COVER PAGE IN ENGLISH
(solo para artículos en español, portugués o francés)

Comparative Evaluation of Gemini and Copilot Performance in University Entrance Exams: A Systematic Analysis Based on Multiple-Choice Questions and Images

Medina Llerena Diego Alonso, Magister[1], Velarde Lam Diego Manuel, Magister[2]
[1]Escuela Académico Profesional de Ingeniería Industrial, Facultad de Ingeniería, Universidad Continental, Av. Los Incas s/n, Arequipa, Perú *dmedinall@continental.edu.pe*
[2]Universidad Tecnológica del Perú, Perú, *c23128@utp.edu.pe*

Abstract*– The objective of this research was to compare the performance of Gemini and Copilot in solving multiple-choice questions, interpreting texts and images, for the entrance exams of a prestigious Peruvian university across its various faculties over the past three years. This study analyzed 838 questions, of which 83 were analyzed as images. The overall results indicate a higher proportion of correct answers for Copilot, at 75% (627/838) versus 67% (561/838) for Gemini. The performance of both AIs was significantly lower in image analysis, with correct answers of 36.1% (30/83) for Gemini and 39.8% (33/83) for Copilot. In conclusion, these findings highlight the need to improve accuracy in image processing, as well as the importance of understanding its current limitations to optimize its performance and integration into the academic field.

Keywords*-- artificial intelligence, chatbot, performance, university, exams.

Evaluación Comparativa del Desempeño de Gemini y Copilot en las Pruebas de Acceso a Educación Universitaria: Análisis Sistemático en base a Preguntas de Opción Múltiple e Imágenes

Medina Llerena Diego Alonso, Magister[1], Velarde Lam Diego Manuel, Magister[2]
[1]Escuela Académico Profesional de Ingeniería Industrial, Facultad de Ingeniería, Universidad Continental, Av. Los Incas s/n, Arequipa, Perú *dmedinall@continental.edu.pe*
[2]Universidad Tecnológica del Perú, Perú, *c23128@utp.edu.pe*

Resumen*– *El objetivo de la investigación fue comparar el desempeño de Gemini y Copilot en la resolución de preguntas de opción múltiple; interpretando textos e imágenes, para los exámenes de admisión de una prestigiosa universidad del Perú en sus distintas facultades durante los últimos tres años. En este estudio se analizaron 838 preguntas de las cuales 83 preguntas fueron analizadas como imágenes. Los resultados globales indican una proporción superior de respuestas correctas para Copilot con un 75% (627/838) frente al 67% (561/838) para Gemini. El rendimiento de ambas IAs fue significativamente menor en el análisis de imágenes con rendimientos de 36.1% (30/83) de respuestas correctas para Gemini y 39.8% (33/83) de respuestas correctas para Copilot. En conclusión, estos hallazgos destacan la necesidad de mejorar la precisión en el procesamiento de imágenes, así como la importancia de comprender sus limitaciones actuales para optimizar su rendimiento e integración en el ámbito académico.

Palabras clave—inteligencia artificial, chatbot, rendimiento, universidad, exámenes.

I. Introducción

La inteligencia artificial (IA) ha experimentado un progreso significativo y se perfila como una tecnología transformadora en diversos aspectos de la vida humana, incluido el aprendizaje estudiantil [1], [2]. En este sentido, se puede argumentar que los avances tecnológicos, particularmente en IA, contribuyen a que los estudiantes adquieran conocimientos de una manera dinámica [3] para desarrollar un pensamiento crítico [4]. Además, la IA tiene el potencial de aumentar la motivación de los estudiantes, captando su atención y promoviendo una participación en el proceso educativo [5], [6]. Por lo tanto, los educadores así como los líderes empresariales desempeñan un papel fundamental en la comprensión y evaluación del impacto de la IA en el futuro de la educación y la economía [7].

Adicional a sus funciones anteriores, los chatbots pueden apoyar a los estudiantes en la redacción de ensayos y otras tareas académicas, planteado preocupaciones en los entornos educativos sobre la originalidad y adecuación de los contenidos generados para fines académicos [8], planteando retos a los educadores para su uso como una forma de plagio [9], además que pueden sentir duda y ansiedad incorporando estas herramientas en su rol educativo [10].

En este contexto, las tecnologías emergentes pueden ser útiles; los chatbots de inteligencia artificial son herramientas compatibles con dispositivos móviles con capacidad de analizar diversas preguntas, encontrando respuestas óptimas para los usuarios [10], [13]. Los chatbots son capaces de responder eficientemente a una amplia gama de preguntas humanas, ofreciendo respuestas detalladas y coherentes [14]. Es así como las IA pueden ser una herramienta valiosa para mejorar el aprendizaje, motivar a los estudiantes y elevar la calidad educativa, aunque también conlleva retos relacionados con su uso, por ello, es crucial examinar sus limitaciones y reflexionar sobre su potencial tanto presente como futuro [6].

Sin embargo, a pesar de que la IA ya está siendo introducida en la educación superior, muchos docentes aún desconocen su verdadero alcance, naturaleza y limitaciones [10]. Es así como, los educadores deben identifican el potencial de los servicios basados en IA para atender de manera más eficaz, a estudiantes con necesidades educativas especiales y estudiantes multilingües [11]. En escenarios con diferentes idiomas, el desempeño de las IA puede variar, lo que destaca la relevancia de un desarrollo constante para brindar una mayor adaptabilidad y personalización en herramientas de aprendizaje digital [12].

Actualmente, además de herramientas de IA generativa y conversacional como ChatGPT, Gemini y Copilot; existen numerosas aplicaciones diseñadas para funciones específicas, que siguen evolucionando y son accesibles en distintos portales web [7]. Las aplicaciones para la educación son diseñadas para un acceso universal, sin embargo; solo un tercio de estas herramientas es de acceso gratuito [8], lo cual genera desigualdad en el acceso [9]. Ahora bien, el contraste que generan las inteligencias artificiales en las evaluaciones en el ámbito educativo a menudo tiene una connotación negativa por ser un camino distinto al proceso de enseñanza tradicional; en el cual las evaluaciones diagnósticas y formativas, proporcionan a los docentes información inmediata sobre las áreas donde los estudiantes necesitan mejorar para comprender mejor el material [10].

No obstante, los educadores deben considerar las limitaciones prácticas y adoptar métodos de evaluación alternativos que prevengan la mala conducta académica y optimicen los beneficios de estas estrategias [18]. Aunque los chatbots elaboran resúmenes científicos que parecen creíbles, se basan en datos generados artificialmente. Aunque son originales y no presentan indicios de plagio, estos resúmenes pueden ser detectados por herramientas especializadas y revisores humanos críticos [19].

De acuerdo con las Naciones Unidas [20] la educación es un derecho fundamental, en el marco para alcanzar un desarrollo social y como herramienta para reducir la pobreza incorporando las tecnologías en la educación para mejorar los sistemas de salud y alcanzar una sociedad estable [21]. Los retos de Latinoamérica 2030 refuerza que es importante fomentar una epistemología que promueva una integración diversa de saberes, orientada a consolidar instituciones educativas sostenibles y productivas, esenciales como pilar central de la sociedad global [22], sin embargo, existen retos que afrontar para el acceso a la educación universitaria en muchas sociedades [23].

La reforma educativa en países como el Perú esperan tener como resultado mejoras en el aprendizaje y una mentalidad enfocada en el aprendizaje [24]. En este contexto, la educación universitaria de acceso público es muy selectiva y exige altas puntuaciones para los exámenes de admisión [25], esta exigencia puede resultar mayor en universidades fuera de la capital, para el examen de admisión 2023 de la Universidad Nacional de San Agustín, de la ciudad de Arequipa, se presentaron un total de 76712 postulantes, de los cuales solo 6315 estudiantes lograron acceder a una vacante con una proporción aproximada de 12 estudiantes para 1 vacante [26].

II. TRABAJOS RELACIONADOS

En Corea, [27] se realizó un estudio con el propósito de comparar el conocimiento y la capacidad de interpretación de ChatGPT comparada con los estudiantes de medicina, en base a un examen de parasitología. El examen consistió en 79 items aplicado a 77 estudiantes de medicina en las mismas condiciones que ha ChatGPT. Se analizaron las respuestas en base a la puntuación, así como las respuestas correctas según el nivel de dificultad de los temas; además de ello se analiza la aceptación de las explicaciones proporcionadas por la IA. El estudio indica que los estudiantes de medicina obtuvieron una calificación de respuestas correctas en el rango de 89% (65/77) y 93.7% (74/77) preguntas correctas en contraste con el rendimiento de la IA fue de 60.8% (48/79). Con esto se concluye que los conocimientos de ChatGPT aun no son comparables al conocimiento de estudiantes de medicina, a pesar de ello es probable que este conocimiento mejore a través del deep learning. De manera similar [28], evalúa el rendimiento de ChatGPT para un examen de 150 preguntas de opción múltiple al estilo de dificultad y contenido de los exámenes del Canadian Royal College el American Board of Radiology. Los resultados indican que ChatGPT tuvo un rendimiento general de 104 preguntas correctas (69%), destacando un desempeño superior en preguntas de gestión clínica que en preguntas de física. Para todas las respuestas, ChatGPT utilizo un lenguaje constante y seguro aun cuando las respuestas fueran erróneas. Se concluye que ChatGPT tuvo un rendimiento aceptable sin tener una capacitación previa, sin embargo, tiene dificultades para resolver preguntas que incluyen descripción de imágenes, clasificaciones, cálculos y aplicación de conceptos.

Adicional a esto, [29] evaluó el rendimiento de ChatGPT en el examen de licencia médica de EE.UU. (USMLE) el cual consistió en un total de 376 preguntas distribuidas en 3 formatos de preguntas: de opción múltiple con justificación, opción múltiple sin justificación y open-ended. Los resultados muestran un rendimiento promedio de 62% para los diversos formatos de preguntas. Se destaca la alta concordancia que muestra ChatGPT en los diversos formatos de preguntas realizadas. En un caso similar, [30] analizó el desempeño de ChatGPT en exámenes de admisión de dos de las universidades más prestigiosas del Perú. Como conclusión, destacaron que para maximizar el potencial de ChatGPT en el ámbito educativo, resulta fundamental realizar evaluaciones periódicas de su rendimiento, proporcionar retroalimentación constante para mejorar su precisión y reducir sesgos, así como realizar ajustes personalizados para optimizar su aplicación en entornos educativos.

III. METODOLOGÍA

Se han seleccionado exámenes de una prestigiosa universidad de Arequipa, Perú correspondiente a los años 2024, 2023 y 2022. Para los años 2023 y 2022 se toman en cuenta los 2 exámenes ejecutados durante el año denominados fase 1 y fase 2. Para cada examen, se toma en cuenta las 3 áreas generales de la universidad para la postulación, las cuales son sociales, ingenierías y biomédicas.

Se tomaron en cuenta todas las preguntas ejecutadas de los exámenes, incluyendo las preguntas que cuentan con imágenes y cuadros. Esto debido a que ambas IA seleccionadas (Gemini y Copilot) tienen la funcionalidad de procesamiento de imágenes. Con ello se pretende determinar el rendimiento de ambas IA para preguntas escritas y procesamiento de imágenes. Las preguntas se organizan en 50 preguntas por examen para el año 2022 y 60 preguntas por examen para los años 2023 y 2024. La data final de preguntas procesadas consistió en 838 preguntas, para las preguntas se excluyeron 2 preguntas por falta de acceso a la información, las preguntas con imágenes fueron incluidas para poner a prueba la funcionalidad de procesamiento de imágenes con la que cuentan ambas IA seleccionadas.

Para la investigación, se utilizó Gemini modelo 1.5 flash en su versión gratuita, a la cual tuvimos acceso mediante el registro de una cuenta de google básica. Para el caso de Copilot, no tiene numeración en cuanto a sus versiones, por

ello se accedió mediante una cuenta Microsoft para el procesamiento de las preguntas.

Las preguntas de los exámenes fueron tomadas de las fuentes oficiales, siendo preguntas de opción múltiple con su respectiva clave de respuestas. Las preguntas basadas en texto fueron ingresadas a las plataformas IA copiando y pegando directamente las preguntas y las opciones que se podía elegir. Las preguntas con imágenes o cuadros fueron ingresadas incluyendo la interacción "¿Cuál es la opción correcta?" al principio de la interacción seguido de la imagen, la cual incluía la pregunta, el gráfico y las opciones de respuesta.

Las evaluaciones para la universidad seleccionada incluyen una diversidad de temáticas, las cuales fueron clasificadas en 6 disciplinas generales y 18 materias específicas según la tabla 1.

TABLA I
CLASIFICACIÓN POR DISCIPLINA Y MATERIA

Disciplina	Materia
Ciencia y Tecnología	Biología Física Química
Ciencias Literarias	Comprensión de textos Lenguaje Literatura del Perú Literatura universal Razonamiento Verbal
Ciencias Sociales	Filosofía Geografía Historia del Perú Historia Universal
Educación cívica	Educación Cívica Ética y ciudadanía Psicología
Idiomas	Ingles
Matemática	Matemática Razonamiento Matemático

IV. RESULTADOS

El rendimiento de la Inteligencia Artificial Gemini de Google y la Inteligencia Artificial Copilot de Microsoft varió en las 15 diferentes pruebas de admisión evaluadas en los años 2022, 2023 y 2024, mostrando patrones discernibles en función del curso y analizando acumulativamente las pruebas nos indican que la proporción de respuestas correctas son significativamente mayor a la proporción de respuestas incorrectas, tanto para la IA Google Gemini con un valor estadístico de $p < 0.001$ y la IA Microsoft Copilot con $p < 0.001$, como se indica en la tabla II.

TABLA II
RENDIMIENTO DE GEMINI Y COPILOT EN LOS EXÁMENES DE ADMISIÓN

Materia	**Cant. Preguntas**	**GOOGLE GEMINI AI**			**MICROSOFT COPILOT AI**		
		Respuesta Correcta	**Resultado**	**Prueba estadistica**	**Respuesta Correcta**	**Resultado**	**Prueba estadistica**
Biología	43	31	72%	< .001	31	72%	< .001
Física	49	30	61%	< .001	37	76%	< .001
Química	60	47	78%	< .001	45	75%	< .001
Comprensión de textos	60	44	73%	< .001	44	73%	< .001
Lenguaje	68	42	62%	< .001	51	75%	< .001
Literatura del Perú	24	12	50%	< .001	16	67%	0.003
Literatura Universal	18	14	78%	0.042	17	94%	0.331
Razonamiento Verbal	42	28	67%	< .001	34	81%	0.003
Filosofía	48	34	71%	< .001	38	79%	< .001
Geografía	16	16	100%	-	16	100%	-
Historia del Perú	37	27	73%	< .001	30	81%	0.006
Historia Universal	15	12	80%	0.082	13	87%	0.164

Educación Cívica	51	36	71%	< .001	40	78%	< .001
Ética y ciudadanía	36	32	89%	0.044	35	97%	0.324
Psicología	35	27	77%	0.003	28	80%	0.006
Ingles	37	29	78%	0.003	33	89%	0.044
Matemática	89	44	49%	< .001	49	55%	< .001
Razonamiento Matemático	110	56	51%	< .001	70	64%	< .001
Total	838	561	67%	< .001	627	75%	< .001

En términos generales, en esta investigación se aglomeran 838 preguntas analizadas para ambas inteligencias artificiales, se identifica con mayor rendimiento de asertividad a la IA Copilot con un acierto en 75% (627/838) (p< 0.001), en comparación a los resultados de la IA Gemini que generó un acierto de 67% (561/838) (p< 0.001). Al comparar el rendimiento de la IA Copilot y Gemini, se observó que ambos alcanzaron una tasa de éxito alta; sin embargo, la IA Copilot muestra un rendimiento 8% mayor que la IA Gemini, como muestra la figura 1.

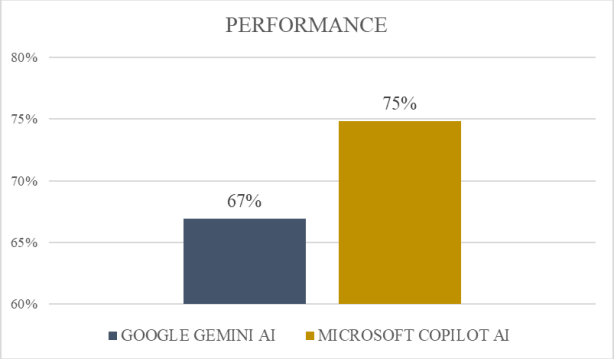

Fig. 1 Rendimiento general de Gemini y Copilot

Las dieciocho materias fueron analizadas individualmente en cada una de las quince evaluaciones y enfocándonos en el alto rendimiento, se identifica en primer lugar a la materia Filosofía, que ha alcanzado un óptimo rendimiento del 100% (16/16) en ambas inteligencias artificiales; la segunda materia con mejor rendimiento alcanzado es la materia Ética y ciudadanía, que ha alcanzado un rendimiento en la IA Copilot del 97% (35/36) (p=0.324) en comparación de la IA Gemini con 89% de rendimiento (32/36) (p=0.044); como tercera materia con mejor rendimiento se identifica a la materia Historia Universal con un rendimiento de respuestas correctas en la IA Copilot del 87% (13/15) (p=0.164) contra el redimiento de la IA Gemini con 80% (12/15) (p=0.082). En la figura 2 identificamos las 3 materias con mejores rendimientos para ambas IAs.

Asimismo, obtuvo un rendimiento aceptable en las 18 materias, en la figura 3 se muestra la eficiencia de ambas IAs ordenadas de manera ascendente con la finalidad de identificar las principales variaciones que tienen ambas IAs en los resultados individuales. Se reconoce que el rendimiento para la IA Copilot como para la IA Gemini son homogéneos en las materias de Biología 72% (31/43) (p< 0.001), Comprensión de texto con 73% (44/60) (p< 0.001) y filosofía 100% (16/16); identificando, que en estas materias, las relaciones entre las respuestas correctas e incorrectas para las IA Gemini y la IA Copilot son idénticas. En la materia de química se puede distinguir una ligera diferencia en el rendimiento entre ambas IAs, aunque se indica que no hay diferencias significativas para ambas, obteniendo resultados de 80% (47/60) (p< 0.001) para la IA Gemini y un rendimiento de 75% (45/60) (p< 0.001) para la IA Copilot.

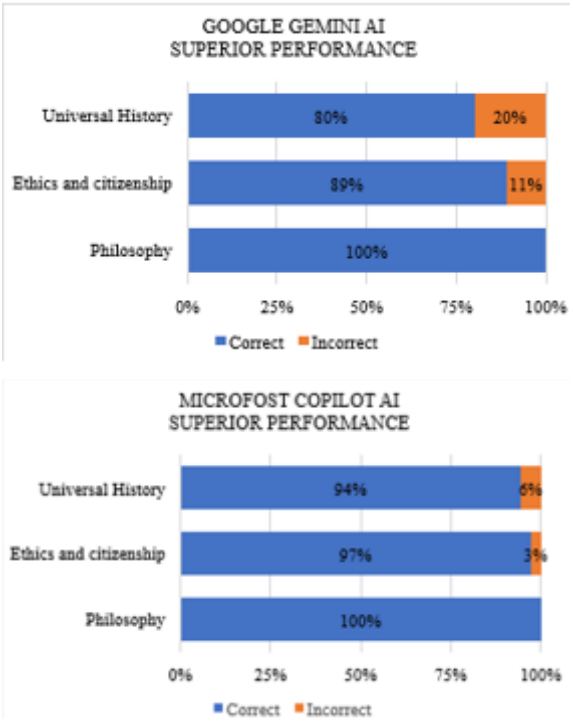

Fig. 2 Las 3 materias con mejor rendimiento para ambas IAs

Además, el rendimiento de IA Copilot fue mayor comparado con la IA Gemini en la materia de Psicología obteniendo un rendimiento de 80% (28/35) (p=0.006) para IA Copilot compara con un 77% (27/35) (p=0.003) para IA Gemini; de igual manera para la materia de Matemáticas, en la cual IA Copilot obtuvo 55% (49/89) (p< 0.001) comparado

con un 49% (44/89) (p< 0.001) para la IA Gemini; de manera similar ocurrió con la materia historia universal en la cual IA Copilot obtuvo 87% (13/15) (p=0.082) que fue mayor al 80% (12/15) (p=0.082) que obtuvo IA Gemini. Un escenario similar se repitió para las materias de educación cívica y política, Ciencias geográficas, historia peruana y Ética y Ciudadanía en las cuales IA Copilot muestra un rendimiento superior a la IA.

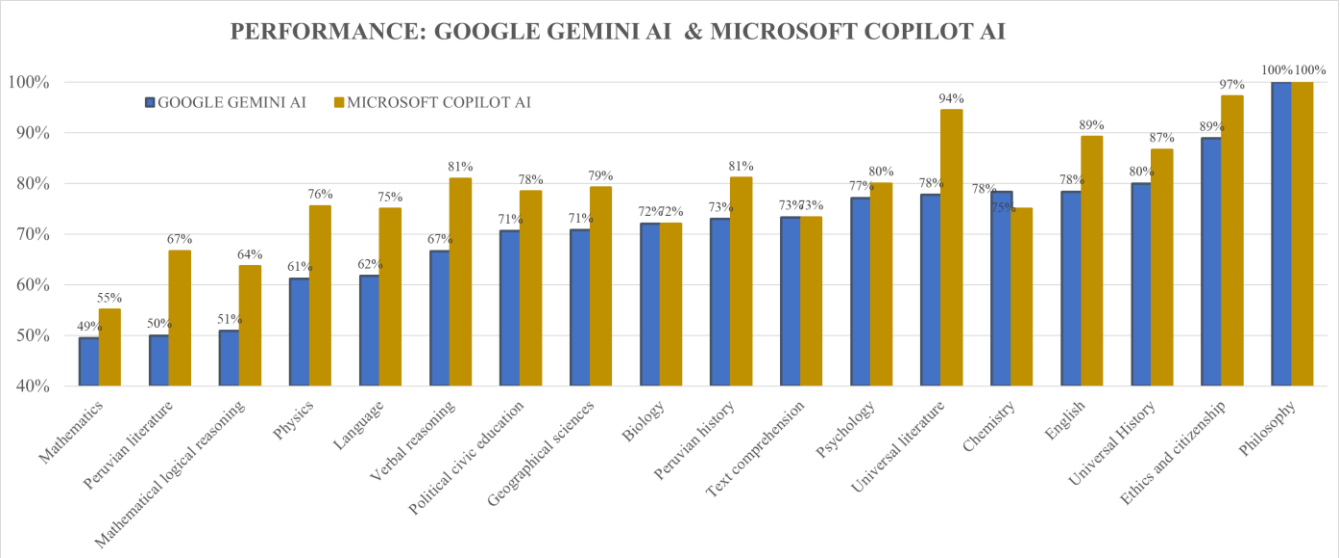

Fig. 3 Rendimiento general de Gemini y Copilot por asignatura en los exámenes de admisión

El análisis muestra que ambas inteligencias artificiales han tenido rendimientos moderados en muchas materias, sin embargo ambas IAs han tenido rendimientos por debajo de lo esperado para las materias de Razonamiento Lógico Matemático, Literatura del Perú y Matemática; por ello se realiza la comparación de los menores rendimientos. En Razonamiento Lógico Matemático ambas IAs se enfrentaron a 110 preguntas de las cuales IA Copilot obtuvo un rendimiento de 64% (70/110) (p<0.001) mientras que IA Gemini obtuvo 51% (56/110) (p<0.001), asimismo el rendimiento en Literatura del Perú fue de 67% (16/24) (p=0.003) para IA Copilot comparado con un rendimiento de 50% (12/24) (p<0.001) para IA Gemini y finalmente para la materia de Matemática tuvo un rendimiento de preguntas correctas de 55% (49/89) (p<0.001) para IA Copilot contra un 49% (44/89) (p<0.001) de IA Gemini.

La selección de preguntas incluyo preguntas que requerían del análisis de imágenes para ser resueltas, estas preguntas estuvieron distribuidas en 6 materias de los exámenes haciendo un total de 83 preguntas que necesitaron del análisis de un análisis combinado.

El rendimiento para el análisis de imagen para ambas IAs fue menor del esperado, con un puntaje de 40% (33/83) para la IA Copilot y un puntaje de 36% (30/83) para la IA Gemini. Entre las 6 materias el menor rendimiento fue para Razonamiento Lógico matemático con un 23% (5/22) (p<0.001) para IA Copilot y un 9% (2/22) (p<0.001) para la IA Gemini.

La IA Copilot obtuvo el mejor puntaje en la materia de física con un 71% (10/14); en cambio la IA Gemini tuvo su mejor resultado para la materia de Química con un puntaje de 75% (9/12). Los resultados de ambas IAs al análisis de preguntas con imágenes se puede analizar en la tabla III.

TABLA III
RENDIMIENTO PARA PREGUNTAS CON IMÁGENES

Materia	Pregunta con analisis de imagen	GOOGLE GEMINI AI			MICROSOFT COPILOT AI		
		Respuesta Correcta	Resultado	Prueba estadistica	Respuesta Correcta	Resultado	Prueba estadistica
Matemática	30	12	40%	< .001	11	37%	< .001
Razonamiento Matemático	22	2	9%	< .001	5	23%	< .001
Fisica	14	5	36%	< .001	10	71%	0.04
Quimica	12	9	75%	0.082	6	50%	0.007
Biologia	5	2	40%	0.07	1	20%	0.016
Total	83	30	36%	< .001	33	40%	< .001

V. DISCUSIÓN

A. Hallazgos principales

El objetivo de la investigación estuvo centrado en comparar el desempeño para la resolución de preguntas de opción múltiple en los exámenes de admisión para una de las principales universidades públicas fuera de la capital del Perú. Para esta evaluación se seleccionaron los exámenes aplicados los últimos 3 años, incluyendo las preguntas que requerían el análisis de imágenes. Las IAs seleccionadas fueron IA Google Gemini y Microsoft Copilot, teniendo en cuenta su nivel de respuesta a las a las exigencias académicas para la educación universitaria y su capacidad para ser adoptada como tutor virtual en ámbitos académicos.

Los resultados obtenidos en esta investigación incluyen el patrón de análisis de imágenes, que ambas plataformas ofrecen dando como resultado un rendimiento adicional a la funcionalidad de la inteligencia artificial, para la variedad de temas tomados en los exámenes de admisión a las diferentes facultades de la universidad seleccionada a lo largo de los últimos años. En primer lugar, cabe mencionar que Copilot obtuvo mejores resultados globales con una tasa de 75% de respuestas correctas, frente a Gemini con una tasa de 67% respuestas correctas. Para el caso de la materia de Filosofía, ambas IAs alcanzaron un rendimiento del 100% (16/16). Estos resultados sugieren un amplio dominio de los contenidos en esta disciplina, lo cual puede deberse a la estabilidad de las teorías filosóficas a lo largo del tiempo, que forman las bases del conocimiento de esta materia.

Por otra parte, analizando específicamente el rendimiento en el análisis de imágenes, se tomó en cuenta las materias de Matemática, Razonamiento Lógico Matemático, Física, Química y Biología para los cuales Copilot obtuvo un rendimiento de 39.8% respuestas correctas frente a 36.1% de Gemini. El desarrollo de esta funcionalidad presenta avances significativos pudiendo reconocer de manera correcta gran parte de los datos ingresados que mezclan tablas, números y textos; sin embargo, es en el desarrollo del problema es donde suele presentar errores disminuyendo la fiabilidad de las IAs frente a esta casuística. Es importante mencionar la complejidad de ciertas preguntas en las cuales se tiene que relacionar datos de cuadros con fórmulas y los requerimientos de la pregunta, adicional a ello, para poder resolver estas preguntas se deben utilizar habilidades inferenciales para luego seleccionar la respuesta correcta de una serie de opciones.

En resumen, los resultados ofrecen información valiosa respecto al rendimiento de la IA para resolver preguntas de texto y de imágenes en distintos cursos universitarios. En las materias de Filosofía, Ética y Ciudadanía e Historia Universal presenta una gran pericia para resolver las preguntas y seleccionar la respuesta correcta de una serie de alternativas. Por otro lado, para el análisis de imágenes suele interpretar los datos, pero aun presenta dificultades para utilizar estos datos, interpretar las opciones a seleccionar y entregar la respuesta correcta.

B. Limitaciones del estudio

Hemos identificado algunas limitaciones en este estudio, entre ellas la cantidad de preguntas que incluyen tablas o imágenes para su interpretación, lo cual puede no ser representativo para otras universidades de Latinoamérica o de otros continentes. Otra limitación se relaciona con el alcance de los exámenes de admisión analizados, que se restringe a la Universidad Nacional de San Agustín y abarca únicamente los periodos desde la Fase I de 2022 hasta la Fase I de 2024. Asimismo, cabe destacar que estas inteligencias artificiales se basan en el aprendizaje automático, dependiente del corpus de datos con el que fueron entrenadas y al que tienen acceso. Esto puede introducir sesgos en el análisis de respuestas y en la interpretación de la información utilizada. Adicional a ello, las bases de datos de ambas IAs fueron cargadas en un idioma distinto al de base de las preguntas, esto nos hace inferir que puede tener sesgos para contestar preguntas especificas en ciertos idiomas, sobre todo para materias que requieren información específica de algunos países de habla hispana.

En resumen, la inteligencia artificial generativa ofrece nuevas oportunidades para el análisis de imágenes; no obstante, aún presenta limitaciones para utilizar la información ingresada con precisión y resolver problemas específicos en contextos educativos.

VI. CONCLUSIONES

En conclusión, los datos presentados en esta investigación muestran una comparación del rendimiento de Gemini y Copilot en los exámenes de admisión para una prestigiosa universidad en Arequipa, Perú, en términos generales se destaca un rendimiento superior de la IA Copilot sobre la IA Gemini sin embargo aun ambas IAs siguen requiriendo una actualización continua de sus bases de datos para mejorar su precisión.

La inteligencia artificial (IA) tiene una presencia creciente en el ámbito educativo, y para maximizar su potencial, es fundamental evaluar de manera continua sus funcionalidades actuales y su evolución, especialmente en áreas como el análisis de imágenes. Esta evaluación constante proporciona información que permite minimizar errores, adaptar respuestas y mejorar la precisión en entornos educativos cada vez más exigentes.

Las IAs analizadas en este estudio destacan por su potencial como herramientas de acceso gratuito, lo cual las posiciona como valiosos recursos de apoyo en la formación académica de estudiantes a nivel mundial. Esto resalta la necesidad de que los educadores estén conscientes de su uso y de los niveles de confiabilidad asociados a estas tecnologías.

REFERENCIAS

[1] R. M. Alali and A. A. Al-Barakat. (2024). “Artificial Intelligence Experts’ Perceptions on the Effective Use of Artificial Intelligence Applications in

University Learning Environments," Journal of Ecohumanism, vol. 3, no. 4, pp. 1780–1793. https://doi.org/10.62754/joe.v3i4.3710
[2] F. J. Hinojo-Lucena, I. Aznar-Díaz, M. P. Cáceres-Reche, and J. M. Romero-Rodríguez. (2019) "Artificial intelligence in higher education: A bibliometric study on its impact in the scientific literature," Educ Sci (Basel), vol. 9, no. 1. https://doi.org/10.3390/educsci9010051
[3] N. H. Haron, R. Mahmood, N. M. Amin, A. Ahmad, and S. R. Jantan (2024) "An Artificial Intelligence Approach to Monitor and Predict Student Academic Performance," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 44, no. 1, pp. 105–119. https://doi.org/10.37934/araset.44.1.105119
[4] R. Suriano, A. Plebe, A. Acciai, and R. A. Fabio. (2024) "Student interaction with ChatGPT can promote complex critical thinking skills," Learn Instr, vol. 95. https://doi.org/10.1016/j.learninstruc.2024.102011
[5] L. Yuan and X. Liu. (2024) "The effect of artificial intelligence tools on EFL learners' engagement, enjoyment, and motivation," Comput Human Behav, vol. 162, p. 108474. https://doi.org/10.1016/j.chb.2024.108474
[6] M. G. Choque-Castañeda, G. Pastor, and M. Romero. (2023) "Impact of using ChatGPT in higher education: A Systematic Review".
[7] S. T. H. Pham. (2023). "Exploring the lived experience of educators and business executives in the phenomenon of artificial intelligence in education," in Phenomenological Studies in Education, IGI Global, pp. 182–206. https://doi.org/10.4018/978-1-6684-8276-6.ch010
[8] M. R. King. (2023) "A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education". https://doi.org/10.1007/s12195-022-00754-8
[9] H. Ibrahim et al. (2023) "Perception, performance, and detectability of conversational artificial intelligence across 32 university courses," Sci Rep, vol. 13, no. 1. https://doi.org/10.1038/s41598-023-38964-3
[10] H. Wang, A. Dang, Z. Wu, and S. Mac. (2024) "Generative AI in higher education: Seeing ChatGPT through universities' policies, resources, and guidelines," Computers and Education: Artificial Intelligence, vol. 7, p. 100326. https://doi.org/10.1016/j.caeai.2024.100326
[11] "Future of Testing in Education: Artificial Intelligence - Center for American Progress." Accessed: Nov. 03, 2024. [Online]. Available: https://www.americanprogress.org/article/future-testing-education-artificial-intelligence/
[12] P. Chatwattana, P. Yangthisarn, and A. Tabubpha. (2024) "The Educational Recommendation System with Artificial Intelligence Chatbot: A Case Study in Thailand," International Journal of Engineering Pedagogy, vol. 14, no. 5, pp. 51–64. https://doi.org/10.3991/ijep.v14i5.48491
[13] B. Guo et al., "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," Jan. 2023, [Online]. Available: http://arxiv.org/abs/2301.07597
[14] Indra, Janani, N. Chauhan, R. Mohan, and A. Kumar, "Understanding the Impact of Artificial Intelligence Applications on Indian Higher Education Sector," in Studies in Big Data, vol. 158, Springer Science and Business Media Deutschland GmbH, 2025, pp. 21–31. https://doi.org/10.1007/978-3-031-70855-8_3
[15] M. Sallam et al.. (2024) "The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: a comparative analysis of English and Arabic responses," BMC Res Notes, vol. 17, no. 1. https://doi.org/10.1186/s13104-024-06920-7
[16] I. Azaiz, O. Deckarm, and S. Strickroth. (2023) "AI-Enhanced Auto-Correction of Programming Exercises: How Effective is GPT-3.5?," International Journal of Engineering Pedagogy (iJEP), vol. 13, no. 8, pp. 67–83. https://doi.org/10.3991/IJEP.V13I8.45621
[17] M. Area-Moreira, A. Del Petre, A. L. Sanabria-Mesa, and B. Sannicolás-Santos. (2024) "NOT ALL AI TOOLS ARE CREATED EQUAL. ANALYSIS OF SMART APPLICATIONS FOR UNIVERSITY TEACHING," Digital Education Review, no. 45, pp. 141–149. https://doi.org/10.1344/der.2024.45.141-149
[18] A. M. Elkhatat. (2023) "Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities," International Journal for Educational Integrity, vol. 19, no. 1. https://doi.org/10.1007/s40979-023-00137-0
[19] C. A. Gao et al.. (2022) "Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers". https://doi.org/10.1101/2022.12.23.521610
[20] "Educación." Accessed: Dec. 04, 2024. [Online]. Available: https://www.bancomundial.org/es/topic/education/overview#1
[21] Y. Dosymov et al. (2023) "Effectiveness of Computer Modeling in the Study of Electrical Circuits: Application and Evaluation," International Journal of Engineering Pedagogy (iJEP), vol. 13, no. 4, pp. 93–112. https://doi.org/10.3991/IJEP.V13I4.34921
[22] J. J. Aldana-Zavala, P. A. Vallejo-Valdivieso, and J. Isea-Argüelles. (2021) "Investigación y aprendizaje: Retos en Latinoamérica hacia el 2030," Alteridad, vol. 16, no. 1, pp. 78–91. https://doi.org/10.17163/ALT.V16N1.2021.06
[23] O. Molina, D. Santa María, and G. Yamada. (2024) "Study for Nothing? Gender and Access to Higher Education in a Developing Country," Econ Dev Cult Change, vol. 72, no. 2, pp. 517–561. https://doi.org/10.1086/721907
[24] J. Saavedra and M. Gutierrez (2020) "Peru: A wholesale reform fueled by an obsession with learning and equity," Audacious Education Purposes: How Governments Transform the Goals of Education Systems, pp. 153–180. https://doi.org/10.1007/978-3-030-41882-3_6/FIGURES/10
[25] J. F. Luesia, I. Benítez, R. Company-Córdoba, I. Gómez-Gómez, and M. Sánchez-Martín. (2023) "Assessing the relevance of academic competencies in college admission tests from a higher-order thinking perspective: A systematic review," Think Skills Creat, vol. 48, p. 101251. https://doi.org/10.1016/J.TSC.2023.101251
[26] "SISADMISION." Accessed: Nov. 04, 2024. [Online]. Available: https://apps.unsa.edu.pe/sisadmision/public/estadistica
[27] S. Huh. (2023) "Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study," J Educ Eval Health Prof, vol. 20. https://doi.org/10.3352/jeehp.2023.20.1
[28] R. Bhayana, S. Krishna, and R. R. Bleakney. (2023) "Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations," Radiology, vol. 307, no. 5, p. e230582. https://doi.org/10.1148/radiol.230582
[29] T. H. Kung et al. (2022) "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models". https://doi.org/10.1101/2022.12.19.22283643
[30] S. Beltozar-Clemente, E. Díaz-Vega, R. Tejeda-Navarrete, and J. Zapata-Paulini. (2024) "We Can Rely on ChatGPT as an Educational Tutor: A Cross-Sectional Study of its Performance, Accuracy, and Limitations in University Admission Tests," International Journal of Engineering Pedagogy, vol. 14, no. 1, pp. 50–60. https://doi.org/10.3991/ijep.v14i1.46787