidad en escenarios médico-educativos.

Objetivo: Comparar los puntajes obtenidos al aplicar los modelos de inteligencia artificial Chat GPT-4o, Gemini Advanced y Copilot al examen del Concurso Nacional de Admisión al Residentado Médico 2017-2024 de Perú.

Material y métodos: Se realizó un estudio comparativo de los resultados obtenidos al aplicar los tres modelos de inteligencia artificial en los exámenes comprendidos en el periodo 2017-2024. Las preguntas fueron clasificadas en seis áreas temáticas: Ciencias Básicas, Medicina Interna, Ginecología y Obstetricia, Pediatría, Cirugía y Salud Pública. Se calcularon porcentajes de aciertos en cada examen por cada modelo de IA, y se realizó un análisis estadístico de varianza (ANOVA) y pruebas post hoc para determinar diferencias significativas.

Resultados*:* El análisis de varianza no identificó diferencias estadísticamente significativas entre los modelos evaluados ($p = 0.188$). Copilot obtuvo un promedio de aciertos y desviación de 89.81% (σ=3.43), mostrando una variabilidad moderada en los porcentajes obtenidos por año. ChatGPT-4º alcanzó un porcentaje de aciertos de 89.01% (σ=4.30) con una alta variabilidad en años específicos. Gemini Advanced alcanzó un porcentaje de aciertos de 86.18% (σ=2.26), presentando una mayor consistencia interanual. En el análisis por bloques temáticos, en los bloques de Ciencias Básicas y Medicina Interna se evidenció altos porcentajes en los aciertos y similitudes entre estos, en contraste con las áreas de Cirugía y Salud Pública, que tuvieron los más bajos porcentajes de aciertos y una mayor diferencia entre estos. Copilot fue el modelo que más destaco en el porcentaje de aciertos al evaluar los bloques temáticos, ocupando el mayor porcentaje en Ciencias Básicas y el menor en Salud Pública.

Conclusiones*:* Nuestro estudio evidenció que no existen diferencias significativas en el rendimiento de los modelos de inteligencia artificial al aplicarlas en los exámenes de Residentado Médico de Perú. Los tres modelos obtuvieron porcentajes satisfactorios de aciertos y demostraron su capacidad de proporcionar explicaciones fundamentadas, resaltando su potencial como herramientas complementarias en la educación médica.

Palabras clave: Inteligencia Artificial; Educación Médica; Residentado Médico.

ABSTRACT

Introduction: Artificial Intelligence (AI) has revolutionized medical education and clinical practice, providing innovative tools for solving complex questions. In the context of the National Medical Residency Admission Competition in Peru, it is necessary to assess and compare the performance of different AI models to determine their applicability and utility in medical-educational scenarios.

Objective: To compare the scores obtained by applying the AI models ChatGPT-4, Gemini Advanced, and Copilot to the National Medical Residency Admission Exam in Peru for the years 2017-2024.

Material and Methods: A comparative study was conducted on the results obtained from applying the three AI models to the exams from the 2017-2024 period. The questions were classified into six thematic areas: Basic Sciences, Internal Medicine, Gynecology and Obstetrics, Pediatrics, Surgery, and Public Health. The percentage of correct answers for each model on each exam was calculated, and an analysis of variance (ANOVA) and post hoc tests were performed to determine significant differences.

Results: The variance analysis did not identify statistically significant differences between the evaluated models ($p = 0.188$). Copilot achieved an average correct answer rate of 89.81% ($\sigma = 3.43$), showing moderate variability in the percentages obtained across years. ChatGPT-4o achieved a correct answer rate of 89.01% ($\sigma=4.30$), with high variability in specific years. Gemini Advanced achieved a correct answer rate of 86.18% ($\sigma=2.26$), demonstrating greater interannual consistency. In the thematic block analysis, high percentages of correct answers and similarities were evident in the Basic Sciences and Internal Medicine blocks, in contrast to the Surgery and Public Health areas, which had the lowest percentages of correct answers and a larger difference between them. Copilot was the model that performed the best in the thematic block evaluation, with the highest percentage in Basic Sciences and the lowest in Public Health.

Conclusions: Our study showed that there were no significant differences in the performance of the AI models when applied to the National Medical Residency Exams in Peru. All three models obtained satisfactory percentages of correct answers and demonstrated their ability to provide well-founded explanations, highlighting their potential as complementary tools in medical education.

Keywords: Artificial Intelligence; Medical Education; Medical Residency.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

INTRODUCCIÓN

En los últimos años, el desarrollo de la inteligencia artificial (IA) ha impactado profundamente diversos campos, incluidos el sector académico y biomédico. Su capacidad para procesar grandes volúmenes de datos y generar respuestas rápidas y precisas ha demostrado ser una herramienta valiosa en el apoyo a procesos complejos, desde la creación de contenidos hasta el diagnóstico médico. Modelos de lenguaje avanzados, como ChatGPT-4, Gemini Advanced y Copilot, destacan por su habilidad para interpretar y responder preguntas de alta complejidad, lo que los posiciona como posibles asistentes en la educación médica y la formación profesional.

El Examen del Concurso Nacional de Admisión al Residentado Médico (ENARM) en Perú es un proceso crítico para los médicos que desean acceder a una especialización. Este examen estandarizado evalúa conocimientos específicos y habilidades clínicas necesarias para desempeñarse en el ámbito de la salud. Sin embargo, el desempeño histórico en este examen ha revelado desafíos significativos, incluidos altos índices de reprobación entre los postulantes, lo que subraya la necesidad de explorar nuevas herramientas que optimicen la preparación académica.

En este contexto, el presente proyecto de tesis busca realizar un análisis comparativo del rendimiento de tres modelos de IA (ChatGPT-4, Gemini Advanced y Copilot) al responder las preguntas del Examen del Concurso Nacional de Residentado Médico en Perú. Creemos que estas herramientas pueden desempeñarse de manera efectiva en un entorno evaluativo complejo, aportando beneficios tanto en la preparación de los estudiantes como en el acceso equitativo a recursos educativos avanzados. A continuación, se describen los elementos clave que guían esta investigación.

Este estudio tiene como objetivo principal Comparar los puntajes obtenidos de los modelos de inteligencia artificial Chat GPT-4o, Gemini Advanced y Copilot al aplicar el examen del Concurso Nacional de Admisión al Residentado Médico 2017-2024 de Perú.

Se realizó un estudio observacional y comparativo, evaluando el rendimiento de los modelos al aplicarles los exámenes oficiales publicados por el CONAREME. El análisis incluirá técnicas estadísticas para establecer diferencias significativas en el desempeño.

Se espera que los modelos de IA logren un desempeño destacado. Los hallazgos podrían sentar las bases para futuras investigaciones sobre el uso de IA en la educación médica y su impacto en el desarrollo de competencias clínicas.

Este proyecto, además de explorar el potencial de la IA, busca motivar el interés en metodologías educativas emergentes que contribuyan a mejorar la formación de profesionales de la salud en Perú y otros contextos similares.

CAPÍTULO I
PLANTEAMIENTO TEÓRICO

1. PLANTEAMIENTO TEÓRICO

1.1. PROBLEMA DE INVESTIGACIÓN

1.1.1. Determinación del Problema

1.1.2. Enunciado del Problema

La creciente adopción de tecnologías basadas en inteligencia artificial (IA) ha generado interés en su potencial para asistir y mejorar la educación médica. Los grandes modelos de lenguaje, como GPT-4, Gemini Advanced y Copilot, tienen la capacidad de procesar grandes cantidades de información y generar respuestas útiles en entornos educativos. Sin embargo, se desconoce con precisión su rendimiento específico en exámenes complejos como el Examen del Concurso Nacional de Admisión al Residentado Médico en Perú. Es necesario evaluar si estas IA pueden desempeñarse de manera efectiva en contextos médicos, y su potencial en contribuir así a la formación médica profesional.

1.1.3. Descripción del Problema

1.1.3.1. Área del Conocimiento

Área General: Ciencias de la Salud.

Área Específica: Medicina Humana

Especialidad: Educación Médica

Línea: Inteligencia Artificial en Educación Médica

1.1.3.3. Análisis y operacionalización de variables e indicadores

Variable	Dimensiones	Indicador	Unidad / Categoría	Escala
Modelo de IA (VD)	Tipo de modelo de Inteligencia Artificial	Tipo de modelo de Inteligencia Artificial	ChatGPT-4.0, Gemini Advanced, Copilot	Cualitativa Nominal
Calificación Obtenida (VI)	Preguntas de los Exámenes por año de publicación	Puntaje obtenido por el modelo al resolver las preguntas de los exámenes	Porcentaje (%) de aciertos	Cuantitativa De Razón
	Preguntas de Bloque Temático de Ciencias Básicas por año de examen			
	Preguntas de Bloque Temático de Medicina Interna por año de examen			
	Preguntas de Bloque Temático de Cirugía por año de examen			
	Preguntas de Bloque Temático de Ginecología y Obstetricia por año de examen			
	Preguntas de Bloque Temático de Pediatría por año de examen			
	Preguntas de Bloque Temático de Salud Pública por año de examen			

1.1.3.4. Tipo de Investigación

Aplicada

1.1.3.5. Diseño de Investigación

Observacional

1.1.3.6. Nivel de Investigación

Exploratorio

1.1.4. Justificación

1.1.4.1. Justificación Científica

La evaluación del rendimiento de modelos de inteligencia artificial avanzados, como GPT-4, Gemini y Copilot, en un examen especializado como el Examen de Residentado Médico tiene un valor científico significativo. Al explorar cómo estas IA pueden procesar y resolver problemas clínicos complejos, se abre un campo de estudio que combina la medicina con la informática. El entendimiento profundo de cómo estas herramientas pueden mejorar el acceso a la información y el aprendizaje en entornos médicos podría contribuir a nuevas metodologías educativas y diagnósticas, basadas en el uso de modelos de lenguaje de gran escala, Además, esta investigación puede proporcionar una base para estudios futuros sobre el uso de IA en la educación médica y el desarrollo de habilidades clínicas.

1.1.4.2. Justificación Humana.

La capacidad de la IA para asistir a los estudiantes de medicina puede aligerar la carga académica y mejorar su comprensión de temas complejos. Estos modelos podrían convertirse en asistentes personalizados para el estudio, que permiten un aprendizaje adaptativo y personalizado. Además, al facilitar el acceso rápido a información precisa, se optimiza el tiempo que los estudiantes dedican a la comprensión profunda y al razonamiento clínico. Finalmente, beneficiaría a los pacientes, al formar médicos mejor preparados.

1.1.4.3. Justificación Social

La integración de IA en la educación médica podría democratizar el acceso a recursos de alta calidad en zonas con limitaciones tecnológicas o educativas. Al permitir que estudiantes de medicina accedan a herramientas avanzadas de aprendizaje. Esta investigación promueve la equidad en la educación. Además, la capacidad de estos modelos para simular entornos clínicos podría aliviar las demandas sobre recursos humanos y económicos en las instituciones educativas.

1.1.4.4. Justificación Contemporánea

En la actualidad, la inteligencia artificial está revolucionando diversos campos, por tal razón es crucial investigar su impacto y utilidad en la

educación médica. La pandemia de COVID-19 demostró la necesidad de métodos educativos flexibles y escalables. La IA no solo responde a esta necesidad, sino que lo hace en un contexto donde las tecnologías emergentes ya están remodelando el futuro de la medicina. Esta tesis aborda un tema relevante y en rápida evolución, que puede tener un impacto en los próximos años en la formación de profesionales de la salud y el ejercicio de la Medicina.

1.2. OBJETIVOS

1.2.1. Objetivo General

Comparar los puntajes obtenidos por los modelos de inteligencia artificial, Chat GPT-4o, Gemini Advanced y Copilot, al aplicarlos al examen del Concurso Nacional de Admisión al Residentado Médico propuesto por el CONAREME de los años 2017-2024 de Perú.

1.2.2. Objetivo Específico

- Determinar el porcentaje de preguntas correctamente resueltas por Chat GPT-4o en el examen del Residentado Médico 2017-2024.
- Determinar el porcentaje de preguntas correctamente resueltas por Gemini Advanced en el examen del Residentado Médico 2017-2024.
- Determinar el porcentaje de preguntas correctamente resueltas por Copilot en el examen del Residentado Médico 2017-2024.

1.3. MARCO TEÓRICO

1.3.1. Conceptos Básicos

1.3.1.1. Inteligencia Artificial y Procesamiento del Lenguaje Natural (PLN)

1.3.1.1.1. La Inteligencia Artificial

La inteligencia artificial (IA) es un campo de la informática que se enfoca en crear sistemas capaces de realizar tareas que, cuando son realizadas por humanos, requieren inteligencia. Estas tareas incluyen el razonamiento, el aprendizaje, la comprensión del lenguaje natural, la percepción visual y la toma de decisiones. La IA se basa en algoritmos y modelos matemáticos que permiten a los procesadores, administrar grandes cantidades de datos, identificar patrones, y generar respuestas o acciones de manera autónoma y rápida.

Russell y Norvig (2020) definen la IA como "cualquier dispositivo que percibe su entorno y toma acciones que maximicen sus posibilidades de éxito en alcanzar sus objetivos"(1). Esta definición refleja la capacidad de los sistemas de IA para adaptarse y mejorar en la resolución de problemas a medida que interactúan con el entorno y aprenden de los datos.

1.3.1.1.2. Procesamiento del Lenguaje Natural (PLN)

El **procesamiento del lenguaje natural (PLN)** es una subárea fundamental de la IA que permite a las máquinas interpretar y generar lenguaje humano. Los modelos de lenguaje grandes (PLN, por sus siglas en español) (LLMs, por sus

siglas en inglés) como ChatGPT, Gemini Advanced y Copilot AI son ejemplos avanzados de sistemas de PLN que pueden procesar grandes volúmenes de información, inclusive data médica, y responder preguntas complejas, replicando habilidades de análisis y síntesis propias a un nivel profesional (2).

ChatGPT, Gemini y Copilot son modelos avanzados de inteligencia artificial diseñados para comprender y generar lenguaje natural (3), con aplicaciones cada vez más prominentes en la medicina. Estas IAs están basadas también en técnicas de aprendizaje profundo y procesamiento del lenguaje natural. Aunque ambos modelos comparten similitudes, como su capacidad para procesar grandes cantidades de datos textuales y generar respuestas informadas, Gemini ha sido diseñado con un enfoque más integrado, combinando habilidades de lenguaje con capacidades avanzadas de razonamiento, resolución de problemas y análisis de datos científicos (4).

1.3.1.1.3. ChatGPT

La arquitectura Generative Pre-trained Transformer (GPT) es uno de los avances más significativos en el procesamiento de lenguaje natural (PLN) en los últimos años. GPT-4o, una de sus versiones más avanzadas, se fundamenta en una red neuronal profunda compuesta por capas de transformadores, los cuales son modelos neuronales que, mediante mecanismos de atención, permiten a la red procesar secuencias de texto de manera eficiente y contextual (5)

a) Red Neuronal Profunda y Mecanismo de Atención

GPT-4o utiliza una red neuronal profunda que, a diferencia de los enfoques tradicionales de redes recurrentes, implementa una estructura de atención que le permite enfocarse en diferentes partes del texto, determinando cuáles palabras o secuencias de palabras son más relevantes para entender el contexto y generar una respuesta adecuada. El mecanismo de **autoatención** es clave en la arquitectura que permite que cada palabra en una secuencia "preste atención" a otras palabras dentro de la misma secuencia, evaluando así la importancia contextual de cada elemento (6).

Por medio de múltiples capas de autoatención y feed-forward, GPT-4o es capaz de aprender y almacenar representaciones semánticas complejas que le permiten:

- **Comprender relaciones contextuales** entre palabras y frases en secuencias largas de texto, algo crucial para analizar preguntas complejas.
- **Adaptarse dinámicamente** a una amplia variedad de temas, desde literatura y ciencias sociales hasta temas técnicos como los de medicina y ciencias de la salud.

b) Entrenamiento con Grandes Volúmenes de Datos Textuales

El entrenamiento de GPT-4o implica el uso de grandes volúmenes de datos textuales de diversas fuentes, que incluyen desde textos generales hasta bases de datos especializadas en temas como medicina, ciencias y tecnología. Este proceso, conocido como preentrenamiento, permite al modelo adquirir una

"comprensión básica" del lenguaje natural y, en particular, desarrollar capacidades avanzadas de generación y comprensión de texto aplicados a múltiples campos clínicos (7).

Durante el preentrenamiento, el modelo se expone a tareas de predicción de palabras en las cuales, dada una secuencia de texto, debe predecir las palabras faltantes o continuar la secuencia de forma coherente. Este procedimiento ayuda a GPT-4o a aprender patrones lingüísticos complejos lo cual es esencial para responder preguntas contextualmente precisas y adaptadas a preguntas de alta especialización, como las que se presentan en un examen médico.

c) Adaptabilidad y Generalización en el Modelo GPT-4o

GPT-4o se caracteriza por su notable capacidad para generar texto a partir de un contexto específico y adaptarse a una amplia variedad de temas de manera precisa y coherente. Esto es posible gracias a su entrenamiento en grandes corpus de texto de diversas disciplinas y fuentes, que le permite comprender términos técnicos y científicos. Su capacidad para generalizar y adaptarse a distintas áreas de conocimiento lo convierte en una herramienta eficaz en el ámbito educativo y en la preparación para exámenes complejos, como el examen de admisión al residentado médico en Perú.

Para resolver preguntas de alto nivel, GPT-4o aprovecha su entrenamiento con datos de contenido médico y general, lo cual incluye textos clínicos, manuales, investigaciones médicas y guías de diagnóstico. Gracias a ello, el modelo es capaz de:

- Interpretar terminología médica y contextualizarla de acuerdo con el tema específico de una pregunta u otra aplicación (8).
- Establecer relaciones causales y secuenciales que le permiten analizar casos clínicos y emitir respuestas alineadas con prácticas clínicas (8).
- Extraer y sintetizar información clave de grandes cantidades de texto, aplicando su conocimiento pre-entrenado para generar respuestas complejas.

d) Habilidades Avanzadas en la Resolución de Preguntas Complejas

Uno de los aspectos más sobresalientes de GPT-4o es su habilidad para resolver preguntas de múltiples niveles de complejidad, gracias a su entrenamiento y a su capacidad para reconocer patrones de pensamiento estructurado y analítico en el lenguaje. Esto incluye habilidades como:

Razonamiento deductivo y lógico: Aunque GPT-4o no tiene conocimiento explícito de "razonamiento" como un humano, su arquitectura le permite inferir secuencias lógicas en preguntas de opción múltiple o de análisis de casos, identificando la respuesta más probable mediante un proceso de eliminación y asociación (9).

Capacidad para realizar conexiones entre conceptos: GPT-4o puede conectar diferentes conceptos en ciencias médicas, como anatomía y farmacología, a través de la información contextual proporcionada en la

pregunta, ofreciendo respuestas complejas que requieren análisis multidimensional (3).

e) Potencial y Limitaciones en la Educación Médica

La arquitectura de GPT-4o permite su aplicación en áreas como la educación médica, brindando un recurso complementario en la preparación para exámenes de admisión como el de residentado médico en Perú. Su capacidad de responder preguntas se ha puesto a prueba en países del primer mundo, demostrando un alto nivel de concordancia y perspicacia en sus explicaciones, considerando su potencial en la educación médica y toma de decisiones clínicas (10,11). A su vez se toma en cuenta los riesgos en cuanto al a veracidad, ética y realidad de la información, ya que su precisión puede ser limitada y puede producir respuestas incorrectas al no tener acceso a la red del internet, respuestas en tiempo real y sesgos utilizados en el entrenamiento de datos amenazando la integridad académica (12); recordando a su vez que estos modelos de inteligencia carecen de criterios éticos y legales, siendo limitados a la base de datos almacenada, por lo que debe guiarse para cualquier acto médico (11).

Sus capacidades lo convierten en un recurso valioso para:

- **Asistir en la resolución de preguntas de práctica**, simulando un entorno de examen en el que los estudiantes pueden recibir respuestas detalladas y explicativas. (13)
- **Ofrecer retroalimentación inmediata** sobre respuestas correctas e incorrectas, lo cual es útil para los estudiantes que buscan entender la lógica detrás de cada opción (13) y una mejor capacitación para la correcta optimización de tiempo en el ámbito académico y/o hospitalario. (14)

Sin embargo, es importante considerar las limitaciones inherentes a GPT-4o, entre las cuales destacan:

- **Sesgo en los datos de entrenamiento**: Dado que GPT-4o es entrenado en grandes cantidades de texto, es posible que ciertas áreas de conocimiento estén mejor representadas que otras, lo cual podría afectar la precisión en temas médicos específicos.
- **Falta de juicio clínico**: Aunque el modelo puede responder preguntas con precisión, carece de experiencia clínica y habilidades de juicio ético, aspectos fundamentales en la toma de decisiones médicas.
- **Dificultades en la interpretación de imágenes y datos no textuales**: Al ser un modelo de lenguaje, GPT-4o no puede analizar ni interpretar imágenes de rayos X, tomografías u otras herramientas de diagnóstico más complejas al mismo nivel que un profesional certificado **(15)**.
- **Carencia de criterios éticos y legales:** Al ser un modelo basado en patrones de texto de grandes cantidades de datos in una comprensión profunda de las implicancias de estos. Siendo un modelo carente de juicio humano, ausencia de los conocimientos

actualizados, consideraciones legales o éticas, precauciones y limitaciones intencionadas o falta de empatía y juicio humano. (11,14,16)

1.3.1.1.4. Gemini Advanced

Gemini Advanced es un modelo avanzado de inteligencia artificial creado por Google DeepMind, diseñado específicamente para entornos de ciencias de la salud. Este modelo utiliza una arquitectura optimizada que emplea redes neuronales profundas para comprender y generar respuestas altamente precisas a consultas clínicas. Es un modelo de IA avanzado especializado en tareas de PLN, que ha sido optimizado para mejorar su rendimiento en términos de **precisión y velocidad** al responder preguntas de opción múltiple y situaciones de examen (17). Basado en el procesamiento de grandes cantidades de datos médicos, Gemini Advanced se destaca en la interpretación de casos complejos, identificando relaciones críticas entre síntomas, patologías y tratamientos. También aplica mecanismos de autoatención que le permiten evaluar datos médicos y extraer patrones clínicos relevantes (18)

a) Arquitectura Especializada y Capacidades de Procesamiento

La arquitectura de Gemini Advanced está diseñada para analizar datos textuales de medicina con una alta precisión contextual y le permite aprender patrones complejos propios de los textos médicos, tales como la relación entre síntomas y diagnósticos, y aplicar esta información en tareas de evaluación y análisis clínico. A diferencia de modelos de IA generales, Gemini Advanced se ha entrenado específicamente en conjuntos de datos médicos, lo que le permite manejar terminología clínica y comprender directrices médicas, optimizando así su rendimiento en pruebas y evaluaciones específicas del área médica (18).

b) Entrenamiento en Datos Especializados en Ciencias de la Salud

Gemini Advanced ha sido entrenado en amplios datos clínicos, incluyendo literatura médica y registros de salud, lo que refuerza su capacidad para responder preguntas complejas con precisión clínica. Este entrenamiento en dominios especializados ha mostrado ser efectivo en tareas donde la precisión es crítica, permitiéndole realizar inferencias diagnósticas preliminares. Investigaciones sugieren que el uso de datos específicos de salud en IA mejora notablemente el rendimiento en aplicaciones clínicas, como diagnósticos y análisis de síntomas (19).

c) Aplicación y Limitaciones en Evaluación Médica

Gracias a su entrenamiento y configuración, Gemini Advanced es capaz de:

- **Analizar y correlacionar síntomas:** Ofrece recomendaciones diagnósticas basadas en la evaluación de múltiples factores clínicos **(20).(21)**
- **Interpretar guías y estándares médicos:** Genera respuestas alineadas con evidencia científica y protocolos clínicos

reconocidos, una ventaja importante en contextos de evaluación de conocimientos médicos **(22)**.

1.3.1.1.5. Copilot

En el ámbito del procesamiento del lenguaje natural y la inteligencia artificial aplicada, el modelo **Copilot** ha demostrado ser una herramienta eficaz para la asistencia en diversas tareas, especialmente en la creación y completación de código y otros procesos técnicos. Desarrollado por OpenAI en colaboración con GitHub, Copilot ha sido entrenado para comprender el contexto y generar sugerencias en tiempo real, lo que lo posiciona como un asistente útil para programadores y profesionales en otros campos **que** requieren respuestas rápidas y precisas a problemas complejos (23)

Copilot, desarrollado por Microsoft impulsado con tecnología de OpenAI y GitHub, es una herramienta de inteligencia artificial (IA) diseñada inicialmente para asistir en la generación de código en entornos de programación. Basado en modelos de lenguaje de gran. Copilot ha demostrado su capacidad para procesar lenguaje humano en tareas técnicas y no técnicas (21). Si bien su aplicación principal ha sido en el ámbito de la programación, su versatilidad lo ha posicionado como una herramienta prometedora para diversos sectores, incluida la medicina. En el ámbito médico, Copilot está siendo explorado como un asistente que puede facilitar la toma de decisiones clínicas, mejorar la documentación médica y optimizar procesos educativos (23).

a) Tecnología

Copilot se basa en un modelo de lenguaje preentrenado diseñado para generar texto a partir de indicaciones contextuales. Su arquitectura utiliza redes neuronales profundas y técnicas de **atención transformacional**, características propias de los modelos GPT, que le permiten procesar y relacionar grandes cantidades de información.

- **Atención transformacional**: La arquitectura subyacente de Copilot se basa en la capacidad de enfocarse en diferentes partes de una secuencia de datos para identificar patrones relevantes, lo que mejora la comprensión de contextos complejos, como preguntas clínicas o escenarios de diagnóstico.
- **Aprendizaje Supervisado y por Refuerzo**: Copilot ha sido ajustado utilizando grandes volúmenes de datos técnicos y especializados, lo que le permite realizar predicciones precisas en áreas de alta especialización, como la medicina.
- **Contextualización Dinámica**: Copilot utiliza información proporcionada en tiempo real para generar sugerencias adaptadas al contexto. En medicina, esto se traduce en respuestas personalizadas basadas en antecedentes médicos, síntomas y guías clínicas

b) Aplicaciones de Copilot en la Medicina

Apoyo en la Educación Médica

En entornos académicos, Copilot puede actuar como una herramienta complementaria para estudiantes y profesionales en formación. Genera explicaciones claras y concisas sobre temas médicos, facilita la resolución de preguntas de examenes y ayuda a comprender conceptos complejos.

- **Generación de Casos Prácticos**: Copilot puede crear escenarios clínicos para la práctica en simulaciones **(24)**.
- **Retroalimentación Personalizada**: Ofrece comentarios inmediatos sobre las respuestas de los estudiantes, ayudando a identificar áreas de mejora **(24)**.

Toma de Decisiones Clínicas

Aunque no sustituye el juicio humano, Copilot puede ser un apoyo en la toma de decisiones médicas al analizar datos de pacientes y sugerir opciones basadas en guías clínicas y evidencia científica.

- **Sugerencias Diagnósticas y Terapéuticas**: Copilot puede recomendar diagnósticos diferenciales y tratamientos basados en los síntomas y antecedentes proporcionados. **(21)**
- **Análisis de Riesgos y Beneficios**: Facilita la evaluación de opciones terapéuticas mediante el análisis de datos disponibles.

1.3.1.2. Examen del Concurso Nacional de Admisión al Residentado Médico

El Examen Nacional de Admisión al Residentado Médico (ENARM) en Perú es un proceso clave para la selección de médicos en programas de especialización. Este examen, gestionado principalmente por el Comité Nacional de Residentado Médico (CONAREME), fue implementado con el objetivo de estandarizar y elevar el nivel de competencia profesional en el área de salud a nivel nacional (25). Desde su creación en los años 90, el ENARM ha evolucionado, adaptándose a cambios en los estándares de calidad y formación médica para satisfacer las demandas del sistema de salud en el país (26).

El examen busca evaluar conocimientos específicos en medicina general y capacidades analíticas y clínicas de los postulantes, brindando una oportunidad para acceder a una educación de especialización con respaldo institucional. La historia de este examen refleja un esfuerzo continuo por mejorar los estándares de la educación médica en Perú, a través de ajustes en su contenido y estructura, lo que ha fortalecido la formación de especialistas que responden adecuadamente a las necesidades del sistema de salud peruano.

1.3.1.2.1. Estructura y Componentes del Examen

El ENARM se compone de preguntas de opción múltiple que cubren áreas médicas fundamentales, tales como medicina interna, pediatría, cirugía general, ginecología y obstetricia, y otras subespecialidades que constituyen la base de la práctica médica en el país. Estas preguntas buscan evaluar no solo el conocimiento teórico de los postulantes, sino también

sus habilidades para el diagnóstico clínico y la resolución de problemas en situaciones complejas.

Diseño de las preguntas: Sigue un formato de opción múltiple, estructuradas para evaluar el conocimiento y la habilidad del postulante para tomar decisiones clínicas. Las preguntas suelen estar basadas en casos clínicos breves, donde se presentan síntomas, antecedentes y exámenes de laboratorio o imágenes relevantes (26). A partir de estos datos, los postulantes deben identificar diagnósticos, tratamientos y decisiones de manejo adecuadas, lo que implica una comprensión profunda de los conceptos médicos y su aplicación práctica en un contexto clínico.

Áreas temáticas: El examen cubre áreas como:

- **Medicina Interna**: Aborda enfermedades y condiciones de órganos internos, diagnóstico diferencial y manejo de patologías crónicas.
- **Pediatría**: Enfocado en el desarrollo, diagnóstico y tratamiento de enfermedades en niños.
- **Ginecología y Obstetricia**: Evaluación de problemas relacionados con el sistema reproductivo femenino y manejo de embarazo.
- **Cirugía**: Conocimientos de técnicas quirúrgicas y manejo postoperatorio.
- **Salud Pública**: Aspectos de salud poblacional y prevención de enfermedades, fundamentales para un enfoque integral en la atención de salud.
- **Ciencias básicas**: Aborda materias las cuales aportan una base para el desarrollo e interpretación de todas las áreas en la medicina.

1.3.1.2.2. Importancia de la Evaluación Estandarizada en el Residentado Médico

La estandarización del examen garantiza que todos los candidatos a médicos especialistas en el Perú tengan una evaluación homogénea y basada en competencias esenciales para el ejercicio profesional. Este proceso permite que las instituciones médicas, hospitales y centros de formación cuenten con personal en formación que cumple con los conocimientos y habilidades mínimas requeridas, asegurando que el futuro especialista pueda adaptarse de manera eficiente a las demandas del sistema de salud (27).

1.3.1.2.3. Aspectos Técnicos en la Evaluación

a) Nivel de Dificultad y Complejidad:

El examen está diseñado para evaluar el dominio de conocimientos básicos y avanzados en medicina. La dificultad se distribuye en preguntas que abarcan desde conceptos fundamentales hasta problemas clínicos de alta complejidad, lo cual asegura que el postulante posea tanto una sólida base teórica como la capacidad para enfrentarse a situaciones reales y críticas de la práctica médica.

b) Peso y Ponderación de las Áreas

Cada área temática tiene una ponderación específica, reflejando la importancia de ciertos temas para la formación de especialistas médicos. La medicina interna y la pediatría suelen tener una mayor cantidad de preguntas en comparación con otras áreas, dado su peso en la práctica general de la medicina y la necesidad de que los médicos especialistas posean un conocimiento sólido en estas disciplinas.

c) Sistemas de Calificación y Criterios de Evaluación

El sistema de calificación del ENARM es objetivo y basado en el número de respuestas correctas, sin penalización por respuestas incorrectas. Esto permite una evaluación precisa de los conocimientos del postulante y promueve un enfoque en la precisión diagnóstica y el razonamiento clínico. La calificación final se determina en base al porcentaje de respuestas correctas, y el puntaje de corte varía cada año según el número de postulantes y el rendimiento general (28).

1.3.2. Revisión de Antecedentes Investigativos

1.3.2.1. Locales

No existen Antecedentes Investigativos a nivel local.

1.3.2.2. Naciónales

Autor: Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al.
Título: "Performance of ChatGPT on the Peruvian National Licensing Medical Examination: Cross-Sectional Study"(29)
Resumen:
El estudio publicado el 09 de abril del 2024, evaluó el rendimiento de ChatGPT (GPT-3.5 y GPT-4) en el Examen Nacional de Medicina (ENAM) de Perú, donde se observó que casi un tercio de los examinados no aprueban. Utilizando 180 preguntas de selección múltiple del ENAM 2022, se comparó el desempeño de ChatGPT con el de 1025 examinados. GPT-4 alcanzó una precisión del 86% y GPT-3.5 del 77%, mientras que los examinados lograron un 55%. Las preguntas de dificultad moderada a alta estuvieron asociadas a respuestas incorrectas en ambos modelos. Tras reformular las preguntas incorrectas con nuevos roles y contexto, se mejoró la precisión de ChatGPT.(29)

Autor: Curioso WH
Título: Inteligencia artificial e innovación para optimizar el proceso de diagnóstico de la tuberculosis
Resumen: Este artículo, publicado el 08 de agosto del 2020, destaca la importancia de la inteligencia artificial como una estrategia clave para enfrentar la tuberculosis a través de un diagnóstico oportuno, especialmente en países de medianos y bajos ingresos. Se analiza la herramienta eRx, que utiliza algoritmos de aprendizaje profundo y redes neuronales convolucionales para el análisis remoto de rayos X en casos

sospechosos de tuberculosis. Además de los aspectos tecnológicos, se resalta la relevancia de factores sociotécnicos, culturales y organizacionales en la implementación de estas soluciones. Las herramientas basadas en inteligencia artificial, como eRx, pueden optimizar el proceso de diagnóstico de la tuberculosis y otras enfermedades transmisibles.(17)

1.3.2.3. Internacionales

Autor: Mihalache A, Huang RS, Popovic MM, Muni RH
Título: "ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination"(30)
Resumen: El estudio, publicado el 15 de octubre del 2023, evaluó el rendimiento de ChatGPT-4 en preguntas de práctica para los exámenes de licencia médica de los Estados Unidos (USMLE) Step 1, Step 2CK y Step 3. ChatGPT-4 respondió correctamente el 88% de las preguntas de Step 1, el 86% en Step 2CK, y el 90% en Step 3. Además, proporcionó explicaciones para todas las preguntas, con tiempos promedio de respuesta entre 23 y 30 segundos por pregunta. Las respuestas correctas fueron más breves que las incorrectas, lo que indica una mayor eficiencia en las respuestas correctas. ChatGPT-4 mostró un rendimiento significativamente mejor que las versiones anteriores del chatbot.(30)

Autor: Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P.
Titulo: "Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports"(31)
Resumen: Este estudio, publicado el 23 de abril del 2024, evaluó la precisión diagnóstica de ChatGPT 4.0 en casos clínicos, además de su capacidad para responder preguntas del examen USMLE Step 2 CK. Se ingresaron 109 preguntas de práctica en ChatGPT 3.5 y 4.0, observando una mejora en la precisión del 47.7% al 87.2% ($p = 0.035$) en ChatGPT 4.0. También se analizaron 63 casos clínicos publicados, y ChatGPT 4.0 generó diagnósticos diferenciales acertados en el 74.6% de los casos. El 70.2% de los diagnósticos correctos fueron la primera opción en la lista de diferenciales. El estudio destaca las mejoras continuas en la precisión diagnóstica y en la capacidad de ChatGPT 4.0 para resolver preguntas estandarizadas del USMLE. (31)

Autor: : Rossettini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, Turolla A, et al
Título: "Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study"(23)
Resumen: Este estudio, publicado el 26 de junio del 2024, evaluó la precisión de tres chatbots de inteligencia artificial (ChatGPT-4, Microsoft Copilot y Google Gemini) en predecir las respuestas correctas del examen estandarizado de ingreso a las ciencias de la salud en Italia (CINECA). Se analizaron 820 preguntas, de las cuales 808 fueron

procesadas por ChatGPT-4 y Google Gemini debido a limitaciones técnicas. Los resultados mostraron diferencias significativas en la precisión entre ChatGPT-4 y Google Gemini, y entre Microsoft Copilot y Google Gemini ($p < 0.001$). ChatGPT-4 y Microsoft Copilot obtuvieron mejores resultados que Google Gemini. La coherencia narrativa fue mayormente lógica en las respuestas correctas (81.5%) y los errores fueron atribuidos a fallos lógicos (88.9%). Aunque los chatbots demostraron una precisión prometedora, se recomienda a los estudiantes utilizarlos como un complemento y no como una fuente principal de estudio. (23)

1.4. HIPÓTESIS

Dado que, los modelos de IA como ChatGPT-4O, Gemini Advanced y Copilot emplean avanzados sistemas de procesamiento de información que le permiten la resolución con alta precisión de escenarios complejos, manejando una gran cantidad de información para llegar a un resultado y que el Examen Nacional de Admisión al Residentado Médico evalúa la capacidad resolución de escenarios que plantean problemáticas existentes en el área de la salud, así como conocimientos básicos de la misma.

Es probable que, exista una diferencia significativa entre el porcentaje de aciertos obtenidos por ChatGPT-4O, Gemini Advanced y Copilot AI al ejecutarlas en el Examen Nacional de Admisión al Residentado Médico.

CAPÍTULO II

PLANTEAMIENTO OPERACIONAL

2. PLANTEAMIENTO OPERACIONAL

2.1. TÉCNICAS, INSTRUMENTOS Y MATERIALES DE VERIFICACIÓN

2.1.1. Técnicas

2.1.1.1. Observación documental:

Utilizada para revisar y analizar los exámenes del Concurso Nacional de Admisión al Residentado Médico de los años 2017 a 2024, identificando las preguntas, su formato y su clasificación por especialidad.

2.1.1.2. Análisis comparativo:

Se llevó a cabo un análisis comparativo de las respuestas generadas por Chat GPT-4o, Gemini Advanced y Copilot, centrándose en el porcentaje de preguntas correctamente respondidas. Esto incluye el análisis estadístico de los datos obtenidos.

2.1.1.3. Evaluación cuantitativa:

Se realizó una evaluación cuantitativa mediante el uso de técnicas estadísticas. Se realizó la prueba de normalidad de las muestras con Shapiro Wilk. Posteriormente se realizó la prueba de Homocedasticidad con la prueba de Levene. Se realizó una prueba de Análisis de Varianzas (ANOVA) para comparar los resultados obtenidos de cada modelo. Además se elaboró una prueba post hoc HSD de Tukey para la comparación entre grupos de IA's.

2.1.2. Instrumentos

Programas de Inteligencia artificial: Los modelos de Inteligencia Artificial Chat-GPT 4o, Gemini Advanced y Copilot.

Cédula para preguntas: Herramienta digital utilizada para estructurar y organizar las preguntas del Examen del Concurso Nacional de Admisión al Residentado Médico (ENARM) por año de examen así como en las diferentes categorías de bloques temáticos.

Ficha de observación: Diseñada específicamente para este estudio, permite registrar y clasificar las respuestas generadas por las inteligencias artificiales (ChatGPT-4o, Gemini Advanced y Copilot). Este instrumento garantiza un registro sistemático de los datos obtenidos

2.2. CAMPOS DE VERIFICACIÓN

2.2.1. Ubicación Espacial

El estudio se llevó a cabo utilizando plataformas digitales para evaluar el rendimiento de los modelos de IA en el examen, sin necesidad de una ubicación física específica

2.2.2. Temporalidad

El análisis se realizó entre Octubre y Diciembre del 2024, evaluando los resultados obtenidos por las IAs en los exámenes de Residentado Médico entre 2017 y 2024

2.2.3. Unidades de Estudio

Las unidades de estudio están constituidas por las preguntas que constituyen los Exámenes del Concurso Nacional de Admisión al Residentado Médico 2017, 2018, 2019, 2020, 2021, 2022, 2023 y 2024.

Población: Exámenes del Concurso Nacional de Admisión al Residentado Médico 2017, 2018, 2019, 2020, 2022, 2023 y 2024.

Para el propósito de este estudio, se trabajó con toda la población.

2.2.3.1. Selección de la Muestra

Criterios de inclusión

Exámenes del Concurso Nacional de Admisión al Residentado Medico de Perú con clave de respuestas oficial del CONAREME.

Exámenes del Concurso Nacional de Admisión al Residentado Medico de Perú de los años 2017, 2018, 2019, 2020, 2022, 2023 y 2024.

Criterios de exclusión

Preguntas que forman parte de los Exámenes del Concurso Nacional de Admisión al Residentado Medico de Perú que durante el proceso exploratorio presente limitaciones para su procesamiento o generación de respuesta por parte del instrumento.

Exámenes del Concurso Nacional de Admisión al Residentado Medico de Perú que no se encuentren publicados en la página web oficial del CONAREME

2.3. ESTRATEGIAS DE RECOLECCIÓN DE DATOS

2.3.1. Organización

Para la recolección de datos, se siguió los siguientes procedimientos:

Obtención del Examen del Concurso Nacional de Admisión al Residentado Médico: Se accederá a una copia del examen del cual utilizaremos el total de preguntas.

Simulación del examen en modelos de IA: Se cargaron el conjunto de preguntas seleccionadas a ambos modelos de IA (GPT-4o, Gemini Advanced y Copilot). Se creó un nuevo chat para cada pregunta para evitar crear memoria en la IA. Las respuestas generadas fueron registradas y clasificadas.

Recolección de Datos: Se recolectó los datos obtenidos a partir de las respuestas de las IA's, las cuales fueron organizadas en nuestra ficha virtual de recolección de datos.

Comparación de respuestas: Las respuestas generadas por los modelos se compararon con las respuestas correctas del examen según clave oficial del CONAREME.

Evaluación del rendimiento: Los resultados de los modelos fueron evaluados en términos de precisión (respuestas correctas e incorrectas) y porcentaje de respuestas correctas.

2.3.2. Recursos

Para realizar el estudio se necesitó lo siguiente:

2.3.2.1. Humanos

- Los investigadores.
- Asesor

2.3.2.2. Materiales

- 2 computadoras personales.
- Chat GPT Plan Plus
- Gemini Advanced
- Copilot
- 1 impresora.

2.3.2.3. Financieros

Autofinanciado

2.3.2.4. Institucionales

Nuestro estudio no requiere uso de instalaciones de nuestra Institución

2.3.3. Validación de Instrumentos

Al ser una ficha de recolección de datos no requiere de validación de expertos.

2.3.4. Criterios para el Manejo de Resultados

2.3.4.1. A nivel de recolección:

Los resultados de las Inteligencias Artificiales fueron registrados de manera sistemática en una ficha de observación digital. Las respuestas de ambos modelos fueron almacenadas en un ambiente digital controlado, y los puntajes obtenidos en el examen se recolectaron inmediatamente después de su finalización, asegurando la integridad de los datos. Además, se realizó una copia impresa de la ficha de recolección de datos en caso de pérdidas de datos digitales.

2.3.4.2. A nivel de sistematización

La información obtenida fue organizada y tabulada en hojas de cálculo utilizando Microsoft Excel, donde se registraron los puntajes individuales de cada modelo para cada pregunta del examen según especialidad y su calificación final, organizando los datos por año de examen y tipo de AI. Además, se crearon tablas resumen para facilitar el análisis comparativo entre los modelos, segmentando los datos según categorías de preguntas por especialidad médica (Ciencias

Básicas, Medicina Interna, Cirugía, Ginecología y Obstetricia, Pediatría, Salud Pública)

2.3.4.3. A nivel de análisis de datos

Análisis Descriptivo:

Se calcularon medidas de tendencia central (media) y dispersión (desviación estándar) para las calificaciones obtenidas por cada modelo por año de examen, así como por bloque temático.

Se elaboraron tablas para visualizar la distribución de las calificaciones.

Pruebas de Hipótesis:

Utilizamos el software de análisis estadístico SPSS para nuestra hipótesis. Se aplicó una Prueba de Normalidad de las muestras con Shapiro Wilk. Posteriormente se realizó la prueba de Homocedasticidad con la prueba de Levene. Se realizó una prueba de Análisis de Varianzas (ANOVA para muestras independientes para comparar las calificaciones promedio entre las IAs. Se hizo una prueba post hoc HSD de Tukey para la comparación entre grupos de IA's.

Se estableció un nivel de significancia del 5% ($p < 0.05$) para determinar si existen diferencias estadísticamente significativas entre los modelos.

CAPÍTULO III

RESULTADOS

3. RESULTADOS

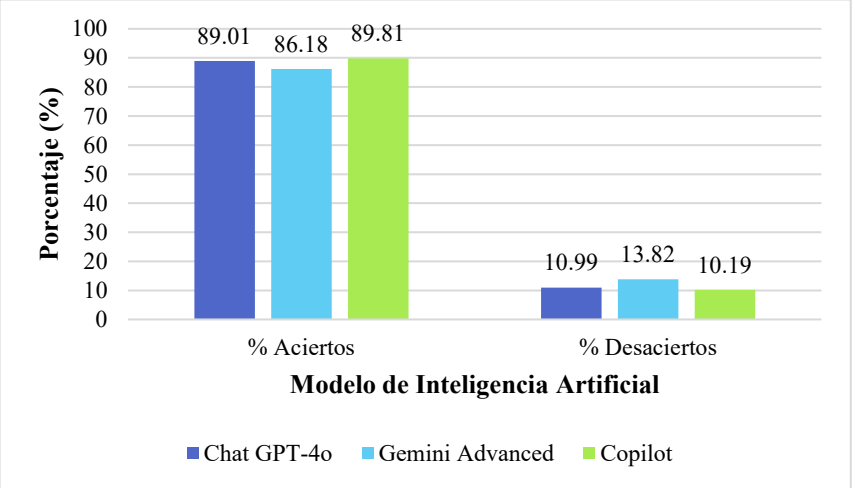
"Análisis comparativo de los puntajes obtenidos por Chat GPT-4o, Gemini Advanced y Copilot al aplicarlas en el examen del Concurso Nacional de Admisión al Residentado Médico, realizado por el Consejo Nacional de Residentado Medico (CONAREME) de los años 2017-2024. Perú."

Tabla 1. Aciertos y Desaciertos según Modelo de Inteligencia Artificial

Modelo de IA	Total de Preguntas	Aciertos n (%)	Desaciertos n (%)	σ
Chat GPT-4o	1380	1229 (89.01)	151 (10.99)	4.30
Gemini Advanced	1380	1190 (86.18)	190 (13.82)	2.26
Copilot	1380	1240 (89.81)	140 (10.19)	3.46

Fuente: Elaboración propia / Matriz de sistematización de datos

Gráfico 1. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial.



Fuente: Elaboración propia / Matriz de sistematización de datos

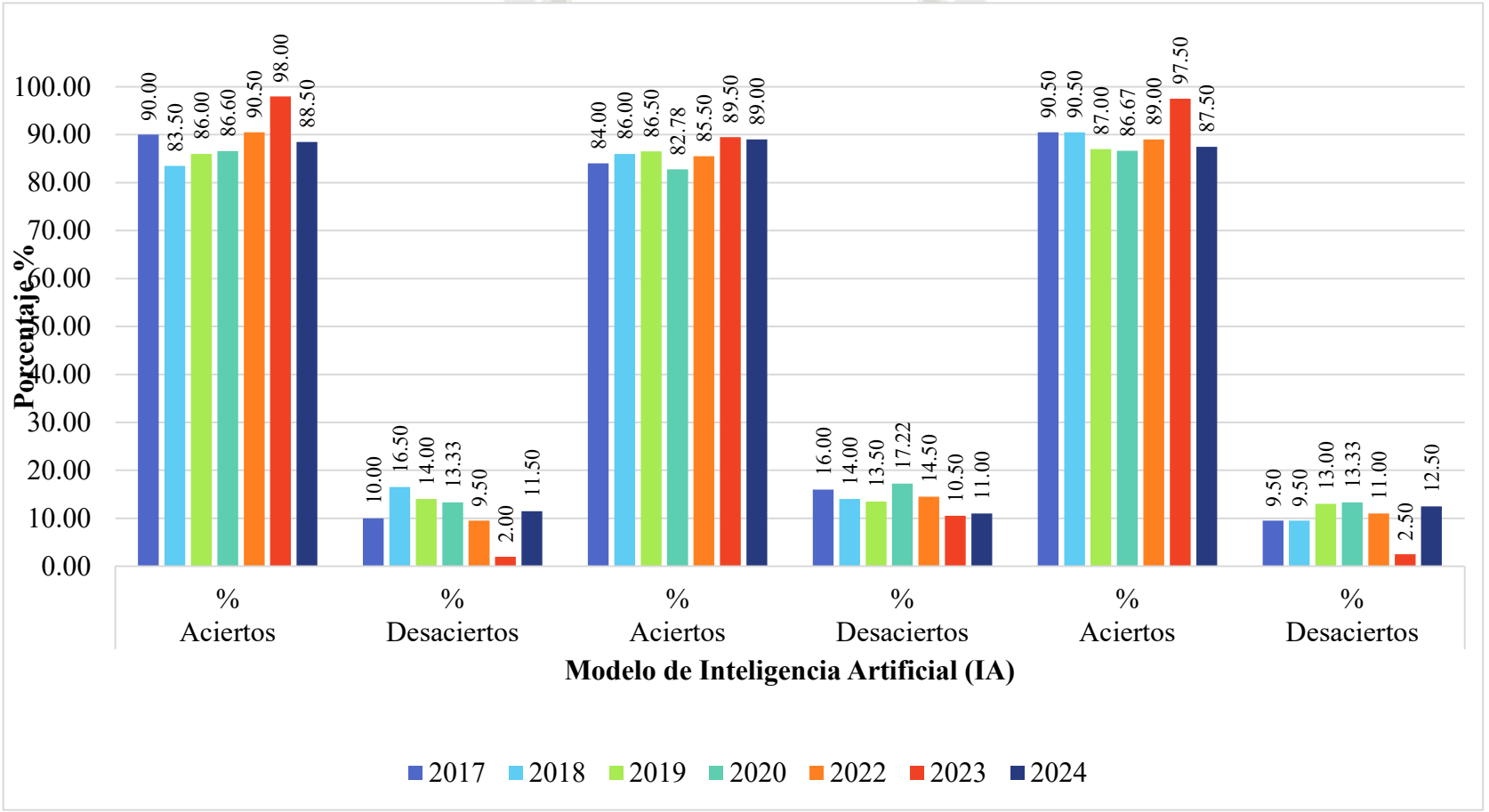
En la **Tabla 1** y **Gráfico 1**, se muestra la comparación en el rendimiento de tres modelos de inteligencia artificial (IA), Chat GPT-4o, Gemini Advanced y Copilot, en un conjunto de 1380 preguntas extraídas de exámenes del Concurso Nacional de Admisión al Residentado Médico (2017-2024). Los resultados revelaron un alto rendimiento general en los tres modelos, con porcentajes de aciertos de 89.01% ($\sigma = 4.30$) para Chat GPT-4o, 86.18% ($\sigma = 2.26$) para Gemini Advanced y 89.8% ($\sigma = 3.46$) para Copilot. Si bien Copilot demostró una ligera superioridad en el porcentaje de aciertos, Gemini Advanced presentó la menor desviación estándar ($\sigma = 2.3$), indicando una mayor consistencia en sus respuestas.

Tabla 2. Aciertos y Desaciertos según Modelo de Inteligencia Artificial (IA) por Año de Examen

Año de Exámen	Total de Preguntas (n)	Modelo de Inteligencia Artificial (IA)					
		Chat GPT-4o		Gemini Advanced		Copilot	
		Aciertos n (%)	Desaciertos n (%)	Aciertos n (%)	Desaciertos n (%)	Aciertos n (%)	Desaciertos n (%)
2017	200	180 (90.00)	20 (10.00)	168 (84.00)	32 (16.00)	181 (90.50)	19 (9.50)
2018	200	167 (83.50)	33 (16.50)	172 (28.00)	28 (14.00)	181 (90.50)	19 (9.50)
2019	200	172 (86.00)	28 (14.00)	173 (86.50)	27 (13.50)	174 (87.00)	26 (13.00)
2020	180	156 (86.67)	24 (13.33)	149 (82.78)	31 (17.22)	156 (87.67)	24 (13.33)
2022	200	181 (90.50)	19 (9.50)	171 (85.50)	29 (14.50)	178 (89.00)	22 (11.00)
2023	200	196 (98.00)	4 (2.00)	179 (89.50)	21 (10.50)	195 (97.50)	5 (2.50)
2024	200	177 (88.50)	23 (11.50)	178 (89.00)	22 (11.00)	175 (87.50)	25 (12.50)
Total de Preguntas (n)	1380	1229	151	1190	190	1240	140
Promedio de Aciertos (%)		89.01	10.98	86.18	13.82	89.81	10.19
σ		4.30		2.26		3.46	

Fuente: Elaboración propia / Matriz de sistematización de datos

Gráfico 2. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial por Año de Examen



Fuente: Elaboración propia / Matriz de sistematización de datos

En la **Tabla 2** y **Gráfico 2**, se muestra la comparación en el rendimiento de tres modelos de IA (Chat GPT-4o, Gemini Advanced y Copilot) en exámenes del Concurso Nacional de Admisión al Residentado Médico (2017-2024), analizando aciertos y desaciertos por año.

Al observar los resultados por año, en 2023, todos los modelos lograron su mejor desempeño. Copilot alcanzó un 97.50% de aciertos, seguido por Chat GPT-4o con un 98% y Gemini Advanced con un 89.50%. Este año resalta como el de mayor desempeño global para los tres sistemas. En contraste, en 2018, se observaron los puntajes más bajos: **Chat GPT-4o** obtuvo un 83.50%, mientras que **Gemini Advanced** y **Copilot** alcanzaron un 86% y 90.50%, respectivamente.

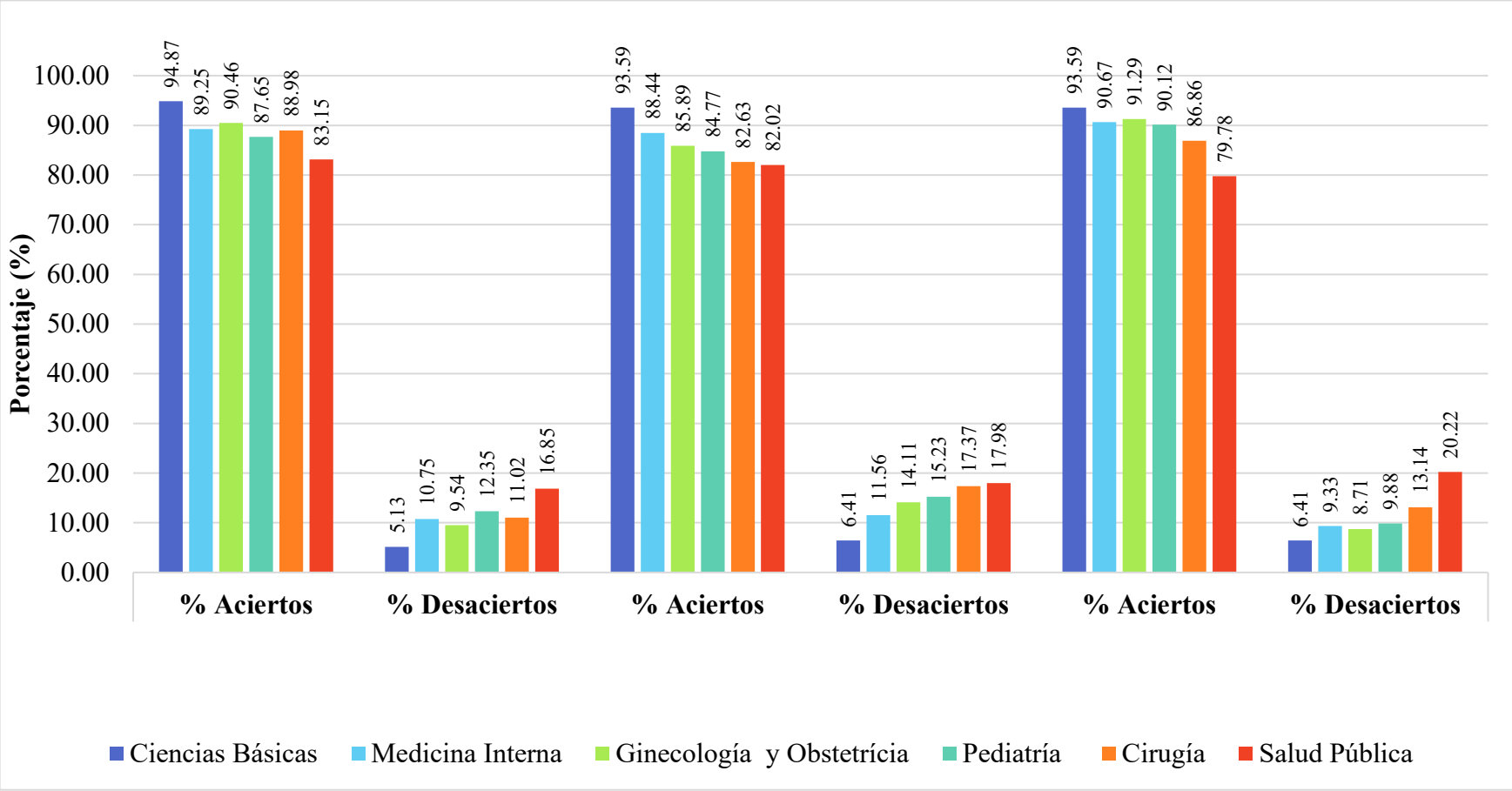
En resumen, los resultados muestran que, aunque los tres modelos son capaces de desempeñarse adecuadamente en el contexto de los exámenes del CONAREME, **Copilot** combina un alto porcentaje de aciertos con una variabilidad moderada, lo que lo posiciona como el modelo más robusto y consistente. **Gemini Advanced**, aunque presenta una mayor estabilidad interanual, tiene un desempeño promedio menor, mientras que **Chat GPT-4o** destaca por su capacidad de alcanzar altos puntajes en años específicos, pero con fluctuaciones más pronunciadas.

Tabla 3. Aciertos y Desaciertos según Modelo de Inteligencia Artificial por Bloque Temático

Bloque Temático	Total de Preguntas	Modelo de Inteligencia Artificial (IA)								
		Chat GPT-4o			Gemini Advanced			Copilot		
		Aciertos n (%)	Desaciertos n (%)	σ	Aciertos n (%)	Desaciertos n (%)	σ	Aciertos n (%)	Desaciertos n (%)	σ
Ciencias Básicas	78	74 (94.87)	4 (5.12)	6.37	73 (93.58)	5 (6.41)	7.51	73 (93.58)	5 (6.41)	5.83
Medicina Interna	493	440 (89.24)	53 (10.75)	7.63	436 (88.43)	57 (11.56)	2.4	447 (90.66)	46 (9.33)	4.21
Ginecología y Obstetricia	241	218 (90.45)	23 (9.54)	5.59	207 (85.89)	34 (14.10)	6.53	220 (91.28)	21 (8.71)	8.46
Pediatría	243	213 (87.65)	30 (12.34)	6.62	206 (84.77)	37 (15.22)	4.83	219 (90.12)	24 (9.87)	5.32
Cirugía	236	210 (88.98)	26 (11.01)	5.36	195 (82.62)	41 (17.37)	6.66	205 (86.44)	31 (13.13)	7.66
Salud Pública	89	74 (83.14)	15 (16.85)	10.43	73 (82.02)	16 (17.97)	8.13	71 (79.77)	18 (20.22)	9.98
Total de Preguntas	**1380**	**1229 (89.00)**	**151 (10.94)**	**4.30**	**1190 (86.23)**	**190 (13.76)**	**2.26**	**1235 (89.49)**	**145 (10.50)**	**3.46**

Fuente: Elaboración propia / Matriz de sistematización de datos

Gráfico 3. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial por Bloque Temático.



Fuente: Elaboración propia / Matriz de sistematización de datos.

En la **Tabla 3** y **Gráfico 3** se muestra la comparación en el rendimiento de tres modelos de inteligencia artificial (ChatGPT-4, Gemini Advanced y Copilot) en la resolución de preguntas divididas en bloques temáticos (Ciencias Básicas, Medicina Interna, Ginecología y Obstetricia, Pediatría, Cirugía y Salud Pública).

Al analizar los bloques temáticos individualmente, se observan diferencias importantes. Gemini Advanced presenta el mayor porcentaje de aciertos en Ciencias Básicas con un 93.59% (σ = 2.3) y un rendimiento relativamente alto en Medicina Interna con un 88.44% (σ = 2.4), aunque superado por los otros dos modelos en esta última área. Chat GPT-40 presenta el mejor desempeño en Pediatría con un 87.65% (σ = 6.6) y un rendimiento similar en Cirugía con 88.98% (σ = 5.4). Copilot, por su parte, destaca en Ginecología y Obstetricia con un 91.29% (σ = 6.6) y también muestra el mejor rendimiento en Medicina Interna con 90.67% (σ = 4.2). En contraste, Copilot presenta los porcentajes más altos de desaciertos en Cirugía con un 13.14% (σ = 7.7) y, especialmente, en Salud Pública con un 20.22% (σ = 10.0), donde los tres modelos muestran el rendimiento más bajo y la mayor variabilidad. Estos resultados evidencian variaciones significativas en la precisión según el modelo y la temática evaluada.

"Análisis comparativo de los puntajes obtenidos por Chat GPT-4o, Gemini Advanced y Copilot al aplicarlas en el examen del Concurso Nacional de Admisión al Residentado Médico, realizado por el Consejo Nacional de Residentado Medico (CONAREME) de los años 2017-2024. Perú."

Tabla 4. Aciertos y Desaciertos según Modelo de Inteligencia Artificial (IA) por Bloque Temático y Año de Examen

Bloque Temático	Modelo de IA	2017			2018			2019			2020			2022			2023			2024			Promedio de Aciertos (%)	σ
		Total de Preguntas	Aciertos n (%)	Desaciertos n (%)	Total de Preguntas	Aciertos n (%)	Desaciertos n (%)	Total de Preguntas	Aciertos n (%)	Desaciertos n (%)	Total de Preguntas	Aciertos n (%)	Desaciertos n (%)	Total de Preguntas	Aciertos n (%)	Desaciertos n (%)	Total de Preguntas	Aciertos n (%)	Desaciertos n (%)	Total de Preguntas	Aciertos n (%)	Desaciertos n (%)		
Ciencias Básicas	Chat GPT-4o	10	9 (90.00)	1 (10.00)	10	10 (100.00)	0 (0.00)	21	21 (100.00)	0 (0.00)	12	12 (100.00)	0 (0.00)	8	7 (87.50)	1 (12.50)	6	5 (83.33)	1 (16.67)	11	10 (90.9)	1 (9.1)	93.11	6.37
	Gemini Advanced		9 (90.00)	1 (10.00)		10 (100.00)	0 (0.00)		21 (100.00)	0 (0.00)		12 (100.00)	0 (0.00)		7 (87.50)	1 (12.50)		5 (83.33)	1 (16.67)		9 (81.82)	2 (18.18)	91.81	7.51
	Copilot		9 (90.00)	1 (10.00)		10 (100.00)	0 (0.00)		20 (95.24)	1 (4.76)		12 (100.00)	0 (0.00)		7 (87.50)	1 (12.50)		5 (83.33)	1 (16.67)		10 (90.9)	1 (9.1)	92.3	5.83
Medicina Interna	Chat GPT-4o	58	56 (96.55)	2 (3.45)	60	46 (76.67)	14 (23.33)	73	59 (80.82)	14 (19.18)	57	48 (84.21)	9 (15.79)	76	72 (94.74)	4 (5.26)	80	78 (97.50)	2 (2.50)	89	81 (91.01)	8 (8.99)	88.79	7.63
	Gemini Advanced		52 (89.66)	6 (10.34)		52 (86.67)	8 (13.33)		63 (86.30)	10 (13.7)		48 (84.21)	9 (15.79)		69 (90.79)	7 (9.21)		73 (91.25)	7 (8.75)		79 (88.76)	10 (11.24)	88.23	2.40
	Copilot		56 (96.55)	2 (3.45)		54 (90.00)	6 (10.00)		61 (83.56)	12 (16.44)		50 (87.72)	7 (12.28)		73 (96.05)	3 (3.95)		73 (91.25)	7 (8.75)		80 (89.89)	9 (10.11)	90.72	4.21
Ginecología y Obstetricia	Chat GPT-4o	43	36 (83.72)	7 (16.28)	40	37 (92.50)	3 (7.50)	33	28 (84.85)	5 (15.15)	31	30 (96.77)	1 (3.23)	33	29 (87.88)	4 (12.12)	32	32 (100.00)	0 (0.00)	29	29 (89.66)	3 (10.34)	90.77	5.59
	Gemini Advanced		34 (79.07)	9 (20.93)		36 (90.00)	4 (10.00)		26 (78.79)	7 (21.21)		29 (93.55)	2 (6.45)		26 (78.79)	7 (21.21)		30 (93.75)	2 (6.25)		29 (89.66)	3 (10.34)	86.23	6.53
	Copilot		36 (83.72)	7 (16.28)		39 (97.50)	1 (2.50)		29 (87.88)	4 (12.12)		30 (96.77)	1 (3.23)		27 (81.82)	6 (18.18)		32 (100.00)	0 (0.00)		27 (93.10)	2 (6.90)	91.54	8.46
Pediatría	Chat GPT-4o	34	30 (88.24)	4 (11.76)	35	28 (80.0)	7 (20.0)	30	27 (90.0)	3 (10.0)	34	26 (76.47)	8 (23.53)	37	34 (91.89)	3 (8.11)	42	41 (97.62)	1 (2.38)	31	27 (87.10)	4 (12.90)	87.33	6.62
	Gemini Advanced		29 (85.29)	5 (14.71)		30 (85.71)	5 (14.29)		28 (93.33)	2 (6.67)		26 (76.47)	8 (23.53)		30 (81.1)	7 (18.9)		36 (85.71)	6 (14.29)		27 (87.10)	4 (12.90)	84.96	4.83
	Copilot		30 (88.24)	4 (11.76)		32 (91.73)	3 (8.57)		28 (93.33)	2 (6.67)		27 (79.41)	7 (20.59)		34 (91.9)	3 (8.1)		41 (97.62)	1 (2.38)		27 (87.10)	4 (12.90)	89.86	5.32
Cirugía	Chat GPT-4o	36	34 (94.44)	2 (5.56)	37	32 (86.49)	5 (13.51)	33	28 (84.85)	5 (15.15)	33	28 (84.85)	5 (15.15)	37	32 (86.49)	5 (13.51)	29	29 (100.00)	0 (0.00)	31	27 (87.10)	4 (12.90)	89.17	5.36
	Gemini Advanced		29 (80.56)	7 (19.44)		29 (78.38)	8 (21.62)		26 (78.79)	7 (21.21)		24 (72.73)	9 (27.27)		33 (89.19)	4 (10.81)		25 (86.21)	4 (13.79)		29 (93.55)	2 (6.45)	82.77	6.66
	Copilot		34 (94.44)	2 (5.56)		33 (89.19)	4 (10.81)		28 (84.19)	5 (15.15)		25 (75.76)	8 (24.24)		31 (83.78)	6 (16.22)		29 (100.00)	0 (0.00)		25 (80.65)	6 (19.35)	86.95	7.66
Salud Pública	Chat GPT-4o	19	15 (78.95)	4 (21.05)	18	14 (77.78)	4 (22.22)	10	9 (90.00)	1 (10.00)	13	12 (92.31)	1 (7.69)	9	7 (77.78)	2 (22.22)	11	11 (100.00)	0 (0.00)	9	6 (66.67)	3 (33.33)	83.35	10.43
	Gemini Advanced		15 (78.95)	4 (12.05)		15 (83.33)	3 (16.67)		9 (90.00)	1 (10.00)		10 (76.92)	3 (23.08)		6 (66.67)	3 (33.33)		10 (90.91)	1 (9.09)		8 (88.89)	1 (11.11)	82.24	8.13
	Copilot		16 (84.21)	3 (15.79)		13 (72.22)	5 (27.78)		8 (80.00)	2 (20.00)		12 (92.31)	1 (7.69)		6 (66.67)	3 (33.33)		10 (90.91)	1 (9.09)		6 (66.67)	3 (33.33)	79.00	9.98

Fuente: Elaboración propia / Matriz de sistematización de datos

En la **Tabla 04,** se muestran los resultados obtenidos por los modelos de inteligencia artificial evaluados en diferentes bloques temáticos y años de examen. Se evidencia diferencias importantes en su desempeño global en los bloques temáticos evaluados. Los modelos considerados (Chat GPT-4o, Gemini Advanced y Copilot) presentan fortalezas y debilidades particulares, reflejadas tanto en sus promedios generales de aciertos como en la variabilidad de sus resultados.

Copilot se posiciona como el modelo con el mejor promedio de aciertos, alcanzando un 90.72% y mostrando una desviación estándar moderada de 4.21. Este balance entre un alto porcentaje de aciertos y una variabilidad controlada refleja una capacidad sólida para abordar de manera consistente los distintos bloques temáticos. Por su parte, **Chat GPT-4o** logró un promedio de aciertos del 88.79%, pero con la mayor desviación estándar entre los modelos (σ=7.63), lo que indica un desempeño más variable. Finalmente, **Gemini Advanced** mostró un promedio similar al de Chat GPT-4o, con un 88.23%, pero con la desviación estándar más baja (σ=2.40), lo que resalta su uniformidad en los resultados.

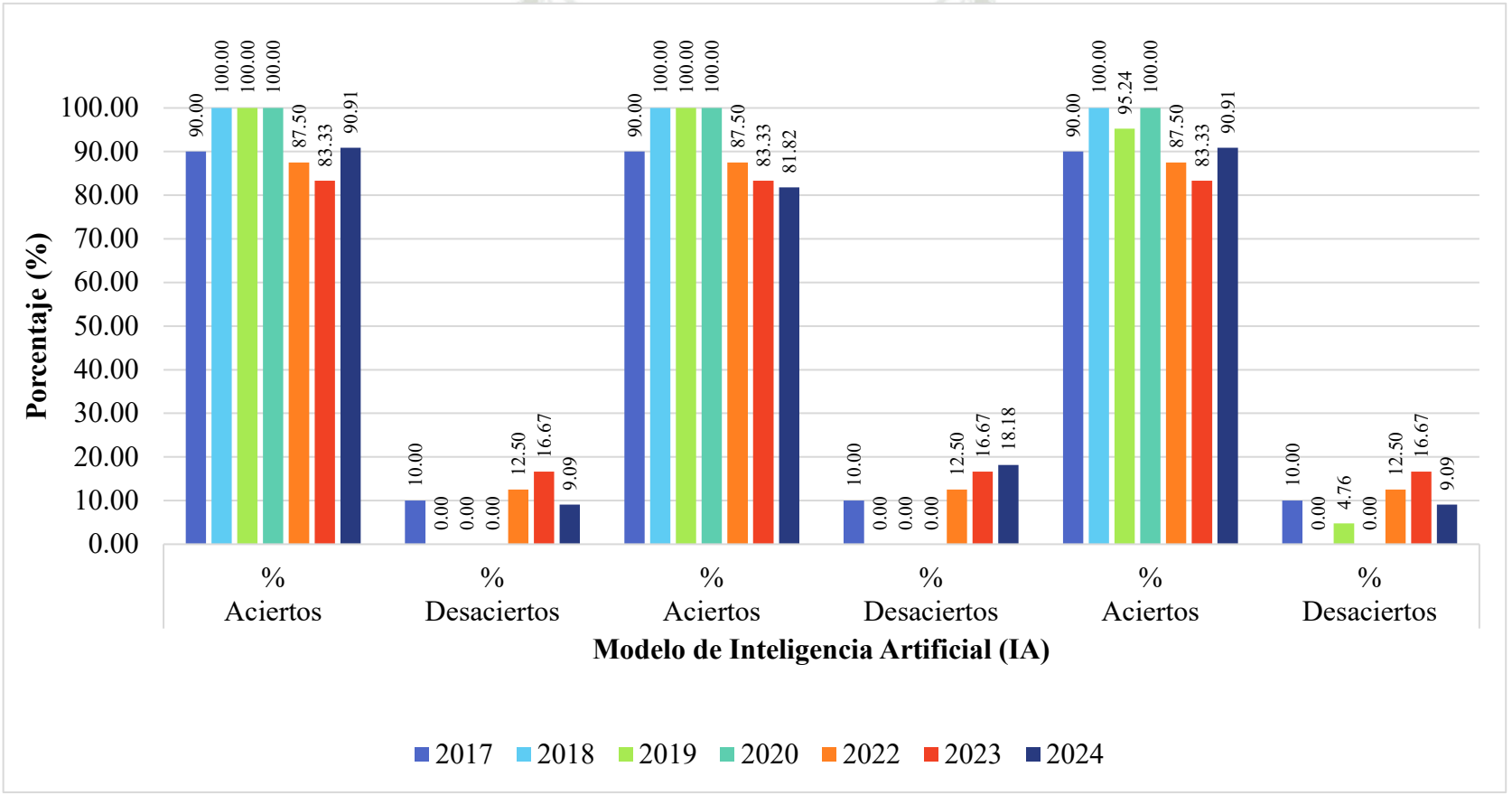
Al considerar el desempeño promedio de los modelos en todos los bloques, se observa que las áreas con mayor consistencia en los aciertos fueron **Salud Pública** y **Cirugía**, donde los tres modelos lograron resultados altos y similares entre sí. Estas áreas parecen representar un desafío menor para los sistemas de IA debido a la naturaleza más estructurada de los conocimientos evaluados.

En contraste, los bloques temáticos de **Pediatría** y **Medicina Interna** presentaron mayores diferencias entre los modelos y menores porcentajes de aciertos globales. Esto sugiere que estos bloques pueden tener un nivel de complejidad mayor o plantear retos específicos para los modelos de IA en términos de análisis contextual o integración de información clínica.

Desde una perspectiva comparativa, los datos reflejan que, aunque **Copilot** obtiene el mejor desempeño promedio general, **Gemini Advanced** ofrece una estabilidad notable que podría ser ventajosa en aplicaciones donde la consistencia de resultados es crucial. Por otro lado, aunque **Chat GPT-4o** destaca en varios bloques específicos, su variabilidad en los resultados puede representar un desafío en entornos donde se requiere uniformidad.

Análisis comparativo de los puntajes obtenidos por Chat GPT-4o, Gemini Advanced y Copilot al aplicarlas en el examen del Concurso Nacional de Admisión al Residentado Médico, realizado por el Consejo Nacional de Residentado Medico (CONAREME) de los años 2017-2024. Perú.

Gráfico 4. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial del Bloque Temático de Ciencias Básicas por Año de Examen.



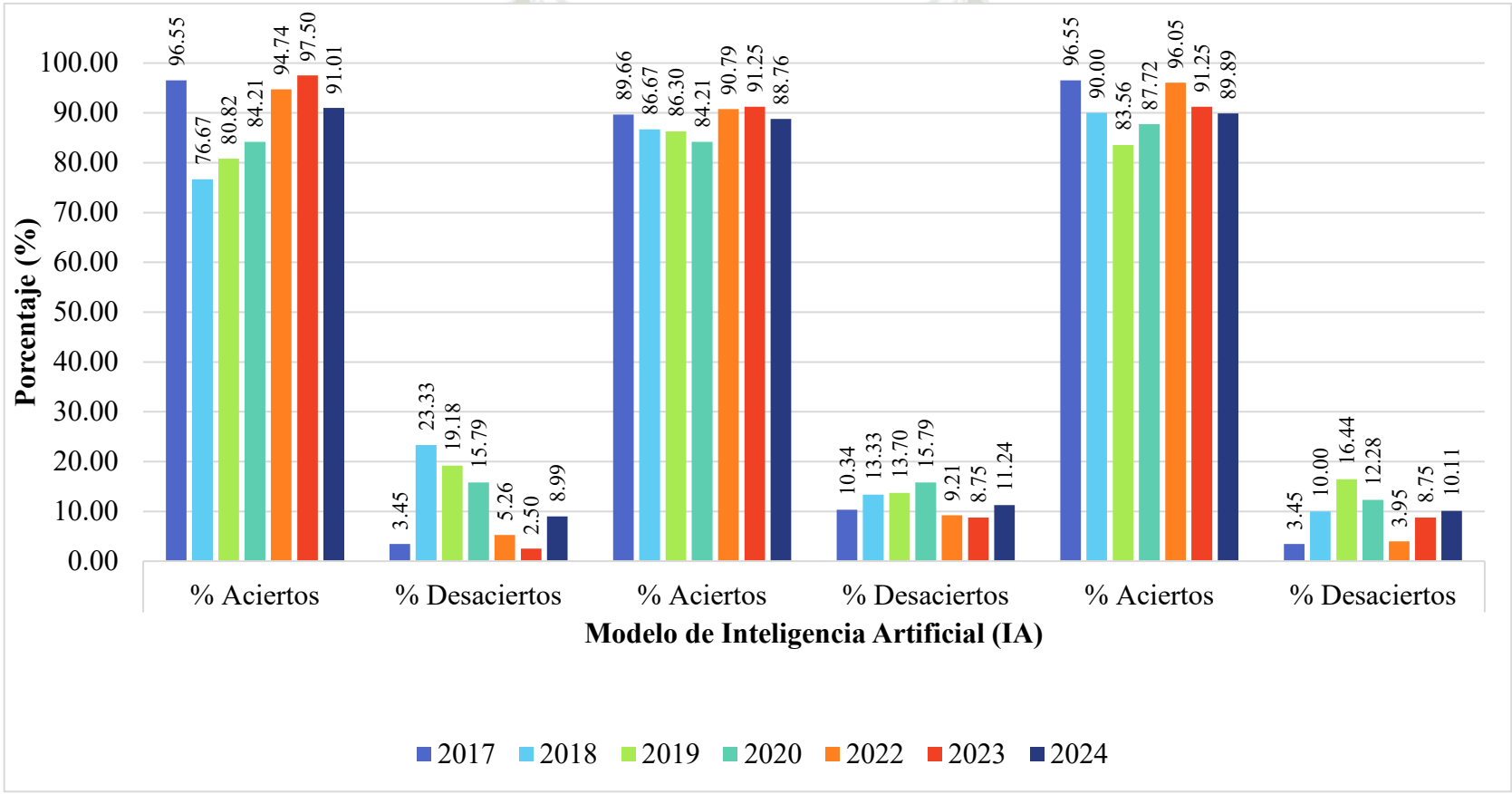
Fuente: Elaboración propia / Matriz de sistematización de datos

En la **Tabla 4** y **Gráfico 4**, se muestra la comparación en el rendimiento de tres modelos de IA (Chat GPT-4o, Gemini Advanced y Copilot) en el bloque temático de Ciencias Básicas.

Chat GPT-4º obtuvo el promedio de aciertos más alto con un 93.11% ($\sigma = 6.37$). Esto sugiere un rendimiento alto y relativamente consistente. Gemini Advanced mostró un promedio de aciertos del 91.81% ($\sigma = 7.51$) y la mayor desviación estándar. Aunque su promedio es ligeramente inferior a los otros dos modelos, presenta una variabilidad notable en su rendimiento. Copilot: Presentó un rendimiento promedio de aciertos del 92.43% ($\sigma = 5.83$). Esto indica una menor variabilidad entre los tres modelos, lo que sugiere una mayor consistencia.

Si bien los promedios de aciertos son similares, Copilot muestra una menor desviación estándar, indicando mayor consistencia interanual. Se observó un porcentaje de aciertos de 100% en varios años para los tres modelos, particularmente en 2018 y 2020, contrastando con un desempeño relativamente menor en 2023. Estos hallazgos sugieren el potencial de la IA como herramienta de apoyo en el estudio de Ciencias Básicas, aunque las variaciones anuales ameritan un análisis más profundo de las características específicas de cada examen.

Gráfico 5. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial del Bloque Temático de Medicina Interna por Año de Examen.



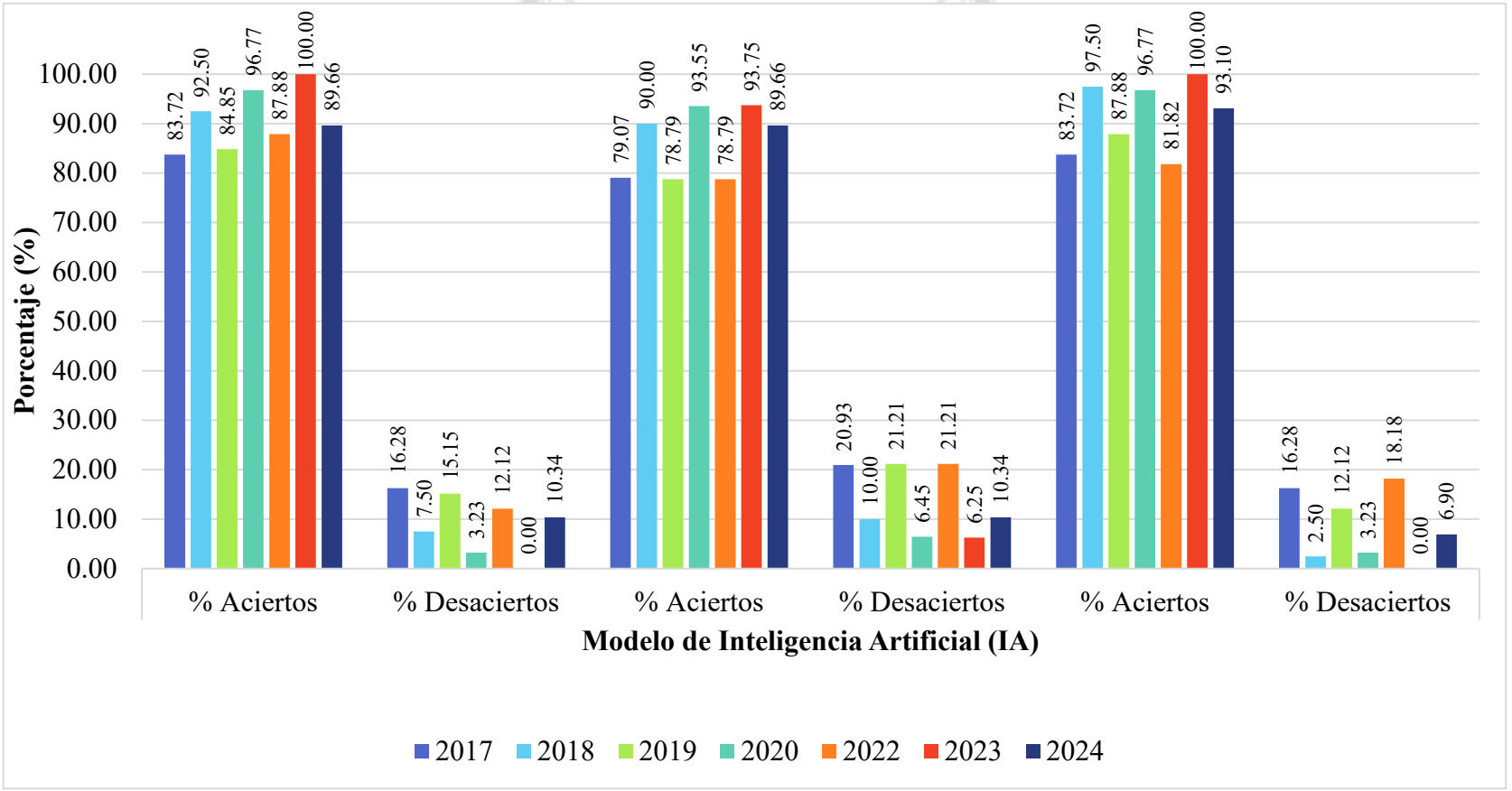
Fuente: Elaboración propia / Matriz de sistematización de datos

En la **Tabla 4** y **Gráfico 5** se muestra la comparación en el rendimiento de tres modelos de IA (Chat GPT-4o, Gemini Advanced y Copilot) en el bloque temático de Medicina Interna de exámenes del Residentado Médico (2017-2024), analizando aciertos y desaciertos anuales.

Chat GPT-4o obtuvo un promedio de aciertos del 88.79% (σ = 7.63). Si bien su promedio es alto, la desviación estándar relativamente mayor indica una cierta variabilidad en su rendimiento a lo largo de los años. Gemini Advanced mostró un promedio de aciertos del 88.23% (σ = 2.40) y la menor desviación estándar, lo que sugiere un rendimiento muy consistente a lo largo de los años, a pesar de tener un promedio ligeramente inferior a los otros dos modelos. Copilot presentó el mejor rendimiento promedio de aciertos de 90.72% (σ = 4.21), lo que sugiere un rendimiento alto y bastante consistente, aunque con una variabilidad ligeramente superior a Gemini Advanced.

Si bien Copilot exhibió un ligero predominio en el promedio de aciertos, Gemini Advanced demostró una mayor consistencia interanual, reflejada en su menor desviación estándar. Se observaron variaciones en el rendimiento a lo largo de los años.

Gráfico 6. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial del Bloque Temático de Ginecología y Obstetricia por Año de Examen.



Fuente: Elaboración propia / Matriz de sistematización de datos

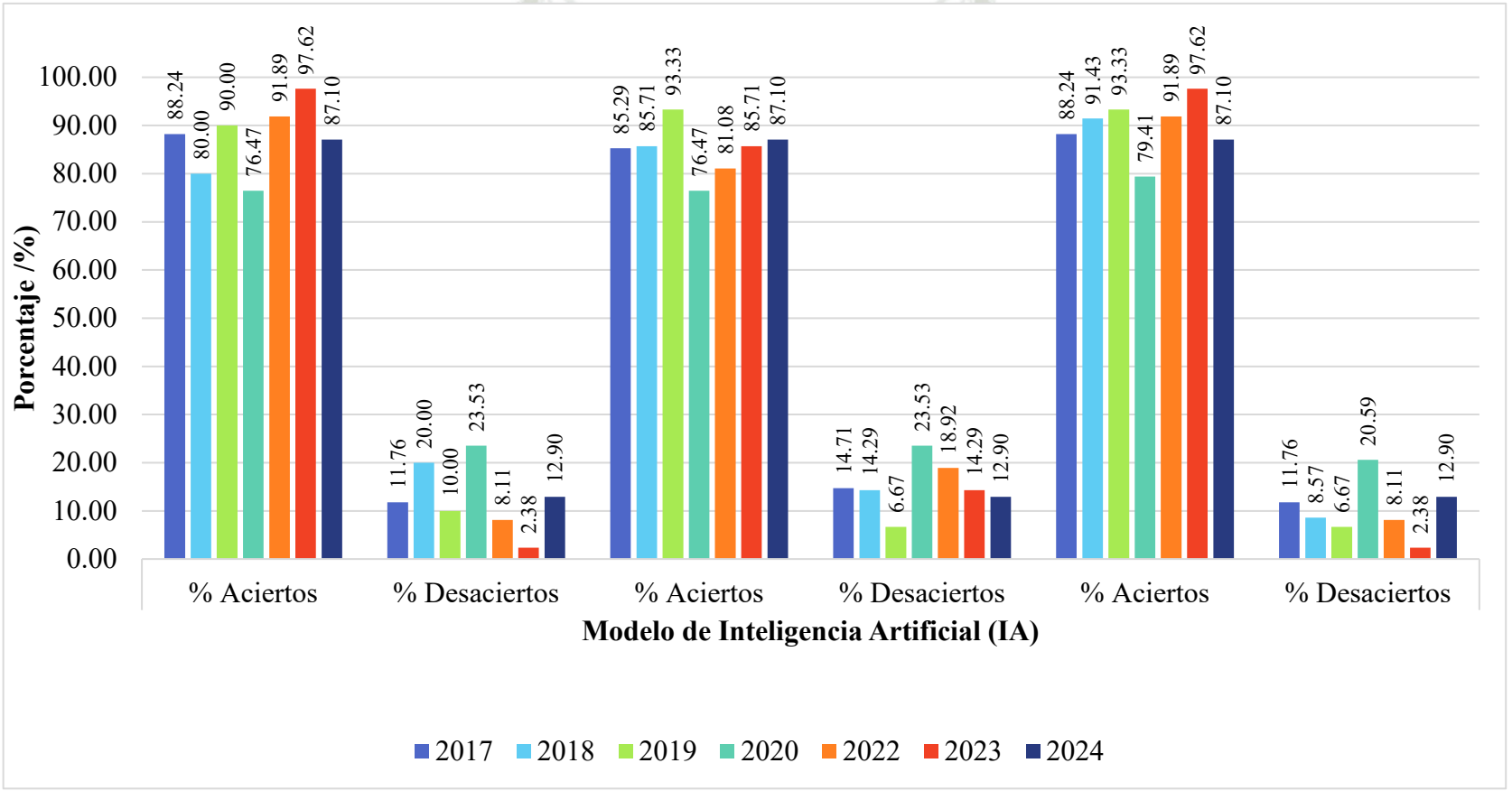
En la **Tabla 4** y **Gráfico 6**, se muestra la comparación en el rendimiento de tres modelos de IA (Chat GPT-4o, Gemini Advanced y Copilot) en el bloque temático de Ginecología y Obstetricia de exámenes del Residentado Médico (2017-2024).

Chat GPT-4o obtuvo el promedio de aciertos más alto (90.77%) y la menor desviación estándar ($\sigma = 5.59$), indicando una alta precisión y consistencia. Copilot, con un promedio de 91.54%, aunque ligeramente superior en promedio a Chat GPT-4o, presentó una mayor desviación estándar ($\sigma = 6.61$), lo que sugiere una menor consistencia en comparación. Gemini Advanced mostró el promedio de aciertos más bajo (86.23%) y la mayor variabilidad ($\sigma = 6.53$).

Si bien hubo variación en el número de preguntas por año, los tres modelos mostraron un rendimiento relativamente alto en 2020 y 2023, con un rendimiento perfecto (100% de aciertos) para Chat GPT-4o y Copilot en 2023. En contraste, el rendimiento tendió a ser menor en 2019 y 2022 para Gemini Advanced y Copilot.

Los resultados revelan que los tres modelos demostraron un rendimiento notablemente alto, con promedios de aciertos que oscilaron entre el 86% y el 92%, lo que sugiere un potencial significativo de la IA. Se destaca a Chat GPT-4o por su consistencia y a Copilot por su ligero mejor promedio general.

Gráfico 7. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial del Bloque Temático de Pediatría por Año de Examen.



Fuente: Elaboración propia / Matriz de sistematización de datos

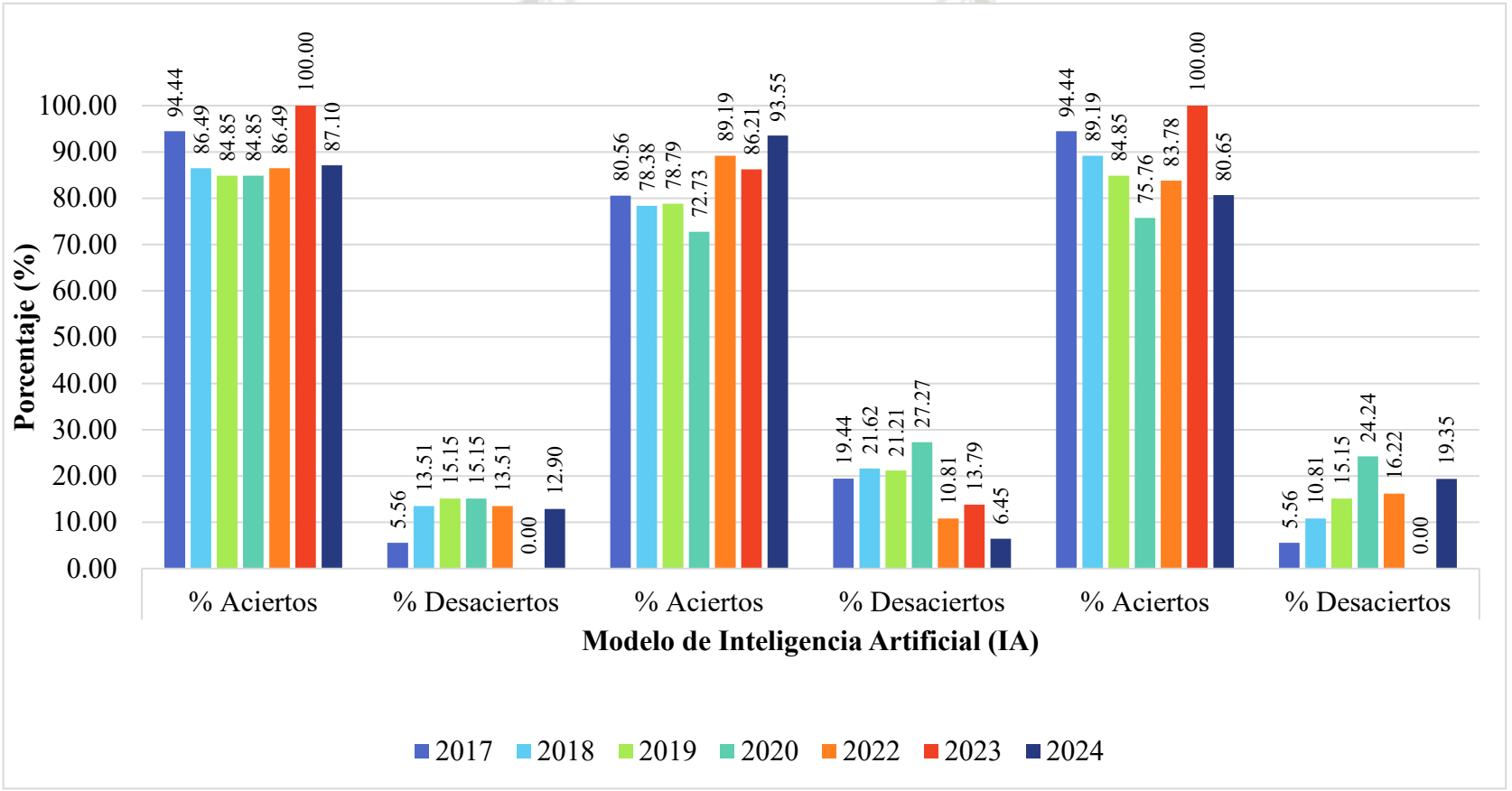
En la **Tabla 4** y **Gráfico 7**, se muestra la comparación en el rendimiento de tres inteligencias artificiales (Chat GPT-4.0, Gemini Advanced y Copilot) en el bloque de Pediatría de los exámenes de Residentado Médico entre 2017 y 2024.

Chat GPT-4.0 presentó un promedio de aciertos del 87.33% ($\sigma = 6.62$), destacándose como el modelo más consistente y preciso, con el mayor porcentaje de aciertos en la mayoría de los años evaluados. Esto indica un rendimiento bueno, aunque con cierta variabilidad a lo largo de los años. Gemini Advanced alcanzó un promedio de aciertos del 84.96% ($\sigma = 4.83$), teniendo el promedio de aciertos más bajo de los tres modelos, pero con una mayor consistencia en su porcentaje de aciertos interanual en comparación con los otros dos modelos. Copilot obtuvo el promedio más alto, con un 89.86% ($\sigma = 5.3$) de aciertos, pero con mayor variabilidad; lo que sugiere un rendimiento alto y relativamente consistente

Si bien Copilot tiene el promedio de aciertos ligeramente más alto, Chat GPT-4o destaca por su mayor consistencia, reflejada en su menor desviación estándar. Gemini Advanced, aunque también muestra un buen rendimiento en algunos años, presenta una mayor fluctuación en sus resultados.

Se puede notar que los tres modelos tienden a mostrar un rendimiento particularmente alto en los años 2019, 2022 y 2023, con Copilot y Chat GPT-4o alcanzando un 97.62% de aciertos en 2023. El año 2020 presenta el rendimiento más bajo para los tres modelos.

Gráfico 8. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial del Bloque Temático de Cirugía por Año de Examen.



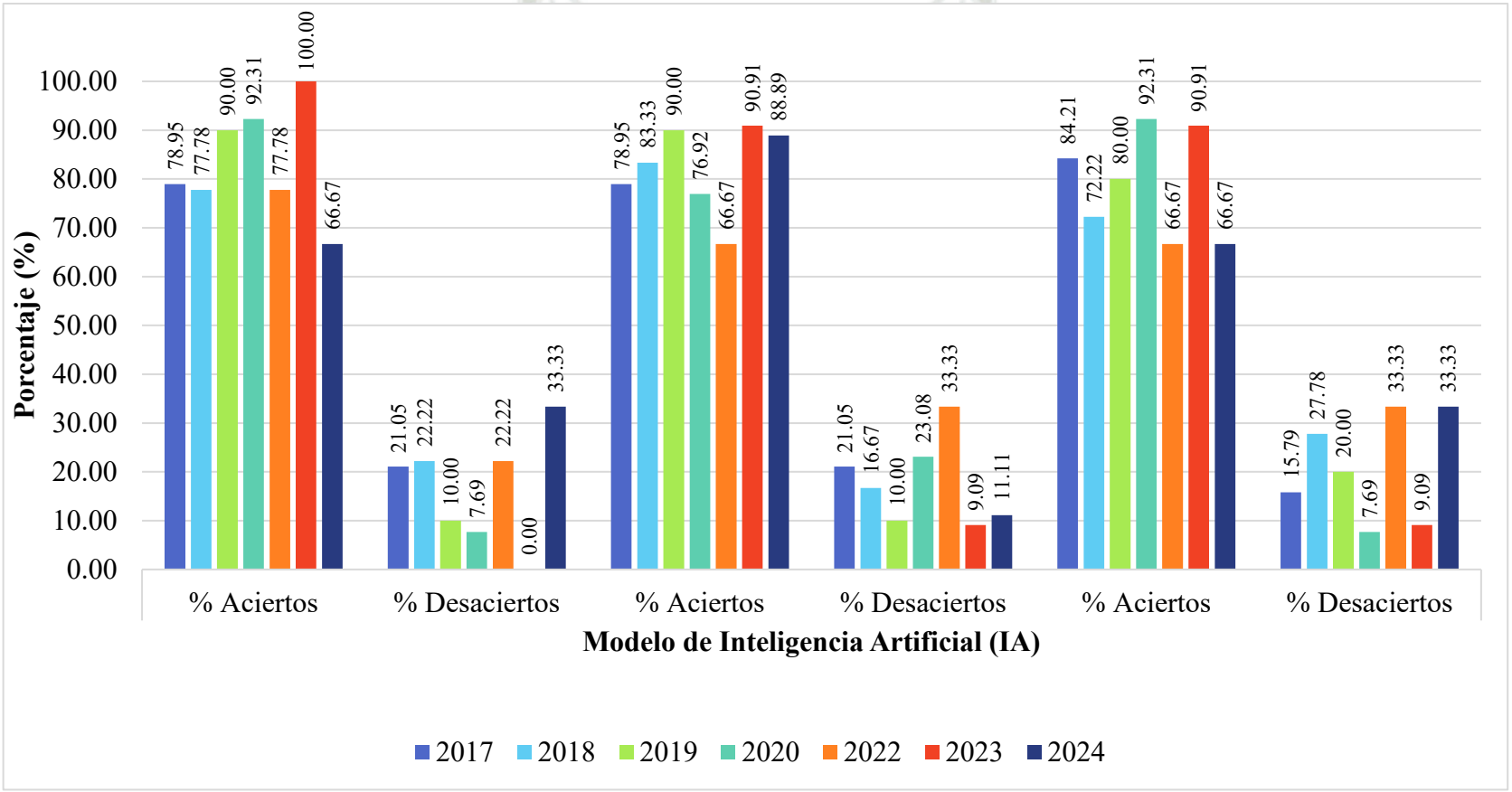
Fuente: Elaboración propia / Matriz de sistematización de datos

En la **Tabla 4** y **Gráfico 8**, se muestra la comparación en el rendimiento de tres modelos de inteligencia artificial (IA) en la predicción de respuestas correctas en el bloque temático de cirugía de los exámenes de residencia médica, evaluados anualmente entre 2017 y 2024.

ChatGPT-4o obtuvo un promedio de aciertos del 89.17% ($\sigma = 5.36$). Esto indica un rendimiento bueno y relativamente consistente a lo largo de los años. Destaca su porcentaje de aciertos perfecto (100%) en el examen de 2023.

Gemini Advanced mostró el promedio de aciertos más bajo de los tres modelos, con un 82.77% ($\sigma = 6.66$). Esto indica un procentaje de aciertos menos consistente y un promedio inferior en comparación con los otros dos modelos. Copilot presentó un rendimiento promedio de aciertos del 86.95% ($\sigma = 7.66$). Si bien su promedio es competitivo, su mayor desviación estándar indica la mayor variabilidad en su rendimiento. Al igual que Chat GPT-4o, obtuvo un rendimiento perfecto en el examen de 2023. Chat GPT-4º ha demostrado consistentemente un mayor porcentaje de aciertos, seguido de Copilot y Gemini Advanced. Si bien los tres modelos han mostrado una capacidad considerable para predecir correctamente las respuestas, ChatGPT-40 ha mantenido un rendimiento superior, especialmente en los exámenes de los últimos años.

Gráfico 9. Porcentaje (%) de Aciertos y Desaciertos según Modelo de Inteligencia Artificial del Bloque Temático de Salud Pública por Año de Examen.



Fuente: Elaboración propia / Matriz de sistematización de datos

En la **Tabla 4** y **Gráfico 9**, se muestra la comparación entre los resultados obtenidos por diferentes modelos de inteligencia artificial (Chat GPT-4.0, Gemini Advanced y Copilot) al responder preguntas del bloque de Salud Pública.

Chat GPT-4O Obtuvo el promedio de aciertos más alto con un 83.35% (σ = 10.43). Esto representa una mayor variabilidad en su porcentaje de aciertos en los diferentes exámenes aplicados. Destaca su porcentaje de aciertos perfecto (100%) en el examen de 2023, pero también presenta el porcentaje más bajo (66.67%) en 2024. Gemini Advanced mostró un promedio de aciertos del 82.24% (σ = 8.13). Su rendimiento es relativamente similar al de Chat GPT-4o en promedio, pero con una menor variabilidad. Copilot presentó el promedio de aciertos más bajo, con un 79.00% (σ = 9.98). Su rendimiento también muestra variabilidad, aunque ligeramente menor que la de Chat GPT-4o.

En comparación con otras áreas (como Cirugía, Pediatría o Ginecología), el rendimiento general en Salud Pública es ligeramente inferior. Esto podría sugerir que las preguntas de Salud Pública presentan un mayor desafío para los modelos de IA, posiblemente debido a la naturaleza más contextual de algunas de estas preguntas. Se observa que Chat GPT-4.0 mantiene consistentemente el mayor número de aciertos y el menor porcentaje de desaciertos a lo largo de los años, destacándose como el modelo más preciso. Copilot evidencia un mayor porcentaje de desaciertos, indicando menor eficacia en esta tarea específica.

Tabla 05. Análisis de Varianza (ANOVA) de los Porcentajes de Aciertos entre los grupos de Modelos de Inteligencia Artificial (IA)

Fuente de Variacion	Suma de cuadrados	gl	Media cuadrática	F	Valor de p (α = 0.05)
Entre grupos	50.88	2	25.44	1.837	0.188
Dentro de grupos	249.23	18	13.85		
Total	300.11	20			

Fuente: Elaboración propia / Matriz de sistematización de datos

En la **Tabla 5,** se muestra los resultados del análisis de varianza (ANOVA) según el porcentaje de aciertos entre y dentro de los grupos de IA's. Los resultados del ANOVA revelaron que no se puede confirmar que existen diferencias estadísticamente significativas entre los grupos de modelos de IA ($F(2, 18) = 1.837$, $p = 0.188$, $\alpha = 0.05$). La suma de cuadrados entre grupos fue de 50.88. La suma de cuadrados dentro de los grupos fue de 249,23. Este patrón indica que la mayor parte de la variabilidad observada en los datos se debe a la variación dentro de los grupos, y no a diferencias entre los grupos de IA. Estos resultados estadísticos son lo suficientemente altos como para concluir que las diferencias observadas probablemente se deben al azar o a la variabilidad inherente dentro de cada grupo, y no a un efecto real del tipo de IA.

La descomposición de la variabilidad total en sus componentes revela que la mayor parte de la variación en los porcentajes de aciertos se explica por la variabilidad *dentro* de cada grupo de modelos de IA (SCD = 249.229), en lugar de la variabilidad *entre* los grupos (SCE = 50.881). Esto refuerza la conclusión de que no existen diferencias significativas entre los grupos. La variabilidad dentro de los grupos podría deberse a diversos factores, como diferencias individuales en las pruebas, errores de medición u otros factores no controlados en el estudio.

Análisis comparativo de los puntajes obtenidos por Chat GPT-4o, Gemini Advanced y Copilot al aplicarlas en el examen del Concurso Nacional de Admisión al Residentado Médico, realizado por el Consejo Nacional de Residentado Medico (CONAREME) de los años 2017-2024. Perú.

Tabla 06. Pruebas post hoc HSD de Tukey para la comparación del porcentaje de aciertos entre grupos de Modelos de Inteligencia Artificial (IA)

Modelo de IA (I)	Modelo de IA (J)	Diferencia de medias (I-J)	Desv. Error	Valor de p	IC al 95%	
					Límite inferior	Límite superior
Chat GPT-4o	Gemini Advanced	2.83	1.99	0.35	-2.24	7.91
	Copilot	-0.80	1.99	0.92	-5.87	4.28
Gemini Advanced	Chat GPT-4o	-2.83	1.99	0.35	-7.91	2.24
	Copilot	-3.63	1.99	0.19	-8.70	1.45
Copilot	Chat GPT-4o	0.80	1.99	0.92	-4.28	5.87
	Gemini Advanced	3.63	1.99	0.19	-1.45	8.70

Fuente: Elaboración propia / Matriz de sistematización de datos

En la **Tabla 6,** presenta los resultados de las pruebas *post hoc* HSD de Tukey para la comparación del porcentaje de aciertos entre grupos de Modelos de Inteligencia Artificial (IA). En todos los casos, los intervalos de confianza al 95% incluyen el cero, lo que indica que no existen diferencias estadísticamente significativas entre ningún par de modelos de IA. En todas las comparaciones realizadas, los valores de p fueron mayores al nivel de significancia ($\alpha = 0.05$), lo que sugiere la ausencia de diferencias significativas en la comparación entre estos grupos.

CAPÍTULO IV
DISCUSIÓN

4. DISCUSIÓN

Los resultados de este estudio revelaron que los modelos de inteligencia artificial evaluados (ChatGPT-4O, Gemini Advanced y Copilot AI) lograron desempeñarse de manera efectiva al resolver preguntas del Concurso Nacional de Admisión al Residentado Médico en Perú. En términos generales, no se encontraron diferencias estadísticamente significativas en el porcentaje de aciertos entre los tres modelos (ANOVA: F = 1.837, p=0.188)). Este hallazgo es consistente con investigaciones previas. En un estudio realizado por investigadores peruanos, se evidencia el desempeño de Chat GPT-3.5 y Chat GPT-4 al resolver el Examen Nacional de Medicina, en el cual dichos modelos obtuvieron una precisión de 86% y 77% respectivamente (29). En otro estudio donde se evalúa la performance de ChatGPT al resolver exámenes similares en los Estados Unidos; ChatGPT-4 respondió 88% de respuestas correctas en el Paso 1 (Step 1) del Examen de Licencia Médica de Estados Unidos (USMLE, United States Medical Licensing Examination), hasta un 86% de aciertos en el Paso 2 de Conocimiento Clínico (Step 2CK); y hasta un 90% de aciertos en el Paso 3 (Step 3)(30). En otra investigación, ChatGPT-4O mostró un rendimiento similar en el USMLE Step 2CK con una tasa de aciertos del 87.2% (31). Aunque hasta nuestra revisión de la literatura no existen revisiones similares entre los tres modelos usados en nuestro estudio para resolver exámenes complejos de Medicina, al no haber diferencias significativas entre los modelos, los otros modelos de Inteligencia Artificial podrían tener un desempeño similar en otros tipos de exámenes.

Al comparar nuestro estudio con otras investigaciones similares, pero que utilizaron versiones anteriores de los modelos de Inteligencia Artificial, observamos una concordancia en los resultados entre las versiones más modernas (32). Esto sugiere que la mejora en el rendimiento de versiones anteriores a las utilizadas en nuestro estudio se debe a un mayor proceso de razonamiento de la IA más que a la aleatoriedad.

En el análisis por áreas temáticas, ChatGPT-4O destacó en bloques como Ciencias Básicas (94.87%) y Medicina Interna (89.24%), resultados que reflejan su entrenamiento extensivo en datos médicos y científicos. Sin embargo, en bloques como Salud Pública, se observó una mayor variabilidad en los resultados, lo cual concuerda con la literatura que reporta que las IA tienen dificultad para interpretar datos poblacionales y epidemiológicos a nivel país (29). Esta similitud con otros estudios podría deberse a que las preguntas son normalmente formuladas en un contexto nacional y de interpretación complejo.

Por otro lado, Copilot mostró un desempeño destacado en bloques como Pediatría (90.12%) y Ginecología y Obstetricia (91.28%). En contraste, Gemini Advanced presentó el promedio de aciertos más bajo (86.18%) y mayor variabilidad en áreas como Cirugía (82.62%). Estas observaciones sugieren que su diseño, aunque adecuado para tareas generales, podría no estar optimizado para responder preguntas de alta especialización médica, lo cual también ha sido señalado en investigaciones previas.

El diseño metodológico de este estudio, basado en la aplicación de los exámenes publicados del Concurso Nacional de Admisión al Residentado Médico en Perú en modelos de IA, los cuales se generaron con una nueva sesión de chat para cada pregunta, permitió evitar sesgos relacionados con la generación de memoria por parte de los

modelos. Esto garantiza que las respuestas sean independientes y contextualmente adecuadas a cada pregunta. Además, el uso de herramientas estadísticas como ANOVA y pruebas post hoc de Tukey permitió realizar un análisis de las diferencias entre los modelos adecuado.

Sin embargo, la metodología presenta algunas limitaciones. Aunque se consideraron preguntas de diferentes años (2017-2024), no se evaluaron posibles cambios en la dificultad del examen ni se ajustaron los datos en función de estas variaciones. Esto podría haber influido en los resultados, particularmente en años donde los tres modelos tuvieron desempeños más bajos, como en 2020.

La hipótesis principal del estudio sugería que existirían diferencias significativas entre los puntajes obtenidos por los tres modelos de IA al resolver las preguntas del Concurso Nacional de Admisión al Residentado Médico. Sin embargo, los resultados del análisis ANOVA demostraron que estas diferencias no fueron estadísticamente significativas ($F = 1.837$, $p=0.188$).

A pesar de esta falta de significancia estadística, los datos descriptivos mostraron que Copilot obtuvo el mejor rendimiento promedio (89.81%), seguido de ChatGPT-4O (89.01%) y Gemini Advanced (86.18%). Estos resultados sugieren que, aunque las diferencias absolutas son pequeñas, podrían reflejar fortalezas específicas de cada modelo en el manejo de ciertos tipos de preguntas o bloques temáticos. Esto es particularmente evidente en bloques como Ciencias Básicas, donde ChatGPT-4O obtuvo el porcentaje más alto de aciertos, y en Pediatría, donde Copilot lideró.

Durante el proceso de evaluación, se registró que cada IA fue capaz de proporcionar respuestas detalladas y fundamentadas ampliamente, no solo en áreas básicas sino también en aspectos clínicos avanzados. Estas herramientas no solo seleccionaron alternativas correctas, sino que también argumentaron sus elecciones y descartaron opciones incorrectas, demostrando una capacidad analítica integral. Este nivel de razonamiento, que incluye la interpretación de exámenes auxiliares y el análisis de contexto clínico, refuerza su potencial como herramientas docentes y de consultoría en el ámbito médico y educativo.

Implicancias

Este estudio tiene varias implicaciones tanto para la educación médica como para la investigación sobre modelos de inteligencia artificial en el contexto del Examen del Concurso Nacional de Admisión al Residentado Médico. En primer lugar, demostramos que ChatGPT-4O, Gemini Advanced y Copilot pueden abordar con eficacia preguntas de alta complejidad clínica, alcanzando niveles de desempeño similares a los de los postulantes médicos en el examen. Aunque nuestro estudio no reproduce exactamente las condiciones del examen real, los puntajes obtenidos por los modelos reflejan un nivel competitivo, especialmente en áreas como Ciencias Básicas y Medicina Interna. Estos hallazgos sugieren que las IA avanzadas podrían utilizarse para simular evaluaciones y personalizar el aprendizaje de acuerdo con las necesidades de los estudiantes, adaptando cada modelo a materias o niveles de dificultad específicos. Por ejemplo, ChatGPT-4O, con su desempeño superior en áreas teóricas y de razonamiento clínico, podría ser ideal para apoyar la preparación de estudiantes avanzados, mientras que Gemini Advanced y Copilot AI podrían ser útiles en etapas iniciales o para preguntas más estructuradas.

En segundo lugar, descubrimos que las áreas temáticas donde los modelos tienden a fallar, como Salud Pública y Cirugía, están asociadas a preguntas con alta especificidad contextual o ambigüedad. Esto abre nuevas oportunidades de investigación para analizar los patrones de error de las IA y explorar las razones detrás de estas fallas. Además, el análisis psicométrico aplicado en este estudio permite identificar las fortalezas y debilidades de los modelos en diferentes bloques temáticos, proporcionando una base sólida para desarrollar futuros estudios

enfocados en mejorar el rendimiento de los modelos en áreas específicas. Estudios posteriores podrían explorar diferentes teorías, como el modelado de diagnóstico cognitivo u otros modelos de clasificación de diagnóstico con conjuntos de datos más grandes, en busca de una comprensión más profunda del proceso de razonamiento de ChatGPT.

En tercer lugar, la incorporación de ajustes en las indicaciones o "prompts" y el uso de preguntas con mayor nivel de contexto podría mejorar significativamente el desempeño de los modelos. Por ejemplo, personalizar las indicaciones para incluir roles clínicos específicos o escenarios detallados podría optimizar la capacidad de las IA para resolver preguntas complejas. Este enfoque sugiere que la ingeniería de indicaciones puede desempeñar un papel crucial en la integración de IA en la educación médica, ya sea en la creación de herramientas de evaluación, el diseño de planes de estudio personalizados o la simulación de interacciones médico-paciente. Asimismo, el desarrollo de modelos de lenguaje entrenados con datos médicos locales y específicos podría mejorar aún más su aplicabilidad en contextos nacionales como el Examen del Concurso Nacional de Admisión al Residentado Médico en Perú.

Por último, aunque los resultados destacan el excelente desempeño de ChatGPT-4O, Gemini Advanced y Copilot AI al aplicarles estos exámenes, la práctica de la medicina requiere habilidades que van más allá de responder correctamente preguntas de opción múltiple; sino, es más bien un proceso continuo de aprendizaje y aplicación de competencias como la comunicación, empatía, la colaboración interdisciplinaria y el pensamiento crítico (16). Es fundamental incluir talleres y capacitaciones sobre el uso ético y eficaz de la inteligencia artificial en el ámbito de educación médica. Esto con el propósito de promover sus usos de manera responsable, promoviendo el pensamiento crítico y el proceso de aprendizaje. Es necesario realizar estudios adicionales que aborden la viabilidad, seguridad y los aspectos éticos del uso de la inteligencia artificial en la medicina. Es esencial garantizar que estas herramientas complementen, en lugar de sustituir, el razonamiento crítico y el juicio clínico de los médicos, promoviendo así un uso equilibrado y ético en la educación y la práctica médica.

Limitaciones del Estudio

Primera. Las respuestas proporcionadas por las IA dependen en gran medida de la claridad y especificidad de las preguntas. Así mismo, la IA es capaz de generar memoria. A pesar de haber utilizado un nuevo chat para cada pregunta, desconocemos si las IA's almacenan los chats eliminados, lo que podría generar variabilidad no atribuible a las capacidades intrínsecas de los modelos.

Segunda. Al no haberse realizado un análisis cualitativo de las respuestas incorrectas, se podría haber proporcionado información valiosa sobre las áreas específicas donde los modelos tienden a fallar y las razones detrás de estos errores, como la ambigüedad en las preguntas o la falta de contexto clínico en los datos de entrenamiento de los modelos.

Tercera. Problemas ocasionales con la conexión a internet o la sobresaturación de datos presentados a las IA pudieron haber afectado la recopilación de respuestas y alterado la consistencia de los resultados.

Cuarta. Los modelos de inteligencia artificial son dinámicos y pueden producir respuestas variables según la formulación de las preguntas o las actualizaciones realizadas en sus sistemas, lo que introduce un factor de incertidumbre en la replicación de los resultados.

Quinta. Aunque el periodo evaluado abarca varios exámenes de diferentes años, no se analizó versiones anteriores de las IA's, limitando la posibilidad de evaluar mejoras significativas en su desempeño.

Estas herramientas tienen un alto potencial para aplicaciones tanto educativas como clínicas, proporcionando razonamientos detallados que pueden enriquecer la enseñanza médica y optimizar la práctica clínica diaria. Sin embargo, su implementación requiere un análisis continuo y optimizaciones adicionales para superar sus limitaciones actuales y maximizar su impacto en el ámbito médico.

CAPÍTULO V
CONCLUSIONES

5. CONCLUSIONES

Primero. No existen diferencias estadísticamente significativas entre los modelos de inteligencia artificial evaluados (ChatGPT-4o, Gemini Advanced y Copilot) en términos de puntajes obtenidos en los exámenes del Concurso Nacional de Admisión al Residentado Médico, según el análisis de varianza (ANOVA, $p = 0.188$).

Segundo. Chat GPT-4o destacó por su consistencia y desempeño sólido, obteniendo un promedio de aciertos del 89.01% en todos los exámenes evaluados.

Tercero. Gemini Advanced obtuvo un promedio general de aciertos del 86.18%, siendo el modelo con mayor variabilidad en su desempeño entre los años y los bloques temáticos.

Cuarto. Copilot fue el modelo con mejor desempeño general, alcanzando un promedio de aciertos del 89.81%.

CAPÍTULO VI

RECOMENDACIONES

6. RECOMENDACIONES

Primera. Dado el potencial de estos modelos como herramientas educativas, se sugiere su incorporación en entornos de aprendizaje para estudiantes de medicina. Esto incluye su uso en la preparación para exámenes, resolución de preguntas prácticas, análisis de casos clínicos, simulaciones y evaluaciones. Además, su capacidad para proporcionar retroalimentación inmediata y fomentar el aprendizaje adaptativo, cuando se utiliza bajo la guía docente, representa una oportunidad significativa para enriquecer la experiencia educativa.

Segunda. Se recomienda la utilización de modelos de inteligencia artificial como Chat GPT-4o, Gemini Advanced y Copilot como herramientas complementarias al trabajo docente en la elaboración, revisión y análisis de preguntas de exámenes de medicina. Estas herramientas pueden optimizar el proceso de diseño y evaluación, asegurando mayor precisión y eficiencia en la preparación de exámenes estandarizados.

Tercera. Recomendamos a las Facultades de Medicina en Perú el desarrollo de espacios de innovación tecnológica que integren estas herramientas de inteligencia artificial en la capacitación docentes y formación de los estudiantes. Este fortalecimiento institucional contribuirá a mejorar la calidad educativa, permitiendo a las instituciones posicionarse como referentes en la adopción de tecnologías emergentes para la educación médica.

Cuarta. Se recomienda fomentar la realización de estudios científicos adicionales que profundicen en el análisis de los modelos de inteligencia artificial evaluados, en diferentes contextos educativos y clínicos para obtener mayor conocimiento de las capacidades y limitaciones de los modelos en el ámbito médico, así como para convertirse en herramientas más eficaces y fiables en la educación médica. Además, es esencial desarrollar e implementar modelos de inteligencia artificial específicos para el entorno educativo médico.

CAPÍTULO VII

REFERENCIAS

REFERENCIAS

1. Russell S, Norvig P. Artificial Intelligence, Global Edition [Internet]. Pearson Deutschland; 2021 [citado 20 de enero de 2025]. Disponible en: https://elibrary.pearson.de/book/99.150005/9781292401171

2. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. Expert Syst Appl. 1 de enero de 2024;235:121186.

3. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 4 de mayo de 2023;6:1169595.

4. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye. septiembre de 2024;38(13):2530-5.

5. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. JMIR Med Inform. 10 de febrero de 2022;10(2):e32875.

6. Kufel J, Bargieł-Łączek K, Kocot S, Koźlik M, Bartnikowska W, Janik M, et al. What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. Diagnostics. 3 de agosto de 2023;13(15):2582.

7. Boers TGW, Fockens KN, van der Putten JA, Jaspers TJM, Kusters CHJ, Jukema JB, et al. Foundation models in gastrointestinal endoscopic AI: Impact of architecture, pre-training approach and data efficiency. Med Image Anal. diciembre de 2024;98:103298.

8. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J Med Internet Res. 22 de agosto de 2023;25:e48659.

9. Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. Nat Hum Behav. septiembre de 2023;7(9):1526-41.

10. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 9 de febrero de 2023;2(2):e0000198.

11. Carrasco JP, García E, Sánchez DA, Porter E, Puente LDL, Navarro J, et al. ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. Rev Esp Educ Médica [Internet]. 16 de febrero de 2023 [citado 20 de enero de 2025];4(1). Disponible en: https://revistas.um.es/edumed/article/view/556511

12. Lo CK. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Educ Sci. abril de 2023;13(4):410.

13. Xu T, Weng H, Liu F, Yang L, Luo Y, Ding Z, et al. Current Status of ChatGPT Use in Medical Education: Potentials, Challenges, and Strategies. J Med Internet Res. 28 de agosto de 2024;26:e57896.

14. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 2024;17(5):926-31.

15. Horiuchi D, Tatekawa H, Oura T, Oue S, Walston SL, Takita H, et al. Comparing the Diagnostic Performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and Radiologists in Challenging Neuroradiology Cases. Clin Neuroradiol. diciembre de 2024;34(4):779-87.

16. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. J R Soc Med. mayo de 2023;116(5):181-2.

17. Curioso WH, Brunette MJ, Curioso WH, Brunette MJ. Inteligencia artificial e innovación para optimizar el proceso de diagnóstico de la tuberculosis. Rev Peru Med Exp Salud Publica. julio de 2020;37(3):554-8.

18. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment - PubMed [Internet]. [citado 18 de noviembre de 2024]. Disponible en: https://pubmed.ncbi.nlm.nih.gov/38615098/

19. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Comparative Study to Evaluate the Accuracy of Differential Diagnosis Lists Generated by Gemini Advanced, Gemini, and Bard for a Case Report Series Analysis: Cross-Sectional Study. JMIR Med Inform. 2 de octubre de 2024;12:e63010.

20. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye. 14 de febrero de 2024;38(8):1412.

21. Kaftan AN, Hussain MK, Naser FH. Response accuracy of ChatGPT 3.5 Copilot and Gemini in interpreting biochemical laboratory data a pilot study. Sci Rep. 8 de abril de 2024;14(1):8233.

22. Baig Z, Lawrence D, Ganhewa M, Cirillo N. Accuracy of Treatment Recommendations by Pragmatic Evidence Search and Artificial Intelligence: An Exploratory Study. Diagn Basel Switz. 1 de marzo de 2024;14(5):527.

23. Rossettini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, Turolla A, et al. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. BMC Med Educ. 26 de junio de 2024;24:694.

24. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLOS ONE. 29 de agosto de 2023;18(8):e0290691.

25. Yzu AO, Coronado RF, Pérez LM, Rodríguez PF. Cumplimiento del programa de formación del residentado en cardiología durante la pandemia COVID-19, Lima-Perú. Arch Peru Cardiol Cir Cardiovasc. 31 de marzo de 2022;3(1):16.

26. Flores-Cohaila J, Rivarola-Hidalgo M, Flores-Cohaila J, Rivarola-Hidalgo M. El desempeño académico previo como predictor del examen nacional de medicina: un estudio transversal en Perú. FEM Rev Fund Educ Médica. 2022;25(6):243-7.

22. Congreso de la República. Ley N° 30453, Ley del Sistema Nacional de Residentado Médico (SINAREME) [Internet]. [citado 18 de noviembre de 2024]. Lima: Congreso de la República, 2012. Disponible en: https://ww2.congreso.gob.pe/Sicr/RelatAgenda/proapro20112016.nsf/ProyectosAprobadosPortal/58CE15772B08B89905257A130076D00F/$FILE/AU00148040612.pdf

23. Consejo Nacional de Residentado Médico (CONAREME). Disposiciones Complementarias del Concurso Nacional de Admisión al Residentado Médico 2024 [Internet]. Consejo Nacional de Residentado Medico (CONAREME) [Internet]. Lima: CONAREME. Recuperado a partir de: https://www.conareme.org.pe/web/Documentos/Admision2024/DISPOSICIONES%20COMPLEMENTARIAS%202024.pdf

29. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian National Licensing Medical Examination: Cross-Sectional Study. JMIR Med Educ. 28 de septiembre de 2023;9:e48039.

30. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. marzo de 2024;46(3):366-72.

31. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep. 23 de abril de 2024;14(1):9330.

32. Mavrych V, Yaqinuddin A, Bolgova O. Claude, ChatGPT, Copilot, and Gemini Performance versus Students in Different Topics of Neuroscience. Adv Physiol Educ. 17 de enero de 2025;

ANEXOS

7. ANEXOS

ANEXO 1.
CEDULA PARA PREGUNTAS DE LOS EXAMENES DE RESIDENTADO MÉDICO

AÑO DE EXAMEN	N° PREG.	BLOQUE TEMÁTICO		PREGUNTA	ALTERNATIVAS					RPTA CORRECTA
		AREA	CÓD.		A	B	C	D	E	

ANEXO 2.

FICHA DE OBSERVACION DE RESPUESTAS DE LAS IA'S

AÑO DE EXAMEN	N° PREG.	BLOQUE TEMÁTICO		RPTA CORRECTA	RPTA. CHAT-GPT 4O	EXPLICACIÓN	V=1 F=0	RPTA. GEMINI	EXPLICACIÓN	V=1 F=0	RPTA. COPILOT	EXPLICACIÓN	V=1 F=0
		AREA	CÓD.										