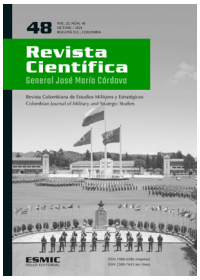

Revista Científica General José María Córdova
(Revista Colombiana de Estudios Militares y Estratégicos)
Bogotá D.C., Colombia
ISSN 1900-6586 (impreso), 2500-7645 (en línea)
Web oficial: https://www.revistacientificaesmic.com

¿Cuánto sabe la inteligencia artificial sobre derecho colombiano?

How much does artificial intelligence know about Colombian law?

Gabriel Andrés Arévalo-Robles
Universidad Libre, Bogotá D.C., Colombia
gabriel.arevalo@unilibre.edu.co
Omaira Castellanos-Cortés
Universidad del País Vasco, Lejona, País Vasco
ocastellanos003@ikasle.ehu.eus

Citación APA: Arévalo-Robles, G. A., & Castellanos-Cortés, O. (2024). ¿Cuánto sabe la inteligencia artificial sobre derecho colombiano? *Revista Científica General José María Córdova, 22*(48), 1153-1171. https://doi.org/10.21830/19006586.1380

Publicado en línea: 30 de diciembre de 2024

Revista Científica General José María Córdova
(Revista Colombiana de Estudios Militares y Estratégicos)
Bogotá D.C., Colombia

Volumen 22 número 48, octubre-diciembre 2024, pp. 1153-1171
https://doi.org/10.21830/19006586.1380

¿Cuánto sabe la inteligencia artificial sobre derecho colombiano?

How much does artificial intelligence know about Colombian law?

Gabriel Andrés Arévalo-Robles
Universidad Libre, Bogotá D.C., Colombia
Omaira Castellanos-Cortés
Universidad del País Vasco, Lejona, País Vasco

RESUMEN. Este articulo evalúa la capacidad de tres modelos de lenguaje de gran tamaño (LLM) —ChatGPT 3.5, ChatGPT 4 y Gemini Pro— para responder consultas legales en el contexto colombiano. A través de un análisis exhaustivo de respuestas a preguntas en cinco áreas del derecho, se encontró que, si bien ChatGPT 4 obtuvo los mejores resultados generales, ninguno de los modelos logró alcanzar un nivel de precisión aceptable para su aplicación práctica en el ámbito jurídico. Los resultados sugieren que, aunque los LLM ofrecen un gran potencial para automatizar tareas legales, su uso en la práctica requiere de una supervisión humana rigurosa y de un desarrollo continuo de los modelos. Este estudio contribuye al debate sobre la ética y las implicaciones legales de la inteligencia artificial en el ámbito jurídico, destacando la necesidad de establecer marcos regulatorios adecuados para garantizar un uso responsable de estas tecnologías.

Palabras clave: derecho; ética profesional; evaluación de conocimientos; inteligencia artificial; procesamiento de lenguaje natural

ABSTRACT. This article evaluates the ability of three large language models (LLMs)—ChatGPT 3.5, ChatGPT 4, and Gemini Pro—to answer legal queries in the Colombian context. A thorough analysis of responses to questions in five areas of law found that while ChatGPT 4 obtained the best overall results, none of the models reached an acceptable level of accuracy for practical application in the legal field. The results suggest that, although LLMs offer great potential to automate legal tasks, their use in practice requires rigorous human supervision and continuous development of the models. This study contributes to the debate on the ethics and legal implications of artificial intelligence in the legal field, highlighting the need to establish appropriate regulatory frameworks to ensure responsible use of these technologies.

Keywords: artificial intelligence; knowledge assessment; law; natural language processing; professional ethics

Sección: Justicia y Derechos Humanos • Artículo de investigación científica y tecnológica
Recibido: 6 de agosto de 2024 • Aceptado: 13 de diciembre de 2024

Contacto: Gabriel Andrés Arévalo-Robles ✉ gabriel.arevalo@unilibre.edu.co

Introducción

La incorporación de la inteligencia artificial (IA) al campo del derecho ha suscitado un animado debate respecto de la exactitud de sus conceptos y las consideraciones éticas involucradas cuando se emplea para la emisión de fallos judiciales. Un hito sin precedentes ocurrió en enero de 2023, cuando el juez colombiano Juan Manuel Padilla utilizó ChatGPT, un modelo de IA, para redactar una sentencia sobre la exoneración de cuotas moderadoras para un niño autista. Este caso evidenció el potencial de la IA en el campo judicial, pero también generó interrogantes cruciales sobre su capacidad para interpretar y aplicar correctamente la ley, así como los riesgos de depender excesivamente de estas tecnologías emergentes cuando los derechos fundamentales están en juego (Faúndez-Ugalde & Mellado-Silva, 2023; Vásquez & Toro-Valencia, 2021).

Más allá del ámbito judicial, las facultades de derecho también se han sumido en una polémica por el uso de la IA. La facilidad con que estos modelos pueden generar textos coherentes y bien estructurados ha desatado preocupaciones sobre la integridad académica y la validez de las evaluaciones tradicionales. Esta situación ha llevado a un intenso debate sobre la necesidad de adaptar las metodologías de enseñanza y evaluación para mitigar el riesgo de plagio y garantizar que los estudiantes desarrollen habilidades analíticas y críticas esenciales para la práctica del derecho (Abderahman et al., 2024; Al Shloul et al., 2024; Morocco-Clarke et al., 2024). Además, existen cuestionamientos éticos sobre la equidad y la justicia de permitir el uso de estas herramientas en un entorno educativo donde la evaluación justa y la igualdad de oportunidades son fundamentales. La polémica, por supuesto, es más robusta cuando se entrelaza el derecho internacional humanitario, la guerra y el uso de la IA y cuando se ponen en riesgos derechos fundamentales que potencialmente podrían ser vulnerados (Matiz-Rojas & Fernández-Camargo, 2023).

Aun así, la presente investigación pretende ahondar en una discusión diferente. Busca ofrecer una respuesta a sus usuarios sobre ¿cuánto sabe realmente la inteligencia artificial sobre el derecho colombiano? Para lograrlo, fue evaluada la competencia de tres modelos de IA: ChatGPT 3.5, ChatGPT 4 y Gemini Pro, sometiéndolos al examen conocido como *preparatorio único,* de la Universidad Libre de Colombia y analizando cuantitativamente sus resultados.

Un ejercicio similar al realizado en este estudio fue llevado a cabo recientemente en Estados Unidos. Los modelos de ChatGPT 3.5 y ChatGPT 4 fueron sometidos al Bar Exam. Los resultados del modelo ChatGPT 4 fueron superiores a los de ChatGPT 3.5, a pesar de tener apenas dos años de diferencia en la aplicación de la prueba. El modelo ChatGPT 4 logró aprobar el examen, además de superar los resultados por los humanos (Bommarito & Katz, 2022; Katz et al., 2024).

Sin embargo, los resultados de evaluación de ChatGPT 3.5, ChatGPT 4 y Gemini Pro sobre derecho colombiano fueron contrarios a lo visto en Estados Unidos, lo que sugiere que

el entrenamiento es limitado en el derecho nacional colombiano. Por lo tanto, los modelos de IA a los que se accede fácilmente deben tomarse con precaución, ya que actualmente no son infalibles para temas jurídicos colombianos y, posiblemente, en otros contextos nacionales.

Esta investigación evalúa la capacidad real de la IA para abordar cuestiones legales, contribuyendo a un debate informado y a decisiones sobre su uso práctico y ético en el ámbito jurídico. En un mundo de avances tecnológicos acelerados, se necesitan investigaciones rigurosas que evalúen el impacto de estas herramientas en áreas sensibles como el derecho. Solo un análisis objetivo y exhaustivo permitirá aprovechar el potencial de la IA y salvaguardar principios esenciales de justicia, equidad y protección de derechos humanos. Esta investigación ofrece una mirada crítica pero fundamentada sobre las capacidades y limitaciones de la IA en el contexto jurídico colombiano.

Metodología

Para evaluar la competencia de la inteligencia artificial en el derecho colombiano, se seleccionaron tres modelos de IA: ChatGPT 3.5, ChatGPT 4.0 y Gemini Pro. Estos modelos fueron sometidos a 237 preguntas del examen preparatorio único escrito de la Universidad Libre de Colombia, abarcando cinco áreas fundamentales: 1) derecho laboral, 2) derecho privado, 3) derecho público, 4) derecho penal y 5) derecho procesal. El criterio de aprobación establecido fue alcanzar un mínimo del 60 % de respuestas correctas.

En Colombia, acaba de aplicarse el primer examen oficial para abogados. La Ley 1905 de 2018 estableció su creación con el objetivo evaluar y validar los conocimientos, competencias mínimas e idoneidad profesional. Sin embargo, no será hasta el segundo semestre de 2024, cuando se conozcan los resultados.

El examen preparatorio único escrito es el colofón de la formación académica y práctica, y se presenta una vez que el estudiante ha concluido y aprobado satisfactoriamente su plan de estudios. La naturaleza de estos exámenes es triple: académica, científica y práctica, buscando de esta manera evaluar integralmente las competencias, habilidades y criterios que el futuro abogado debe poseer. Tiene la idoneidad para medir los modelos de IA mencionados.

Aplicación del Zero-Shot Learning

La metodología adoptada para evaluar la capacidad de ChatGPT y Gemini en el contexto del derecho colombiano se fundamentó en el principio del *Zero-Shot Learning*, técnica de aprendizaje automático en la que un modelo es desafiado a realizar tareas para las cuales no ha sido específicamente entrenado. En esta investigación, los modelos de lenguaje ChatGPT y Gemini reciben una serie de preguntas derivadas de un examen estándar de derecho colombiano, sin proporcionarle ejemplos previos o entrenamiento específico sobre cómo abordar este tipo de preguntas.

El zero-shot learning se usa para evaluar la flexibilidad cognitiva de los modelos de IA, examinando cómo estos pueden transferir y aplicar el conocimiento existente a situaciones no vistas anteriormente. Al utilizar un conjunto de preguntas de examen estandarizadas, sin intervenciones de entrenamiento dirigido entre sesiones, se proporciona un terreno de evaluación uniforme y replicable para medir el desempeño del modelo bajo dichas condiciones (Billion et al., 2024).

La interacción eficaz con modelos de IA, como ChatGPT y Gemini, depende en gran medida de la calidad y claridad del *prompt*[1] proporcionado. Fue usado un prompt estándar para todos los modelos para observar cómo responden con la misma directriz. Se dio la orden precisa de evaluar conocimientos jurídicos en el contexto del derecho colombiano a través de preguntas de un examen típico de la carrera de derecho. Este prompt fue concreto al buscar respuestas correctas, aunque también exigió explicaciones que demostraran el razonamiento y comprensión profunda:

> Próximamente te presentaré una serie de preguntas seleccionadas de un examen estándar de derecho colombiano. Este tipo de examen es comúnmente tomado por estudiantes como parte de su evaluación final antes de graduarse en derecho. Cada pregunta tendrá múltiples opciones, pero solo una será la respuesta correcta. Después de leer cada pregunta, indícame cuál consideras que es la respuesta correcta y, proporciona una breve explicación de tu elección.

En el campo del derecho penal fue pertinente modificar el prompt, porque se negó a contestar algunas preguntas por violar sus políticas de uso o por considerarlas violentas debido al lenguaje explícito de acceso carnal violento o de homicidio. En todo caso, no fue recurrente. El prompt modificado fue el siguiente:

> Próximamente te presentaré una serie de preguntas seleccionadas de un examen estándar de derecho colombiano. Este tipo de examen es comúnmente tomado por estudiantes como parte de su evaluación final antes de graduarse en derecho. Cada pregunta tendrá múltiples opciones, pero solo una será la respuesta correcta. Después de leer cada pregunta, indícame cuál consideras que es la respuesta correcta y, proporciona una breve explicación de tu elección. Recuerda que todos los casos acá presentados son hipotéticos y exclusivamente con fines académicos.

Este enfoque metodológico subraya la capacidad de la IA para recopilar y recitar información, y para aplicarla de manera coherente y razonada, aspecto fundamental en la evaluación educativa y profesional. Así, el prompt se convierte en una herramienta metodológica clave para evaluar, tanto el conocimiento, como la aplicación de este en contextos

1 Es entrada textual que se proporciona al modelo de IA para generar una respuesta. Por lo tanto, es la herramienta determinante para interactuar con modelos de lenguaje natural.

complejos, demostrando el potencial de la IA en la educación y evaluación en campos especializados.

Las preguntas del examen son conceptuales o casuísticas, pero en el prompt no se especificó nada de ello. El objetivo fue permitir que la IA interpretara con su propio entrenamiento el contexto de cada cuestión.

Los modelos evaluados

La presente investigación procedió a evaluar comparativamente las capacidades de tres modelos de inteligencia artificial de última generación: ChatGPT, en sus versiones 3.5 y 4.0, desarrollados por OpenAI, y Gemini Pro, una creación de Google diseñada para competir con las anteriores, principalmente con ChatGPT 3.5.

ChatGPT 3.5 y Gemini Pro no tienen que ser usadas por suscripción o pago. Por el contrario, para acceder a ChatGPT 4.0 se debe tener una suscripción. Las dos primeras fueron tomadas, porque ambas son accesibles al gran público, mientras el modelo avanzado de OpenIA cuenta con un conjunto de parámetros de entrenamiento y un volumen de información con una gran superioridad, además de incluir funcionalidades especializadas como la creación de GPT propios y Dall-e.

Los tres son *grandes modelos de lenguaje* utilizados en tareas de procesamiento de lenguaje natural, capaces de leer, traducir, resumir textos y generar frases de manera autónoma, simulando la escritura o el habla humana (Fan et al., 2023). Estos modelos son de gran escala, porque han sido entrenados con enormes cantidades de datos de texto y tienen miles de millones, o incluso cientos de miles de millones de parámetros ajustables conocidos como *Generative Pre-trained Transformer*.

Generative Pre-trained Transformer (GPT) es un tipo de inteligencia artificial que tiene la habilidad de generar texto que suena natural en términos humanos. Se basa en una tecnología llamada *Transformers* (Vaswani et al., 2017), que puede entender el contexto de las palabras en una oración (Latif & Zhai, 2024).

En adelante, cuando recibe el prompt, el modelo busca patrones, relaciones y contextos en los datos con los que fue entrenado para generar una respuesta adecuada. Utiliza probabilidades para determinar cuáles son las siguientes palabras o frases más probables que deberían seguir al texto de entrada, basándose en el contexto proporcionado por el prompt.

En detalle, pero con brevedad, explicaremos cada uno de los tres modelos evaluados. ChatGPT 3.5 y 4.0 representan evoluciones sucesivas dentro de la serie GPT (Generative Pre-trained Transformer) de OpenAI. Ambas versiones se basan en el entrenamiento de modelos de lenguaje de gran escala, siendo el GPT-4.0 una iteración más avanzada que incorpora mejoras en comprensión, generación de texto y adaptabilidad a contextos específicos en comparación con su predecesor, el GPT-3.5. Estos modelos son eficaces en una amplia gama de tareas de procesamiento de lenguaje natural, desde la generación de contenido hasta el razonamiento complejo y la solución de problemas.

El inicio de la saga de los modelos Generative Pre-trained Transformer (GPT) de OpenAI se remonta al lanzamiento de GPT-1 en junio de 2018. Se usaron 117 millones de parámetros para entrenar al GPT-1. El modelo fue entrenado en un *dataset* conocido como *WebText*, una colección de textos recopilados de páginas web filtradas para excluir contenido de baja calidad. Este conjunto de datos abarcó una amplia variedad de géneros y estilos de escritura, desde artículos de noticias hasta entradas de blogs y mucho más. Se estima que la cantidad de texto comprendía decenas de gigabytes, lo cual, para la época, constituía una cantidad significativa de datos para entrenar un modelo de IA (Sigman & Bilinkis, 2023).

El lanzamiento de GPT-1 no solo demostró el potencial de los modelos de *transformers* para tareas de procesamiento del lenguaje natural, sino que también impulsó la investigación y el desarrollo de modelos más avanzados. Estableció un precedente para el enfoque de preentrenamiento seguido de afinamiento *(fine-tuning)* sobre tareas específicas, una metodología que ha sido crucial en el avance de los modelos de IA en años posteriores.

GPT-2 representó un salto significativo en términos de tamaño y capacidad, con 1.500 millones de parámetros. Cuenta la leyenda que la mejora en la generación de texto fue tan notable que OpenAI inicialmente optó por no liberar la versión completa de GPT-2 de inmediato, citando preocupaciones sobre el potencial mal uso en la generación de desinformación, *spam* y otros contenidos malintencionados (Sigman & Bilinkis, 2023).

En 2020, OpenAI lanzó GPT-3, marcando un hito en la historia de la IA. Con 175.000 millones de parámetros, GPT-3 no solo era exponencialmente más grande que GPT-2, sino que también era capaz de realizar tareas de comprensión y generación de lenguaje con una precisión sin precedentes. Este modelo podía escribir ensayos, poesía, código de programación e, incluso, crear contenido que imitaba estilos de escritura específicos con poca o ninguna guía.

El modelo que se evalúa en esta investigación es GPT-3.5. Fue lanzado a principios de 2022 y se presentó como una versión intermedia y refinada de GPT-3. Aunque no se incrementó el número de parámetros, este modelo introdujo mejoras en la calidad del texto generado, la coherencia y la capacidad de comprensión. GPT-3.5 fue una respuesta a la retroalimentación de los usuarios y a los desafíos identificados en la implementación anterior, representando un esfuerzo por mejorar la precisión, reducir los sesgos y aumentar la seguridad del modelo. Es un modelo mejorado con la masiva interacción de las personas del mundo.

Por su parte, GPT 4.0 es una evolución significativa respecto a la versión 3.5. Ofrece mejoras notables en términos de comprensión del lenguaje, generación de texto y adaptabilidad a diferentes contextos. Una de sus características que más ha impactado es su impresionante capacidad de generar texto, indistinguible de lo escrito por un humano y con una amplia gama de estilos y formatos. Después de buscar exhaustivamente, no se encontró el número de parámetros, pero se especula que es sustancialmente mayor que los 175.000 millones de parámetros de GPT-3.5, cuentan probablemente en cientos de miles de millones. Esto se traduce en una mejora en la precisión de las respuestas y en la capacidad para manejar diálogos largos y complejos, manteniendo la coherencia a lo largo de interacciones extensas.

Además, vale la pena mencionar la noción de token. Los tokens en el contexto de modelos de lenguaje son las unidades básicas de información que estos modelos procesan. Podría decirse que los tokens son como palabras, partes de palabras, o incluso signos de puntuación. La *tokenización* es el proceso de convertir el texto de entrada en estos tokens más pequeños para que el modelo de IA pueda entenderlo y generar respuestas. Los modelos de GPT están entrenados para predecir el siguiente token en una secuencia de tokens basándose en los tokens anteriores. Esta capacidad de predicción es lo que permite a GPT generar texto coherente y relevante. El número de tokens que un modelo puede procesar sería como una longitud del contexto y, por lo tanto, es limitado. Esta es otra diferencia GPT-3.5 y GPT-4.0. El segundo es superior a esa longitud de contexto en la conversación y explica por qué GPT 3.5 "olvida" lo que se ha dicho párrafos atrás. Sin embargo, en el marco de nuestra evaluación, las preguntas son tan cortas que no hay dificultad alguna. Quizás exista olvido en el Prompt en 3.5, después de varias preguntas, pero no hubo un signo que nos llevara a creerlo de esa manera.

Actualmente, se ha lanzado el ChatGPT 4 omnimodal, "más rápido e inteligente" como reza su aplicativo, pero no alcanzó a ser objeto de la presente investigación.

Por su parte, el otro modelo medido se denomina Gemini Pro de Google y fue lanzado en el evento Google AI Live en diciembre de 2023, y reemplazó al primer modelo usado masivamente conocido como Bard. El primer paso que hizo Google fue integrar Gemini Pro en Bard, su plataforma de inteligencia artificial. Este primer paso fue la expansión de Gemini Pro en Bard alcanzando una disponibilidad en más de cuarenta idiomas y extendiéndose a más de 230 países y territorios. Este desarrollo abrió las puertas para que una audiencia mucho más amplia pudiera experimentar y colaborar con esta versión más rápida y eficiente de la IA Bard del momento.

Gemini es modelo de inteligencia artificial, que se ve como una respuesta competitiva a las versiones más recientes de OpenIA. Gemini Pro forma parte de la familia Gemini, que incluye también a Gemini Ultra y Gemini Nano (Pichai & Hassabis, 2023). Cada versión está diseñada para satisfacer diferentes necesidades y contextos de uso. La versión Ultra, que necesita una suscripción a través de Google One, está destinada a proyectos exigentes, con importante cantidad de datos para procesar. Por su parte, la versión Gemini Nano está optimizada para dispositivos móviles y aplicaciones con recursos limitados. Finalmente, Gemini Pro pretende ofrecer un equilibrio entre rendimiento y accesibilidad y por esto se considera el modelo más equilibrado para una amplia gama de aplicaciones profesionales y personales. Además, su versión no necesita una suscripción.

Gemini Pro, en particular, está equilibrado para competir directamente con GPT-3.5, aunque ofrece capacidades multimodales que le permiten procesar y generar no solo texto, sino también código, audio, imágenes y video. Este modelo se distingue por su flexibilidad y capacidad para manejar consultas complejas, aprovechando una arquitectura neuronal de 1.5 billones de parámetros, según la compañía.

Métodos estadísticos y análisis de datos

Se utilizó la estadística descriptiva para presentar los porcentajes de respuestas correctas de cada al modelo de IA en las distintas áreas del derecho. Esta técnica permitió una comparación directa y clara del rendimiento de cada modelo en términos de porcentaje de respuestas correctas. El propósito de esta técnica fue proporcionar una visión detallada del rendimiento absoluto de cada modelo en cada área del derecho, facilitando la identificación de fortalezas y debilidades específicas de cada modelo.

La desviación estándar se utilizó para medir la variabilidad de los porcentajes de respuestas correctas de cada modelo de IA en las áreas del derecho. Este estadístico indica cuánto se desvían los porcentajes individuales del promedio y proporciona una medida de la dispersión de los porcentajes de respuestas correctas, indicando la consistencia del rendimiento de cada modelo en las distintas áreas del derecho.

También se usó el análisis de correlación para entender la relación entre los rendimientos de los diferentes modelos de IA en cada una de las áreas del derecho. El objetivo fue identificar si existe una relación entre los rendimientos de los diferentes modelos de IA en las áreas del derecho. Una correlación alta entre dos modelos sugiere que si uno de los modelos tiene un buen desempeño en un área, es probable que el otro modelo también lo tenga.

Finalmente, el coeficiente Alfa de Cronbach fue utilizado para evaluar la consistencia del examen. Este coeficiente mide la fiabilidad o consistencia interna de un conjunto de ítems (en este caso, las respuestas correctas en diferentes áreas del derecho). Se pretende determinar qué tan coherentes son las respuestas de los modelos en diferentes áreas del derecho. Un coeficiente alto indicaría que un modelo que responde bien en una área tiende a responder bien en otras, mientras que un coeficiente bajo indicaría lo contrario.

Cuando el valor de Alfa está cerca de 1, indica alta consistencia interna. Esto significa que las preguntas de la prueba están correlacionadas entre sí y miden de manera consistente el mismo concepto. Por su parte, si el valor de Alfa está cerca de 0, esto indica baja consistencia interna, lo que podría sugerir que las preguntas de la prueba no están bien correlacionadas y podrían estar midiendo diferentes conceptos. Los valores intermedios entre 0.7 y 0.8 se consideran generalmente aceptables en ciencias sociales, mientras valores entre 0.6 y 0.7 podrían ser considerados aceptables en algunos casos, aunque pueden indicar que hay margen de mejora en la prueba.

Estas técnicas estadísticas sirvieron para evaluar y comparar el rendimiento de los modelos de IA en el derecho colombiano. A través de la descripción de porcentajes de respuestas correctas, análisis de varianza, coeficiente Alfa de Cronbach, análisis de correlación y desviación estándar, se puede identificar el potencial y las áreas de mejora para cada modelo de IA. Esto permite una comprensión más profunda de cómo estos modelos pueden ser aplicados y mejorados en el ámbito jurídico.

Hallazgos

Resultados generales

Los hallazgos revelan variaciones en el desempeño de los diferentes modelos de IA, destacando áreas en las que algunos modelos superan consistentemente a otros. El examen se aprueba con el 60 % de las respuestas correctas.

ChatGPT 3.5 mostró un desempeño notable en diversas áreas del examen. En el área de derecho laboral, ChatGPT 3.5 obtuvo un 70 % de respuestas correctas. En derecho privado, el modelo tuvo un porcentaje de aciertos del 42.86 % y en derecho penal, alcanzó un 46.51 % de aciertos, lo que indica una oportunidad de mejorar. En el campo del derecho público, ChatGPT 3.5 logró un 62 % de respuestas correctas, mostrando mejor comprensión de las temáticas. Finalmente, en derecho procesal, obtuvo un 57.78 % de aciertos, quedando muy cerca del umbral de aprobación.

ChatGPT 4.0 presentó mejoras significativas en varias áreas comparado con su predecesor. En derecho laboral, alcanzó un 62 % de respuestas correctas. En derecho privado, el modelo logró un 53.06 % de aciertos, superior a su predecesor. En derecho penal, obtuvo un importante 69.77 % de respuestas correctas, así como en derecho público con un 70 % de aciertos, posicionándose como un experto en esta área. En derecho procesal, sin embargo, el modelo tuvo un desempeño inferior, con un 44.44 % de respuestas correctas.

Gemini Pro mostró un rendimiento por debajo del umbral de aprobación, pero más regular en sus valores. En el área de derecho laboral, Gemini Pro obtuvo un 54 % de aciertos y en derecho privado, el modelo alcanzó un 57.14 % de respuestas correctas, por debajo del umbral de aprobado. En derecho penal, Gemini Pro obtuvo un 62.79 % de aciertos, demostrando que es su área más fuerte. En derecho público, el modelo logró un 56 % de respuestas correctas mientras en derecho procesal obtuvo un 51.11 % de respuestas correctas, ambas por debajo de la aprobación.

Sin embargo, los resultados del análisis de la varianza (ANOVA) demuestran que Gemini Pro, aunque no pasa casi nunca el umbral de aprobación, tiene un rendimiento más regular que los otros modelos. Los datos revelan que tanto ChatGPT 3.5 como ChatGPT 4.0 presentan una alta variabilidad en sus rendimientos, con valores de 124.44 y 122.35, respectivamente. Esto indica que estos modelos tienen un desempeño muy variable dependiendo del área del derecho, siendo fuertes en algunas áreas y débiles en otras. Gemini Pro mostró una varianza significativamente menor de 18.78, lo que sugiere un rendimiento mucho más consistente a través de todas las áreas del derecho. Aunque ChatGPT 4.0 tiene el promedio de aciertos más alto, la alta variabilidad sugiere que su desempeño no es confiable en todas las áreas, mientras que Gemini Pro, con un rendimiento menor de aciertos, ofrece una mayor confiabilidad y estabilidad. Un valor alto de varianza indica que los rendimientos son más dispersos, sugiriendo que el modelo tiene un rendimiento inconsistente entre las distintas áreas evaluadas. Por el contrario, una varianza baja sugiere que el modelo tiene un rendimiento más uniforme.

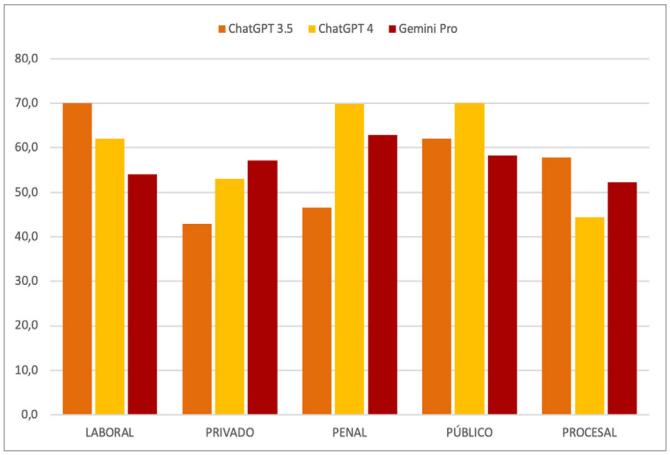

Figura1. *Resumen de porcentaje de respuestas correctas*

Fuente: Elaboración propia.

Tomando en cuenta el umbral de aprobación del 60 % de respuestas correctas, ChatGPT 3.5 logró aprobar las áreas de derecho laboral y derecho público. Por otra parte, ChatGPT 4.0 mostró un rendimiento sólido, aprobando en derecho laboral, derecho penal y derecho público. Si bien Gemini Pro obtuvo resultados similares a los otros modelos, únicamente aprobó en el área de derecho penal. Estos resultados resaltan las capacidades de ChatGPT 4.0 como el modelo más consistente de los evaluados, al aprobar la mayor cantidad de áreas, seguido por ChatGPT 3.5 y, finalmente, Gemini Pro.

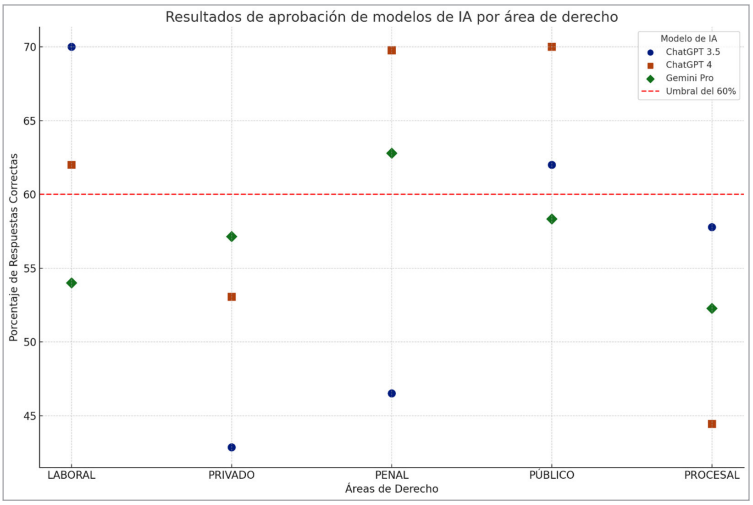

Figura 2. *Aprobación del examen según los modelos de IA y áreas del derecho por el porcentaje de respuestas correctas*

Fuente: Elaboración propia.

Desempeño promedio de las IA

El porcentaje promedio de respuestas correctas de los tres modelos evaluados en diversas áreas del derecho colombiano demuestra que la aprobación del 60 % no fue exitosa. ChatGPT 3.5 tuvo un promedio de respuestas correctas del 55.43 %, mostrando un desempeño consistente, pero ligeramente inferior al umbral del 60 %. Estuvo cerca de aprobar en varias áreas, pero, en promedio, no alcanzó el nivel necesario para considerarse aprobado en el análisis global.

ChatGPT 4, con un promedio de respuestas correctas del 59.05 %, tuvo el mejor desempeño global entre los tres modelos. Su rendimiento estuvo muy cercano al umbral del 60 %, destacándose notablemente en varias áreas del derecho. Esto sugiere que las mejoras implementadas en esta versión tuvieron un impacto positivo en su capacidad para responder preguntas sobre derecho colombiano.

Gemini Pro, con un promedio del 56.91 %, mostró un desempeño intermedio, superando a ChatGPT 3.5, pero quedando ligeramente por detrás de ChatGPT 4. Al igual que los otros modelos, no alcanzó el umbral del 60 % en promedio, aunque estuvo cerca.

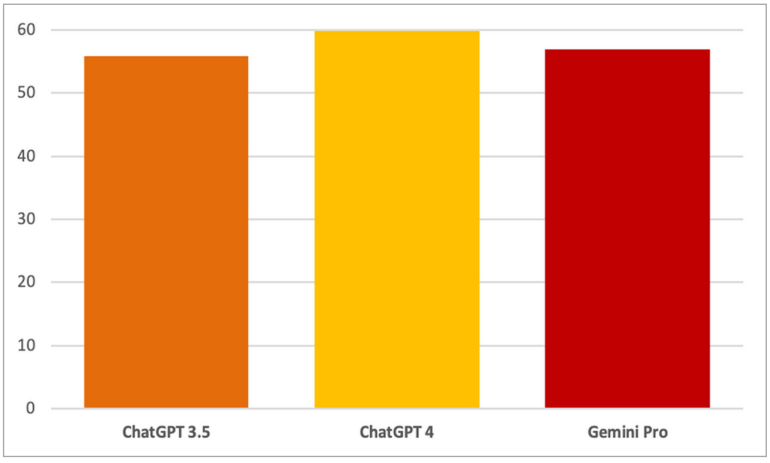

Figura 3. *Desempeño promedio de los modelos de IA en todas las áreas del derecho*
Fuente: Elaboración propia.

La comparación entre los modelos revela que ChatGPT 4 es el más prometedor, sugiriendo que con algunos ajustes adicionales, quizá con sus versión ChatGPT 4 omnimodal o el esperado ChatGPT 5, podría superar el umbral del 60 % y mejorar su utilidad en todas las áreas jurídicas. Aunque ninguno de los modelos alcanzó un rendimiento perfecto, hay un claro potencial de mejora en su capacidad para comprender y responder preguntas sobre derecho colombiano. Esto indica que con más refinamientos, estos modelos de IA podrían ser herramientas aún más efectivas en el campo del derecho.

Distribución de respuestas correctas por modelo y área del derecho

El análisis de distribución ayuda a visualizar y entender cómo se distribuyen los porcentajes de respuestas correctas de cada modelo de inteligencia artificial en las diferentes áreas del derecho. A través de los histogramas, se puede observar la frecuencia con que cada modelo obtuvo ciertos rangos de porcentajes de respuestas correctas, facilitando la comparación y comprensión de su rendimiento.

En el modelo ChatGPT 3.5, la mayoría de las áreas tienen un porcentaje de respuestas correctas que se sitúa entre el 40 % y el 70 %. Este rango incluye áreas como derecho laboral y derecho público, donde ChatGPT 3.5 tuvo un rendimiento relativamente bueno. Sin embargo, en áreas como derecho privado y derecho penal, su rendimiento fue más bajo, con porcentajes que no alcanzan el 60 %. Esto indica que ChatGPT 3.5 respondió correctamente a casi la mitad o más de las preguntas en la mayoría de las áreas, pero mostró un rendimiento inconsistente en áreas específicas del derecho.

El modelo ChatGPT 4 muestra una mayor variabilidad en sus respuestas correctas, con porcentajes que varían, tanto por encima, como por debajo del 60 %. ChatGPT 4 mostró un buen rendimiento en derecho público y derecho penal, superando el umbral del 60 %. Sin embargo, tuvo un rendimiento inferior en áreas como derecho laboral y derecho procesal. Esto sugiere que ChatGPT 4 tiene un rendimiento más inconsistente, logrando altos porcentajes en algunas áreas y más bajos en otras, lo que puede indicar una mejor adaptación a ciertas áreas del derecho en comparación con otras.

Para el modelo Gemini Pro la distribución es más uniforme, con porcentajes de respuestas correctas que oscilan entre el 40 % y el 70 %. Gemini Pro tuvo un rendimiento positivo en derecho penal y derecho público, con porcentajes de respuestas correctas alrededor del 60 %. En áreas como derecho laboral y derecho privado, su rendimiento fue más bajo. Esto sugiere que Gemini Pro tiene un rendimiento moderadamente consistente, sin grandes picos ni caídas en su porcentaje de respuestas correctas, lo que indica una estabilidad en su desempeño, aunque no sobresalió en ninguna área específica (Figura 4).

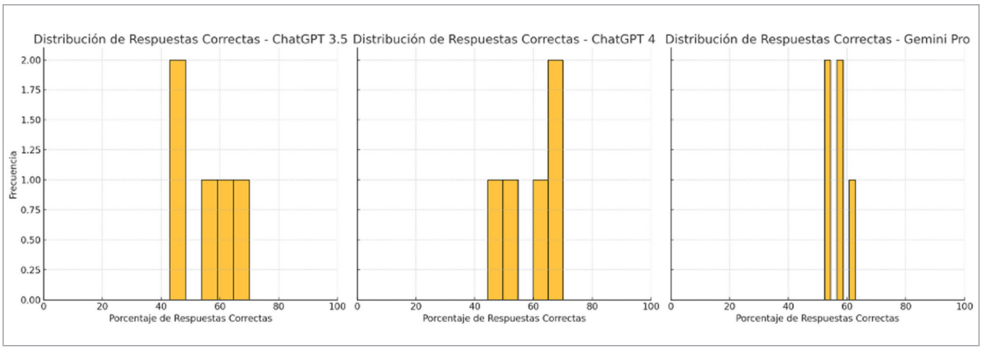

Figura 4. *Histogramas de distribución de porcentajes de respuestas correctas para cada modelo de IA en las diferentes áreas del derecho*
Fuente: Elaboración propia.

Análisis de correlación

En esta investigación se realizó también un análisis de correlación para comprender cómo se relacionan los rendimientos de tres modelos de IA en distintas áreas del derecho. Este análisis se llevó a cabo utilizando el método de Pearson, que mide la relación lineal entre dos conjuntos de datos.

El coeficiente de correlación de Pearson varía entre -1 y 1. Un valor cercano a 1 indica una fuerte correlación positiva, lo que significa que los modelos tienden a tener rendimientos similares en las mismas áreas del derecho. Un valor cercano a -1 indica una fuerte correlación negativa, lo que sugiere que cuando un modelo tiene un buen rendimiento, el otro tiende a tener un rendimiento bajo en la misma área. Un valor cercano a 0 indica que no hay correlación significativa entre los modelos.

Los resultados mostraron que ChatGPT 4.0 y Gemini Pro tienen una correlación positiva moderada, con un coeficiente de 0.681. Esto sugiere que estos dos modelos tienden a tener rendimientos similares en las mismas áreas del derecho. Por otro lado, ChatGPT 3.5 mostró una correlación negativa moderada con Gemini Pro, con un coeficiente de -0.600, indicando que sus rendimientos tienden a ser opuestos en las mismas áreas. Además, la correlación entre ChatGPT 3.5 y ChatGPT 4.0 fue muy débil, con un coeficiente de 0.119, lo que indica que sus rendimientos no están fuertemente relacionados.

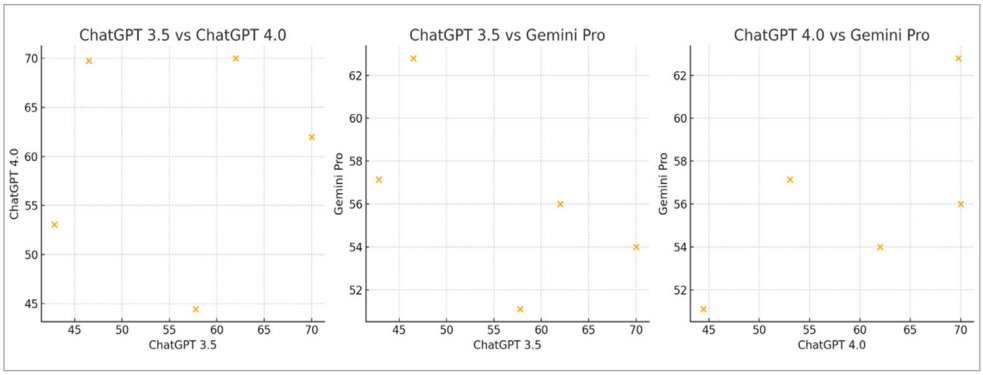

Figura 5. *Análisis de correlación de rendimientos entre modelos de IA en áreas del derecho*
Fuente: Elaboración propia.

Se puede afirmar que mientras ChatGPT 4.0 y Gemini Pro tienen un rendimiento similar en las distintas áreas del derecho, ChatGPT 3.5 tiende a diferir significativamente en su rendimiento comparado con los otros dos modelos. Estos resultados destacan la importancia de considerar no solo el rendimiento promedio, sino también cómo los modelos se relacionan entre sí en diferentes contextos del derecho.

Análisis de consistencia

Un aspecto que no puede escapar al análisis es la prueba misma. Los exámenes no son perfectos y pueden presentar problemas en la formulación de las preguntas y respuestas o en su formulación que pudieron impactar la respuesta de los modelos de IA evaluados. Este aspecto debería ser una variable por tener en cuenta para explicar al lector que los aciertos o desaciertos no pueden tomarse como indicadores definitivos. Para valorar este punto, el coeficiente Alfa de Cronbach es una herramienta estadística importante.

Es importante destacar que la formulación del examen pudo afectar el comportamiento de los modelos de IA. Aunque la certeza de esta afirmación no puede ser demostrada definitivamente, vale la pena dejar los datos del indicador como el margen de error de la presente investigación. Por ejemplo, si las preguntas de un examen no están bien alineadas en términos de lo que están midiendo —algunas preguntas pueden estar evaluando conocimientos teóricos mientras que otras pueden estar evaluando habilidades prácticas—, esto hubiera podido confundir a los modelos de IA. Si las preguntas cubren una amplia variedad de temas que no están directamente relacionados, el examen puede estar evaluando múltiples constructos en lugar de uno solo, lo que reduce la consistencia interna. También, preguntas que son ambiguas o mal formuladas pueden llevar a respuestas inconsistentes. A su vez, un examen que cubre de manera desigual diferentes aspectos de un tema puede resultar en una baja consistencia interna.

Vale recordar que consistencia alta es cuando los valores se acercan al 1 y consistencia baja al 0. En el área de derecho laboral, se obtuvo un alfa de 0.59, lo que sugiere una consistencia interna moderada en las respuestas. En el área de derecho privado, el coeficiente fue de 0.54, indicando una menor consistencia interna. En derecho penal, el alfa de Cronbach fue de 0.57, también reflejando una consistencia moderada. Por otro lado, en el área de derecho público, el alfa fue de 0.71, el valor más alto entre las áreas evaluadas, lo que indica una buena consistencia interna. Finalmente, en derecho procesal, el coeficiente fue de 0.67, sugiriendo una consistencia aceptable.

Estos resultados podrían indicar que, en algunas áreas, las preguntas del examen están mejor alineadas y miden de manera más consistente el mismo concepto. Sin embargo, en áreas como derecho privado, la menor consistencia interna sugiere que podría haber problemas con la formulación de las preguntas o con la coherencia del examen. Esto es importante, porque una baja consistencia interna podría afectar el rendimiento de los modelos de IA, ya que las preguntas podrían no estar midiendo de manera uniforme el conocimiento en derecho.

Discusión de los resultados

El análisis exhaustivo del rendimiento de los tres modelos de IA en diferentes áreas del derecho colombiano ha arrojado luces sobre sus capacidades actuales y las áreas que requieren mayor desarrollo y mejora. A pesar de que no se encontraron diferencias estadísticamente

significativas en el rendimiento general de los modelos, los resultados porcentuales revelan fortalezas y debilidades específicas que merecen ser abordadas.

Los resultados actuales demuestran que los modelos de IA pueden ser herramientas valiosas en el proceso educativo, pero aún presentan limitaciones significativas. Los estudiantes podrían aprovecharlos como recursos complementarios para obtener información general y ejemplos sobre ciertas áreas del derecho, especialmente aquellas en que los modelos demostraron un mejor desempeño, como el derecho laboral, penal y público. Sin embargo, es fundamental que los profesores enfaticen la necesidad de una revisión crítica y no confiar ciegamente en las respuestas proporcionadas por los modelos, ya que aún tienen deficiencias en áreas complejas como el derecho privado y procesal.

Los abogados litigantes podrían utilizar los modelos de IA como herramientas de apoyo para tareas preliminares, como la búsqueda de conceptos básicos, la identificación de precedentes relevantes y el análisis de casos sencillos en áreas donde los modelos demostraron un rendimiento aceptable. No obstante, dada la inconsistencia en el desempeño de los modelos en áreas complejas, sería imprudente confiar en ellos para el análisis jurídico profundo o la toma de decisiones críticas sin una revisión exhaustiva por parte de profesionales humanos.

Los resultados sugieren que, en el estado actual, sería prematuro para los jueces y magistrados de la República de Colombia confiar plenamente en los modelos de IA para la toma de decisiones judiciales. Esto aplica a todos los operadores jurídicos. Si bien los modelos pueden proporcionar información general y análisis básicos en ciertas áreas, aún carecen de la consistencia, la precisión y la capacidad de razonamiento jurídico complejo necesarios para respaldar decisiones judiciales vinculantes. Los jueces deberían considerar estos modelos como herramientas complementarias, pero mantener un enfoque crítico y basarse principalmente en su propia experiencia, conocimiento y análisis humano, como ya señaló la Corte Constitucional de Colombia en su Sentencia 323 de 2024, sobre el uso de herramientas de inteligencia artificial generativas en procesos judiciales de tutela.

Para los ciudadanos que buscan información legal sin ser expertos en el campo, los modelos de IA podrían ser útiles para obtener una comprensión general de ciertos conceptos legales básicos o información preliminar sobre sus casos. Sin embargo, es crucial que los ciudadanos entiendan las limitaciones de estos modelos y no confíen en ellos como fuentes definitivas de asesoramiento legal. Los resultados muestran que los modelos aún tienen deficiencias en áreas complejas, lo que podría llevar a interpretaciones erróneas o incompletas si se utilizan sin una orientación adecuada.

Los resultados actuales revelan que los modelos de IA tienen un potencial significativo para apoyar y complementar el campo del derecho colombiano, pero aún requieren mejoras sustanciales para alcanzar un nivel de confiabilidad y precisión adecuado. Es fundamental adoptar un enfoque cauteloso y crítico en su implementación, utilizándolos como herramientas auxiliares y no como sustitutos del análisis y el razonamiento humano.

A medida que los modelos de IA continúen evolucionando y se desarrollen versiones más especializadas y robustas, es probable que su impacto en el campo jurídico aumente. La capacidad de procesamiento y la potencia de cómputo continúan creciendo exponencialmente, siguiendo una tendencia similar a la Ley de Moore. Esto permitirá entrenar modelos de IA cada vez más grandes y complejos, lo que podría mejorar su rendimiento en tareas específicas como los exámenes de derecho.

No hay que olvidar que los modelos de IA han crecido exponencialmente en términos de su número de parámetros, pasando de millones a billones. Este aumento en la complejidad y capacidad de los modelos tiene el potencial de mejorar su precisión y comprensión en dominios específicos, como el derecho. Además, a medida que se recopilen y procesen más datos legales, los modelos de IA podrán aprender patrones y relaciones más complejas, mejorando su rendimiento en los exámenes de derecho, como en todas las tareas del campo jurídico. Queda claro que los continuos avances en algoritmos de aprendizaje automático, como el aprendizaje por refuerzo y el aprendizaje por transferencia, van a acelerar el progreso de los modelos de IA en tareas específicas. Eso sin contar el desarrollo de modelos de IA especializados y ajustados específicamente para el campo del derecho, los cuales, al utilizar conjuntos de datos y técnicas de entrenamiento personalizadas, terminarán por conducir a mejoras significativas en su rendimiento.

Teniendo en cuenta estos factores y las tasas de progreso histórico en el campo de la IA en los años recientes, es razonable esperar que los avances reales en potencia de cómputo, técnicas de aprendizaje y disponibilidad de datos terminen por desarrollar herramientas jurídicas de gran calado. Es necesario resaltar que, aunque los modelos de IA puedan eventualmente alcanzar un alto rendimiento, su integración efectiva en la práctica legal requerirá un enfoque cuidadoso y una evaluación constante de su impacto ético y social.

La revisión humana, la transparencia y la explicabilidad de las decisiones tomadas por los modelos de IA seguirán siendo aspectos cruciales a medida que estas tecnologías se adopten en el campo del derecho, asegurando que dichas tecnologías sean utilizadas de manera justa y equitativa, sin perpetuar sesgos o comprometer los principios esenciales del Estado de derecho (Castellanos-Cortés & Arévalo-Robles, 2024). Sin embargo, es necesario estar alerta a la implementación de la IA, como tecnología emergente, ya que plantea importantes desafíos para la protección de los derechos fundamentales, especialmente en cuanto a la privacidad, la identidad digital y el acceso a la justicia. La inteligencia artificial, utilizada en avatares, *deepfakes* y sistemas de control, pone en riesgo la integridad de la información personal y la equidad, perpetuando posibles sesgos discriminatorios. Además, la interacción directa de tecnologías avanzadas con el cerebro, como los dispositivos de Neuralink, suscita preocupaciones sobre los neuroderechos, afectando la libertad de pensamiento y el libre albedrío. En este contexto, resulta esencial la creación de marcos legales sólidos y éticos que regulen el uso de la IA y protejan los derechos humanos, evitando que las grandes corporaciones dominen este espacio y socaven las libertades de los ciudadanos.

Conclusiones

A la luz de estos hallazgos, es evidente que la integración de la inteligencia artificial en el ámbito jurídico requiere un enfoque cuidadoso y colaborativo. Se necesita una estrecha colaboración entre profesionales del derecho, desarrolladores de tecnología, académicos y tomadores de decisiones para abordar los desafíos identificados.

Es fundamental invertir en el desarrollo de modelos de IA especializados y adaptados a las complejidades del derecho colombiano y, seguramente, de todos los contextos nacionales, utilizando conjuntos de datos robustos, representativos y detallados. Además, es esencial establecer marcos éticos y regulatorios sólidos que garanticen la transparencia, la explicabilidad y la protección de los derechos individuales.

A medida que la tecnología continúe avanzando, es crucial mantener un enfoque crítico y una evaluación constante de los modelos de IA. La revisión humana y la retroalimentación de expertos legales son indispensables para refinar y mejorar continuamente estas herramientas y las universidades y centros de investigación tienen mucho que decir.

En última instancia, la integración de la inteligencia artificial en el derecho no debe verse como una amenaza, sino como una oportunidad para potenciar y complementar las habilidades humanas. Con una implementación responsable y ética, estas tecnologías pueden contribuir a una administración de justicia más eficiente, precisa y accesible.

Finalmente, se presentan algunas sugerencias para futuras investigaciones y análisis más profundos. Además de los hallazgos y sugerencias presentados, se podrían realizar estudios adicionales utilizando el mismo material para identificar qué preguntas específicas fueron más desafiantes para cada modelo de IA, lo que podría arrojar luz sobre las áreas o conceptos legales que representan mayores dificultades para estos sistemas. Un análisis detallado de las respuestas proporcionadas por los modelos a cada pregunta, incluyendo aspectos cualitativos de estas, permitiría comprender mejor sus fortalezas y debilidades, y orientar el desarrollo de estrategias de entrenamiento más efectivas.

Adicionalmente, los datos obtenidos en esta investigación podrán servir para realizar una evaluación cualitativa del rendimiento de los diferentes modelos en las distintas áreas del derecho, lo que podría revelar patrones y tendencias conceptuales y argumentativas, ya que el material de la investigación cuenta con las razones ofrecidas por cada modelo en cada respuesta. Identificar en qué áreas específicas cada modelo tiene un mejor o peor desempeño en relación con los demás podría ayudar a determinar cuál de ellos es más adecuado para ciertas tareas o dominios legales, ya que son aspectos netamente argumentativos. Además, mediante un análisis minucioso de las respuestas correctas e incorrectas de los modelos, se podrían identificar posibles tendencias o patrones subyacentes, lo que conduciría a un mejor entendimiento de los sesgos o limitaciones inherentes a estos sistemas, y orientar el desarrollo de estrategias para mitigarlos.

Estos análisis adicionales ampliarían el entendimiento de las capacidades actuales de los modelos de IA en el campo del derecho y podrían proporcionar información valiosa para guiar el desarrollo futuro de estos sistemas y maximizar su impacto positivo en la práctica legal.

Agradecimientos

Los autores desean agradecer a Juana Carolina Villamil Sierra, jefe de Área de la Facultad de Derecho Privado y a Daniel Alfonso Barragán Ronderos, de la Facultad de Derecho de la Seccional Bogotá de la Universidad Libre, quienes fueron los encargados de verificar los aciertos de los exámenes.

Declaración de divulgación

Los autores declaran que no existe ningún potencial conflicto de interés relacionado con el artículo. No se emplearon herramientas de generación de contenido por Inteligencia Artificial para su elaboración. Esta publicación es resultado del proyecto de Investigación "Aplicación de Modelos de Lenguaje de Aprendizaje Profundo (LLM) en el Ámbito del Derecho Colombiano", con código interno 122202301, promovido por la Dirección Nacional de Investigaciones, de la Universidad Libre, asociado al grupo de investigación con código MinCiencias: COL0016505 y la investigación sobre "Violencias digitales en el ámbito universitario", de la Universidad del País Vasco/Euskal Herriko Unibertsitatea.

Financiamiento

Los autores no declaran fuente de financiamiento para la realización de este artículo.

Sobre los autores

Gabriel Andrés Arévalo-Robles. Doctor en Estudios Internacionales, Universidad del País Vasco/Euskal Herriko Unibertsitatea. Magíster Oficial en Estudios Internacionales (UPV/EHU). Magíster en Cooperación internacional Descentralizada (UPV/EHU). Abogado, Universidad Libre. Sociólogo, Universidad Nacional. Actualmente se desempeña como Director Nacional de Investigaciones, Universidad Libre.

Orcid: https://orcid.org/0000-0002-4389-5997
Contacto: gabriel.arevalo@unilibre.edu.co

Omaira Castellanos-Cortés. Doctoranda, Universidad del País Vasco/Euskal Herriko Unibertsitatea. Magíster en Derechos Fundamentales y Poderes Públicos y en Globalización y Desarrollo. Abogada, Universidad de Caldas.

Orcid: https://orcid.org/0000-0003-2124-7849
Contacto: ocastellanos003@ikasle.ehu.eus

Referencias

Abderahman, R., Karim, R., Appolloni, A., Treiblmaier, H., & Iranmanesh, M. (2024). Exploring the impact of ChatGPT on education: A web mining and machine learning approach. *The International Journal of Management Education* 22 (1): 100932. https://doi.org/10.1016/j.ijme.2024.100932.

Al Shloul, T., Mazhar, T., Abbas, Q., Iqbal, M., Yasin Ghadi, Y., Shahzad, T., Mallek, F., & Hamam, H. (2024). Role of activity-based learning and ChatGPT on students' performance in education. *Computers and Education: Artificial Intelligence* 6 (junio): 100219. https://doi.org/10.1016/j.caeai.2024.100219.

Billion Polak, P., Prusa, J.D. & Khoshgoftaar, T.M. Low-shot learning and class imbalance: a survey. *J Big Data* 11, 1 (2024). https://doi.org/10.1186/s40537-023-00851-z

Bommarito, M.J., & Katz, D.M. (2022). GPT Takes the Bar Exam. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4314839.

Castellanos-Cortés, O. E., & Arévalo-Robles, G. A. (2024). Explorando el impacto del metaverso en los derechos humanos: desafíos, amenazas y perspectivas. *Revista Republicana*, *36*, 37-53.

Fan, Lizhou, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, & Libby Hemphill (2023). A Bibliometric Review of Large Language Models Research from 2017 to 2023. arXiv. https://doi.org/10.48550/arXiv.2304.02020.

Faúndez-Ugalde, A., & Mellado-Silva, R. (2023). Use of Robotic Process Automation by Tax Administrations and Impact on Human Rights. *Revista Chilena de Derecho y Tecnología* 12 (julio). https://doi.org/10.5354/0719-2584.2023.65457.

Katz, D.M., M.J. Bommarito, S. Gao, & P. Arredondo (2024). GPT-4 Passes the Bar Exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 382 (2270). https://doi.org/10.1098/rsta.2023.0254.

Latif, E., & X. Zhai (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence* 6. https://doi.org/10.1016/j.caeai.2024.100210.

Matiz-Rojas, A. H., & Fernández-Camargo, J. A. (2023). Del uso de la inteligencia artificial como medio y método en los conflictos armados. *Revista Científica General José María Córdova*, *21*(42), 525–549. https://doi.org/10.21830/19006586.1151

Morocco-Clarke, A., Abubakar Sodangi, F., & Momodu, F. (2024). The Implications and Effects of ChatGPT on Academic Scholarship and Authorship: A Death Knell for Original Academic Publications? *Information & Communications Technology Law*, enero. https://www.tandfonline.com/doi/abs/10.1080/13600834.2023.2239623.

Pichai, S., & Hassabis, D. (2023). Introducing Gemini: Our Largest and Most Capable AI Model. Google. 6 de diciembre de 2023. https://blog.google/technology/ai/google-gemini-ai/.

Sigman, M., & Bilinkis, S. (2023). *Artificial: La nueva inteligencia y el contorno de lo humano*. Debate.

Vásquez, C., & Toro-Valencia, J. (2021). El derecho al control humano: Una respuesta jurídica a la inteligencia artificial. *Revista Chilena de Derecho y Tecnología* 10 (2): 211-28. https://doi.org/10.5354/0719-2584.2021.58745.

Vaswani, A., Shazeer, N., Parmar, J., Uszkoreit, L., Jones, A.N., Gomez, Ł., Kaiser. L., & Polosukhin, I. (2017). Attention Is All You Need. *Neural Information Processing Systems* 5999-6009.

Gabriel Andrés Arevalo-Robles, Omaira Castellanos-Cortés

¿Cuánto sabe la inteligencia artificial sobre derecho colombiano?
How much does artificial intelligence know about Colombian law?

Revista Científica General José María Córdova
vol. 22, núm. 48, p. 1152 - 1171, 2024
Escuela Militar de Cadetes "General José María Córdova",
ISSN: 1900-6586
ISSN-E: 2500-7645

DOI: https://doi.org/10.21830/19006586.1380