Revista Científica
DE LA FACULTAD DE DERECHO
Y CIENCIAS SOCIALES
Y POLÍTICAS - UNNE

Alucinaciones de los grandes modelos de lenguaje y la importancia del control humano en el ámbito jurídico. Recomendaciones de uso a partir del caso "Roberto Mata vs. Avianca Airlines Inc."

Hallucinations of Large Language Models and the Importance of Human Oversight in the Legal Field. Recommendations for Use Based on the "Roberto Mata v. Avianca Airlines Inc." Case

Darío Saúl Navarro
Consejo de la Magistratura de la provincia de Corrientes, Argentina
darionavarro85@gmail.com
https://orcid.org/0000-0001-5876-1863
Especialista en Derechos Humanos, Universidad de Bologna, Italia
Especialista en Derecho Laboral, Facultad de Derecho y Ciencias Sociales y Políticas, Universidad Nacional del Nordeste, Argentina
Especialista en Teoría y Técnica de losProcesos Judiciales, Facultad de Derecho y Ciencias Sociales y Políticas, Universidad Nacional del Nordeste
Secretario del Consejo de la Magistratura, provincia de Corrientes

Recepción: 18 de junio de 2025
Aceptación: 3 de septiembre de 2025

Resumen

Este artículo explora el fenómeno de las alucinaciones generadas por los grandes modelos de lenguaje y su posible impacto en el campo jurídico. A partir del análisis del caso *Roberto Mata vs. Avianca Airlines. Inc*., en el que se introdujo una alucinación producida por ChatGPT a un proceso judicial, se argumenta sobre la necesidad del control humano para promover un uso provechoso de estas herramientas. Como corolario, se sugieren pautas para evitar la inclusión de alucinaciones generadas por LLM en documentos jurídicos.

Palabras clave: Inteligencia Artificial Generativa, deontología profesional, estudio de casos.

Abstract

This article explores the phenomenon of hallucinations generated by *large language*

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

models and their potential impact on the legal field. Through the analysis of the case *Roberto Mata v. Avianca Airlines Inc.*, in which a hallucination produced by ChatGPT was introduced into a judicial proceeding, the article argues for the necessity of human oversight to ensure the beneficial use of these tools. As a corollary, it proposes guidelines to prevent the inclusion of LLM-generated hallucinations in legal documents.

Keywords: Generative Artificial Intelligence, professional ethics, case study

1. Introducción

Los Grandes Modelos de Lenguaje o *Large Language Models* (LLM) son un tipo de inteligencia artificial (IA)[1] que permiten el funcionamiento de aplicaciones como ChatGPT (desarrollada por OpenAI[2]) o Gemini (desarrollada por Alphabet[3]). 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, s.f. -a), 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, s.f. -b) 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 analizar las implicancias de las alucinaciones de los LLM en la praxis judicial, mediante una revisión de literatura jurídica e informática, junto al análisis de regulaciones internacionales sobre IA. Además, con el propósito de fomentar un aprovechamiento responsable de los LLM en el campo del Derecho, se sugieren una serie de directrices para evitar o mitigar los riesgos derivados de alucinaciones, con base en el control humano.

La estructura del artículo se divide en tres secciones principales: en la primera, se realiza un abordaje técnico para explicar el funcionamiento de los LLM y la aparición de alucinaciones. En la segunda parte, se analizará el caso *Roberto Mata vs. Avianca Airlines. Inc.*

1. Dado que los LLM representan un subcampo dentro de los programas de Inteligencia Artificial (IA), este artículo utilizará ambos términos de manera indistinta.
2. OpenAI tiene una organización mixta. Está dirigida por una organización sin fines de lucro, pero recibe financiación privada con un modelo de ganancias limitadas. Su objetivo declarado es que la inteligencia artificial general beneficie a toda la humanidad.
3. Alphabet es la empresa matriz de Google. A partir de los desarrollos de una empresa subsidiaria, Google Deep Mind, se lanzó el LLM conocido como Gemini, en diciembre de 2023.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

con el objeto de ilustrar los efectos de la introducción de una alucinación generada por IA en un proceso judicial. En la última parte, se abordará la importancia del control humano en la IA, y su regulación en instrumentos internacionales, para finalmente sugerir pautas concretas que puedan servir de guía para un uso ético de la IA en la práctica jurídica.

1.1 Metodología

El presente artículo se desarrolla bajo un enfoque de investigación cualitativo, combinando el estudio de caso con la revisión documental y el análisis normativo. El objetivo es analizar las implicancias de las alucinaciones de los LLM en la práctica judicial y proponer directrices para su uso responsable.

2. Los grandes modelos de lenguaje o large language models(LLM)

2.1. Los LLM como subcampo de la IA

Según Russell y Norvig (2004)

> La IA abarca en la actualidad una gran variedad de subcampos, que van desde áreas de propósito general, como el aprendizaje y la percepción, a otras más específicas como el ajedrez, la demostración de teoremas matemáticos, la escritura de poesía y el diagnóstico de enfermedades. La IA sintetiza y automatiza tareas intelectuales y es, por lo tanto, potencialmente relevante para cualquier ámbito de la actividad intelectual humana. En este sentido, es un campo genuinamente universal. (p. 1)

Entre los subcampos de la IA, el procesamiento del lenguaje natural (NLP) se ocupa de analizar, entender y generar lenguaje humano. El objetivo de estos sistemas es simular el procesamiento del lenguaje (en el sentido de dar los mismos resultados), pero en ningún caso se persigue que *imiten* (funcionen del mismo modo) lo que hace la mente humana. En este sentido, Russel y Norving han ilustrado este enfoque, explicando que la búsqueda de un ingenio que "volara artificialmente" tuvo éxito cuando los hermanos Wright, entre otros, dejaron de imitar a los pájaros y comprendieron los principios de la aerodinámica (2004).

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

2.2 Caracterización de los LLM

Los LLM son modelos de propósito general que son capaces de generar texto coherente y relevante para una multitud de tareas empleando el lenguaje natural, en lugar de estar capacitados para una tarea específica, como el análisis de sentimientos, el reconocimiento de entidades nombradas o el razonamiento matemático (Wei et al., 2022). Son sistemas entrenados con aprendizaje profundo que pueden reconocer, resumir, traducir, predecir y generar texto basado en el conocimiento obtenido de conjuntos de datos masivos.
El diálogo entre el operador y los LLM se enmarca en la denominada "ventana de contexto", lo que comprende tanto los *prompts* (instrucciones) como las respuestas. Para analizar la instrucción del usuario, estos sistemas simplifican los textos, y remueve artículos, pronombres, palabras usuales y se queda con las palabras definitorias del sentido. En base a esto, hace un cálculo probabilístico para definir la respuesta más adecuada de acuerdo a las palabras introducidas por el operador. Sus respuestas son el producto de complejas operaciones probabilísticas, sustentadas en una enorme base de datos y su entrenamiento reforzado por la retroalimentación humana.

De acuerdo a Corvalán y Carro (2023), la limitación central de los LLM es que solo realizan correlaciones estadísticas, pero no tienen una comprensión del mundo real.

2.3 La tecnología detrás de los LLM

El tamaño o escala de los LLM es fundamental, ya que se ha demostrado que su incremento puede conducir a un mejor rendimiento en una variedad de tareas (Kaplan *et al*, 2020). De acuerdo a Bommasani (2022), el salto tecnológico actual se ha impulsado a partir de tres motores: 1) mejoras en el hardware de computadoras, como el rendimiento de GPU y la memoria, que han aumentado 10 veces en los últimos cuatro años; 2) el desarrollo de la arquitectura del modelo "Transformer", que aprovecha el paralelismo del hardware para entrenar modelos mucho más expresivos que antes; 3) la disponibilidad de muchos más datos de entrenamiento.

2.3.1 Los *transformers*

Son un tipo de arquitectura de modelos de aprendizaje automático o *machine learning*. Fueron introducidos por primera vez en el artículo Attention is All You Need por Vaswani *et al*. en 2017. Según Merritt (2022), se trata de un tipo de red neuronal que aplica técnicas matemáticas, llamadas atención o atención propia, para detectar formas sutiles de interacción y dependencia entre los elementos de datos de una serie (en este caso, palabras). Antes de que llegaran los *transformers*, los desarrolladores tenían que entrenar

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

redes neuronales con grandes conjuntos de datos etiquetados por personas que eran costosos y lentos de producir. Al encontrar patrones entre elementos matemáticamente, los *transformers* eliminan esa necesidad, ya que están disponibles los billones de imágenes y *petabytes* de datos de texto en la Web y en las bases de datos corporativas.

2.3.2 Potencia de cómputo

La matemática que usan los *transformers* aprovecha el procesamiento paralelo de los procesadores gráficos conocidos como GPU (*Graphics Processing Unit*), para que estos modelos puedan ejecutarse rápidamente. Antes el procesamiento de datos era secuencial, un dato detrás de otro, por lo que con estos sistemas se puede procesar varios datos por distintos caminos simultáneos.

2.3.3 Disponibilidad de datos

Los LLM necesitan de grandes volúmenes de datos. Se entrenan con conjuntos de datos lo suficientemente grandes como para incluir casi todo lo que se ha escrito en Internet durante un período de tiempo prolongado. Por ejemplo, ChatGPT 3.5 se ha nutrido de Common Crawl[4], una base de datos que comprende más de 50.000 millones de páginas web, y Wikipedia, que tiene aproximadamente 57 millones de páginas (Merritt, 2022).

En la etapa de entrenamiento, cantidades masivas de texto se suministran al modelo mediante el aprendizaje no supervisado, lo que implica que se le proporciona un conjunto de datos sin instrucciones explícitas sobre qué hacer con él. A través de este método, un LLM aprende palabras, así como las relaciones entre ellas y los conceptos detrás de ellas. De este modo, puede aprender a diferenciar los significados de una palabra en función de su contexto (Lee, 2023).

Además, modelos como ChatGPT se han optimizado para el diálogo utilizando el Aprendizaje por Refuerzo con Retroalimentación Humana o *Reinforcement Learning with Human Feedback* (RLHF), un método que utiliza demostraciones humanas y comparaciones de preferencias para guiar al modelo hacia el comportamiento deseado (OpenAI s.f.-c).

2.4 Las alucinaciones

A pesar de la naturaleza convincente y realista de los textos generados por LLM, una preocupación creciente es que los modelos pueden generar información ficticia con confianza (Manakul, 2023). La expresión “alucinación” es una metáfora para denominar

4. Common Crawl es una organización sin fines de lucro que cuenta con un banco de datos de millones de páginas web y ofrece libremente sus archivos al público.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

al contenido generado por una IA que carece de sentido o no es fiel al contenido fuente proporcionado (Ji, *et al*., 2023).

El uso de esta expresión ha sido objeto de críticas, porque la palabra "alucinar" sugiere la experiencia de percepciones sensoriales en ausencia de estímulos externos. Algunos investigadores proponen el término de "confabulación", que en psicología humana, ocurre cuando la memoria de alguien tiene un vacío y el cerebro llena convincentemente el resto sin la intención de engañar a los demás (Edwards, 2023).

De cualquier manera, es importante tener en claro que las expresiones "alucinación" o "confabulación", son metáforas antropomórficas para denominar un error de procesamiento en un sistema informático. Pero de ninguna manera estos resultados indeseados se pueden equiparar a un estado mental equivalente al del cerebro humano.

2.4.1 Causas

Las alucinaciones generadas por la inteligencia artificial (IA) pueden atribuirse a diversos motivos. En primer lugar, la complejidad del lenguaje humano presenta desafíos para los modelos de lenguaje, que pueden interpretar de manera incorrecta el contexto o las sutilezas lingüísticas. Además, problemas como la calidad deficiente de los datos de entrada, el diseño limitado de los modelos y el sesgo en los conjuntos de entrenamiento pueden influir en la generación de resultados inesperados (Spinak, 2023).

En el plano técnico Mishra (2023), alude a las siguientes causas:

> 1.- Sobre-optimización: Durante el entrenamiento, modelos como GPT-3 se optimizaron para producir texto coherente y contextualmente relevante. Esta optimización a veces los lleva a inventar información que se ajusta al contexto, incluso si no es verdadera.
>
> 2.- Ausencia de Verificación Externa: Puede ocurrir si un LLM carece de la capacidad para verificar información de fuentes externas, como internet.
>
> 3.- Inferencia Contextual: Los modelos de lenguaje infieren el contexto a partir del texto precedente, pero pueden interpretarlo o extrapolarlo incorrectamente, lo que lleva a alucinaciones.

Además, resulta relevante el hecho que los LLM son entrenados para producir texto fluido y similar al humano. Sin embargo, esta expresividad también permite generar texto detallado y convincente, aunque el contenido sea ficticio o sin sentido. Asimismo, estos modelos están entrenados para detectar relaciones estadísticas entre palabras y conceptos en texto, pero no razonan sobre la plausibilidad de la respuesta. Como resultado, pueden generar *outputs* detallados que carecen de consistencia lógica. Por otro lado, los LLM

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

carecen de conocimiento del mundo real, lo que le impide confrontar la verosimilitud de sus respuestas. (Ghosh, 2023).

A pesar de estas vulnerabilidades, los LLM no están programados para inventar respuestas, sino que están optimizados para "maximizar la respuesta". Esto implica la generación de respuestas basadas en la probabilidad estadística derivada de los patrones en los datos de entrenamiento. Entonces, si no encuentran una respuesta clara o precisa, generan la mejor respuesta posible basada en esos patrones, pero no están programados para reconocer y admitir sus limitaciones o desconocimiento (San Juan, 2024).

3. El caso *Mata vs. Avianca Airlines Inc.*

La elección de este caso se funda en el hecho de que fue el primer caso de introducción de una alucinación a un proceso judicial que tomó dominio público a nivel internacional y como tal, se convirtió en un caso paradigmático del peligro de la IA en relación a un proceso judicial[5] .

A los fines del presente artículo, entenderemos como "caso" a una situación fáctica, donde distintas partes (particulares, personas jurídicas, el estado, etc.) aparecen con intereses encontrados, en razón de los cuales se ha generado un conflicto. Este conflicto es presentado ante una autoridad judicial que tiene la potestad de resolverlo mediante algún tipo de pronunciamiento, que ha de estar fundado en derecho y dotado de racionalidad, respaldado por la amenaza de emplear la coacción en caso de que fuera necesaria para hacer efectivo lo resuelto (Peñalva, 2021).

Para efectuar el análisis del caso *Mata vs. Avianca Airlines Inc*., se adoptan como fuentes la resolución del 22 de junio de 2023 emitida por el juez P. Kevin Castel, del Tribunal de Distrito Sur de Nueva York sobre la cuestión de fondo; y la resolución del mismo magistrado y de la misma fecha, que abordó la responsabilidad disciplinaria de los abogados del Sr. Mata.

3.1 Contexto judicial

De acuerdo al Sr. Mata, el 27 o 28 de agosto de 2019, entre las 11 a.m y la 1 p.m sufrió una lesión, cuando un asistente de vuelo lo golpeó en la rodilla izquierda con un carrito

5. Este caso fue difundido por el New York Times, en un artículo titulado "Here's What Happens When Your Lawyer Uses ChatGPT" (Esto es lo que pasa cuando tu abogado usa ChatGPT) (Weiser, 2023), y con el mismo tono crítico fue replicado en distintos diarios del mundo. Para advertir la notoriedad del caso, se recomienda buscar en motores de búsqueda como *Google* utilizando términos relevantes, como [Chatgpt+ abogado+alucinación].

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

de servicio de metal, durante un vuelo desde El Salvador hasta el Aeropuerto John F. Kennedy en Nueva York.

En base a esta plataforma fáctica, presentó una demanda por daños contra Avianca Inc.[6] (Avianca), el 28 de julio de 2020. El 20 de noviembre del mismo año, uno de los abogados del Sr. Mata tuvo conocimiento que Avianca había solicitado la quiebra en mayo de dicho año. Según el U.S. Code[7], la quiebra suspende todos los actos y procedimientos en contra del deudor y la propiedad del deudor (título 11, sección 362). Después que Avianca emergió de la quiebra en enero de 2022, el Sr. Mata desistió de su reclamo anterior y presentó una nueva acción el 2 de febrero del mismo año, la cual dio origen al caso en estudio[8].

El Sr. Mata planteó que al caso le era aplicable una norma estatal, que preveía un plazo de prescripción de 3 años. En forma subsidiaria, argumentó que la acción fue presentada a tiempo porque el proceso de quiebra de Avianca había suspendido el período de prescripción estipulado en la Convención de Montreal.

La aerolínea solicitó la desestimación del reclamo, argumentando que la demanda se había presentado fuera del plazo establecido por la Convención para la Unificación de Ciertas Reglas para el Transporte Aéreo Internacional (Convención de Montreal). Esta convención, establece un período de dos años para presentar reclamaciones relacionadas con el transporte aéreo internacional (art. 35 inc. 1). Bajo esta legislación, la prescripción de la acción habría operado el 28 de agosto de 2021.

3.2 La introducción de una alucinación al proceso judicial

Steven A. Schwartz, abogado del Sr. Mata, le solicitó a ChatGPT 3.5 que le brinde citas de decisiones judiciales, para respaldar los fundamentos de la oposición al planteo de la parte demandada.

Schwartz incluyó los casos generados por el LLM en el escrito de oposición al planteo de prescripción, sin verificar su veracidad. Luego se lo remitió a un colega, Peter Lo Duca, para que lo presente en nombre del Sr. Mata, debido a que Schwartz no estaba matriculado para intervenir ante el fuero federal de Nueva York.

Lo Duca tampoco verificó la existencia de los casos citados y presentó el documento

6. Avianca Inc. es la filial estadounidense de la corporación Avianca Group International Limited. La empresa tuvo su origen en Colombia, con el nombre de Aerovías del Continente Americano.
7. El U.S Code es una codificación por materia de las leyes generales y permanentes de los Estados Unidos. Está elaborado por la Oficina del Consejero de Revisión Legal de la Cámara de Representantes de los Estados Unidos. Su contenido se puede consultar en https://uscode.house.gov/.
8. La acción se presentó ante la Corte Suprema del Estado de Nueva York, pero a instancias de la demandada, el caso se trasladó a una Corte Federal del Distrito Sureste de Nueva York, en el entendimiento que se trataba de un caso federal.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

de oposición en la Corte. Una vez notificados de la oposición al planteo de prescripción, los abogados de la aerolínea señalaron al juez que no encontraban los casos citados por la parte actora. El magistrado tampoco pudo encontrar aquellos precedentes, por lo que los abogados del Sr. Mata debieron dar explicaciones ante la interpelación del tribunal (Weiser, 2023).

De acuerdo al historial de conversaciones suministrado por Steven A. Schwartz al declarar ante el juez interviniente[9], su primer *prompt* fue: "argumente que el plazo de prescripción es suspendido por la quiebra del demandado de conformidad con el convenio de Montreal". ChatGPT respondió con descripciones amplias del Convenio de Montreal, los plazos de prescripción y la suspensión federal de la quiebra, aconsejó que "[l]a respuesta a esta pregunta depende de las leyes del país en el que se presente la demanda". El Sr. Schwartz insistió, solicitando: "proporciona jurisprudencia que respalde que el plazo de prescripción es suspendido por la quiebra del demandado bajo el convenio de Montreal", "muéstrame decisiones específicas en casos federales donde el plazo de prescripción fue suspendido debido a la quiebra de la aerolínea", "muéstrame más casos" y "dame algunos casos en los que la convención de Montreal permitió la suspensión del plazo de prescripción debido a la quiebra".

En base a estas instrucciones el LLM citó media docena de precedentes con nombres verosímiles, con indicación de fecha y tribunal.

Los casos generados por ChatGPT 3.5 fueron[10]:

- Varghese v. China Southern Airlines Co., Ltd., 925 F.3d 1339 (11th Cir. 2019).
- Shaboon v. Egyptair, 2013 IL App (1st) 111279-U (Ill. App. Ct. 2013).
- Peterson v. Iran Air, 905 F. Supp. 2d 121 (D.D.C. 2012).
- Martínez v. Delta Airlines, Inc., 2019 WL 4639462 (Tex. App. Sept. 25, 2019).
- Estate of Durden v. KLM Royal Dutch Airlines, 2017 WL 2418825 (Ga. Ct. App. June 5, 2017).
- Ehrlich v. American Airlines, Inc., 360 N.J. Super. 360 (App. Div. 2003).
- Miller v. United Airlines, Inc., 174 F.3d 366, 371-72 (2d Cir. 1999).

9. Traducción al español efectuada con el LLM Claude el día 21/3/2024. https://claude.ai/chat/17bd20b3-82e6-40da-829d-4c079bd94d85

10. Este listado es una cita textual de los precedentes generados por ChatGPT de acuerdo a lo reseñado en la resolución que abordó la responsabilidad disciplinaria de los abogados del Sr. Mata.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

La veracidad de las citas jurisprudenciales es una exigencia deontológica para los abogados, y su incumplimiento es una violación a un deber profesional, ya sea que se inserten por descuido o de manera maliciosa. Por ejemplo, el Código de Ética de los abogados de la provincia de Corrientes, en su art. 24 establece que: "Es falta de ética efectuar citas doctrinarias o jurisprudenciales inexistentes, o exponerlas en forma tal que falseen la opinión o el fallo invocado, o realizar falsas transcripciones de resoluciones judiciales".

En su declaración ante el juez, Schwartz reconoció que los precedentes judiciales inexistentes los había solicitado a ChatGPT, ya que supuso que se trataba de un motor de búsqueda y que los incluyó en el documento sin revisar su veracidad. En su defensa, le dijo al juez que no tenía la intención de engañar al tribunal ni a la aerolínea y que nunca antes había utilizado ChatGPT, por lo que no tenía conocimiento que pudiera generar información falsa. Finalmente, Schwartz y su colega, fueron sancionados con una multa de $5.000 dólares y que informen a su cliente y a los jueces cuyos nombres fueron invocados incorrectamente sobre las sanciones impuestas. En cuanto al litigio de fondo, el juez consideró que el Convenio de Montreal se adoptó para promover la aplicación uniforme entre los Estados signatarios, y que la aplicación de los principios de suspensión locales podría socavar ese objetivo, por lo que desestimó la demanda planteada por el Sr. Mata.

3.3 Impacto de la alucinación en el proceso judicial

Recapitulando sobre el caso, es posible dividir la conducta del Sr. Schwartz en dos etapas, la primera en relación la generación del output y la segunda vinculada a su introducción al proceso.

De acuerdo a la transcripción de los *prompts* suministrados por el abogado, en un principio, el LLM adoptó una postura cautelosa al abordar una cuestión legal. Recién cuando el abogado insistió a ChatGPT para que le brindara precedentes judiciales en un sentido determinado, el sistema accedió, generando el nombre de casos inexistentes. Aunque la alucinación fue promovida por el usuario al suponer que interactuaba con un motor de búsqueda avanzado, el LLM no tuvo la capacidad de advertir que no podía emitir una respuesta fiable sobre esa temática.

Lo anterior no importa soslayar que la introducción de una alucinación al proceso judicial tuvo como causa exclusiva a la conducta negligente de Steven A. Schwartz. En efecto, el abogado reconoció que no verificó la veracidad de la información del output. Incluso, en el marco de la investigación sobre su conducta, el Sr. Schwartz omitió expresar si tenía experiencia previa en casos vinculados a la Convención de Montreal o sobre suspensiones por quiebras. Además, es importante resaltar que aunque se utilizó un LLM para una tarea específica, como la citación de precedentes, se trataba de una cuestión sensible, con potencial de decidir la suerte del proceso.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

En los Estados Unidos, rige el sistema del *Common Law*, las sentencias de tribunales edifican el Derecho, y mediante la doctrina del *stare decisis*, los fallos de los tribunales superiores son obligatorias (Amaya, 2015). El análisis de las sentencias tiene una importancia capital, se analizan diversas sentencias que contienen el mismo principio, visto desde diversos casos, para extraer el criterio que se aplicará al caso en estudio (Torres Zárate y García Martínez, 2008). Incluso si una decisión judicial no proviene de un tribunal superior, los precedentes influyen en la decisión en cada caso, en la medida en que le da a un juez una forma de pensar sobre el problema que debe decidir (Coney Barrett, 2013).

Ante este contexto, el deber de controlar un output con citación de casos judiciales resultaba imprescindible, no sólo por la responsabilidad profesional, sino porque la citación representaba un argumento fundamental para sustentar la pretensión del justiciable.

4. La utilización de los LLM bajo supervisión humana o *human in the loop*

En su acepción original, la expresión *Human in the loop* (HITL), en español "humano en el bucle o en el proceso" se aplica al aprendizaje automático o *machine learning*. HITL describe el proceso en el que un sistema informático no puede resolver un problema y necesita intervención humana en las etapas de entrenamiento y prueba de construcción del algoritmo, para crear un bucle de retroalimentación continua que permita al algoritmo proporcionar resultados cada vez mejores (Bisen, 2020). ·

Siguiendo a Cohen, *et al.* (2023), utilizaremos la sigla HITL de manera amplia, para referirnos a la operatoria con IA en la cual los humanos tienen un papel determinante en el proceso de toma de decisiones. En este artículo, el concepto de HITL se centra en involucrar a los operadores jurídicos para validar el output de un LLM y tomar una decisión final respecto a su utilización.

En la Argentina, todavía no existe una norma que regule la utilización de IA por abogados litigantes ni por integrantes de la administración de justicia. Ante este escenario, resulta conveniente recurrir a fuentes extranjeras que hayan previsto una regulación general vinculada al uso de la IA.

Por tal motivo, a continuación, se efectúa un análisis de dos normativas de alcance internacional pioneras en la regulación de la IA. Por un lado, la Recomendación sobre la ética de la inteligencia artificial de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO); y por otro, el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de la Unión Europea (EU).

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

Estos documentos contienen regulaciones amplias sobre el desarrollo, utilización y efectos de la IA, por lo que el presente análisis se circunscribe a la incidencia del control humano por parte de los operadores de LLM.

4.1 Recomendación sobre la ética de la IA de la UNESCO

En noviembre de 2021, la UNESCO elaboró la primera norma mundial sobre la ética de la IA, denominada "Recomendación sobre la ética de la inteligencia artificial", adoptada por los 193 Estados miembros, entre ellos la Argentina. El documento considera a la ética como una base dinámica para la evaluación y la orientación normativas de las tecnologías de la IA, tomando como referencia la dignidad humana, el bienestar y la prevención de daños. La Recomendación interpreta a la IA como aquellos sistemas con capacidad para procesar datos de forma similar a un comportamiento inteligente. Se adoptó esta concepción debido a que el rápido ritmo del cambio tecnológico podría dejar obsoleta de forma repentina cualquier definición fija y estrecha (UNESCO, s.f).

El documento diferencia entre valores y principios. Los primeros son ideales que postulan un comportamiento deseable y representan los fundamentos de los principios. Estos últimos revelan los valores subyacentes de manera más concreta, de manera que puedan aplicarse más fácilmente en las declaraciones de política y en las acciones.

Los valores mencionados en la Recomendación son: 1) Respeto, protección y promoción de los derechos humanos, las libertades fundamentales y la dignidad humana; 2) Prosperidad del medio ambiente y los ecosistemas; 3) Garantía de la diversidad y la inclusión; 4) Vivir en sociedades pacíficas, justas e interconectadas.

Los principios son: 1) Proporcionalidad e inocuidad; 2) Seguridad y protección; 3) Equidad y no discriminación; 4) Sostenibilidad; 5) Derecho a la intimidad y protección de datos; 6) Supervisión y decisión humana; 7) Transparencia y explicabilidad; 8) Responsabilidad y rendición de cuentas; 9) Sensibilización y educación; 10) Gobernanza y colaboración adaptativas y de múltiples partes interesadas.

A los fines de este artículo, nos centraremos en tres principios: Proporcionalidad e inocuidad; Supervisión y decisión humana; Responsabilidad y rendición de cuentas.

4.1.1 Proporcionalidad e inocuidad

Según este documento, la decisión de utilizar sistemas de IA y la elección del método de IA debería justificarse de las siguientes maneras: a) el método de IA elegido debería ser adecuado y proporcional para lograr un objetivo legítimo determinado; b) el método de IA elegido no debería constituir una violación o un abuso de los derechos humanos;

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

y c) el método de IA elegido debería ser adecuado al contexto y basarse en fundamentos científicos rigurosos.

Desde la perspectiva del derecho, la decisión humana implica decidir en primer lugar si se utiliza un LLM en la elaboración de un documento jurídico, evaluando la conveniencia y riesgos implicados.

Además, esta normativa indica que en los casos en que se entienda que las decisiones tienen un impacto irreversible o difícil de revertir, la decisión final debería ser adoptada por un ser humano. En el ámbito procesal, puede entenderse que todos los actos son "irreversibles" en función del principio de preclusión, por el que una vez consumado el acto (demanda, contestación, recursos, etc), las partes no pueden volver a modificarlo. Por lo tanto, en este ámbito, el control humano debería abarcar cualquier acto procesal en el que se aplique un LLM.

4.1.2 Supervisión y decisión humana

En estas Recomendaciones se admite la posibilidad de recurrir a los sistemas de IA en la adopción de decisiones y en la ejecución de tareas por razones de eficiencia. Al respecto, Chumbita señala que un LLM puede optimizar desde la confección de un escrito, a la reformulación de un texto para expresarse con mayor claridad (2023).

No obstante, el documento proclama que el uso de un sistema de IA nunca podrá desligarse del control humano. En sintonía, la política de uso de la IA en la administración de justicia de España, establece que la IA nunca debe reemplazar la toma de decisiones humanas. En este marco normativo se prevé que los sistemas de IA se utilicen como herramientas para ayudar a los profesionales del derecho en sus funciones, aunque la responsabilidad final de tomar decisiones legales debe recaer en jueces y magistrados (CTEAJE, 2024).

Además, se aclara que la supervisión humana se refiere, no solo a la supervisión humana individual, sino también a la supervisión pública[11].

4.1.3 Responsabilidad y rendición de cuentas

La responsabilidad ética y la obligación de rendir cuentas de las decisiones y las acciones basadas de alguna manera en un sistema de IA siempre deberían ser atribuibles, en última instancia, a los desarrolladores y usuarios de la IA.

11. El control público podría configurarse mediante la formación de comités especiales para evaluar la utilización ética de la IA en el Derecho, ya sea mediante órganos estatales o por los Colegios de Magistrados o de Abogados.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

Cobbe *et al.* (2023) explican que la rendición de cuentas implica que un actor proporcione explicaciones sobre sus actividades, lo que impone consecuencias para corregir su conducta si es necesario. En consecuencia, argumentan que quienes son responsables de los diversos aspectos de la producción, distribución y uso de los sistemas de IA deben ser identificados correctamente para que la responsabilidad pueda asignarse en consecuencia.

No obstante, como lo demuestra la sanción aplicada a los abogados del reclamante en el caso Mata, la falta de regulación específica no impide que la rendición de cuentas se haga efectiva en la actualidad, ya que los tradicionales deberes de veracidad y buena fe pueden permitir la evaluación de actos procesales elaborados con intervención de un LLM.

En línea con la necesidad de establecer mecanismos de responsabilidad y rendición de cuentas, el Convenio Marco del Consejo de Europa sobre IA (Consejo de Europa, 2024), exige que los Estados aseguren remedios accesibles y efectivos frente a violaciones de derechos resultantes del uso de sistemas de IA.

Concretamente, el art. 14 del Convenio requiere: (i) documentar información pertinente sobre los sistemas de IA con potencial de afectar derechos; (ii) proveerla a autoridades competentes y, cuando corresponda, ponerla a disposición de las personas afectadas; y (iii) que esa información sea suficiente para impugnar decisiones adoptadas o sustancialmente informadas por IA, incluyendo la posibilidad efectiva de reclamar ante autoridades competentes.

De forma complementaria, el art. 15 dispone que existan garantías y derechos procesales efectivos cuando un sistema de IA impacte de manera significativa en el goce de derechos. Esta norma instaura un remedio procesal específico como garantía de acceso a la justicia, que podría aplicarse en el supuesto de que una alucinación cause la violación de algún derecho.

4.2 Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo

En agosto de 2024, entró en vigor el primer marco jurídico integral sobre IA en todo el mundo, con el nombre de Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, publicado el 12 de julio de 2024. Es conocido popularmente como "Ley de IA" o "*AI Act*" (en adelante RIA). El RIA establece normas armonizadas en materia de inteligencia artificial y modifica los Reglamentos (CE) nº 300/2008, (UE) nº 167/2013, (UE) nº 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828. Esta norma proclama como objetivo proporcionar a los desarrolladores e implementadores de IA requisitos y obligaciones claros con respecto a los usos específicos de la IA, garantizando el respeto de los derechos fundamentales, la seguridad y los principios éticos (UE, 2024).

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

Al ser la primera sistemática y exhaustiva sobre el tema, la UE podría establecer un estándar regulatorio mundial. Por ejemplo, la UE lo consiguió con su Reglamento General de Protección de datos (RGPD), que fue imitado en distritos fuera de Europa[12], como en Estados Unidos o la India (Ryan Mosley, 2024).

No obstante, Novelli, *et al*. (2024) advierten que la regulación europea todavía parece no estar adecuadamente preparada para afrontar los desafíos actuales y futuros, ya que las normas se diseñaron antes de la explosión de los LLMs. En este sentido, señalan que el desarrollo de una normativa enteramente coherente es una cuestión pendiente. Por lo tanto, resulta esperable que estas estas regulaciones sufran adaptaciones junto al desarrollo de la tecnología.

4.2.1 Enfoque normativo basado en el riesgo

El RIA asume un enfoque basado en el riesgo. En el art. 3 se define al "riesgo" como: "la combinación de la probabilidad de que se produzca un daño y la gravedad de ese daño".

El marco regulador define cuatro niveles de riesgo para los sistemas de IA: Riesgo inaceptable, Alto riesgo, Riesgo limitado y Riesgo mínimo. Entre los sistemas de alto riesgo, se incluyen aquellos que sirvan de asistentes a una decisión judicial.

El art. 6 ap. 2, considera como sistemas de alto riesgo a los enumerados en el Anexo III. En el punto 8 de este último, se incluye a:

> Sistemas de IA destinados a ser utilizados por una autoridad judicial o en su nombre para asistir a una autoridad judicial en la investigación e interpretación de hechos y de la ley y en la aplicación de la ley a un conjunto concreto de hechos o utilizados de forma similar en la resolución alternativa de litigios.

No obstante, el ap. 3 del mismo artículo aclara que los sistemas de IA no se considerarán de alto riesgo si no supone un riesgo significativo de perjuicio para la salud, la seguridad o los derechos fundamentales de las personas físicas, incluso si no influye significativamente en el resultado de la toma de decisiones. Esto se aplica cuando el sistema de IA está destinado a realizar una tarea de procedimiento limitada o el sistema de IA está destinado a mejorar el resultado de una actividad humana previamente realizada.

12. Anu Bradford, catedrática de Derecho de la Universidad de Columbia, ha llamado a esto el "efecto Bruselas", para denominar el fenómeno por el cual la UE pretende influenciar a los mercados globales mediante la introducción de marcos regulatorios sobre fenómenos novedosos.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

Si se toma en cuenta esta disposición, la sola búsqueda de precedentes judiciales para respaldar una posición elaborada previamente por un abogado, como ocurrió en el caso Mata, no representaría una actividad de alto riesgo. Aunque esta tarea pueda considerarse *a priori* como de riesgo limitado, en dicho caso la falta de control humano tuvo un impacto negativo, tanto para el abogado que utilizó los *prompts* que contenía alucinaciones, como para el reclamante de la indemnización.

4.2.2 Regulación de los LLM

Los LLM como ChatGPT y Gemini resultan encuadrados en la definición de "modelo de IA de propósito general" (GPAI). De acuerdo al art. 3, inc. 63, se tratan de:

> … un modelo de IA, incluso cuando dicho modelo se entrena con una gran cantidad de datos utilizando autosupervisión a escala, que muestra una generalidad significativa y es capaz de realizar de forma competente una amplia gama de tareas distintas independientemente de la forma en que se comercialice el modelo y que puede integrarse en diversos sistemas o aplicaciones posteriores, excepto los modelos de IA que se utilizan para actividades de investigación, desarrollo o creación de prototipos antes de su comercialización.

De acuerdo a esta regulación, en principio los GPAI no son considerados como sistemas de alto riesgo, pero podrían ser incluidos como IA con riesgo sistémico, dependiendo del número de parámetros del modelo o incluso del número de usuarios finales registrados (art. 51).

4.2.3 Supervisión humana

La Supervisión humana está expresamente regulada en el art. 14, ap. 2, señalando que:

> La supervisión humana tendrá por objeto prevenir o reducir al mínimo los riesgos para la salud, la seguridad o los derechos fundamentales que puedan surgir cuando un sistema de IA de alto riesgo se utilice de acuerdo con su finalidad prevista o en condiciones de uso indebido razonablemente previsible.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

El ap. 4 de este artículo promueve la capacitación y comprensión adecuadas de las capacidades y limitaciones del sistema de IA para garantizar una supervisión efectiva.

Además, se postula que las personas que utilicen los modelos de IA deben permanecer en alerta ante posibles anomalías, disfunciones o resultados inesperados, para ser capaces de interpretar correctamente los resultados producidos por el sistema.

En este sentido, la normativa señala el peligro que los operadores habituales de IA incurran en el "sesgo de automatización", por el cual pueden tender a confiar demasiado en los resultados de los sistemas de IA. En consonancia, OpenAI (2024 -d) advierte que las alucinaciones pueden volverse más peligrosas a medida que los modelos se vuelven más veraces, ya que los usuarios depositan su confianza en el modelo cuando proporciona información veraz en áreas con las que tienen cierta familiaridad.

Asimismo, Levitina (2024) advierte que para lograr una estrategia de gobernanza efectiva, se deberían incluir revisiones periódicas del rendimiento, mecanismos de retroalimentación continua y desarrollo profesional permanente para las personas supervisoras.

Por último, esta norma dispone que las personas deben contar con la autoridad y los medios para decidir no utilizar el sistema en determinadas situaciones o para intervenir en su funcionamiento, e incluso para detenerlo por motivos de seguridad.

4.3 Directrices para prevenir la introducción de alucinaciones en documentos jurídicos

En esta última sección, en base a los contenidos desarrollados en los pasajes anteriores, se articulan una serie de recomendaciones prácticas para la utilización de LLM en la elaboración de documentos jurídicos. También 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 han sido adaptadas a nuestra propuesta de supervisión humana para prevenir alucinaciones en la elaboración de documentos jurídicos.

4.3.1 Asumir un control directo en todas las etapas de uso de un LLM

Al interactuar con un LLM, la estrategia más eficiente es la de asumir el control directo, definiendo para qué y cómo se va usar, y finalmente, controlar sus 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.

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

En el caso de ChatGPT, el modelo es sensible a cambios en la formulación de los *pompts* o al intentar la misma solicitud múltiples veces. Por ejemplo, dada una formulación de una pregunta, puede afirmar que no conoce la respuesta, pero dada una leve reformulación, puede responder correctamente (OpenAI, 2024 -d). Incluso, se han registrado respuestas completamente contradictorias a dos instrucciones idénticas (Corvalán, *et al*., 2023).

Además, los outputs de los LLM, incluso los más avanzados como GPT-4 pueden contener errores de razonamiento (OpenAI, s.f. b), por lo que resulta fundamental que los abogados no confíen sin reservas en las inferencias que pueda efectuar el sistema.

Por último, no se puede soslayar la revisión de eventuales sesgos de género, raza, orientación sexual u otros, que consisten en una incorrecta o injusta representación de una población o fenómeno, que tiene origen en la base de datos del sistema o por sesgos del entrenamiento por refuerzo humano (Ortiz de Zárate *et al*., 2024)

4.3.2 Utilizar los *outputs* para ámbitos en los que se tenga suficiente *expertise*

Antes de aplicar un *output* para la elaboración de un documento, es necesario evaluar su idoneidad en relación al uso concreto que se pretende darle (OpenAI, s.f. -e). En el ámbito jurídico, el operador debería contar con suficientes conocimientos jurídicos sobre la cuestión abordada por el LLM, para formular correctamente el *prompt*, y luego evaluar la veracidad y utilidad de las respuestas del modelo. En el caso Mata, el abogado Schwartz le pidió a ChatGPT 3.5 que le brinda citas de casos en los que se haya decidido que "el plazo de prescripción es suspendido por la quiebra del demandado bajo el convenio de Montreal"; pero al investigarse su responsabilidad profesional, el abogado no demostró contar con experiencia en casos vinculados al Convenio de Montreal ni sobre suspensión por quiebras.

4.3.3 Evitar fundar una decisión exclusivamente en un output de un LLM

En los términos de uso de ChatGPT, 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 (2024-f). Entendemos que 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.

4.3.4 Dividir tareas complejas en otras más simples

Para las tareas enviadas a un modelo de lenguaje OpenAI recomienda descomponer un sistema complejo en un conjunto de componentes modulares, ya que las tareas complejas

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

tienden a tener tasas de error más altas que las tareas más simples. Además, se indica que las tareas complejas a menudo 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. Eventualmente, se podrían unificar los *outputs* anteriores para integrarlos en un documento coherente.

4.3.5 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).

Si tomamos como referencia el caso Mata, se podría pedir una opinión a un LLM para que emita una opinión fundada respecto a la prescripción del plazo para demandar. En ese caso, el usuario debería indicar la fecha de ocurrencia del accidente, los plazos de prescripción previstos por las normas posiblemente aplicables y si es posible, una síntesis de uno o más casos judiciales análogos.

4.3.6 Evitar los LLM como única fuente de investigación

OpenAI (s.f. -g) 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.

4.3.7 Verificar la auto consistencia de los *outputs*

Según 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.

Para graficar esta circunstancia, los investigadores señalan que durante el preentrenamiento, un LLM se entrena con la predicción de la siguiente palabra sobre cantidades masivas de datos textuales. Si se tiene en cuenta el *prompt* "Lionel Messi es un _" es probable que el token o palabra "futbolista" tenga una alta probabilidad, ya que, al ser un

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

futbolista famoso, puede haber aparecido múltiples veces en el preentrenamiento. Sin embargo, para una entrada diferente como "John Smith es un _", el sistema estará inseguro de la continuación, lo que aumenta la probabilidad de la generación de una alucinación factual.

En el estudio, se afirma que, con una muestra de múltiples respuestas de un LLM, se puede medir la consistencia de la información entre las diferentes respuestas y determinar si las afirmaciones son fácticas o alucinadas. Si bien a mayor número de ejemplos se obtienen mejores rendimientos, los investigadores han logrado obtener resultados consistentes a partir de 4 muestras.

De tal manera, una opción válida podría ser la de confrontar al menos cuatro respuestas sobre el mismo *prompt* a fin de verificar la consistencia de estos *outputs*.

5. Conclusiones

Los LLM representan un avance tecnológico disruptivo con el potencial de transformar de forma rotunda a la actividad jurídica. Si bien ofrecen un potencial notable para procesar textos y elaborar argumentos, también pueden generar información falsa o inconsistente, con apariencia verosímil.

Por esta razón, resulta imperativo que los operadores jurídicos que opten por valerse de los LLM, asuman un control directo y riguroso en todas las etapas de su uso. Para ello, resulta imprescindible la alfabetización digital, lo que en este caso implica conocer las capacidades y límites de esta tecnología, y desarrollar la habilidad de redactar *prompts* eficaces.

Como se puede colegir a partir del estudio del caso Mata vs. Avianca inc, la falta de diligencia profesional y el uso indolente de las innovaciones tecnológicas es un cóctel peligroso, que puede comprometer la responsabilidad disciplinaria de los profesionales y el derecho de los justiciables. Pero renunciar al potencial y beneficios de una tecnología tan potente como los LLM, tampoco resulta una elección razonable; máxime en un contexto en el que la cantidad y complejidad de causas sobrepasan los nunca bien ponderados esfuerzos de la gran mayoría de operadores del sistema de justicia.

Referencias bibliográficas

Amaya, J. (2015). *Control de constitucionalidad*. (2.ª ed.). Astrea.

Bisen, V. (2020). *What is human in the loop machine learning: Why & how used in AI?* Medium. https://medium.com/vsinghbisen/what-is-human-in-the-loop-machine-learning-why-how-used-in-ai-60c7b44eb2c0

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., vonArx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S.,

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). On the Opportunities and Risks of Foundation Models. *Center for Research on Foundation Models*. [Preprint] arXiv. https://doi.org/10.48550/arXiv.2108.07258

Chumbita, S. (2023). Legal prompt engineering Una nueva habilidad para la era de la Inteligencia Artificial conversacional multimodal y multipropósito. *Revista La Ley*, 2023-C.

Cobbe, J., Veale, M., & Singh, J. (2023, junio 12–15). Understanding accountability in algorithmic supply chains. *Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. https://doi.org/10.1145/3593013.3593980

Cohen, I., Babic, B., Gerke, S., Xia, Q., Evgeniou, T., & Wertenbroch, K. (2023). How AI can learn from the law: Putting humans in the loop only on appeal. npj Digital Medicine, 6, 160. https://doi.org/10.1038/s41746-023-00906-8

Coney Barrett, A. (2013).Precedent and jurisprudential disagreement. *Texas Law Review*, *91*,1711-1738. https://scholarship.law.nd.edu/law_faculty_scholarship/293

Consejo de Europa. (2024). *Convenio Marco sobre Inteligencia Artificial, los Derechos Humanos, la Democracia y el Estado de Derecho*. CETS n.º 225. https://rm.coe.int/1680afae3c

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

Corvalán, J., Estévez E., Le Fevre C., Enzo M., Schapira, D., & Simari, G. (2023) *ChatGPT vs. GPT4: ¿Imperfecto por diseño?: explorando los límites de la inteligencia artificial conversacional*. La Ley.

Corvalán J. y Carro M., (2023) Mas allá del juego de la imitación: las principales limitaciones de la inteligencia artificial. En Corvalán J., *Tratado de Inteligencia Artificial y Derecho*, (2da ed., t. I). La Ley.

Edwards, B. (2023, 4 de junio). Why ChatGPT and Bing Chat are so good at making things up. *Ars Technica*. https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/

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

Ghosh, B. (2023, 11 de noviembre). Tackling Hallucinations in LLMs. *Medium*. https://medium.com/@bijit211987/tackling-hallucinations-in-llms-f2d7cbf35e72

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Chan, H. S., Dai, W., Madotto, A., & Fung, P. (2023). *Survey of hallucination in natural language generation*. *arXiv*. https://doi.org/10.1145/3571730

Kaplan, J., McCandlish, S., Henighan, T., Brown, T., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020) *Scaling Laws for Neural Language*

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

Models. *arXiv.* https://doi.org/10.48550/arXiv.2001.08361

Lee, A. (2023, 29 de marzo). ¿Para qué se Utilizan los Grandes Modelos de Lenguaje? *NVIDIA Latinoamérica*. https://la.blogs.nvidia.com/2023/03/29/para-que-se-utilizan-los-grandes-modelos-de-lenguaje/#:~:text=Un%20grande%20modelo%20de%20lenguaje,de%20conjuntos%20de%20datos%20masivos

Levitina, A. (2024). Los seres humanos en la toma de decisiones automatizada en el marco del RGPD y la Ley de IA. *Revista CIDOB d'Afers Internacionals,* 138(*3*), 121-138. https://doi.org/10.24241/rcai.2024.138.3.121

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SELFCHECKGPT: *Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*. University of Cambridge - ALTA Institute. https://doi.org/10.48550/arXiv.2303.08896

Merritt, R. (2022, 19 de abril). ¿Qué es un Modelo Transformer? *NVIDIA Latinoamérica.* https://la.blogs.nvidia.com/2022/04/19/que-es-un-modelo-transformer/

Mishra, A. N. (2023, 22 de octubre). Hallucination in Large Language Models. *Medium* Recuperado de https://medium.com/@asheshnathmishra/hallucination-in-large-language-models-2023-f7b4e77855ae

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024, marzo 15). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity (Working Paper). *arXiv*. https://arxiv.org/abs/2401.07348

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). *What is ChatGPT.* Recuperado el 1 de abril de 2024 de https://help.OpenAI.com/en/articles/6783457-what-is-chatgpt

OpenAI (2024-d) GPT-4 Technical Report (v6). https://doi.org/10.48550/arXiv.2303.08774

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

OpenAI (2024 -f). *Terms of use*. https://openai.com/es/policies/terms-of-use

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

Ortiz de Zarate, J., Dias J., Avenburg A., González Quiroga J. (2024, 25 de marzo). Sesgos algorítmicos y representación social en los modelos de lenguaje generativo (LLM). *Fundar*. https://fund.ar/publicacion/sesgos-algoritmicos-y-representacion-social-en-los-modelos-de-lenguaje-generativo/

Peñalva G.(2021). Lo difícil de los casos difíciles. *Anales de la Facultad de Ciencias*

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC

Jurídicas y Sociales, 18*(51),* 785–812. https://doi.org/10.24215/25916386e103

Russell, S., & Norvig, P. (2004). *Inteligencia artificial: Un enfoque moderno* (2.ª ed.). Pearson.

Ryan Mosley, T. (2024, 23 de enero). Vuelta al mundo por las regulaciones de la IA en 2024. *Mit Technology Review*. https://www.technologyreview.es//s/16069/vuelta-al-mundo-por-las-regulaciones-de-la-ia-en-2024

Spinak, E. (2023, 20 de diciembre). ¿Es que la Inteligencia Artificial tiene alucinaciones? *Blog SciELO*. https://blog.scielo.org/es/2023/12/20/es-que-la-inteligencia-artificial-tiene-alucinaciones/

UNESCO (s.f.). Ética de la inteligencia artificial. Recuperado el 7 de abril de 2024 https://www.unesco.org/es/artificial-intelligence/recommendation-ethics

Unión Europea (2024). *Ley de IA*. Configurar el futuro digital de Europa. Recuperado el 17 de marzo de 2024. https://digital-strategy.ec.europa.eu/es/policies/regulatory-framework-ai

San Juan, E. (2024, 6 de enero). Alucinaciones en IA, qué son y cómo detectarlas. *Community.es*.https://community.es/cursos/alucinaciones-en-ia-que-son-y-como-detectarlas/?trk=article-ssr-frontend-pulse_little-text-block

Secretaría General del Comité Técnico Estatal de la Administración Judicial (CTEAJE). (2024, junio). *Política de uso de la inteligencia artificial en la administración de justicia*. CTEAJE.

Torres Zárate, F. y García Martínez, F. (2008). Common law: una reflexión comparativa entre el sistema inglés y el sistema estadounidense. *Revista Alegatos, (68-69).* 71-100. https://alegatos.azc.uam.mx/index.php/ra/article/view/446

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser,L.,&Polosukhin,I.(2017).*Attention Is All You Need*. En *31st Conference on Neural Information Processing Systems*. *arXiv*. [Preprint]. https://doi.org/10.48550/arXiv.1706.03762

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2206.07682

Weiser, B. (2023, May 27). Here's What Happens When Your Lawyer Uses ChatGPT. *The New York Times*. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

Resoluciones judiciales

Tribunal del Distrito Sur de Nueva York. (2023, junio 22). Mata v. Avianca, Inc., Opinión y orden (Doc. 55).

Tribunal del Distrito Sur de Nueva York. (2023, junio 22). *Mata v. Avianca, Inc.*, Opinión y orden de sanciones (Doc. 54).

Revista Científica de la Facultad de Derecho y Ciencias Sociales y Políticas - UNNE, Vol. 4. Núm.2, 2025. E-ISSN 2953-4232
DOI: https://doi.org/10.30972/rcd.428685 - CC BY-NC