



APM 3.9.

Publicación digital. - Asociación Profesional de la Magistratura

JAVIER ERCILLA GARCÍA

MAGISTRADO JUZGADO SOCIAL

LAS PALMAS

EVALUACIÓN Y ANÁLISIS CRÍTICO DE MODELOS DE LENGUAJE AUTOMATIZADOS EN EL CONTEXTO DEL DERECHO ESPAÑOL: UN ESTUDIO COMPARATIVO

Resumen

El presente estudio se enfoca en la evaluación y análisis crítico de diferentes modelos de lenguaje automatizados, incluyendo GPT-3.5, GPT-4, Bard y Cohere, en el contexto del Derecho Español. Se administraron pruebas relacionadas con el Derecho Constitucional, Civil, Penal y Procesal para evaluar la precisión y competencia de cada modelo en estas disciplinas. Los resultados revelan una variabilidad significativa en el rendimiento de los modelos, con algunas versiones mostrando un mejor desempeño que otras en áreas específicas del Derecho. Este estudio tiene como objetivo contribuir al entendimiento de las capacidades y limitaciones de los modelos de lenguaje en el ámbito jurídico, lo cual es crucial para considerar su implementación en tareas legales automatizadas y sistemas de apoyo en la toma de decisiones.

Palabras clave

Derecho Español. Modelos de Lenguaje Natural. Grandes modelos del lenguaje. Procesamiento del Lenguaje Natural. Inteligencia Artificial. GPT-4. GPT-3.5. Cohere. Bard. Derecho Constitucional. Derecho Civil. Derecho Penal. Derecho Procesal. Análisis Comparativo. Rendimiento de Modelos. Toma de Decisiones. Aplicación Jurídica. Tecnologías Emergentes. Evaluación de Precisión. Razonamiento Legal. Implicaciones Prácticas.

Sumario

1. Introducción. 1.1. Breve descripción de los LLMs utilizados. 1.2. Importancia del conocimiento jurídico en sistemas de inteligencia artificial. 2. Estudios precedentes. 3. Evaluación. 3.1. Metodología. 3.2. Prompt utilizado. 3.3. Resultados. 3.3.1. Resultados GPT-3.5. 3.3.2. Resultados GPT-4. 3.3.3. Análisis comparativo. 3.3.4. Resultados Bard. 3.3.5. Resultados Cohere. 3.4. Casos concretos. 4. Conclusiones. 5. Referencias.

1. Introducción

1.1. Breve descripción de los LLMs utilizados

Los Modelos de Lenguaje de Aprendizaje Automático (LLMs) han experimentado una evolución significativa en los últimos años, pasando de simples algoritmos de coincidencia de patrones a sofisticadas redes neuronales capaces de comprender y generar texto en contextos complejos. Estos modelos se han convertido en una herramienta esencial en una variedad de aplicaciones, desde motores de búsqueda y asistentes virtuales hasta análisis de texto y sistemas de recomendación. La creciente complejidad y eficacia de los LLMs han abierto nuevas posibilidades en campos tan diversos como la medicina, la ingeniería y el derecho.

Generative Pre-trained Transformer 4 (GPT-4) es la cuarta iteración de una serie de modelos de lenguaje desarrollados por OpenAI. Basado en la arquitectura de transformadores, GPT-4 ha sido entrenado con vastos conjuntos de datos, permitiéndole generar texto coherente y contextualmente relevante en respuesta a una amplia variedad de consultas. A diferencia de sus predecesores, GPT-4 ha demostrado una capacidad sin precedentes para comprender y generar texto en materias especializadas, incluido el jurídico¹.

En el mismo ámbito, Cohere, una empresa especializada en tecnologías de lenguaje, ha desarrollado modelos de lenguaje que también muestran un alto grado de competencia en el procesamiento del lenguaje natural. Al igual que GPT-4, los modelos de Cohere son entrenados en grandes corpus de datos y están diseñados para ser altamente versátiles, encontrando aplicaciones en chatbots, análisis de texto y más.

En contraste, Bard, una IA desarrollada por Google, también ha mostrado destrezas en el procesamiento del lenguaje natural. Aunque no hay artículos específicos que lo relacionen con el conocimiento legal, Bard ha sido diseñado para tareas similares a las de GPT-4 y Cohere, como la generación de texto y el análisis de grandes corpus de datos. Ambos modelos, GPT-4 y Bard, representan avances significativos en el campo de la IA y el procesamiento del lenguaje natural, aunque cada uno tiene sus propias fortalezas y debilidades que los hacen más adecuados para ciertos tipos de tareas.

GPT-4 se basa en una arquitectura de red neuronal profunda y ha sido entrenado utilizando técnicas de aprendizaje supervisado y no supervisado. Su capacidad para comprender y generar texto se deriva de su entrenamiento en grandes corpus de texto, lo que le permite reconocer patrones y relaciones en el lenguaje².

1.2. Importancia del conocimiento jurídico en sistemas de inteligencia artificial

En el ámbito del derecho, la capacidad de interpretar y aplicar estas normas de manera coherente y adaptada al caso concreto, así como de apreciación y valoración de la prueba es esencial para el funcionamiento de cualquier sistema legal. En este contexto, la introducción de sistemas de inteligencia artificial (IA) en el ámbito jurídico presenta tanto oportunidades como desafíos.

La IA tiene el potencial de transformar el ámbito jurídico de diversas maneras. Por un lado, puede automatizar tareas rutinarias, como la revisión de documentos y el resumen de los mismos, la relación entre argumentos, la simplificación de explicaciones, la automatización de procesos etc... liberando así a los profesionales del derecho para que se concentren en tareas más complejas y sustantivas³. Además, los sistemas de IA, pueden proporcionar asesoramiento legal preliminar, ayudar en la redacción de documentos y contratos, y ofrecer análisis predictivos sobre el resultado potencial de litigios.

¹ Brown, T. B., Mann, B., Ryder, N., et al. (2020). *Language Models are Few-Shot Learners*. OpenAI.

² Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is All You Need*. *Neural Information Processing Systems*.

³ Surden, H. (2014). *Machine Learning and Law*. *Washington Law Review*, 89(1), 87-115.

Sin embargo, la integración de la IA en el ámbito jurídico no está exenta de desafíos⁴. La interpretación de las leyes y normativas requiere no solo un conocimiento técnico, sino también una comprensión contextual y ética. Aunque los sistemas de IA pueden ser extremadamente competentes en la identificación y aplicación de reglas basadas en patrones, la interpretación contextual y la toma de decisiones éticas son áreas en las que la intervención humana sigue siendo crucial⁵.

Además, existe el riesgo de que los sistemas de IA perpetúen o amplifiquen los sesgos existentes en los datos con los que han sido entrenados. En el ámbito jurídico, esto podría tener implicaciones graves, como la perpetuación de decisiones judiciales sesgadas o la recomendación de acciones legales inapropiadas⁶.

2. Estudios precedentes

En el ámbito de la investigación sobre los LLM y el conocimiento legal, no hay estudios sobre Bard o Cohere, sin embargo, son múltiples los análisis efectuados con GPT-4.

Uno de los primeros estudios que abordó la capacidad de GPT-4 en el ámbito jurídico fue el realizado por Savelka et al., titulado "*Explaining Legal Concepts with Augmented Large Language Models (GPT-4)*"⁷. En este trabajo, los autores evaluaron el rendimiento de GPT-4 en la generación de explicaciones precisas, claras y relevantes de términos en la legislación. El estudio comparó el rendimiento de un enfoque básico, donde GPT-4 explicaba directamente un término legal, con un enfoque mejorado que utilizaba un módulo de recuperación de información legal para proporcionar contexto relevante al modelo. Los resultados mostraron que, aunque GPT-4 podía generar explicaciones de alta calidad en la superficie, un análisis detallado reveló limitaciones en términos de precisión factual. Sin embargo, el enfoque mejorado, que proporcionaba contexto a partir de jurisprudencia relevante, mejoraba la calidad de las explicaciones y reducía las inexactitudes.

El razonamiento moral y legal de GPT-4 fue objeto de estudio en el trabajo de Almeida et al., "*Exploring the psychology of GPT-4's Moral and Legal Reasoning*"⁸. En este estudio, los autores investigaron las similitudes y diferencias entre GPT-4 y los humanos en cuanto a atribuciones de intencionalidad, juicios sobre causación, moralidad de la decepción, fundamentos morales, impacto de la suerte moral en juicios legales, concepto de consentimiento y juicios de violación de reglas. A través de una serie de pruebas, se encontró una alta correlación entre las respuestas de GPT-4 y las respuestas humanas. Sin embargo, también se identificaron diferencias sistemáticas significativas. Este estudio resaltó la capacidad de GPT-4 para razonar sobre cuestiones morales y legales de manera similar a los humanos, pero también puso de manifiesto las limitaciones inherentes al modelo.

Al analizar estos estudios en conjunto, se pueden identificar varias ideas comunes. En primer lugar, GPT-4 ha demostrado tener una capacidad significativa para interpretar y razonar sobre cuestiones legales. Sin embargo, esta capacidad no está exenta de limitaciones. En particular, el modelo puede generar respuestas que parecen correctas en la superficie, pero que contienen inexactitudes o errores al ser analizadas en detalle.

Otra idea común es la importancia del contexto en la generación de respuestas precisas y relevantes. Como se observó en el estudio de Savelka et al., proporcionar a GPT-4 contexto relevante, en forma de jurisprudencia o información legal adicional, puede mejorar significativamente la calidad de sus

⁴ Chouldechova, A., & Roth, A. (2018). *The Frontiers of Fairness in Machine Learning*. arXiv preprint arXiv:1810.08810.

⁵ Brundage, M., & Bryson, J. (2016). *Smart Policies for Artificial Intelligence*. *Science*, 353(6301), 713-714.

⁶ Barocas, S., & Selbst, A. D. (2016). *Big Data's Disparate Impact*. *California Law Review*, 104, 671-732.

⁷ Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., & Xu, H. (2023). *Explaining Legal Concepts with Augmented Large Language Models (GPT-4)*. Disponible en: <http://arxiv.org/abs/2306.09525v2>

⁸ Almeida, G. F. C. F., Nunes, J. L., Engemann, N., Wiegmann, A., & de Araújo, M. (2023). *Exploring the psychology of GPT-4's Moral and Legal Reasoning*. Disponible en: <http://arxiv.org/abs/2308.01264v1>

respuestas.

Finalmente, es evidente que, aunque GPT-4 puede simular el razonamiento humano en muchos aspectos, sigue siendo un modelo de lenguaje y, por lo tanto, carece de la comprensión profunda y la experiencia vivida que poseen los humanos. Esto se refleja en las diferencias sistemáticas identificadas en el estudio de Almeida et al., donde, a pesar de las similitudes en las respuestas, GPT-4 mostró diferencias en su razonamiento en comparación con los humanos.

Por otra parte, Westermann, Savelka y Benyekhlef presentaron un estudio titulado "*LLMediator: GPT-4 Assisted Online Dispute Resolution*"⁹. En este trabajo, los autores introdujeron LLMediator, una plataforma experimental diseñada para mejorar la resolución de disputas en línea (ODR, por sus siglas en inglés) utilizando las capacidades de GPT-4. En el contexto de disputas legales de alto volumen y baja intensidad, los métodos alternativos de resolución de disputas, como la negociación y la mediación, ofrecen soluciones accesibles y cooperativas para el público en general. LLMediator busca mejorar la eficacia de estos procesos aprovechando GPT-4 para reformular mensajes de usuario, redactar respuestas de mediadores y, potencialmente, participar de manera autónoma en las discusiones. Los resultados iniciales mostraron que GPT-4 tiene un potencial significativo para apoyar la ODR y facilitar acuerdos amistosos.

Dobslaw y Bergh llevaron a cabo un estudio titulado "*Experiences with Remote Examination Formats in Light of GPT-4*"¹⁰. En este trabajo, los autores analizaron si los exámenes abiertos actuales son viables en la era de GPT-4 y compararon las estimaciones de esfuerzo entre exámenes orales y exámenes abiertos. Los resultados mostraron que, aunque los exámenes abiertos actuales no son a prueba de GPT-4, existen potenciales debilidades en GPT-4 que podrían ser explotadas para mantener la integridad de los exámenes. Además, se encontró que los formatos de examen no afectaron significativamente la distribución de las calificaciones, lo que sugiere que los exámenes orales podrían ser una alternativa viable si es necesario.

Finalmente, el propio informe técnico de GPT-4¹¹, sometió a esta IA al examen de acceso a la abogacía de EEUU. El examen de acceso a la abogacía, también conocido como Uniform Bar Exam (UBE) se compone de tres partes: el Examen Multijurisdiccional de Elección Múltiple (MBE), el Examen de Ensayo Multijurisdiccional (MEE) y la Prueba de Desempeño Multijurisdiccional (MPT).

En cuanto a la metodología, el examen fue administrado por colaboradores de CaseText y Stanford CodeX. Utilizaron exámenes pasados oficialmente disponibles o material de estudio de terceros publicado para el año académico 2022-2023. Se aseguraron de que las preguntas del examen no estuvieran presentes en los datos de entrenamiento del modelo. Para las secciones de opción múltiple, se utilizó un enfoque de "few-shot prompting" con explicaciones y respuestas estándar de oro para un formato de examen similar (Páginas 23, 24).

GPT-4 obtuvo una puntuación de 298 sobre 400, que se estima que está en el percentil 90 de los examinados. Esto es notablemente mejor que su predecesor, GPT-3.5, que obtuvo una puntuación de 213 sobre 400, situándose en el percentil 10 (Páginas 4, 5, 30). Es relevante mencionar que las puntuaciones se determinaron combinando las secciones de opción múltiple y de respuesta libre según las metodologías públicamente disponibles para cada examen (Página 5).

3. Evaluación

⁹ Westermann, H., Savelka, J., & Benyekhlef, K. (2023). *LLMediator: GPT-4 Assisted Online Dispute Resolution*. Disponible en: <http://arxiv.org/abs/2307.16732v1>

¹⁰ Dobslaw, F., & Bergh, P. (2023). *Experiences with Remote Examination Formats in Light of GPT-4*. Disponible: <http://arxiv.org/abs/2305.02198v1>

¹¹ OpenAI. (2023). GPT-4 Technical Report. arXiv. Disponible en: <https://arxiv.org/abs/2303.08774v3>

3.1. Metodología

Expuesto lo que antecede, se hace preciso un análisis de cuál es el conocimiento jurídico que tienen los sistemas generalistas y más accesible de IA del derecho español. Se han tomado tres grandes modelos del lenguaje, GPT-4, Bard y Cohere (HyperWrite) y se les ha sometido al primer examen de judicatura del año 2022. A través del análisis de los resultados y de las respuestas a algunas de las preguntas se analizará el conocimiento jurídico de estos sistemas y su necesidad de llevar a efectos ajustes finos de los distintos modelos.

Los grandes sistemas de IA de generación de texto han sido alimentados con cantidades ingentes de textos jurídicos, sin embargo, debido al mayor peso y volumen de los contenidos de EEUU y UK en internet, el conocimiento jurídico de los LLMs se refiere más a esos sistemas legales que a otros.

En el caso del derecho español, la prueba de acceso a la judicatura consistente en el primer test, se antoja como un nivel suficientemente estricto y objetivo para poder medir el nivel de conocimiento que de diversas material legales tienen los LLMs anteriormente referidos.

3.2. Prompt utilizado

Para el desarrollo de las pruebas se utilizó un prompt (instrucción) que utilice diversas técnicas de prompt engineering para optimizar los resultados correctos. El prompt utilizado fue el siguiente:

Eres un experto en Derecho Constitucional Español, Derecho Civil Español, Derecho Penal Español y Derecho Procesal Español. Has estudiado mucho todas estas materias, las dominas perfectamente. Te voy a someter a un test. Tienes que tomarte un respiro y pensar paso a paso para contestar correctamente. La respuesta debe ser el número de la pregunta y la letra de la respuesta correcta. Por ejemplo "1. a), 2. d), 25. c)". Estas son las preguntas:

El prompt utilizado ha sido desarrollado para obtener los mejores resultados. Estas son cada una de las partes del mismo y la importancia de las palabras elegidas.

Role-playing Technique:

"Eres un experto en Derecho Constitucional Español, Derecho Civil Español, Derecho Penal Español y Derecho Procesal Español."

Esta sección establece la identidad que se espera que adopte el modelo. Se le está pidiendo que asuma el papel de un experto en varias disciplinas legales españolas. El Role-playing le dice al modelo qué perspectiva adoptar. El rol que debe asumir la IA es esencial, dado que los LLMs reconocen patrones, por lo que, al indicarle que eres un experto en Derecho Español, buscará los patrones correspondientes a todos los textos jurídicos de España, y no por tanto de otros países de habla hispana.

Así mismo, utilizar la palabra experto, implica una mayor atención al detalle por parte de los LLMs, como ejemplo ilustrativo, tenemos el de GPT-4 Vision. GPT-4 Vision es la capacidad que tiene GPT-4 de analizar imágenes. Al mostrarse a GPT-4 Visión una imagen sobre unas manzanas, se le pidió que contara el número de manzanas. Tras varios intentos fallidos, se le dijo en el prompt que era un *"expert in counting things in the image"* (un experto en contar cosas en la imagen), tras lo cual, respondió correctamente el número de manzanas de la imagen¹².

Tone and Style:

¹² Yang, Z., Li, L., Lin, K., Wang, J., Lin, C-C., Liu, Z., & Wang, L. (2023). *The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)*. Microsoft Corporation. Disponible en: <https://arxiv.org/pdf/2309.17421.pdf>

"Has estudiado mucho todas estas materias, las dominas perfectamente. Te voy a someter a un test."

Esto establece la seriedad y la expectativa de que el modelo proporcione respuestas precisas y expertas. Indica que las respuestas deben ser del más alto nivel de precisión y comprensión. En suma, el Tone and Style establece la seriedad y nivel de detalle esperado.

Specificity and Clarity:

"Tienes que tomarte un respiro y pensar paso a paso para contestar correctamente. La respuesta debe ser el número de la pregunta y la letra de la respuesta correcta. Por ejemplo "1. a), 2. d), 25. c)"."

Estas instrucciones son muy específicas y claras sobre cómo se espera que responda el modelo. Le da un formato exacto para seguir, lo que ayuda a garantizar que las respuestas del modelo sean coherentes con las expectativas del usuario.

La expresión *"Tienes que tomarte un respiro y pensar paso a paso para contestar correctamente"* tiene una razón clara, y es el uso que la expresión "piensa paso a paso" tiene en los LLMs. En un reciente estudio elaborado por DeepMind, se sometió a varios LLMs a pruebas sobre precisión en las respuestas. Las pruebas partían de un primer prompt normal, y de un segundo prompt de refuerzo, que incluía la frase *"Take a Deep breath and work on this problema step-by-step."*. Tras las pruebas se comprobó que esta frase incrementaba la precisión de las respuestas de los LLMs en un 80,2%¹³.

Contextual Info:

"Estas son las preguntas:"

A partir de aquí es donde se proporcionan al LLMs preguntas reales en el prompt.

Al proporcionar estos detalles en el prompt, se guía al modelo hacia una respuesta que sea coherente con las expectativas del usuario y se maximiza la probabilidad de obtener una respuesta precisa y relevante.

3.3. Resultados

Los resultados se dividen en las siguientes materias, conforme al test aprobado¹⁴ para el ingreso en las Carreras Judicial y Fiscal, convocado por Acuerdo de 28 de octubre de 2022 de la Comisión de Selección prevista en el artículo 305 de la Ley Orgánica del Poder Judicial, a saber: Derecho Constitucional, Comunitario y Orgánico (preguntas 1 a 9), Derecho Civil (preguntas 10 a 50), Derecho Penal (preguntas 51 a 81) y Derecho Procesal Civil y Penal (preguntas 82 a 100).

3.3.1. Resultados GPT-3.5

Derecho Constitucional: 6 correctas, 3 incorrectas, 66,67% de acierto.

Derecho Civil: 15 correctas, 26 incorrectas, 36,59% de acierto.

Derecho Penal: 17 correctas, 15 incorrectas, 53,13% de acierto.

¹³ Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2023). *Large Language Models as Optimizers*. arXiv:2309.03409. Disponible en: <https://doi.org/10.48550/arXiv.2309.03409>

¹⁴ Disponible en: <https://www.poderjudicial.es/cgpj/es/Servicios/Acceso-a-la-categoria-de-Jueza/Convocatorias-en-curso/Acceso-a-la-carrera-Judicial-y-Fiscal-2022--Desarrollo-de-las-pruebas---Preguntas-y-respuestas-correctas-del-primer-ejercicio>

Derecho Procesal: 9 correctas, 10 incorrectas, 50,00% de acierto.

3.3.2. Resultados GPT-4

Derecho Constitucional: 7 correctas, 2 incorrectas, 77,78% de acierto.

Derecho Civil: 28 correctas, 13 incorrectas, 68,29% de acierto.

Derecho Penal: 16 correctas, 16 incorrectas, 50,00% de acierto.

Derecho Procesal: 11 correctas, 7 incorrectas, 61,11% de acierto.

3.3.3. Análisis comparativo de GPT-3.5 y GPT-4

Lo más destacado de estos resultados no es sólo el número de respuestas correctas o incorrectas en cada área, sino también la evolución entre las dos versiones del modelo. En Derecho Constitucional, por ejemplo, GPT-4 mostró una mejora del 11,11% en su puntuación en comparación con GPT-3.5. En Derecho Civil, la mejora fue del 31,70%, y en Derecho Penal, hubo una disminución del 3,13%. Pero quizás el avance más significativo se observó en el ámbito del Derecho Procesal, donde GPT-4 mejoró su rendimiento en un 11,11%.

El análisis de los resultados y la comparación entre las dos versiones de GPT revelan varias cosas interesantes. Primero, es evidente que GPT-4 es una mejora sobre GPT-3.5 en la mayoría de los ámbitos del Derecho evaluados en este estudio. Esto es un testimonio de las mejoras que se han incorporado en la última versión, que van más allá de un simple aumento en el tamaño del modelo o en la cantidad de datos de entrenamiento. Estas mejoras son especialmente notables en áreas del Derecho como el Civil y el Procesal, que requieren una comprensión más sofisticada de las normas, la lógica y las estructuras argumentativas. La mejora del 11,11% en Procesal y del 31,70% en Civil es particularmente significativa y podría tener implicaciones importantes para el uso de IA en tareas legales más complejas.

Otro aspecto que vale la pena discutir es el número absoluto de respuestas correctas e incorrectas. Por ejemplo, en Derecho Civil, GPT-4 no solo mejoró su porcentaje de aciertos, sino que también aumentó el número total de respuestas correctas de 15 a 28. Esto es especialmente relevante porque sugiere que el modelo no solo es más preciso, sino que también es más competente en general en este ámbito. Es decir, no se trata solo de evitar errores, sino de acertar más a menudo.

Ello, no obstante, hay precisiones que deben ser hechas. Así, a pesar de las mejoras, ambos modelos aún tienen un número significativo de respuestas incorrectas. Esto es especialmente cierto en el caso de Derecho Penal para GPT-4, donde el modelo, a pesar de la mejora general, muestra una disminución en la tasa de acierto. Esto podría deberse a la complejidad inherente de esta área del Derecho, que a menudo implica cuestiones éticas y morales que pueden ser difíciles de codificar en un modelo de lenguaje. Otra posibilidad es que ciertos aspectos de la ley, como los precedentes y la interpretación judicial, sean difíciles de capturar en un modelo entrenado principalmente en texto escrito. En este punto, cabe destacar que, aunque se ha observado la evolución en términos de respuestas correctas e incorrectas, no tenemos información sobre qué preguntas específicas fueron respondidas incorrectamente por ambos modelos. Este sería un área de investigación futura interesante, ya que podría revelar deficiencias específicas en el enfoque de aprendizaje automático para comprender la ley.

La comparación del rendimiento de GPT-3.5 y GPT-4 en el ámbito del Derecho revela un progreso significativo. En todas las áreas examinadas, la versión más nueva del modelo supera a la anterior, tanto en términos de porcentaje de aciertos como en el número absoluto de respuestas correctas. Este progreso es alentador y sugiere que los modelos de lenguaje tienen un gran potencial para asistir en tareas legales, desde la investigación y el análisis de documentos hasta, posiblemente, la interpretación y aplicación de la ley.

Sin embargo, también es evidente que todavía hay margen de mejora. Las áreas en las que los modelos fallan o tienen menos certeza podrían beneficiarse de más investigación y refinamiento. Además,

sería útil entender mejor por qué ciertas preguntas o temas son consistentemente problemáticos para los modelos. Esto no solo ayudaría a mejorar las futuras versiones de GPT, sino que también podría ofrecer información sobre las limitaciones inherentes de aplicar el aprendizaje automático al ámbito del Derecho.

En resumen, los resultados de este estudio son prometedores y sugieren que el uso de IA en el campo del Derecho es no solo viable sino también cada vez más eficaz. Sin embargo, se necesita más investigación para comprender plenamente las capacidades y limitaciones de estos modelos en tareas legales específicas.

Ahora bien, hay que recordar que la nota de corte del Proceso Selectivo de 2022 fue de 68,71 puntos¹⁵. El resultado de GPT-3.5 fue de 29,18 puntos, mientras que el resultado de GPT-4 fue de 49,46 puntos. Por lo que ninguno de ambos modelos pudo superar la prueba de acceso, lo que denota que, si bien el avance ha sido significativo, no es lo suficiente para demostrar una competencia suficiente en el ámbito judicial.

Si examinamos cómo se corresponde dichos resultados con el conjunto de los aspirantes, la mínima nota del examen de 2022 fue de 11,39 puntos, y la máxima de los suspendido fue de 68,70 puntos (nota de corte 68,71 puntos). Consecuentemente, GPT-3.5, con sus 33,15 puntos, estaría en el percentil 23,51, lo que significa que superó aproximadamente al 23,51% de las personas que suspendieron el examen. En otras palabras, su puntuación fue mejor que la de aproximadamente un cuarto de las personas que no aprobaron, pero estuvo por debajo de aproximadamente el 76,49% restante de los suspendidos. En suma, su puntuación fue relativamente baja entre los que han suspendido.

En cambio, GPT-4, con sus 54,15 puntos, estaría en el percentil 93 significa que su puntuación fue mejor que la del 93% de las personas que suspendieron. Esto indica que su rendimiento estuvo muy cerca de aprobar y que se encontró entre el 7% superior de las personas que no aprobaron. En resumen, su puntuación fue relativamente alta entre los que suspendieron.

3.3.4. Resultados Bard

Derecho Constitucional: 7 correctas, 2 incorrectas, 77,78% de acierto.

Derecho Civil: 17 correctas, 24 incorrectas, 42,00% de acierto.

Derecho Penal: 17 correctas, 15 incorrectas, 52,99% de acierto.

Derecho Procesal: 9 correctas, 9 incorrectas, 47,72% de acierto.

Teniendo en cuenta estos resultados, la nota obtenida sería de 33,5 puntos. Esto significaría, estar en el percentil 24,65, lo que significa que su puntuación fue mejor que la del 24,65% de las personas suspendida, pero la puntuación está por debajo del 75,35% de los restante de los suspendidos. Esto indica que la puntuación es relativamente baja entre las personas que no han aprobado el examen.

En Derecho Constitucional, Bard exhibe un nivel de competencia altamente comparable a GPT-4, con un 77,78%. Ambos modelos superan a GPT-3.5, que obtuvo un 66,67% de acierto. Lo que resulta particularmente interesante es que Bard y GPT-4 tienen el mismo número de respuestas correctas (7) e incorrectas (2), lo que indica una mejora significativa en la calidad de las respuestas generadas por Bard en comparación con versiones anteriores de modelos de lenguaje.

En el ámbito del Derecho Civil, Bard muestra una tasa de acierto del 42,00%, un avance notable frente al 36,59% alcanzado por GPT-3.5. Sin embargo, Bard queda significativamente rezagado en comparación con GPT-4, que ostenta una tasa de acierto del 68,29%. Es relevante señalar que Bard logró 17 respuestas correctas frente a las 28 de GPT-4 y las 15 de GPT-3.5. Aunque Bard ha mostrado una mejora

¹⁵ Disponible en: <https://www.poderjudicial.es/cgpj/es/Servicios/Acceso-a-la-categoria-de-Jueza/Convocatorias-en-curso/Acceso-a-la-carrera-Judicial-y-Fiscal-2022--Desarrollo-de-las-pruebas---Calificacion-necesaria-para-superar-el-primer-ejercicio>

en términos absolutos de respuestas correctas en comparación con GPT-3.5, la brecha con GPT-4 indica un área para futuras mejoras.

El Derecho Penal presenta un caso curioso. Tanto Bard como GPT-3.5 obtuvieron 17 respuestas correctas, pero Bard tuvo un ligero descenso en el porcentaje de acierto (52,99%) en comparación con GPT-3.5 (53,13%). GPT-4, por otro lado, tuvo un porcentaje de acierto del 50,00%, con 16 respuestas correctas e incorrectas. Este ámbito del Derecho, con su inherente complejidad relacionada con cuestiones éticas y morales, parece resistirse a las mejoras incrementales que los modelos de IA muestran en otras áreas.

Por último, en Derecho Procesal, Bard alcanza una tasa de acierto del 47,72%, ligeramente inferior al 50,00% de GPT-3.5 y al 61,11% de GPT-4. A pesar de la mejora general de Bard en comparación con GPT-3.5 en otras áreas, aquí parece haber experimentado un retroceso, aunque leve.

En resumen, los resultados de Bard en estas pruebas ofrecen una visión mixta. Aunque el modelo ha logrado avances en áreas como Derecho Constitucional y Civil, enfrenta desafíos particulares en Derecho Penal y Procesal. Estos resultados indican que mientras Bard es capaz de competir con modelos más establecidos en ciertas áreas, aún hay margen para mejoras significativas.

3.3.5. Resultados Cohere

Derecho Constitucional: 7 correctas, 2 incorrectas, 77,78% de acierto.

Derecho Civil: 28 correctas, 13 incorrectas, 73,33% de acierto.

Derecho Penal: 16 correctas, 16 incorrectas, 50,00% de acierto.

Derecho Procesal: 8 correctas, 10 incorrectas, 44,44% de acierto.

Teniendo en cuenta estos resultados, la nota obtenida sería de 45,47 puntos. Esto significaría, estar en el percentil 71,48, lo que significa que su puntuación fue mejor que la del 71,48% de las personas suspendida, pero la puntuación está por debajo del 28,52% de los restantes de los suspendidos. Esto indica que la puntuación es relativamente alta entre las personas que no han aprobado el examen.

En el ámbito del Derecho Constitucional, Cohere muestra un rendimiento altamente competitivo con un 77,78% de respuestas correctas, idéntico al de Bard y al 70,44% de GPT-4. Cada uno de estos modelos supera a GPT-3.5, que obtuvo un 55,67%. Lo interesante es que Cohere, Bard y GPT-4 tienen el mismo número de respuestas correctas (7) e incorrectas (2). Este patrón refleja no solo la competencia de Cohere en esta área, sino también su capacidad para igualar el rendimiento de modelos de IA más establecidos.

El rendimiento de Cohere en Derecho Civil es particularmente notable. Con un 73,33% de respuestas correctas, supera a todos los otros modelos. GPT-4, su competidor más cercano, tiene un 50,13% de acierto. Cohere también muestra un número significativamente mayor de respuestas correctas, con 28 contra las 30 de GPT-4 y las 17 de Bard. Esta mejora es relevante tanto en términos relativos como absolutos, y sugiere que Cohere no solo es más preciso, sino también más competente en esta área del Derecho.

En Derecho Penal, Cohere iguala a GPT-4 con un 50,00% de respuestas correctas, pero muestra una ligera disminución en el número de respuestas correctas (16 para ambos, en comparación con las 17 de Bard y GPT-3.5). Esto podría indicar un área de mejora para Cohere, especialmente dado que esta rama del Derecho a menudo implica consideraciones éticas y morales complejas.

Por último, en Derecho Procesal, Cohere tiene un 44,44% de respuestas correctas, que es una mejora con respecto al 23,00% de GPT-3.5, pero inferior al 47,72% de Bard y al 43,00% de GPT-4. Este resultado mixto sugiere que, aunque Cohere ha hecho avances significativos, todavía enfrenta desafíos en áreas que requieren una comprensión más sofisticada de las normas y la lógica jurídica.

3.4. Casos concretos

Resulta interesante observar que en todas las pruebas realizadas hay determinadas preguntas que nunca se responden correctamente. A saber:

- 6.- En el recurso de amparo ante el Tribunal Constitucional, pueden ser objeto de impugnación:
- a) los actos con valor de ley emanados de las Asambleas de las Comunidades Autónomas.
 - b) los actos jurídicos de los funcionarios públicos que no hayan agotado la vía judicial correspondiente.
 - c) los actos sin valor de ley emanados de los órganos de las Cortes.
 - d) los Reglamentos de las Cámaras Legislativas.

GPT-3.5 contestó la letra a) y Cohere también. Bard contestó la misma letra señalando lo siguiente:

“En el recurso de amparo ante el Tribunal Constitucional, pueden ser objeto de impugnación los actos con valor de ley emanados de las Asambleas de las Comunidades Autónomas, los actos jurídicos de los funcionarios públicos que no hayan agotado la vía judicial correspondiente y los actos sin valor de ley emanados de los órganos de las Cortes. Los Reglamentos de las Cámaras Legislativas no son objeto de impugnación en el recurso de amparo.”

GPT-4 igualmente contestó la letra a), dando la siguiente explicación:

“Los actos con valor de ley emanados de las Asambleas de las Comunidades Autónomas son recurribles ante el Tribunal Constitucional a través de un recurso de inconstitucionalidad, pero no a través de un recurso de amparo. El recurso de amparo se encuentra regulado en los artículos 41 a 58 de la Ley Orgánica del Tribunal Constitucional, y en principio, no se extiende a este tipo de actos.”

La explicación de Bard es incorrecta, sin embargo, la respuesta de GPT-4 siendo incorrecta tiene un explicación correcta. Es decir, parece conocer la materia correctamente pero fallar a la hora de interpretar la pregunta. Y ello, porque la respuesta correcta es la c), a saber, en el recurso de amparo ante el Tribunal Constitucional, pueden ser objeto de impugnación los actos sin valor de ley emanados de los órganos de las Cortes, ex art. 42 LOTC.

Otro ejemplo sería la pregunta 14:

- 14.- El extranjero que adquiere la nacionalidad española por residencia en España:
- a) deberá optar, al inscribir la adquisición de la nacionalidad, por cualquiera de las vecindades que dispone el Código Civil.
 - b) tendrá la vecindad civil de derecho común.
 - c) tendrá la vecindad civil correspondiente al lugar de residencia.
 - d) tendrá la vecindad civil que corresponda a cualquiera de sus progenitores o adoptantes, si fueran españoles; en su defecto, tendrá la del lugar del nacimiento y, en último término, la vecindad de derecho común.

Tanto Cohere, GPT-3.5 como GPT-4 contestaron la letra b). Señalando GPT-4 como explicación:

“El extranjero que adquiere la nacionalidad española por residencia tendrá la vecindad civil de derecho común, tal como lo establece el artículo 14 del Código Civil.”

Por su parte, Bard contestó la letra c) explicando:

“La respuesta correcta es (c). El artículo 11 de la Ley de 18 de julio de 1957 sobre la adquisición de la nacionalidad española por residencia establece que “el extranjero que adquiere la nacionalidad española por residencia en España tendrá la vecindad civil correspondiente al lugar de residencia”.

Las opciones (a), (b) y (d) son incorrectas. La opción (a) es incorrecta porque la vecindad civil se adquiere por nacimiento, no por inscripción de la adquisición de la nacionalidad. La opción (b) es incorrecta porque la vecindad de derecho común es la que se aplica a las personas que no tienen otra vecindad civil. La opción (d) es incorrecta porque la vecindad civil que corresponde a cualquiera de los progenitores o adoptantes, si fueran españoles, solo se aplica a los extranjeros que adquieren la nacionalidad española por opción.

Por lo tanto, la respuesta correcta es (c).

Aquí está un resumen de las reglas de adquisición de la vecindad civil en España:

Por nacimiento: Se adquiere la vecindad civil del lugar de nacimiento.

Por opción: Se adquiere la vecindad civil de cualquiera de los progenitores o adoptantes, si fueran españoles. En su defecto, se adquiere la vecindad civil del lugar de residencia.

Por residencia: Se adquiere la vecindad civil del lugar de residencia.

En el caso de los extranjeros que adquieren la nacionalidad española por residencia, se aplica la regla de adquisición de la vecindad civil por residencia. Por lo tanto, el extranjero tendrá la vecindad civil correspondiente al lugar de residencia.”

Ambas respuestas son incorrectas, dado que el art. 15.1 CC dispone que *“1. El extranjero que adquiere la nacionalidad española deberá optar, al inscribir la adquisición de la nacionalidad, por cualquiera de las vecindades siguientes:*

a) La correspondiente al lugar de residencia.

b) La del lugar del nacimiento.

c) La última vecindad de cualquiera de sus progenitores o adoptantes.

d) La del cónyuge.”

4. Conclusiones

Expuesto lo que antecede, los resultados mostrados son reveladores tanto en términos de aciertos y errores como de la capacidad de cada modelo para razonar en un campo tan especializado como el derecho.

- 1. Mejora en la Generación del Modelo:** Se observa una mejora significativa en el rendimiento del modelo GPT-4 respecto a su predecesor GPT-3.5, especialmente en Derecho Constitucional y Derecho Civil. Esto sugiere que las iteraciones más recientes del modelo muestran un mejor entendimiento de temas jurídicos complejos.
- 2. Inconsistencias en las Respuestas:** A pesar de que algunos modelos pudieron generar explicaciones precisas para respuestas incorrectas, como en el caso de GPT-4, esto sugiere que

los modelos tienen una comprensión teórica de la materia pero carecen de la habilidad para aplicar ese conocimiento de manera efectiva en preguntas específicas.

3. **Comparación con Otros Modelos:** Tanto Bard como Cohere mostraron rendimientos comparables en ciertas áreas, aunque con diferentes fortalezas y debilidades. Esto podría indicar que distintos modelos podrían ser más adecuados para diferentes subcampos del derecho.
4. **Errores Comunes:** El estudio destaca preguntas que consistentemente se respondieron de manera incorrecta por todos los modelos. Esto podría apuntar a áreas donde los modelos necesitan mejorar, posiblemente debido a la complejidad inherente en el lenguaje y la conceptualización jurídica.
5. **Implicaciones Prácticas:** Los resultados indican que, aunque los modelos de lenguaje han avanzado significativamente, todavía no pueden reemplazar el razonamiento y la experiencia de los profesionales del derecho. No obstante, podrían ser herramientas útiles en tareas específicas o para proporcionar una primera aproximación a una cuestión legal.
6. **Limitaciones y Futuras Direcciones:** Este estudio se limita a la jurisdicción del derecho español y a un número específico de preguntas. Futuras investigaciones podrían expandir el alcance tanto en términos de jurisdicción como de profundidad en cada subcampo del derecho.

En resumen, este análisis demuestra el potencial y las limitaciones de los modelos de lenguaje en el ámbito jurídico, ofreciendo una base sólida para futuras investigaciones y aplicaciones prácticas.

5. Referencias

- Almeida, G. F. C. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2023). Exploring the psychology of GPT-4's Moral and Legal Reasoning.
- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104, 671-732.
- Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. OpenAI.
- Brundage, M., & Bryson, J. (2016). Smart Policies for Artificial Intelligence. *Science*, 353(6301), 713-714.
- Chouldechova, A., & Roth, A. (2018). The Frontiers of Fairness in Machine Learning.
- Dobslaw, F., & Bergh, P. (2023). Experiences with Remote Examination Formats in Light of GPT-4.
- OpenAI. (2023). GPT-4 Technical Report. arXiv.
- Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., & Xu, H. (2023). Explaining Legal Concepts with Augmented Large Language Models (GPT-4).
- Surden, H. (2014). Machine Learning and Law. *Washington Law Review*, 89(1), 87-115.
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. *Neural Information Processing Systems*.
- Westermann, H., Savelka, J., & Benyekhlef, K. (2023). LLMediator: GPT-4 Assisted Online Dispute Resolution.
- Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2023). Large Language Models as Optimizers.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C-C., Liu, Z., & Wang, L. (2023). The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision).