XX

JORNADAS DE COMUNICACIONES CIENTÍFICAS DE LA FACULTAD DE DERECHO Y CIENCIAS SOCIALES Y POLÍTICAS - UNNE

2024

2 décadas de ciencia compartida: raíces hacia nuevos horizontes

FACULTAD DE DERECHO
Y CIENCIAS SOCIALES Y POLÍTICAS

XX Jornadas de Comunicaciones Científicas de la Facultad de Derecho y Ciencias Sociales y Políticas

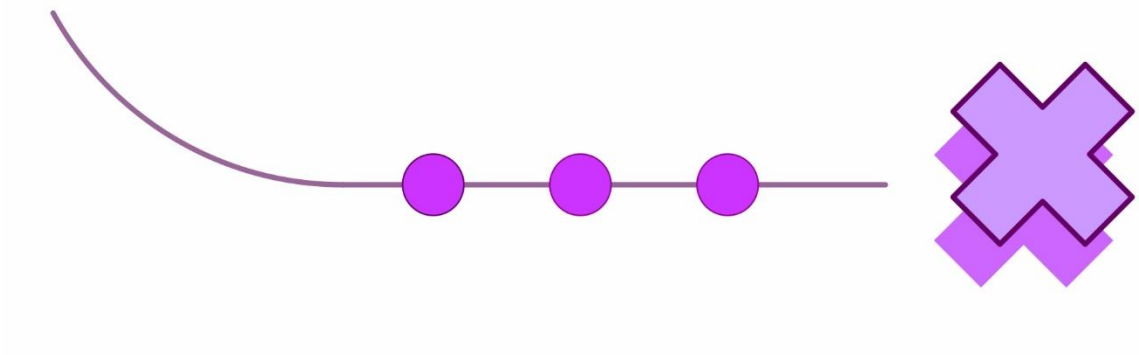
UNNE

2024

Dos décadas de ciencia compartida:

raíces hacia nuevos horizontes

Corrientes - Argentina



Dirección General
Dr. Mario R. Villegas

Dirección Editorial
Dra. Lorena Gallardo

Coordinación editorial y compilación
Esp. Martín M. Chalup
Abg. M. Benjamin Gamarra

Asistentes – Colaboradores
Lic. Agustina M. Bergadá

Edición
Secretaría de Ciencia y Transferencia
Facultad de Derecho y Ciencias Sociales y Políticas
Universidad Nacional del Nordeste
Salta 459 • C.P. 3400
Corrientes • Argentina

Villegas, Mario R.
XX Jornadas de Comunicaciones Científicas de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE / Mario R. Villegas ; Lorena Gallardo ; Martín Miguel Chalup ; compilación de Martín Miguel Chalup ; Mauro Benjamín Gamarra ; coordinación general de Lorena Gallardo ; director Mario R. Villegas ; Lorena Gallardo ; prólogo de Claudia Diaz. - 1a edición especial - Corrientes : Universidad Nacional del Nordeste. Facultad de Derecho y Ciencias Sociales y Políticas, 2024.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-631-6623-05-8

1. Legislación. 2. Normas. 3. Regulación. I. Chalup, Martín Miguel, comp. II. Gamarra, Mauro Benjamín, comp. III. Gallardo, Lorena, coord. IV. Villegas, Mario R., dir. V. Gallardo, Lorena, dir. VI. Diaz, Claudia, prolog. VII. Título.
CDD 340

DIRECTRICES PARA PREVENIR LA INTRODUCCIÓN DE ALUCINACIONES EN DOCUMENTOS JURÍDICOS ELABORADOS CON GRANDES MODELOS DE LENGUAJE

Navarro, Dario S.

darionavarro85@gmail.com

RESUMEN

En este artículo se formulan una serie de recomendaciones prácticas para prevenir la generación de alucinaciones por parte de Grandes Modelos de Lenguaje (LLM) en la elaboración de documentos jurídicos. Para su desarrollo se toma como referencia a la documentación publicada por OpenAI como guía de uso de ChatGPT, por su grado de desarrollo y por resultar extensibles al uso de otros LLM. Estas recomendaciones se enmarcan dentro del paradigma de supervisión humana para prevenir alucinaciones en la elaboración de documentos jurídicos.

PALABRAS CLAVE

Tecnología, alucinaciones, control humano

INTRODUCCIÓN

Los Grandes Modelos de Lenguaje o *Large Language Models* (LLM) son un tipo de inteligencia artificial (IA) que permiten el funcionamiento de aplicaciones como *ChatGPT*, *Gemini o Claude*. Gracias a una interfaz intuitiva y un sistema de acceso gratuito, estas aplicaciones democratizaron el uso de la IA. Los actuales LLM tienen la capacidad de realizar tareas generales vinculadas al lenguaje, y pueden responder preguntas abiertas, analizar textos o redactar documentos completos (OpenAI, 2024a), lo que abre un abanico de nuevas posibilidades en el campo jurídico, así como múltiples desafíos y riesgos (Corvalán & Caparrós, 2023).

Entre los riesgos asociados al uso de LLM, se encuentra la posibilidad de que generen outputs (respuestas) verosímiles con información incorrecta o engañosa. Este fenómeno es conocido como "alucinaciones" por la literatura especializada (OpenAI, 2024b) y podría constituir un motivo para limitar o excluir la utilización de la IA en los procesos judiciales.

En base a esta problemática, este artículo tiene por objeto proponer una serie de directrices para evitar o mitigar los riesgos derivados de alucinaciones, con base en el control humano.

MÉTODOS

Diseño de la investigación: El estudio adoptó un enfoque cualitativo y exploratorio, centrado en la revisión y análisis de literatura especializada y documentos normativos relevantes.

Métodos empleados: 1) Revisión sistemática de literatura, 2) Análisis documental. 3) Síntesis interpretativa.

Técnicas de recolección de datos: 1) Búsqueda y selección de fuentes bibliográficas relevantes. 2) Recopilación de documentos normativos y regulatorios. 3) Extracción y categorización de información pertinente.

Fuentes de datos: 1) Literatura jurídica especializada. 2) Literatura técnica relacionada con IA y LLM. 3) Regulaciones internacionales sobre el uso ético de la IA. 4) Documentación técnica y recomendaciones de uso desarrolladas por OpenAI.

Proceso de análisis: 1) Revisión crítica de la literatura y documentación recopilada. 2) Identificación de principios y prácticas recomendadas para el uso de IA en contextos jurídicos. 3) Adaptación y síntesis de la información para proponer directrices específicas al campo del Derecho. 4) Elaboración de recomendaciones para el uso responsable de IA generativa en la creación de documentos jurídicos.

RESULTADOS y DISCUSIÓN

La expresión Human in the loop (*HITL*), en español "humano en el bucle" describe el proceso en el que un sistema informático necesita intervención humana en las etapas de entrenamiento y prueba, para proporcionar mejores resultados (Bisen, 2020). ·

Las presentes recomendaciones se basan en este paradigma, en el que los humanos tienen un papel determinante en el proceso de operatoria con una IA.

1) Asumir un Control Directo en Todas las Etapas de Uso de un LLM.

Para un uso seguro de la IA, la estrategia más eficiente es la de asumir el control directo, definiendo para qué, cómo se va usar, y finalmente, controlar los resultados antes de aplicarlos. Por ejemplo, un juez primero debería analizar si por razones de eficiencia es conveniente valerse de un LLM para hacer más clara la redacción de una sentencia. Luego, debería decidir si utilizará la IA para redactar todo el documento o un pasaje en especial. Una vez obtenido un output, debería controlar la veracidad y fidelidad de la respuesta emitida, para finalmente decidir su inclusión al acto procesal.

Para esto, es necesario conocer el funcionamiento de un LLM, en especial su funcionalidad y limitaciones. En este sentido, el art. 14, ap. 4 de la Ley de Inteligencia Artificial de la Unión Europea promueve la capacitación y comprensión adecuadas de las capacidades y limitaciones del sistema de IA para garantizar una supervisión efectiva.

2) Utilizar los Outputs para Ámbitos en los que se Tenga Suficiente Expertise.

Antes de aplicar un output para la elaboración de un documento, es preciso evaluar su idoneidad en relación al uso concreto que se pretende darle (OpenAI, 2024c). En el ámbito jurídico, el operador debería contar con suficientes conocimientos jurídicos sobre la cuestión abordada por el LLM, para formular correctamente el prompt (instrucción), y luego evaluar la veracidad y utilidad de las respuestas del modelo. Esto es relevante porque dada una formulación de una pregunta, *ChatGPT* puede afirmar que no conoce la respuesta, pero dada una leve reformulación, puede responder correctamente (OpenAI, 2024d).

3) Evitar Fundar una Decisión Exclusivamente en un Output de un LLM.

En los términos de uso de *Chat GPT*, OpenAI efectúa un deslinde de responsabilidad al manifestar que no se debería utilizar un output para ningún fin que pueda tener un impacto significativo o consecuencias legales sobre una persona (2024e). Esto no implica descartar la aplicación de la IA en el ámbito jurídico, sino entender que el output de un LLM no debería fundar por sí mismo una decisión vinculada al derecho u obligación de una persona.

4) Dividir Tareas Complejas en otras más Simples.

Para mejorar la eficiencia del output de un LLM, las tareas complejas pueden redefinirse como tareas más simples en las que las salidas de las tareas anteriores se utilicen para construir las entradas de las tareas posteriores (2024f).

Verbigracia, si en un prompt se requiere a un LLM que redacte una demanda completa por daños, de este modo se incrementa la posibilidad de errores o alucinaciones. Por ello, sería recomendable pedirle que elabore determinados argumentos o secciones, y eventualmente, unificar los segmentos.

5) Brindar Contexto a los Prompts.

Los LLM no están entrenados de forma especializada para realizar tareas jurídicas. Para minimizar la posibilidad de obtener respuestas incorrectas, resulta conveniente dotar a los prompts de un contexto claro y preciso. De lo contrario, el LLM puede interpretar la petición de diferentes maneras y, como resultado, generar una contestación que no sea certera (García Sánchez, 2023). Esto puede incluir detalles sobre el tema, la audiencia, el propósito de la interacción y definir el tono y el estilo esperados, así como las restricciones que se deben adoptar.

6) Evitar los LLM Como Única Fuente de Investigación.

OpenAI (2024f) advierte que sus LLM pueden inventar cosas como citas o referencias, por lo que no lo recomienda como única fuente de investigación. Un uso responsable de los LLM en el proceso de investigación jurídica implica verificar los resultados proporcionados en sitios webs oficiales y/o bibliografía especializada.

7) Verificar la autoconsistencia de los outputs. De acuerdo a un estudio de la Universidad de Cambridge (Manakul, 2023), cuando un LLM ha sido entrenado en un concepto dado, las respuestas son similares y contienen hechos consistentes. Sin embargo, para los hechos alucinados, las respuestas divergen y pueden contradecirse entre sí. Si bien a mayor número de ejemplos se obtienen mejores rendimientos, se afirma que a partir de 4 outputs se puede medir la consistencia de la entre los diferentes outputs y apreciar para indicios de alucinación.

Conclusión: Debido a la falibilidad de los LLM, resulta imperativo que los operadores jurídicos que opten por valerse de esta tecnología, asuman un control directo y riguroso en todas las etapas de su uso. Para ello, resulta imprescindible la alfabetización digital, lo que implica conocer las capacidades y límites de los LLM, así como desarrollar la habilidad de redactar prompts eficaces.

REFERENCIAS BIBLIOGRÁFICAS

Bisen, V. (20 de mayo de 2020). *What is Human in the Loop Machine Learning: Why & How Used in AI?* Medium. Recuperado el 1 de abril de 2024 de: https://medium.com/vsinghbisen/what-is-human-in-the-loop-machine-learning-why-how-used-in-ai-60c7b44eb2c0.

Corvalán y Caparrós (2023). *Guía de directrices para el uso de ChatGPT e IA generativa de texto en la Justicia*. Ed La Ley. https://ialab.com.ar/wp-content/uploads/2023/11/Guia-de-directrices-usos-de-ChatGPT-e-IA-generativa-en-la-justicia.pdf

García Sánchez, M. (2023) El abordaje de ChatGPT: el "Rinoceronte Gris" de la IA conversacional. *IUS ET SCIENTIA*, 9(1). https://dx.doi.org/10.12795/IETSCIENTIA

OpenAI (s.f. a). *Text generation models*. Recuperado el 1 de abril de 2024 de https://platform.OpenAI.com/docs/guides/text-generation

OpenAI (s.f. b). *Does ChatGPT tell the truth?* Recuperado el 1 de abril de 2024 de:https://help.OpenAI.com/en/articles/8313428-does-chatgpt-tell-the-truth.

OpenAI (s.f. c). *Introducing ChatGPT*. Recuperado el 1 de abril de 2024 de https://OpenAI.com/blog/chatgpt.

OpenAI (2024d). *GPT-4 Technical Report,* (V. 6). https://doi.org/10.48550/arXiv.2303.08774.

OpenAI (2024, 24 de noviembre e). *Terms of use*. Recuperado el 1 de abril de

2024 de https://openai.com/es/policies/terms-of-use

OpenAI (s.f f). *Prompt engineering*. Recuperado el 1 de abril de 2024 de https://platform.OpenAI.com/docs/guides/prompt-engineering

EJE TEMÁTICO DE LA COMUNICACIÓN

Derecho Y Nuevas Tecnologías

FILIACIÓN

AUTOR 1: Otros Roles Que No Se Encuentran Especificados En Las Opciones Anteriores - PEI-FD 2024/001