Versión 1.1 - 15 de diciembre de 2025

Marco normativo y estratégico para la adopción responsable de la **Inteligencia Artificial Generativa** en el Ministerio Público de la provincia de Buenos Aires

Ministerio Público
Provincia de Buenos Aires

SIED Secretaría de Innovación y Experiencia Digital

Versión 1.1 - 15 de diciembre de 2025

Marco normativo y estratégico para la adopción responsable de la **Inteligencia Artificial Generativa** en el Ministerio Público de la provincia de Buenos Aires

I. Introducción

La Inteligencia Artificial Generativa (IAGen) constituye una de las innovaciones más disruptivas de la actualidad. Se trata de una tecnología avanzada capaz de producir contenido en múltiples formatos —texto, imágenes, audio, entre otros—, con niveles crecientes de sofisticación y eficiencia, lo que la convierte en una herramienta sumamente versátil y de creciente relevancia. Su potencial transformador trasciende lo estrictamente tecnológico: según estimaciones recientes, podría generar un valor económico global de entre 2.6 y 4.4 billones de dólares (McKinsey, 2023), redefiniendo en los próximos años tanto el panorama económico como la estructura misma de los sistemas productivos y sociales.

Sin embargo, su adopción no está exenta de riesgos. En particular, surgen preocupaciones relacionadas con la posible difusión de desinformación, la exposición indebida de información sensible, la afectación de los derechos de propiedad intelectual (UNESCO, s.f.) y la amplificación de sesgos presentes en los datos de entrenamiento. Asimismo, su condición “fundacional” en-

traña la posibilidad de que surjan riesgos emergentes o de que se revelen otros aún no previstos en el estado actual del conocimiento.

En este marco, la incorporación de la IAGen en el ámbito del Ministerio Público de la Provincia de Buenos Aires (MPBA) abre una amplia ventana de posibilidades para fortalecer la labor judicial y optimizar el servicio de justicia. No obstante, su implementación exige un abordaje responsable, guiado por una visión humanista y sustentado en tres pilares: el establecimiento de un marco normativo que promueva un uso ético y seguro; el compromiso activo de los agentes del organismo en la utilización crítica y responsable de esta tecnología; y el desarrollo de una plataforma tecnológica interna de IAGen que fortalezca el control, la seguridad informática y su adecuada integración en los procesos del MPBA.

Sobre esta base se elaboró el presente marco normativo y estratégico para la adopción responsable de la IAGen en el MPBA, de cumplimiento obligatorio para todos los integrantes del organismo que, en el ejercicio de sus funciones, interactúen con esta tecnología en cualquier dispositivo o plataforma. Su finalidad es promover un uso equilibrado, ético y seguro, en consonancia con las leyes vigentes, la Constitución Nacional y Provincial, y los tratados internacionales de derechos humanos aplicables.

Asimismo, el documento ofrece a los agentes una comprensión preliminar de los fundamentos técnicos de la IAGen: repasa su evolución, introduce nociones clave y explica el valor de la ingeniería de *prompts* para orientar resultados útiles y seguros. Identifica los principales riesgos y sus implicancias para el ámbito judicial y, a partir de este diagnóstico, establece los principios que deben guiar su uso en el MPBA y las directrices operativas para un empleo ético, seguro y eficaz.

Finalmente, la Procuración General, a través de la Secretaría de Innovación y Experiencia Digital, presenta el Plan Estratégico de Implementación de IAGen en el MPBA, apoyado en el desarrollo de una plataforma propia —ChatIA— concebida como interfaz web para interactuar con modelos de inteligencia artificial bajo estrictos criterios de control y seguridad.

II. Fundamentos técnicos de la IAGen

En el presente capítulo se examinan, de manera sucinta, los elementos técnicos esenciales de la IAGen. El objetivo no es ofrecer un abordaje exhaustivo ni especializado, sino brindar a los integrantes del organismo una comprensión preliminar de cómo funcionan los modelos que sustentan esta tecnología. Si bien se trata de un campo en permanente evolución, contar con un conocimiento preliminar resulta indispensable para comenzar a interactuar con estas herramientas de manera crítica y consciente.

Para ello, se presenta una breve revisión de la evolución histórica de la IAGen, se introducen conceptos centrales como los modelos fundacionales y la arquitectura Transformer, y se describen las principales estrategias de inferencia que guían la generación de texto.

a) Orígenes del Aprendizaje Automático

Uno de estos pilares es el Aprendizaje Automático (*Machine Learning*), una subdisciplina de la Inteligencia Artificial que permite a los sistemas aprender o inferir patrones a partir de datos mediante un conjunto de instrucciones denominado “algoritmo”.

El término Aprendizaje Automático fue introducido por Arthur Samuel en 1959 para describir sistemas informáticos capaces de modificar su comportamiento en función de la “experiencia”, sin requerir una programación explícita para cada situación. Es decir, sistemas que replican procesos que, en humanos, serían considerados aprendizaje.

Cuando estos métodos de aprendizaje se implementan a través de redes neuronales profundas, pasan a formar parte del campo del Aprendizaje Profundo (*Deep Learning*), una subcategoría del Aprendizaje Automático que ha resultado decisiva para el desarrollo de la IAGen.

Estas redes están inspiradas en la estructura del cerebro humano y se componen de múltiples capas de procesamiento interconectadas, lo que les permite modelar relaciones complejas y no lineales entre los datos de entrada y salida. Cuanto mayor es la cantidad de capas intermedias, mayor es el poder de representación de la red, y por ende, su capacidad para abordar tareas de alta complejidad.

Las redes neuronales profundas no son simplemente una mejora técnica: son el cimiento sobre el cual se han construido los modelos generativos más avanzados disponibles en la actualidad.

b) Modelos fundacionales

La definición del concepto de "modelo fundacional" es relevante en este marco dado que la IAGen se sustenta precisamente en aquellos. Esta expresión, acuñada en 2021 por Bommasani et al., subraya su papel como cimientos arquitectónicos —análogos a las bases de una construcción—, que proporcionan una estructura para el desarrollo de aplicaciones futuras.

Su característica distintiva es que no están diseñados para una tarea única, sino que funcionan como una base flexible que puede adaptarse a múltiples aplicaciones: desde la comprensión y generación de lenguaje natural, hasta la creación de imágenes, el procesamiento de audio o el análisis de datos no estructurados. Para ello, son entrenados con decenas o incluso cientos de miles de millones de parámetros.

Sin embargo, la noción de modelos fundacionales también refleja las incertidumbres significativas que estos conllevan. En este sentido, los propios autores destacan que todavía no es posible determinar con claridad si la fundación sobre la que descansan puede considerarse plenamente confiable. Esto plantea desafíos críticos en términos de calidad, seguridad y confianza.

Bommasani et al. ilustran esta dualidad al señalar que "*un modelo fundacional es en sí mismo incompleto, pero sirve como la base común desde*

la cual se construyen muchos modelos específicos para tareas mediante adaptación" (2021, p. 7).

Sobre esta plataforma conceptual, resulta pertinente distinguir dos grandes clases de modelos fundacionales. Por un lado, los Modelos de Lenguaje de Gran Tamaño (*Large Language Model* o LLM), diseñados para procesar y generar texto. Por otro, los modelos multimodales, capaces de operar con distintos tipos de datos además del texto, como imágenes, audio o video[1].

Dentro de los LLM, existen modelos de código cerrado (como GPT, Gemini o DeepSeek), cuyos procesos de entrenamiento no suelen detallarse por razones de competencia, y modelos de código abierto (como LLaMA o Mistral), que permiten un mayor nivel de personalización.

c) Entrenamiento de LLM

Los LLM se destacan por su capacidad de predecir, en fracciones de segundo, cuál es la palabra o token[2] más probable que siga a otra dentro de una secuencia logrando producir textos que simulan razonamientos complejos.

Su rendimiento depende de múltiples factores entre los que se destacan el número de parámetros; la capacidad de cómputo; la cantidad, calidad y diversidad de los datos; la arquitectura empleada y su alineación ética.

El entrenamiento de un LLM comprende dos etapas principales. La primera es el pre-entrenamiento, cuyo objetivo es que el modelo aprenda patrones generales del lenguaje. Para ello, se lo expone a volúmenes masivos de texto que abarcan libros, artículos académicos, repositorios de código y una amplia variedad de recursos disponibles en línea. Este proceso incluye la lim-

1. Un ejemplo reciente de esta última categoría es Sora, el modelo de OpenAI concebido para generar videos a partir de instrucciones en lenguaje natural.

2. Sin perjuicio de que el término técnicamente correcto es "*token*", en el presente texto se lo utilizará de manera indistinta junto con "*palabra*", por considerarse un vocablo menos abstracto y, en ciertos pasajes, más adecuado para facilitar la comprensión del proceso de generación de texto en los LLM.

pieza de datos —para eliminar duplicados o información de baja calidad— y la selección de fuentes confiables que aseguren una base sólida de conocimiento.

Sin embargo, este insumo no es infinito: estimaciones recientes advierten que el stock efectivo de texto público generado por humanos podría agotarse entre 2026 y 2032 (Villalobos et al., 2024), lo que constituye un desafío central para el futuro del entrenamiento de esta tecnología.

Ahora bien, aprender el "lenguaje en general" no es suficiente: el modelo debe ajustarse para responder de manera útil, segura y alineada con las expectativas humanas. Aquí entra en juego la segunda etapa, el post-entrenamiento.

Durante esta segunda fase se aplican diversas técnicas. Una de ellas es el Ajuste fino supervisado (*Supervised Fine Tuning o* SFT), mediante el cual el modelo se entrena con ejemplos explícitos preparados por expertos, que incluyen tanto comportamientos deseados como indeseados. De esta manera, aprende a imitar las respuestas correctas y a evitar reproducir las incorrectas, ajustándose a los criterios definidos para el dominio o caso de uso específico.

A su vez, es complementado por el Aprendizaje por refuerzo con retroalimentación humana (*Reinforcement Learning From Human Feedback o* RLHF). A diferencia de la técnica anterior, en el RLHF las personas no redactan ejemplos de referencia, sino que eligen cuál entre distintas respuestas generadas por el LLM es más adecuada. Con esas comparaciones se entrena un modelo de recompensa, que luego guía, mediante una técnica de aprendizaje por refuerzo, el ajuste del LLM para producir respuestas más alineadas con las de los evaluadores humanos.

Más recientemente surgió la Optimización de Preferencias Directas (*Direct Preference Optimization* o DPO), que incorporó de manera directa los datos de preferencias en el proceso de optimización, simplificando y haciendo más eficiente el ajuste del modelo.

Por último, al integrar los LLM en sus operaciones, las organizaciones disponen de diversas técnicas para realizar un ajuste posterior al entrenamiento base, como el *fine-tuning* o la Generación Aumentada con Recuperación (*Retrieval-Augmented Generation o* RAG). Esta última técnica permite aumentar dinámicamente el contexto del modelo con información específica y actualizada de una organización.

d) Arquitectura Transformer

Entre las distintas variantes de redes neuronales profundas desarrolladas antes del surgimiento de la IAGen, una de las más relevantes fueron las redes neuronales recurrentes (*Recurrent Neural Networks* o RNN).

Las RNN están diseñadas para trabajar con secuencias de información, incorporando un estado interno que se actualiza en cada paso temporal. Este estado funciona como una "memoria" que permite que cada nuevo procesamiento considere no solo el dato ingresado en ese momento, sino también la información ya analizada, lo que las hizo especialmente útiles en tareas donde el orden de los elementos es decisivo, como la traducción automática, el reconocimiento de voz o la predicción de la palabra siguiente en un texto. En términos simples, las RNN "leen" la información paso a paso, reteniendo lo anterior para dar sentido a lo que viene después.

No obstante, las RNN tradicionales presentaban limitaciones para mantener información a largo plazo. Ante ello, surgió la arquitectura Memoria larga a corto plazo (*Long Short-Term Memory* o LSTM) como una evolución de las RNN, que incorpora mecanismos de control mediante compuertas que permiten regular cuánta información se retiene, se actualiza o se descarta en cada paso. Esto permitió manejar dependencias de largo alcance con mayor eficacia.

Ahora bien, aun con las mejoras introducidas por las LSTM, estas arquitecturas mantenían limitaciones importantes: su naturaleza secuencial impedía procesar en paralelo y evaluar de forma global todo el contexto de secuen-

cias extensas. Esto generaba además un cuello de botella, ya que toda la información debía atravesar un único canal paso a paso, ralentizando tanto el entrenamiento como la inferencia. A su vez, su memoria seguía siendo limitada, lo que dificultaba, por ejemplo, predecir con precisión la siguiente palabra cuando tenía que analizar otra palabra que se encontraba muy alejada en el texto.

Esta situación condujo al desarrollo de la arquitectura Transformer, diseñada por Vaswani et al. en 2017, que marcó un punto de inflexión en el procesamiento del lenguaje natural.

La estructura del Transformer se encuentra internamente conformada por dos grandes bloques, el *encoder* y el *decoder*. El *encoder* es la parte del modelo que procesa el texto de entrada a través de una serie de pasos encadenados, diseñados para que la máquina pueda "entender" mejor el significado de cada palabra en su contexto.

Primero, cada palabra o fragmento es convertida en un número, un proceso conocido como *tokenización*. Después, esos tokens se transforman en vectores, denominados *embeddings*, que son representaciones matemáticas entrenadas durante la fase de pre-entrenamiento. Estos embeddings no se calculan en cada uso, sino que forman parte de los parámetros del modelo.

Gracias a ellos, el sistema puede representar de manera matemática distintas propiedades lingüísticas de las palabras y calcular qué tan cercanas o relacionadas están las palabras entre sí y organizar una representación o una suerte de mapa del vocabulario que facilita la comprensión de los matices del lenguaje. Por ejemplo, probablemente, la palabra "imputado" se ubicará más cerca de "acusado" que de "contrato".

El siguiente paso es aplicar el mecanismo de atención, que permite procesar todas las palabras del texto de entrada en paralelo —a diferencia de las RNN, que las analizan de a una—. Para lograrlo, cada palabra se transforma en tres vectores distintos: consulta (*query)*, clave (key) y valor (value).

Los vectores consulta y clave se combinan para generar un puntaje de atención, que indica en el caso concreto cuál es la relevancia contextual de cada palabra respecto de las demás. Dicho de otro modo, en este caso no se mide la cercanía semántica global de las palabras (como hacen los *embeddings*), sino que se calcula qué tan conectadas están dentro de esa oración específica, en función de patrones sintácticos y semánticos que el modelo aprendió durante su entrenamiento.

Esto le permite, por ejemplo, relacionar palabras que aparecen distantes en el texto, pero que están contextualmente vinculadas. En la frase: *"El Ministerio Público goza de la autonomía e independencia que le otorga la Constitución para el debido cumplimiento de sus funciones"*, el modelo puede detectar fácilmente que *"cumplimiento de sus funciones"* se refiere al "*Ministerio Público"*, aunque esas palabras estén separadas por otras.

Una vez detectada esa relación, la información se transporta a través del vector valor, generando una nueva representación de cada palabra que incorpora el contexto de las demás. Esa representación es la que el modelo tiene en cuenta al construir el sentido de la oración de entrada. De esta manera, se supera la limitación de las RNN, ya que ahora cada palabra puede contextualizarse con cualquier otra de la frase, incluso si están muy alejadas entre sí.

Con posterioridad, el *encoder* aplica capas adicionales —como normalización, conexiones residuales y transformaciones lineales— que refinan estas representaciones, dotándolas de mayor consistencia, fidelidad y profundidad.

El segundo bloque del Transformer es el *decoder*, la parte del modelo encargada de generar el texto de salida. Su funcionamiento se basa en calcular la atención entre las palabras que ya fueron generadas, sin acceder a las que aún no se han producido. Esta restricción, conocida como "enmascaramiento", es fundamental para que el modelo escriba de manera secuencial, prediciendo una palabra por vez, al igual que lo haría una persona al redactar. Así, si el *decoder* ya produjo "El" y "Ministerio", todavía no puede anticipar las pa-

labras que siguen. O sea, aunque los cálculos internos se realicen en paralelo gracias a la arquitectura Transformer, la lógica de generación permanece secuencial: cada nueva palabra depende necesariamente de las anteriores.

Luego, el *decoder* incorpora un mecanismo de atención cruzada que le permite "mirar" no solo lo que ya escribió, sino también el texto original procesado por el *encoder*. De este modo, puede mantener la coherencia entre la entrada y la salida, asegurando que lo generado guarde relación directa con el mensaje original.

A continuación, el *decoder* transforma esa información en una representación numérica que recorre todo el vocabulario posible. En cada paso, el modelo calcula cuál es la palabra más probable que debería seguir, tomando como referencia lo que ya generó previamente. Este procedimiento se repite de manera iterativa, generando una palabra por vez, hasta alcanzar la longitud máxima establecida o producir una palabra especial de fin de secuencia.

Finalmente, antes de generar cada palabra, el *decoder* realiza una serie de operaciones adicionales que refinan las representaciones y aseguran mayor coherencia en la salida.

e) Estrategias de inferencia o decodificación en la fase de generación

Las estrategias de decodificación que se utilizan para estimar una distribución de probabilidad sobre todas las palabras posibles y seleccionar la siguiente son un aspecto fundamental de los LLM.

Estas estrategias pueden ser deterministas, lo que significa que la misma entrada siempre produce el mismo resultado, o estocásticas, lo que incorpora cierto grado de aleatoriedad.

Algunos de las estrategias más utilizadas son las siguientes:

- El algoritmo *Greedy Search* es una técnica determinista que siempre elige el token con mayor probabilidad en cada paso. Su principal ventaja es la simplicidad y rapidez, aunque puede conducir a soluciones que no son óptimas porque no considera alternativas que, en conjunto, podrían dar lugar a secuencias más coherentes o naturales.

- El *Beam Search* también es un algoritmo determinista, pero, en lugar de limitarse al token con mayor probabilidad en cada paso, mantiene en paralelo los candidatos más prometedores (beams). Para cada candidato, evalúa la probabilidad acumulada de la secuencia completa hasta ese punto. Esto significa que no se toma únicamente la probabilidad del token actual, sino que se combina con las probabilidades de todos los tokens anteriores, de manera que cada secuencia candidata se puntúa en función de su coherencia global y no solo de la última elección. Como contrapartida, requiere mayor poder de cómputo y memoria.

- El muestreo *Top-K* selecciona el siguiente token de manera aleatoria dentro de un conjunto preestablecido y limitado de candidatos: los “k” más probables según la distribución. Esto lo convierte en una técnica estocástica, ya que con la misma entrada no siempre se obtiene el mismo resultado. Sin embargo, si el valor de “k” se limita a un candidato, el método se reduce al *Greedy Search*, volviéndose determinista.

- El muestreo *Top-P* no fija un número de tokens, sino que selecciona dinámicamente el conjunto de candidatos cuya probabilidad acumulada alcanza un umbral predeterminado, comenzando por los más probables. Luego, el modelo elige aleatoriamente entre esos tokens. Dicho de otro modo, el siguiente token se selecciona al azar dentro del “núcleo” de opciones que suman la probabilidad establecida. En casos donde la distribución está muy concentrada —por ejemplo, si un token supera por sí solo el umbral—, el conjunto se reduce a una sola opción, y el resultado se vuelve determinista de hecho.

- La temperatura es un parámetro que regula el nivel de aleatoriedad con el que el modelo genera texto, tanto si se aplica *Top-K* como *Top-P*. Este parámetro se aplica antes del muestreo para modificar la forma de la distribución de probabilidades de los tokens, pudiendo configurarla para que sea más concentrada o más plana.
    - Con una temperatura baja, las probabilidades de los tokens más probables se amplifican, lo que hace que el modelo sea más predecible.
    - Con una temperatura alta, las diferencias entre tokens se reducen, lo que abre el abanico de posibilidades y vuelve la respuesta más variada, aleatoria o creativa.

f) Ingeniería de *Prompts*

El *prompt* constituye la interfaz textual mediante la cual los usuarios transmiten sus deseos o instrucciones al modelo de IAGen (Amatriain, 2024). En términos técnicos, en lo aquí interesa, se trata de la entrada que recibe el *encoder* del LLM, por lo que es el punto de partida que orienta cómo el sistema procesa la información y qué tipo de salida producirá.

Schulhoff et al. (2025) señalan que existen diversas clases de *prompt*: la directiva, que se formula como instrucción o pregunta —incluso de manera implícita—; el ejemplo, que actúa como demostración para guiar al modelo en la tarea; el formato de salida, que define la estructura en que debe presentarse la respuesta; y las instrucciones de estilo, orientadas a ajustar el tono y la forma del texto. Asimismo, el rol o persona aporta una perspectiva específica desde la cual el modelo debe responder, mientras que la información adicional incorpora datos complementarios o de contexto que permiten obtener resultados más precisos, relevantes y adecuados.

Asimismo, la Ingeniería de *prompts* es la disciplina orientada a diseñar y optimizar instrucciones para guiar al modelo hacia respuestas más útiles y ali-

neadas con los objetivos del usuario. La formulación de la entrada resulta decisiva para obtener resultados precisos, coherentes y relevantes (Schulhoff et al., 2025). La calidad de esas respuestas depende, en gran medida, de cuatro factores clave: comprender las capacidades y limitaciones de la IAGen; considerar el contexto en el que se aplica; contar con conocimiento del dominio específico de la consulta; y aplicar un enfoque metódico que permita adaptar los *prompts* a diferentes situaciones (Amatriain, 2024).

La Guía de Ingeniería de *Prompt* de Open AI (2025) recomienda ser específico y conciso, usar comandos directos, definir el formato o estilo de salida y, de ser pertinente, incluir ejemplos relevantes. Del mismo modo, es necesario balancear detalle y claridad, adaptando el *prompt* al contexto y refinándolo de manera sistemática hasta lograr la respuesta más adecuada. Por otro lado, también es posible configurar los parámetros previamente descriptos para mejorar la calidad de la respuesta del modelo.

g) Conclusión preliminar

En suma, la IAGen se apoya en modelos fundacionales, que, a través del pre y post-entrenamiento, aprenden primero patrones generales del lenguaje y luego se ajustan a las expectativas humanas mediante técnicas como SFT, RLHF y DPO. Su desempeño depende de la interacción entre datos, cómputo y parámetros, articulados a través de la arquitectura Transformer. En la fase de inferencia, las estrategias de decodificación permiten modular la aleatoriedad y coherencia de los textos generados. En conjunto, los LLM representan el núcleo de la IAGen y explican su potencial disruptivo.

III. Riesgos asociados a la IAGen

Los LLM presentan diversos riesgos legales, éticos y sociales que requieren ser identificados y mitigados para garantizar un uso seguro y responsable de la IAGen (Bender et al., 2021; Bommasani et al., 2021; Dinan et al., 2021; Kenton et al., 2021). A continuación, se destacan algunos de ellos:

a) Desinformación

Como se ha señalado anteriormente, los LLM se encuentran diseñados para predecir la siguiente palabra en un contexto dado, basándose en la probabilidad estadística según patrones aprendidos en los datos de entrenamiento. Aunque este enfoque es altamente eficaz para muchas tareas, también puede llevar a generar respuestas conocidas como “alucinaciones”.

El inglés ha incorporado las nociones vinculadas con la IAGen con mayor rapidez que otros idiomas[3]. En esa línea, el diccionario norteamericano *Merriam-Webster* define la “alucinación” como una respuesta que, aunque parece plausible, resulta incorrecta o engañosa y es producto de un algoritmo de inteligencia artificial.

Así como ocurre con las alucinaciones humanas, estas respuestas pueden ser difíciles de identificar como falsas a primera vista[4]. Estas “alucinaciones” reflejan una limitación inherente de los LLM: su capacidad para generar texto no garantiza la veracidad de las afirmaciones.

Algunas de las causas detrás de este fenómeno incluyen la calidad de los datos de entrenamiento, los patrones estadísticos utilizados, la ambigüedad léxica y la falta de comprensión contextual (Weidinger et al., 2021). Es decir, la capacidad de los LLM para ofrecer respuestas objetivas y precisas depende de múltiples factores. Entre ellos, se destacan la calidad de los datos utilizados durante su entrenamiento, su habilidad para interpretar los *prompts* proporcionados por los usuarios, la precisión de los *prompts* y su capacidad para establecer correlaciones relevantes entre los datos. Además, posibles limitaciones en los mecanismos de atención pueden llevar a errores en la generación de contenido, como la omisión o malinterpretación de información clave.

3. Hasta el momento, el Diccionario de la Real Academia Española no ha incorporado esta acepción del término “alucinación”.

4. Tal como señalaron Boyd & Crawford (2012) respecto del Big Data, los grandes volúmenes de datos pueden generar la percepción errónea de que sus resultados son siempre objetivos, verdaderos o precisos.

En este contexto, la desinformación podría ser también producto de una vulnerabilidad informática. Según el NIST (2024), uno de los principales riesgos de ciberseguridad en sistemas de IAGen es el aumento de la superficie de ataque, lo que facilita prácticas maliciosas como la inyección de *prompts* o el envenenamiento de datos. Este último implica manipular los *prompts* para alterar las respuestas o el funcionamiento del modelo[5] (por ejemplo, para obtener información confidencial) o introducir "datos envenenados" para producir resultados erróneos, amplificando los riesgos asociados a las salidas generadas por el sistema.

b) Recolección y uso de datos

Como se mencionó anteriormente, los LLM se entrenan utilizando vastos conjuntos de datos que contienen miles de millones de parámetros. Este enfoque, basado en la premisa de que una mayor cantidad de datos mejora la calidad del entrenamiento (Kaplan et al., 2020), impulsa la recopilación de datos de manera continua para optimizar la utilidad y el rendimiento de sus productos.

Una fuente de información para este propósito puede ser el usuario. En particular, en las versiones gratuitas, los datos proporcionados por los usuarios son utilizados por algunas plataformas para entrenar y mejorar los modelos, lo que genera un riesgo potencial de exposición de información confidencial o sensible.

Al respecto, la consultora Gartner (s.f.) sostiene que: "*Los usuarios deben asumir que cualquier dato o consulta que ingresen en ChatGPT o en sus competidores podría convertirse en información pública (...)*".

Este riesgo no solo afecta a los usuarios individuales, sino que también plantea serias amenazas para las organizaciones que implementan estas herra-

5. Un ejemplo de ello es el *jailbreak* de IA, que consiste en la manipulación del sistema para forzarlo a ejecutar acciones que se encuentran restringidas.

mientas. La posibilidad de que información confidencial sea revelada o explotada maliciosamente representa un desafío crítico.

Un ejemplo ilustrativo es el caso de empleados de Samsung, quienes utilizaron una reconocida herramienta de IAGen para solucionar problemas con su código fuente, introduciendo datos sensibles que posteriormente fueron filtrados públicamente (Ray, 2023). Este incidente llevó a Samsung a restringir temporalmente el uso de dicha herramienta, adoptando medidas adicionales para garantizar un entorno más seguro en futuras interacciones.

Además, si los datos de entrenamiento contuvieran material protegido por derechos de propiedad intelectual, los LLM podrían generar contenido que vulnere dichos derechos. En ese escenario, el uso de la IAGen podría facilitar la reproducción no autorizada de contenido protegido (derechos de autor, patentes, entre otros), amplificando los riesgos para los derechos de propiedad intelectual.

Por otra parte, el volumen masivo de datos no asegura la diversidad ni la representatividad de la información (Bender et al., 2021). Los sesgos inherentes a los datos utilizados para entrenar modelos fundacionales tienden a reflejarse en las respuestas generadas, lo que podría derivar en la perpetuación de estereotipos sociales y prácticas de discriminación injusta. La escala y el alcance de los modelos fundacionales amplifican considerablemente el impacto de este riesgo, especialmente debido a su uso generalizado y a la complejidad de las tareas que son capaces de abordar.

Este tipo de riesgos plantea un desafío especialmente complejo en el ámbito judicial, donde la información suele involucrar datos sensibles cuya divulgación indebida podría poner en riesgo derechos fundamentales como la vida, la intimidad o el honor de las personas.

Por último, otro aspecto relevante a considerar es que las prácticas de las herramientas de IAGen podrían, eventualmente, entrar en conflicto con la regulación nacional de protección de datos personales.

c) Otros riesgos

La naturaleza de esta tecnología conlleva la probabilidad de que se registren riesgos emergentes o de que se manifiesten otros aún no previstos en el estado actual del conocimiento. Se trata de un campo en constante evolución, con una dinámica compleja y cambiante que dificulta anticipar de manera exhaustiva todas sus posibles implicancias.

IV. Principios que rigen el uso de IAGen en el MPBA

En este apartado se establecen los principios que regulan el uso de IAGen en el MPBA. Su fundamento parte de la premisa de que los usuarios deben ejercer una responsabilidad activa y consciente, ponderando las consecuencias sociales, éticas y legales de sus decisiones y evitando toda subordinación acrítica frente a la tecnología (Dicasterio para la Doctrina de la Fe & Dicasterio para la Cultura y la Educación, 2025).

Estos principios son de observancia obligatoria para todos los agentes del MPBA y actúan como criterios rectores destinados a asegurar que el empleo de la IAGen se mantenga en plena consonancia con los valores institucionales, los derechos fundamentales y el marco jurídico vigente.

- **Dignidad**: El uso de IAGen deberá regirse por el respeto irrestricto de la dignidad humana y del bien común.

- **Responsabilidad:** Su utilización no exime de las obligaciones funcionales, legales y éticas vigentes. Los agentes del MPBA serán plenamente responsables de asegurar en todos los casos su adecuación al Código de Ética (Res. PG N.° 32/19), así como a la Constitución, las leyes y los tratados de derechos humanos aplicables.

- **Supervisión Humana:** La utilización de IAGen deberá estar sujeta a un control humano activo y permanente. Queda prohibida la delegación de decisiones sustantivas en la herramienta.

- **Confidencialidad de la información:** Deberá garantizarse la protección de los datos personales (Ley N.° 25.326), la preservación de la confidencialidad y el resguardo de los datos sensibles, evitando cualquier tratamiento que comprometa derechos fundamentales o la legalidad de los procesos.

- **Prohibición de contenidos inapropiados:** Queda vedada toda utilización que derive en la generación o difusión de contenidos ofensivos, discriminatorios, inapropiados o ilegales.

- **Ciberseguridad:** La utilización de IAGen deberá sustentarse en la preservación de la seguridad informática y la protección de la información institucional, garantizando la integridad, disponibilidad y confidencialidad de los datos en todo momento.

- **Uso exclusivo de la plataforma institucional**: La interacción con sistemas de IAGen deberá realizarse exclusivamente a través de *ChatIA*, la plataforma oficial desarrollada por el MPBA. Queda prohibida la utilización de herramientas externas o no autorizadas, a fin de garantizar el control, la seguridad y la trazabilidad de la información institucional.

- **Precaución frente a Riesgos Emergentes**: El empleo de IAGen deberá regirse por el principio de precaución, anticipando y limitando los impactos derivados de riesgos emergentes o aún no previstos, propios de la naturaleza evolutiva de esta tecnología.

V. Plan estratégico de implementación de IAGen en el MPBA

La implementación de la IAGen en el MPBA se concibe como un proceso estratégico y progresivo, sustentado en un enfoque humanista que reconoce a las personas como centro de la organización. Su finalidad es optimizar el desempeño institucional, fortaleciendo las capacidades de los agentes y orientando sus funciones hacia actividades de mayor valor agregado.

En este marco, la Procuración General, a través de la Secretaría de Innovación y Experiencia Digital (SIED), desarrolló el sistema "ChatIA", concebido como la única interfaz del MPBA autorizada para que los agentes judiciales interactúen con IAGen.

El sistema integra, en un entorno seguro, múltiples capacidades: gestión centralizada de múltiples modelos de IAGen, administración de usuarios, control de acceso basado en roles, monitoreo y auditoría de las interacciones, encriptación de comunicaciones y gestión presupuestaria de tokens para modelos externos.

Asimismo, el sistema se encuentra integrado localmente con el modelo Ollama, lo que permite mayor autonomía operativa. Considerando el carácter estratégico de la información institucional, la política de adopción de IAGen promoverá el uso prioritario de modelos locales.

La estrategia de implementación se desarrollará en tres etapas sucesivas:

Primera etapa: prueba piloto

La experiencia inicial se llevará a cabo con los integrantes de la Comisión de Inteligencia Artificial del MPBA, bajo la coordinación de la Secretaría de Innovación y Experiencia Digital. Dicha Comisión se encuentra integrada por un representante de cada una de las siguientes áreas o dependencias: Fiscalía ante el Tribunal de Casación, Fiscalías de Cámara de Bahía Blanca y Mercedes, Defensoría Departamental de Moreno - General Rodríguez y Asesoría de Menores e Incapaces N° 2 del Departamento Judicial Lomas de Zamora.

El alcance de la prueba piloto incluirá la experimentación con ChatIA en tareas de apoyo al agente judicial, tales como la redacción, edición y resumen de documentos. A su vez, se pondrá en marcha durante esta etapa una solución destinada a facilitar la búsqueda de precedentes judiciales relevantes en el ámbito de la Fiscalía ante el Tribunal de Casación. El objetivo de la prue-

ba piloto será evaluar la pertinencia, seguridad y utilidad de la herramienta en entornos controlados.

En paralelo, se prevé la incorporación también bajo la modalidad de prueba piloto de IAGen al asistente virtual "Vicky", implementado en 2019 como el primer sistema de este tipo en el Poder Judicial a nivel nacional. "Vicky" brinda información permanente sobre los derechos de las personas y las funciones del Ministerio Público, y fue concebido con el fin de mejorar la calidad de la atención y ofrecer un servicio más cercano y efectivo. A partir de octubre del presente año, su funcionamiento se reforzará mediante la integración de capacidades de IAGen. Para un aprovechamiento adecuado del mismo, los usuarios deberán tener presente este documento.

Profundización y ampliación de la prueba piloto

En la segunda fase de la prueba piloto, cada área interviniente podrá explorar al menos un caso de uso específico definido en función de sus necesidades funcionales, características operativas y tipo de intervención institucional.

En el marco de la ampliación del alcance, se habilitará de manera gradual y bajo condiciones estrictamente controladas la posibilidad de trabajar con información vinculada a causas en trámite, exclusivamente a través del Modelo de Lenguaje de Gran Tamaño (LLM) interno alojado en la infraestructura tecnológica institucional.

A su vez, se propiciará el diseño de una biblioteca institucional de prompts seguros o "potentes" para optimizar el uso de la IAGen y mitigar el riesgo de posibles alucinaciones. Además, con el objetivo de alinear las expectativas de los usuarios, se incorporarán recomendaciones o buenas prácticas por cada clase de Modelo de Gran Tamaño.

En el mediano plazo, se prevé avanzar en la integración progresiva de ChatIA con los sistemas institucionales del MPBA a fin de facilitar flujos de trabajo asistidos por IAGen. Las eventuales integraciones se desarrollarán de mane-

ra gradual, conforme a la madurez de los casos de uso identificados, la evaluación de riesgos y la factibilidad técnica.

Segunda etapa: expansión progresiva

Una vez verificada la confiabilidad y precisión en el piloto, se ampliará el uso de ChatIA a otros agentes del MPBA, extendiéndose eventualmente a casos de uso adicionales. Esta fase se acompañará de instancias de capacitación para los usuarios y de protocolos de supervisión.

Tercera etapa: evaluación y retroalimentación institucional.

Se implementará un sistema de monitoreo continuo para medir el impacto de la IAGen en el desempeño del MPBA, a partir de indicadores de éxito y de riesgo. En particular, se evaluará: (i) la eficiencia operativa —nivel de adopción y tiempo promedio de trabajo liberado—; (ii) la calidad del trabajo —tasa de correcciones posteriores y nivel de satisfacción de los usuarios—; y (iii) la seguridad —incidentes reportados y detección de sesgos o salidas problemáticas—. Los resultados serán analizados por la SIED, con la intervención del Departamento de Control de Gestión de la Secretaría de Desarrollo y Seguimiento del Proyecto Institucional, a fin de garantizar la mejora continua del proyecto.

Finalmente, la SIED podrá considerar la incorporación de nuevos modelos o el diseño de agentes adicionales dentro de los ya existentes, en función de la experiencia acumulada y las necesidades institucionales.

VI. Directrices para el uso de IAGen en el MPBA

En consonancia con los principios previamente enunciados, y considerando las características técnicas de la plataforma tecnológica adoptada por el MPBA, se establecen las siguientes directrices de uso obligatorio:

a) Acceso y uso general de ChatIA

1. Uso de la plataforma: La IAGen deberá emplearse únicamente en los casos de uso autorizados y siempre dentro del marco de las funciones laborales del agente. Queda expresamente prohibido todo uso con fines personales o ajenos a las competencias institucionales. Asimismo, los agentes deberán evaluar la pertinencia de su aplicación en función de la complejidad, sensibilidad y finalidad de la tarea, así como los casos para los cuales el modelo ha sido entrenado, privilegiando su utilización solo cuando aporte un valor agregado respecto de métodos tradicionales.

2. Acceso autenticado: El acceso a ChatIA se realizará exclusivamente mediante autenticación centralizada y bajo el rol asignado a cada usuario. Queda prohibido el uso de credenciales de terceros o eludir los mecanismos de control de acceso.

3. Gestión responsable de recursos: Cada usuario será responsable del uso de los tokens presupuestados, debiendo administrar su consumo conforme a criterios de eficiencia y proporcionalidad.

4. Auditoría: Los contenidos generados en ChatIA podrán ser auditados.

b) Sobre la información de entrada

5. Protección de la información: Se prohíbe incorporar: (i) información de causas en trámite; (ii) datos sensibles o anonimizados cuando exista riesgo de reidentificación; (iii) imágenes personales; (iv) claves, credenciales personales, tokens u otra clase de información privada; (v) material protegido por derechos de propiedad intelectual, salvo que medie licencia, autorización o excepción legal aplicable; y (vi) cualquier otro contenido que comprometa la confidencialidad, la seguridad o la legalidad.

La información ingresada deberá ser estrictamente aquella que sea necesaria para la tarea. En línea con ello, la opción de adjuntar archivos se encuentra deshabilitada.

Cuando se requieran ejemplos o formatos de salida para orientar a la IAGen, deberá utilizarse información ficticia o genérica, o datos anonimizados de manera robusta que impidan toda reidentificación.

Se debe tener en cuenta que los datos ingresados podrían ser almacenados, procesados o, en algunos casos, compartidos por la IAGen, lo que aumenta el riesgo de exposición no deseada.

La Secretaría de Innovación y Experiencia Digital podrá evaluar su incorporación, como un nuevo caso de uso, siempre dentro de la infraestructura institucional y bajo estrictas condiciones de seguridad.

6. Ingeniería de *Prompts*: Los *prompts* deberán formularse de manera clara, específica y proporcional a la finalidad institucional perseguida, evitando expresiones ambiguas o excesivamente generales que puedan inducir a errores, sesgos o alucinaciones en los resultados. Cuando corresponda, se deberá incluir información relevante sobre la tarea, precisiones de contexto, extensión y tono de la respuesta, así como ejemplos o formatos deseados. En las tareas complejas, se deberá privilegiar el uso de estrategias de inferencia deterministas (ver inciso “e” del punto II). Los usuarios deberán ajustar iterativamente los *prompts* hasta obtener resultados adecuados y mantenerse actualizados en buenas prácticas de ingeniería de prompting.

c) Sobre la salida de la herramienta

7. Validación de resultados: Toda salida deberá ser verificada de manera efectiva por el usuario, garantizando objetividad, ética y precisión. Se deberá poner especial atención a riesgos de alucinaciones y sesgos propios de los datos y del diseño del modelo.

Por tanto, el agente deberá contar con conocimiento suficiente para interpretar los resultados, contrastarlos con fuentes confiables y realizar una revisión exhaustiva antes de su aplicación. En ningún caso la salida de la IAGen podrá ser utilizada como fundamento único para la adopción de decisiones sustantivas, sin validación humana independiente. Por regla, los resultados deberán ser contrastados preferentemente con fuentes oficiales o de reconocida reputación, sin que ello releve al usuario de efectuar un control exhaustivo e independiente. Siempre que sea técnicamente posible, deberán examinarse los procesos y mecanismos empleados por la herramienta para generar la información. La responsabilidad del uso adecuado de la IAGen y la decisión final de incorporar una respuesta generada por ChatIA recaen siempre en el usuario.

8. Uso Ético y no discriminatorio: En caso de detectar la generación de un contenido ofensivo, discriminatorio, inapropiado o ilegal, el usuario estará obligado a descartarlo de inmediato y reportar el incidente a la Subsecretaría de Informática o a su delegación departamental, según corresponda.

9. Revisión de Sesgos y Calidad: En la validación de resultados, los usuarios deberán prestar especial atención a la detección de sesgos, omisiones relevantes o errores de coherencia que pudieran comprometer la equidad, la precisión o la objetividad del contenido generado.

d) Seguridad informática y gestión de riesgos

10. Ciberseguridad: La IAGen deberá emplearse exclusivamente en entornos que garanticen medidas adecuadas de ciberseguridad y resguardo de la información. Los agentes deberán observar en todo momento las buenas prácticas en la materia, siendo obligatorio informar de inmediato cualquier incidente a la Subsecretaría de Informática o a su delegación departamental, según corresponda.

11. Adecuada gestión de riesgos emergentes: Todo comportamiento anómalo, salida inusual o riesgo emergente detectado deberá informarse de in-

mediato a la Subsecretaría de Informática o a su delegación departamental, según corresponda.

e) Capacitación y mejora continua

12. Aprendizaje continuo: El uso de la IAGen exige actualización permanente y participación en instancias de capacitación sobre su utilización ética, segura y eficiente. Los agentes deberán mantenerse informados acerca de las mejores prácticas y consultar de manera periódica recursos especializados, como el *AI Risk Atlas* de IBM[6] u otros equivalentes, para identificar y mitigar riesgos. La falta de conocimiento técnico no exime de responsabilidad ética ni legal.

13. Actualización del documento: El presente documento podrá ser objeto de revisiones y actualizaciones por la SIED. Cada modificación dará lugar a una nueva versión numerada, cuya referencia constará en el encabezado. A todos los efectos, se considerará vigente únicamente la última versión publicada en Confluence, informada a través del SIMP Portal.

f) Declaración de conocimiento y aceptación

14. Declaración obligatoria: Todo agente judicial que acceda y utilice el sistema ChatIA deberá suscribir una declaración expresa de conocimiento y aceptación del presente documento, condición necesaria para el uso de la plataforma.

VII. Bibliografía

- Amatriain, X. (2024). Prompt Design and Engineering: Introduction and Advanced Methods. arXiv. https://arxiv.org/abs/2401.14423
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmit-

6. https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas

chell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. Association for Computing Machinery.

- Biden, J. R. Jr. (2023, October 30). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. The White House.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., … Liang, P. (2021). *On the opportunities and risks of foundation models*. Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University. arXiv. https://arxiv.org/abs/2108.07258.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.
- Courts and Tribunals Judiciary. (2023, diciembre 12). *Artificial intelligence (AI): Guidance for judicial office holders*. Reino Unido.
- Corte Constitucional de Colombia. (2024). *Sentencia T-323 de 2024*. Sala Segunda de Revisión. Expediente T-9.301.656.

- Dicasterio para la Doctrina de la Fe & Dicasterio para la Cultura y la Educación. (2025, enero 28). *ANTIQUA ET NOVA: Nota sobre la relación entre la inteligencia artificial y la inteligencia humana*. Vaticano. https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_sp.html

- Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. (2021). *Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling*. arXiv. https://arxiv.org/abs/2107.03451.
- Delua, J. (2021, March 12). Supervised versus unsupervised learning: What's the difference?. IBM. https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning.
- European Commission for the Efficiency of Justice (CEPEJ). (2024). *Use of generative artificial intelligence (AI) by judicial professionals in a work-related context*. CEPEJ Working Group on Cyberjustice and Artificial Intelligence (CEPEJ-GT-CYBERJUST). Estrasburgo, Francia.
- Gartner. (s.f.). Gartner experts answer the top generative AI questions for your enterprise. https://www.gartner.com/en/topics/generative-ai.
- Houde, S., Liao, V., Martino, J., Muller, M., Piorkowski, D., Richards, J., Weisz, J., & Zhang, Y. (2020). *Business (mis)use cases of generative AI*. arXiv. https://arxiv.org/abs/2003.07679.
- IBM. (s.f.). Machine Learning. https://www.ibm.com/es-es/topics/machine-learning.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models*. arXiv. https://arxiv.org/abs/2001.08361.

- Llama. (s.f.). *How-to guides: Prompting*. https://www.llama.com/docs/how-to-guides/prompting/.
- McKinsey (2023). The economic potential of generative AI. The next productivity frontier.
- Merriam-Webster. (s.f.). Hallucination. In *Merriam-Webster.com* dictionary. https://www.merriam-webster.com/dictionary/hallucination.

- National Institute of Standards and Technology. (2024). *Trustworthy and responsible AI: NIST AI 600-1. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. U.S. Department of Commerce.
- OpenAI. (2025). *Ingeniería de Prompt*. https://www.promptingguide.ai/es
- OpenAI. (s.f.). *Best practices for prompt engineering with the OpenAI API: How to give clear and effective instructions to OpenAI models*. OpenAI Help Center. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api.
- Ray, S. (2023, mayo 2). *Samsung Bans ChatGPT Among Employees After Sensitive Code Leak*. Forbes Middle East. https://www.forbesmiddleeast.com/innovation/artificial-intelligence-machine-learning/samsung-bans-chatgpt-among-employees-after-sensitive-code-leak
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal*, *July*.
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., Da Costa, H., Gupta, S., Rogers, M. L., Goncearenco, I., Sarli, G., Galynker, I., Peskoff, D., Carpuat, M., White, J., Anadkat, S., Hoyle, A., & Resnik, P. (2025). The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. arXiv. https://arxiv.org/pdf/2406.06608.
- UNESCO. (s.f.). Cómo gobernar correctamente la IA. UNESCO. https://www.unesco.org/es/forum-ethics-ai?hub=32618.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023).

Attention is all you need. En *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*.

- Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). *Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data*. EPOCH AI. https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from language models*. arXiv. https://arxiv.org/abs/2112.04359v1.