REVISTA CIENTÍFICA
CONECTIVIDAD
Instituto Superior Tecnológico Universitario Rumiñahui

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

Artículo Científico

Innovaciones en Visión Artificial: Evaluación de ChatGPT, Gemini y Copilot para el Análisis de Imágenes

Innovations in Computer Vision: Evaluation of ChatGPT, Gemini, and Copilot for Image Analysis

Pablo Minango Negrete[1] , Marcelo Zambrano Vizuete[2] , Juan Minango Negrete[3] , César Minaya Andino[4] , Carlos León Galeas[5]

[1]Instituto Superior Tecnológico Rumiñahui, pablo.minango@ister.edu.ec
[2]Instituto Superior Tecnológico Rumiñahui, marcelo.zambrano@ister.edu.ec
[3]Instituto Superior Tecnológico Rumiñahui, juancarlos.minango@ister.edu.ec
[4]Instituto Superior Tecnológico Rumiñahui, cesar.minaya@ister.edu.ec
[5]Instituto Superior Tecnológico Rumiñahui, carlosj.leon@ister.edu.ec

Autor para correspondencia: pablo.minango@ister.edu.ec

Citas
Minango, P., Zambrano, M., Minango, J., Minaya, C., & Leon, C. (2025). Innovaciones en Visión Artificial: Evaluación de ChatGPT, Gemini y Copilot para el Análisis de Imágenes. CONECTIVIDAD, 6(2). https://doi.org/10.37431/conectividad.v6i2.284

RESUMEN
En los últimos años los Modelos de Lenguaje de Gran Escala (LLM) han tenido un crecimiento exponencial evolucionado rápidamente, desde sus inicios cuando fueron concebidos bajo la premisa de simples herramientas que comprendían texto hasta nuestros tiempos que se han convertido en sistemas multimodales capaces de generar contenido creativo y complejo. Esta innovación se ha impulsado por los grandes avances en arquitecturas de redes neuronales y ha eso sumarle la disponibilidad de grandes conjuntos de datos. En este estudio, se tiene como objetivo principal comparar tres LLMs más usados que son: ChatGPT, Gemini y Copilot, en la ejecución de la tarea de convertir imágenes en texto (I2T). Se evaluó la capacidad que tiene cada modelo para describir de manera detallada y precisa diferentes tipos de imágenes, entre las cuales se evaluó pinturas artísticas, escenas urbanas e imágenes con instrucciones. Los resultados obtenidos muestran que los tres modelos poseen un alto nivel de desempeño, el modelo de Gemini sobresale gracias a que mostro habilidad para integrar información visual y textual de manera más eficiente.

Los resultados del estudio muestran que los LLMs continúan evolucionando, con lo que podemos esperar ver avances aún más significativos en su capacidad para comprender y generar lenguaje natural. Así mismo, se espera que esta evolución permita a estos modelos verse más aplicados en la vida cotidiana de todas las personas, automatizando procesos y ayudando a mejorar el desarrollo de asistentes virtuales.

Palabras clave: ChatGPT; Gemini; Copilot; IA; Procesamiento de lenguaje natural.

ABSTRACT
In recent years, Large Scale Language Models (LLM) have had an exponential growth and have evolved rapidly, from their beginnings when they were conceived under the premise of simple tools that understood text to our times when they have become multimodal systems capable of generating creative and complex content. This innovation has been driven by the great advances in neural network architectures and, in addition, the availability of large data sets. In this study, the main objective is to compare three most used LLMs: ChatGPT, Gemini and Copilot, in the execution of the task of converting images to text (I2T). The capacity of each model to describe in a detailed and precise way different types of images was evaluated, among which artistic paintings, urban scenes and images with instructions were evaluated. The results obtained show that the three models have a high level of performance, the Gemini model stands out thanks to its ability to integrate visual and textual information more efficiently.

The results of the study show that LLMs continue to evolve, so we can expect to see even more significant advances in their ability to understand and generate natural language. It is also expected that this evolution will allow these models to be more widely applied in the daily lives of all people, automating processes and helping to improve the development of virtual assistants.

Key words: AI, Natural Language Processing; ChatGPT; Gemini; Copilot.

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

1. INTRODUCCIÓN

Los Modelos de Lenguaje de Gran Escala (LLM), han revolucionado el campo de la Inteligencia Artificial (IA), en especial desde el Procesamiento de Lenguaje Natural (NLP). Desde los primeros desarrollos, los LLM han crecido desde su complejidad de aprendizaje hasta el desarrollo de nuevas capacidades, permitiendo no solo la generación y comprensión de texto a niveles sin precedentes, sino que también su aplicación en áreas más amplias como lo es la visión computacional (Achiam y otros, 2023).

El origen de LLM se da con avances en la tarea de NLP, donde modelos como ELMo y BERT, que son modelos que inicialmente estuvieron diseñados para comprensión del lenguaje, demostraron el poder del uso de arquitecturas de redes neuronales profundas para capturar el contexto semántico en el texto. A medida que la tecnología avanza, las arquitecturas de LLM se vuelven más robustas, como lo fue con la introducción de modelos como GPT-2 y GPT-3, que no solo pueden generar texto coherente, sino que también realizan tareas más complejas de manera autónoma (Devlin y otros, 2018).

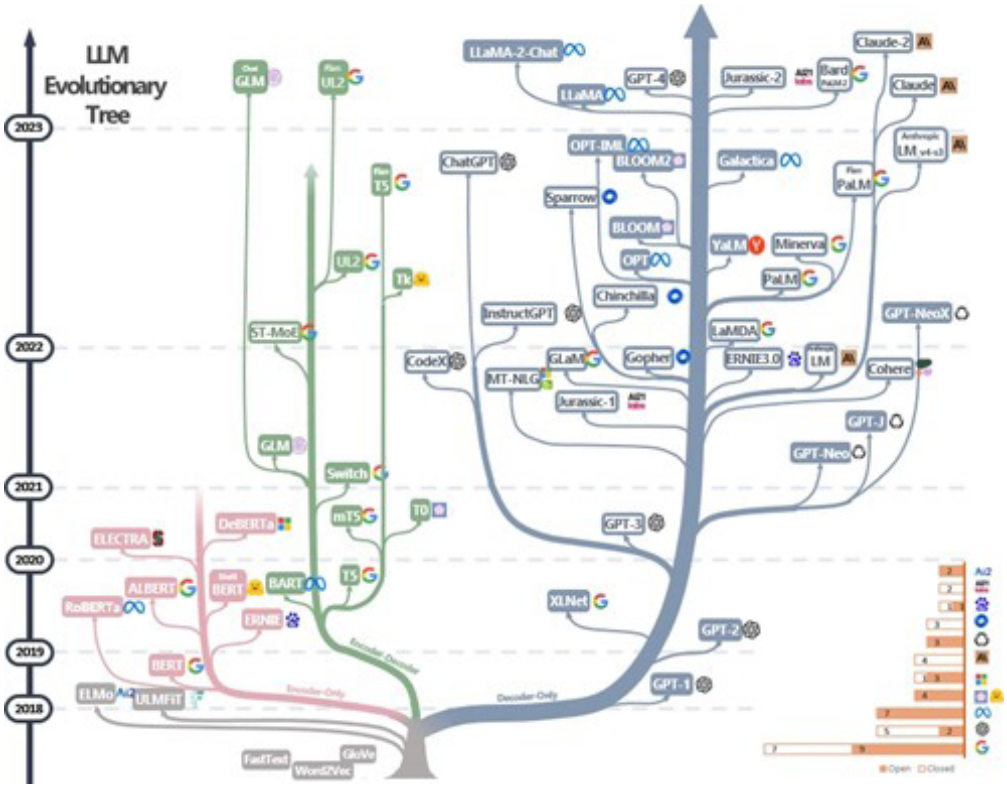
La Figura. 1., muestra el diagrama del árbol evolutivo de los modelos LLM (Yang y otros, 2023) en donde se puede apreciar un crecimiento exponencial en la capacidad y aplicaciones, donde podemos destacar los siguientes hitos importantes:

- **2018 – 2020: Fundación de los LLM**, el modelo BERT y sus derivados como RoBERTa (Liu et al., 2019) y ALBERT (Lan et al., 2019), marcaron el inicio utilizando modelos bidireccionales para entender el contexto completo de una palabra dentro de la oración. Estos modelos fueron los pioneros en establecer el marco para la generación de texto y comprensión textual.
- **2020 – 2022: Expansión de Capacidades,** en esta nueva etapa evolutiva GPT-3 amplió las capacidades de generación de lenguaje, permitiendo no solo respuestas textuales, sino también la generación de código y otras tareas especializadas (Open AI, 2020). Codex, una derivación de GPT-3, se especializa en aplicaciones de programación siendo adaptada para comprender y generar códigos de programación (Open AI, Anthropic AI, Zipline, 2021) siendo esta la base de herramientas para GitHub Copilot.
- **2022 – 2023: Integración y Especialización,** dentro de esta etapa Google desarrolló LaMDA, modelo orientado a mantener conversaciones coherentes y contextualizadas en múltiples tareas (Google, 2022), dando paso a Gemini que integra dichas capacidades con aplicaciones en visión por computadora, permitiendo la interpretación simultánea de datos textuales y visuales (Gemini Team y otros, 2024). Adicional se tiene el desarrollo de GPT-4 el cual posee mejoras en la capacidad de generación de texto, comprensión semántica, y además ofrece una integración de datos visuales dentro de sus análisis (Open AI, 2023).

Innovaciones en Visión Artificial: Evaluación de ChatGPT, Gemini y Copilot para el Análisis de Imágenes. pp. 251 - 262 / Volumen 6, número 2 / DOI: https://doi.org/10.37431/conectividad.v6i2.284

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

Figura. 1. Diagrama del árbol evolutivo de los modelos LLM.



Fuente: Yang, J. y otros., 2023

Estas etapas mostradas en el árbol evolutivo muestran como los LLM han evolucionado desde simples modelos de NLP hasta herramientas complejas que permiten abordar tareas de visión por computadora, marcando así un cambio de paradigma en la forma que la IA puede interpretar y generar contenido tanto textual como visual.

Bajo esta premisa, Gemini (qué está basado en LaMDA), ChatGPT (GPT-4), y Copilot (qué está basado en Bard) simbolizan el pico de la evolución, ya que estos modelos han demostrado una gran capacidad en la generación de texto, análisis de imágenes logrando interpretar descripciones en lenguaje natural y proveer análisis contextuales, la Tabla. 1. muestra una comparativa más detallada de estos tres modelos, los cuales comparten características comunes al ser LLM, su principal distintivo se debe a aplicaciones, enfoques, y capacidades.

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

Característica	ChatGPT	Gemini	Copilot
Modelo Base	GPT-4	Bard (basado en LaMDA y otras tecnologías)	Codex (basado en GPT-3)
Desarrollador	Open AI	Google	Open AI y GitHub
Entrenamiento	Preentrenado en un gran corpus de texto	Preentrenado en datos multilingües y diálogos	Preentrenado en datos de código fuente
Aplicaciones	Generación de texto, chat, asistencia.	Búsqueda de información, chat, asistencia.	Asistencia en programación, autocompletado de código.
Capacidades	Respuestas contextuales y coherentes, generación de texto creativo.	Respuestas de búsqueda, integración con otros servicios de Google.	Completar código, sugerir fragmentos, generación de código.
Adaptabilidad	Alta, con capacidades de personalización y fine-tuning.	Alta, integración con servicios de Google.	Alta, enfocada en entornos de desarrollo.
Interfaz de Usuario	API, aplicaciones integradas, plataformas web.	Integrado en productos de Google, API.	Integrado en IDEs, API.

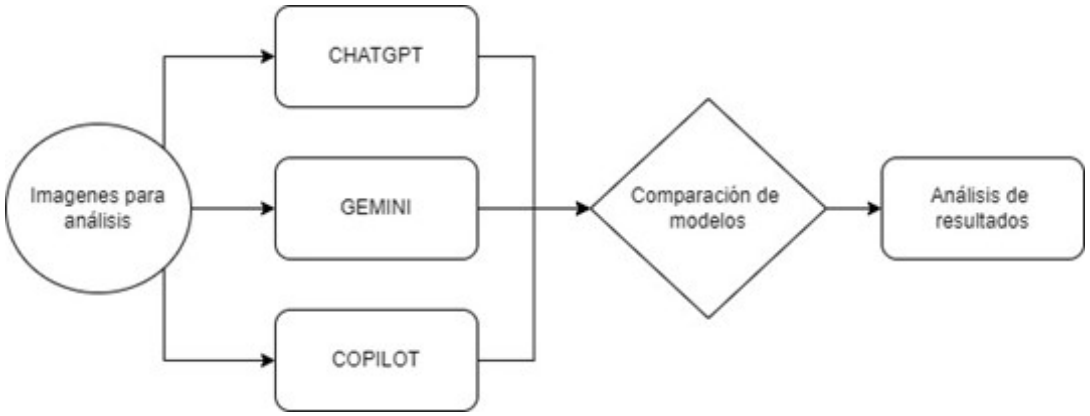
2. MATERIALES Y MÉTODOS

El objetivo principal del presente trabajo investigativo se centra en evaluar el rendimiento de tres tecnologías avanzadas de LLM para IA en la resolución de tareas de tipo I2T. Las tecnologías escogidas para ser evaluadas son: ChatGPT, Gemini y Copilot. Nuestro análisis se enfoca en el comportamiento de cómo cada modelo maneja la interpretación de imágenes y cuales son cada una de sus capacidades aplicadas en el campo de la visión artificial. La Figura 2. muestra el esquema metodológico implementado para evaluar el rendimiento de estos modelos propuestos.

Innovaciones en Visión Artificial: Evaluación de ChatGPT, Gemini y Copilot para el Análisis de Imágenes. pp. 251 - 262 / Volumen 6, número 2 / DOI: https://doi.org/10.37431/conectividad.v6i2.284

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

Figura. 2. Metodología para el análisis de imágenes con modelos de LLM.



2.1 Imágenes para Analizar

Se seleccionaron tres imágenes distintas con el propósito de evaluar el desempeño de los modelos de LLM en diferentes figuras:

- **Pintura artística:** Dentro del prompt de cada modelo se solicita proporcionar una descripción detallada de la obra artística subida, lo que se espera es que cada modelo capture elementos visuales y posibles interpretaciones del contenido.
- **Escena Urbana:** Dentro de esta evaluación se buscó imágenes que incluyan edificios, vehículos y personas, en el prompt de cada modelo se le solicita que efectué un conteo de los vehículos que se encuentran presentes en la imagen, evaluando como es la capacidad del modelo para contar e identificar objetos en un entorno urbano complejo.
- **Imagen con instrucciones:** Se generó una imagen la cual contiene instrucciones específicas que se deben seguirse. Para el prompt de cada modelo se proporciona una imagen que tiene indicaciones escritas sobre cómo realizar una tarea, se pedirá a cada modelo que interprete y siga las instrucciones que se encuentran contenidas en la imagen, con eso se evalúa la capacidad para entender y ejecutar directrices basadas en informaciones visuales.

Cada una de estas figuras son enviadas a los tres modelos de LLM propuestos con solicitudes específicas y que sean las mismas en el prompt de cada modelo, con el propósito de evaluar como cada uno de los modelos maneja diferentes situaciones en la tarea I2T.

2.2 Comparación de modelos

Las descripciones que son generadas por cada uno de los modelos se evalúan empleando los siguientes criterios:

- **Precisión:** La medida en la que la descripción generada refleja con exactitud los elementos y detalles presentes en la imagen. Evaluar si los modelos logran captar todos los aspectos relevantes y específicos de las imágenes.
- **Detalles:** Analizar cuantas características importantes de la imagen el modelo logra identificar, considerando si las descripciones proporcionan información completa y

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

minuciosa sobre los elementos visuales expuestos en la imagen.

- **Coherencia:** La descripción de cada modelo es lógica, en este apartado se verifica si las descripciones generadas poseen un flujo lógico y poseen un componente comprensible en el contexto de la imagen.
- **Comparación:** Comparar entre las informaciones generadas por los tres modelos para identificar cual modelo ofrece una descripción más completa, precisa y detallada. Esto consiente en identificar que modelo posee un mayor desempeño en la tarea asignada de I2T.

2.3 Análisis de resultados

- **Análisis Cualitativo:** La evaluación cualitativa está enfocada en las descripciones que proporcione cada modelo para identificar las fortalezas y debilidades al momento de generar texto a partir de imágenes. Se examina las diferencias en la calidad de profundidad y precisión en las descripciones.
- **Discusión:** Se discute cómo cada modelo generó texto descriptivo a partir de un promt con instrucciones y el anexo de imágenes, se considera las alcances de los resultados para aplicaciones prácticas en visión artificial, como la automatización de procesos para describir imágenes y la mejora de sistemas de accesibilidad.

3. RESULTADOS Y DISCUSIÓN

En esta sección, se analizan los resultados obtenidos a partir de la evaluación de los tres modelos de LLM propuestos, los cuales son: (ChatGPT, Gemini y Copilot) al efectuar una tarea de I2T. Este análisis tiene como propósito analizar la capacidad que tiene cada modelo para interpretar, profundizar y describir imágenes en diversos entornos.

3.1 Pintura artística

Al revisar los resultados obtenidos en los tres modelos de LLM (Figuras 3(a), 3(b) y 3(c)) con una misma solicitud de descripción de una pintura colocada en el prompt, podemos evidenciar una notable capacidad de todos los modelos para generar textos detallados y precisos. Sin embargo, al analizar a profundidad cada uno de los resultados con mayor detenimiento, se puede aprecian ciertas diferencias en la complejidad de las descripciones.

Como primer caso, los modelos de ChatGPT y Copilot ofrecen una visión general de la pintura, destacando los elementos visuales más prominentes como la composición, el color, y los objetos representados dentro de la pintura. Sin embargo, el modelo de Gemini, no solo proporciona esta descripción detallada de los elementos, sino que adicional se adentra en una interpretación más subjetiva de la obra de arte, es decir este modelo demuestra una mayor capacidad para reconocer relaciones, patrones, y simbolismos implícitos en la pintura, lo que le permite generar descripciones más profundas y significativas basados en el contexto de la pintura artística.

Innovaciones en Visión Artificial: Evaluación de ChatGPT, Gemini y Copilot para el Análisis de Imágenes. pp. 251 - 262 / Volumen 6, número 2 / DOI: https://doi.org/10.37431/conectividad.v6i2.284

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

Figura. 3(a). Resultado en la descripción de Pintura artística con ChatGPT.

Figura. 3(b). Resultado en la descripción de Pintura artística con Gemini.

Figura. 3(c). Resultado en la descripción de Pintura artística con Copilot.

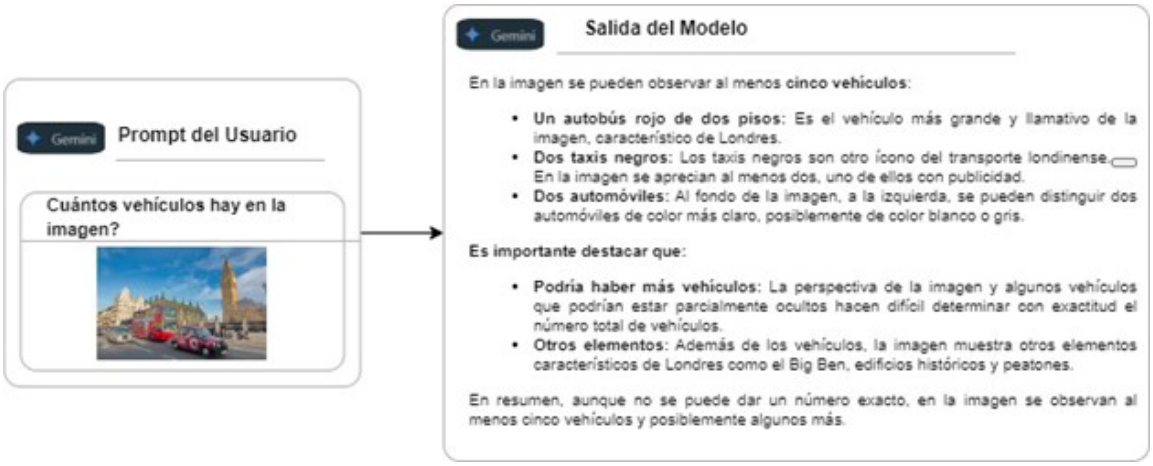
Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

3.2 Escena urbana

Para evaluar las capacidades de visión computacional en las tres tecnologías de modelos de LLM, se les asignó en cada prompt la tarea de contar vehículos en una imagen urbana. Los resultados (Figuras 4(a) y 4(b)) indican que ChatGPT y Gemini destacaron con respecto a Copilot en la detección de vehículos, ya que lograron identificar 5 de los 7 vehículos presentes en la imagen. Adicional, Gemini mostró mayor robustez al detectar la posible existencia de vehículos adicionales. Los tres modelos de LLM emplearon técnicas de visión computacional con el propósito de segmentar y clasificar los objetos en la imagen, obteniendo resultados que no solo incluyen el conteo inicial, sino que adicionalmente proporcionaron informaciones sobre el color, tipo de publicidad y ubicación de los vehículos, por otro lado, Gemini y Copilot lograron identificar la ciudad de Londres en la imagen.

Figura. 4(a). Resultado en la descripción de escena urbana con ChatGPT.

Figura. 4(b). Resultado en la descripción de escena urbana con Gemini.



Innovaciones en Visión Artificial: Evaluación de ChatGPT, Gemini y Copilot para el Análisis de Imágenes. pp. 251 - 262 / Volumen 6, número 2 / DOI: https://doi.org/10.37431/conectividad.v6i2.284

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

Figura. 4(c). Resultado en la descripción de escena urbana con Copilot.

3.3 Imagen con instrucciones

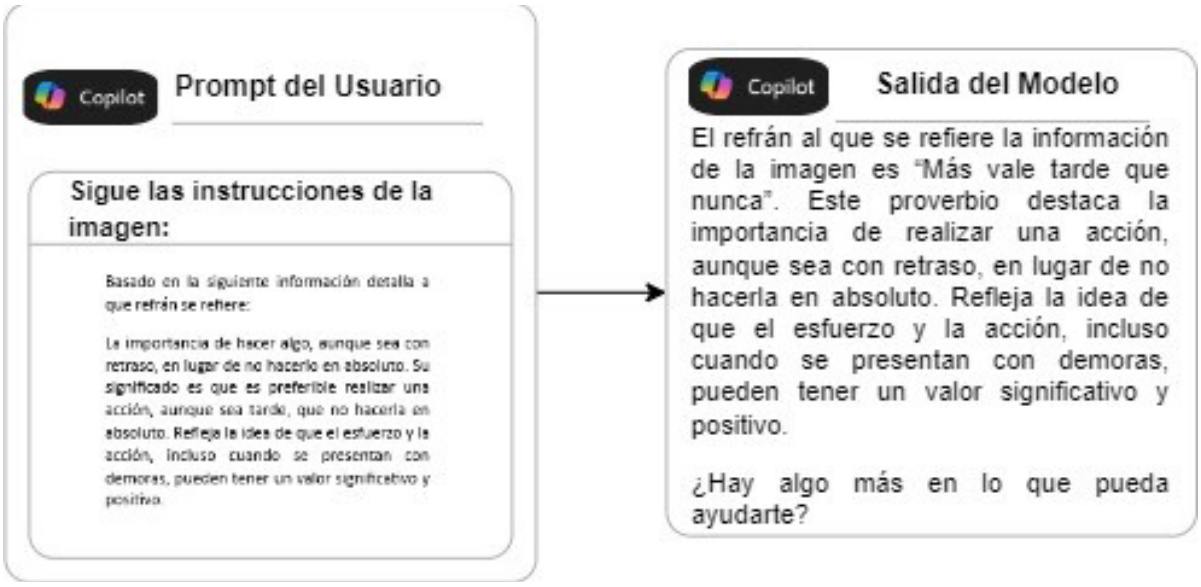
Finalmente, para evaluar la capacidad de los tres modelos de LLM integrar diferentes técnicas de procesamiento de información, se subió una imagen la cual contiene instrucciones textuales. Los tres modelos (Figuras 5(a), 5(b) y 5(c)) mostraron una gran destreza para interpretar el contenido textual dentro de la imagen, logrando identificar correctamente el refrán que fue escrito como referencia en la imagen. Sin embargo, el modelo de Gemini fue el que más se diferenció ya que generó un análisis más profundo, incluyendo una transcripción precisa de todo el texto, así como una justificación más detallada de la interpretación del refrán y una exploración de posibles alternativas sobre refranes referentes. Esta amplia capacidad del modelo de Gemini para combinar visión por computadora y procesamiento del lenguaje natural de manera eficiente lo posiciona como una de las herramientas más poderosas para tareas que requieren un alto nivel de comprensión y extracción.

Figura. 5(a). Resultado en la descripción de instrucciones vía texto en imagen con ChatGPT.

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

Figura. 5(b). Resultado en la descripción de instrucciones vía texto en imagen con Gemini.

Figura. 5(c). Resultado en la descripción de instrucciones vía texto en imagen con Copilot.



Los resultados obtenidos en este estudio muestran una creciente optimización de los modelos de LLM en las tareas de comprender texto, procesar información visual y textual. Al realizar la evaluación de la capacidad de los modelos de ChatGPT, Gemini y Copilot en las tareas de interpretar imágenes y seguir instrucciones, pudimos evidenciar un gran avance en el campo de la IA que está enfocado en modelos LLMs.

Finalmente, el modelo de Gemini mostró una mayor capacidad descriptiva integrando diferentes técnicas de procesamiento de información, como lo son: visión por computadora y el procesamiento del lenguaje natural, en términos generales su habilidad para interpretar imágenes a mayor profundidad le permite extraer informaciones relevantes y generar respuestas relacionadas y más comprensivas a la imagen subida, por tal razón, lo posicionan como una de las herramientas más competentes para varias aplicaciones, como por ejemplo, puede ser para su uso desde la atención al cliente hasta la investigación científica.

Sin embargo, es importante recalcar que estos tres modelos evaluados aún presentan limitaciones

Innovaciones en Visión Artificial: Evaluación de ChatGPT, Gemini y Copilot para el Análisis de Imágenes. pp. 251 - 262 / Volumen 6, número 2 / DOI: https://doi.org/10.37431/conectividad.v6i2.284

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

como, por ejemplo, el desempeño puede verse afectado por la calidad de la imagen, la complejidad del texto o la presencia de ambigüedades en la información que se le solicite mediante el prompt. Además, es importante que abordemos cuestiones éticas relacionadas al uso de estas tecnologías, como son la privacidad de los datos y la posibilidad de que los modelos generen contenido engañoso.

4. CONCLUSIONES

Finalmente, el estudio comparativo entre los modelos de LLMs analizados que son: ChatGPT, Gemini y Copilot, muestran una notable evolución de las tecnologías de IA al procesar información para el campo de visión artificial específicamente en las tareas de I2T. Cada uno de estos modelos han demostrado una gran capacidad para comprender y generar lenguaje natural, así como la interacción e interpretación con contenido visual.

ChatGPT, Gemini y Copilot comprueban diversas fortalezas de las cuales podemos destacar las siguientes:

- ChatGPT: Destaca en la generación de textos coherente y creativos, demostrando entendimiento de los patrones del lenguaje humano.
- Gemini: Sobresale en la integración de múltiples modalidades, combinando de manera efectiva el procesamiento de lenguaje natural con la visión por computadora.
- Copilot: De igual manera muestra buenos resultados al momento de describir imágenes e identificar texto.

Sin embargo, estos avances significativos en los modelos de LLM aún existen desafíos que se deben tener mucha precaución como son los sesgos en los resultados, la interpretación incorrecta de información confusa o contradictoria, la generación de contenido desordenado y la dificultad para comprender textos complejos. Es importante tomar en cuenta las cuestiones éticas relacionadas con la privacidad de los datos, la transparencia y la responsabilidad en el desarrollo y en el uso de la IA.

Estos modelos de lenguaje de gran tamaño ChatGPT, Gemini y Copilot, representan un amplio avance en el desarrollo de IA. No obstante, es esencial que el desarrollo e investigación sean realizados de manera responsable y ética, para maximizar sus beneficios y mitigar posibles riesgos.

REFERENCIAS

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F., y McGrew, B. (2023). Gpt-4 technical report. *arXiv e-prints*. https://doi.org/arXiv:2303.08774

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prin*ts. https://doi.org/10.48550/arXiv.1810.04805

Fecha de recepción: 18 / 07 / 2024
Fecha de aceptación: 17 / 10 / 2024
Fecha de publicación: 16 / 05 / 2025

Gemini Team, Georgiev, P., Lei, V., Burnell, R., Bai, L., ..., y Vinyals, O. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* e-prints. https://doi.org/10.48550/arXiv.2403.05530

Google. (2022). LaMDA: Language Models for Dialog Applications. *arXiv e-prints*. *https://doi.org/10.48550/arXiv.2201.08239*

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., y Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv e-prints*. https://doi.org/10.48550/arXiv.1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-print*s. https://doi.org/10.48550/arXiv.1907.11692

Open AI. (2020). Language Models are Few-Shot Learners. *arXiv e-p*rints. https://doi.org/10.48550/arXiv.2005.14165

Open AI. (2023). *GPT-4V(ision) System Card. https://openai.com/index/gpt-4v-system-card/*

Open AI, Anthropic AI, Zipline. (2021). Evaluating Large Language Models Trained on Code. *arXiv e-prin*ts. https://doi.org/10.48550/arXiv.2107.03374

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., . . . Hu, X. (2023). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. https://doi.org/10.48550/arXiv.2304.13712