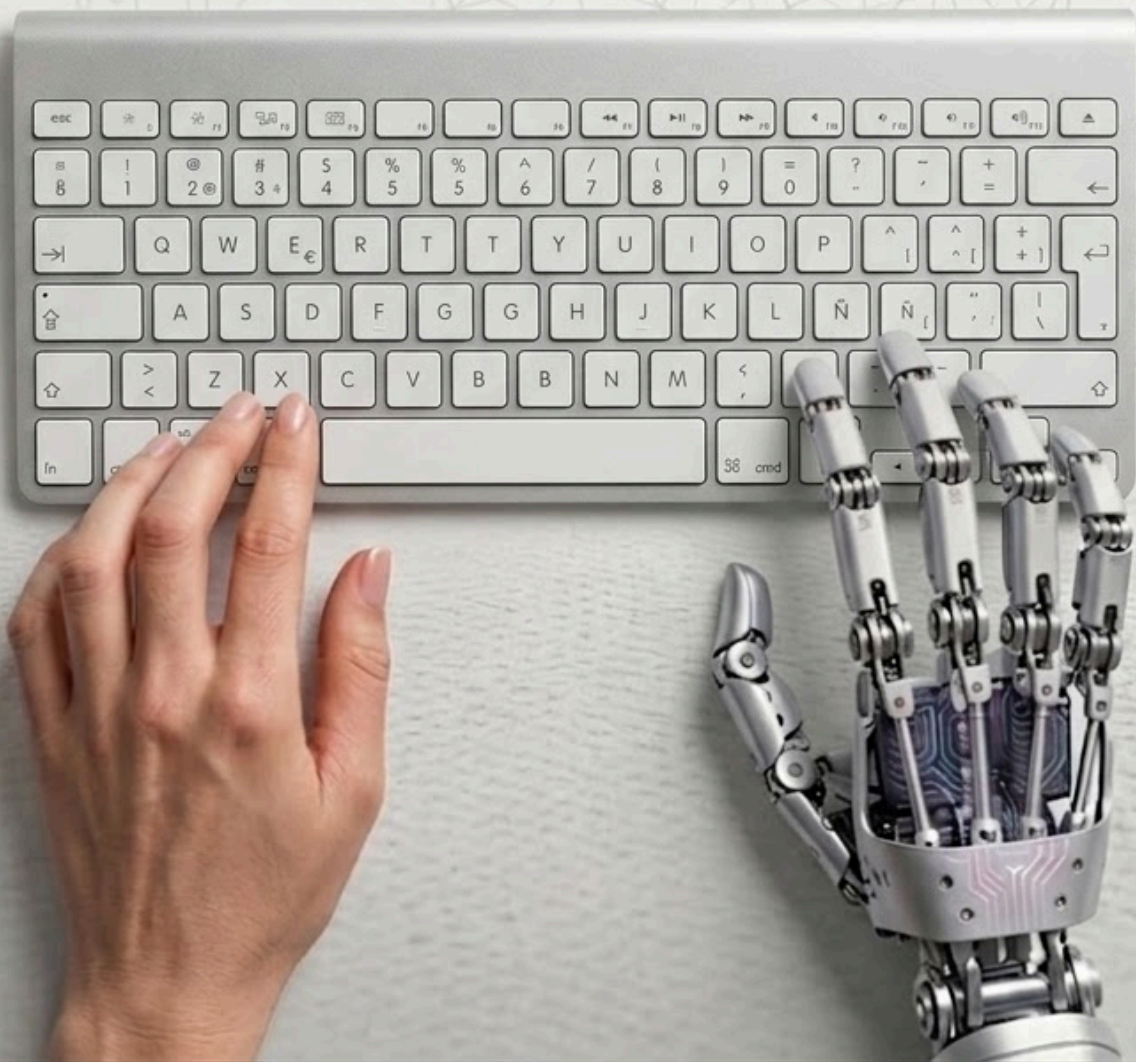




# ¿Lo escribió un humano o la IA?

Evaluación de autoría en textos generados o asistidos por inteligencia artificial





## **Dirección General**

Juan G. **Corvalán**

## **Autoría**

Mariana **Sánchez Caparrós**,  
Martina **Nuccitelli**

## **Investigación**

Carolina **Martín**, Gisel **Alvarado**,  
Lola **Ramos Pereyra**

## **Diseño**

Sofía **Rolleri**



IALAB

# Contenido

<b>Resumen Ejecutivo</b>	<b>03</b>
<hr/>	
<b>A. Introducción</b>	<b>05</b>
<hr/>	
<b>B. Marco de evaluación multinivel</b>	<b>06</b>
1. Enfoque general	06
2. Dimensiones del marco de validación	06
<b>NIVEL 1: Detección algorítmica automatizada (uso instrumental)</b>	<b>07</b>
Estado del arte de la detección algorítmica automatizada	08
Clasificación binaria en entornos controlados	08
Persistencia de huellas estilísticas	09
Degradación de la detección ante intervención humana	09
Evaluación exhaustiva con modelos de vanguardia	10
Conclusiones del estado del arte	11
Evaluación de herramientas realizada por IALAB	12
Primera fase: detección sobre textos originales (100 inputs)	13
Segunda fase: textos humanizados y revalidados (67 inputs)	14
Análisis comparativo e interpretación	14
<b>NIVEL 2: Trazabilidad técnica - metadatos y marcas de agua</b>	<b>15</b>
Alcance del análisis técnico y principio probatorio	16
Metadatos documentales: naturaleza, categorías y persistencia	16
Limitaciones del nivel	16

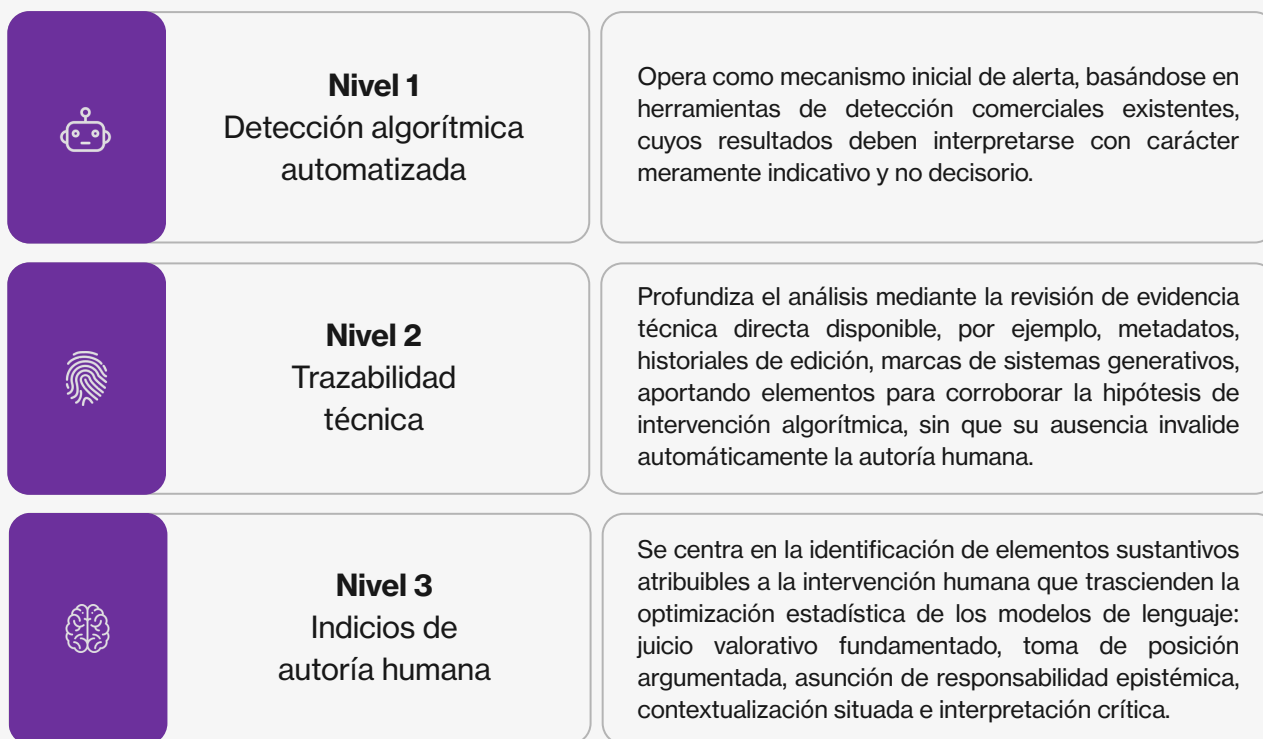
<b>NIVEL 3: Identificación de huellas humanas irreductibles</b>	<b>18</b>
Rendición de cuentas, juicio situado y legitimidad decisoria en contextos de escritura asistida por IA	18
Enfoque metodológico	18
Dimensiones de análisis	19
A. Juicio evaluativo explícito	19
B. Toma de posición frente a alternativas	19
C. Asunción de responsabilidad	20
D. Contextualización situada	20
E. Grado de interpretación	20
F. Experiencia humana o institucional reconocible	21
G. Ética aplicada al caso concreto	21
H. Creatividad funcional orientada a un fin	21
<hr/>	
<b>C. Reflexiones finales</b>	<b>22</b>
<hr/>	
<b>D. Alcances, limitaciones y líneas de trabajo futuro</b>	<b>24</b>
<hr/>	
<b>E. Cierre</b>	<b>26</b>
<hr/>	
<b>Anexo I - Matriz de evaluación - Nivel 3: Huellas humanas irreductibles</b>	<b>27</b>
Instrucciones de uso	27
Criterio de valoración final	28
Observaciones metodológicas importantes	29
1. La evaluación no es mecánica	28
2. Contextualización por tipo de documento	29
3. Principio de prevalencia jerárquica	29

## Resumen ejecutivo

Dadas las limitaciones documentadas de las herramientas automatizadas de detección de escritura generada por IA (ej. GPTZero, YouScan, Quillbot, My Detector, entre otras), y en particular, su problemática tasa de falsos positivos que puede resultar en señalamientos injustificados contra autores humanos, presentamos un marco metodológico de asistencia destinado a quienes necesiten realizar este tipo de evaluación en contextos académicos, profesionales o institucionales.

Nuestro enfoque reconoce la creciente coexistencia de prácticas de autoría humano-máquina y se distancia deliberadamente de lógicas punitivas o de sospecha automática. En su lugar, propone un marco de análisis que trasciende la dimensión meramente técnica, apoyada en sistemas de detección automatizada, para ofrecer una perspectiva más amplia, que considera elementos adicionales, aplicable con las precauciones necesarias según el contexto y los objetivos específicos de cada caso..

El marco se compone de tres niveles complementarios que permiten avanzar progresivamente desde la identificación de señales preliminares hasta una evaluación sustantiva del contenido:



Este enfoque de niveles complementarios y jerarquizados (ver cuadro) no busca detectar la intervención de IA como un fin en sí mismo, sino orientar la evaluación del alcance y relevancia de dicha intervención según el contexto y propósito de análisis.

**Principio operativo:** El marco no opera por acumulación mecánica de señales, sino por jerarquía inversa: la presencia consistente de indicios de autoría humana (NIVEL 3) prevalece sobre cualquier resultado de detección automatizada (NIVEL 1).

	NIVEL 1	NIVEL 2	NIVEL 3
Dimensión	Detección algorítmica automatizada	Trazabilidad técnica	Indicios de autoría humana
Objeto de análisis	Texto analizado	Documento y evidencia técnica asociada	Contenido, razonamiento y contexto
Herramientas / insumos	Detectores de IA existentes (p. ej. Turnitin AI, GPTZero)	Metadatos, historial de versiones, marcas de agua, trazas técnicas	Análisis cualitativo
Qué se evalúa	Compatibilidad estadística del texto con patrones de IA	Evidencia técnica directa de intervención algorítmica	Juicio, toma de posición, responsabilidad epistémica, contextualización situada, interpretación crítica
Tipo de señal	Probabilística	Técnica–objetiva (cuando existe)	Sustantiva–cualitativa
Alcance	Alerta inicial	Refuerza o debilita hipótesis	Permite decisión final
Límites principales	Falsos positivos y negativos. Opacidad algorítmica	No es universal. La ausencia no prueba autoría humana	No automatizable. Requiere análisis situado. Es subjetiva
Resultado posible	Indicio preliminar	Evidencia técnica parcial o inexistente	Acreditación de intervención humana sustantiva
Peso decisorio	Bajo	Medio	Alto
Rol en el marco	Habilita el análisis	Aporta trazabilidad	Define la conclusión

## A. Introducción

La expansión y progresiva masificación de plataformas de inteligencia artificial generativa (ej. ChatGPT, Gemini, Claude, Grok, entre otras) y su capacidad para redactar una amplia variedad de documentos, ha dado lugar a diversos debates en torno a la producción escrita contemporánea.

Entre ellos, ocupa un lugar central la discusión sobre la necesidad y viabilidad de detectar la intervención de inteligencia artificial en la elaboración de textos, particularmente en ámbitos académicos, administrativos y jurídicos, donde persisten criterios tradicionales de atribución de autoría y responsabilidad basados en el supuesto de producción predominante o totalmente humana.

En ese contexto, se ha dado un debate específico sobre las posibilidades técnicas de detectar escritura producida por inteligencia artificial. Este debate usualmente es planteado en términos binarios, es decir, texto humano versus texto artificial, y se apoya en el uso de herramientas de detección automatizada cuyos resultados presentan limitaciones significativas de fiabilidad y explicabilidad.

El marco aquí propuesto no pretende zanjar ese debate, ni ofrecer soluciones definitivas al problema de la detección como tal. En cambio, parte de una premisa distinta: el análisis de autoría en contextos de trabajo con asistencia de inteligencia artificial no puede reducirse a una clasificación dicotómica ni a la determinación absoluta del origen del texto, objetivo técnicamente inviable en el estado actual del arte.

Consideramos, en cambio, que la cuestión requiere un enfoque gradual, contextual y prudente, orientado a evaluar:



**El grado de intervención humana sustantiva**



**La trazabilidad técnica disponible**



**La legitimidad del contenido producido**

Siempre en el contexto de la situación particular que ha requerido llevar a cabo esa evaluación. Sobre esta base, se propone una metodología estructurada y proporcionada, basada en el uso responsable de herramientas de análisis automático existentes, el análisis de evidencias técnicas objetivas y, de manera central, en la identificación de huellas humanas irreductibles. Estas últimas funcionarán como criterio decisorio principal para la evaluación de la autoría en escenarios donde la asistencia algorítmica ocupa un lugar legítimo en las prácticas contemporáneas de producción de documentos de texto.

## B. Marco de evaluación multinivel

### 01. Enfoque general

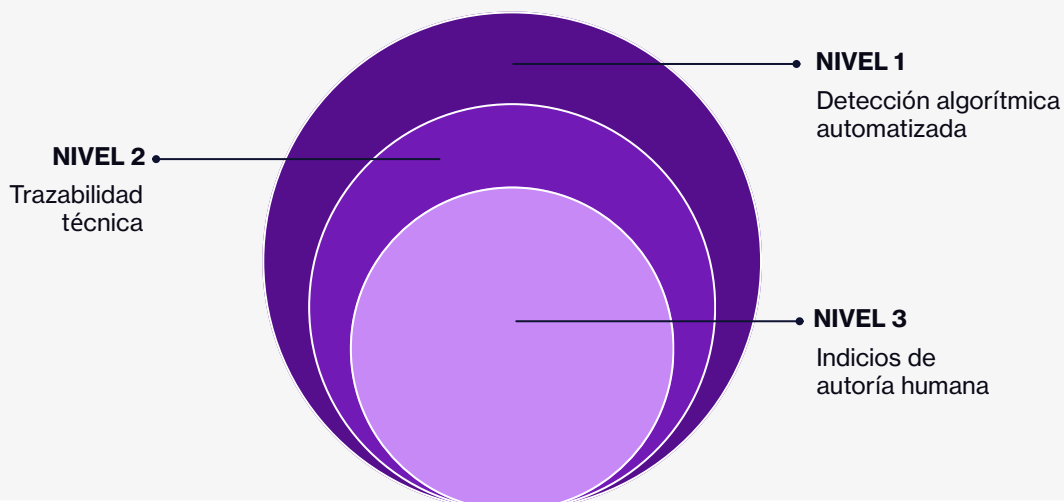
En primer lugar, es importante señalar que el marco propuesto opera bajo una lógica de jerarquía inversa: el NIVEL 3 (evidencia de autoría humana sustantiva) prevalece sobre el NIVEL 1 (detección automatizada) y el NIVEL 2 (trazabilidad técnica). Esta estructura permite refutar la hipótesis de autoría algorítmica, incluso, ante resultados positivos que arrojan herramientas de detección automatizada.

Esta estructura permite refutar la hipótesis de autoría algorítmica, incluso, ante resultados positivos que arrojan herramientas de detección automatizada. En ese marco, el principio rector es claro: ninguna señal aislada resulta suficiente para atribuir autoría a la IA, ni tampoco para inferir conclusiones definitivas sobre el grado de intervención humana en la construcción del texto.

El análisis debe integrar distintos indicios y evidencias para converger hacia una inferencia razonable y contextual, que considere el caso concreto y el objetivo perseguido con la detección (por ejemplo, la detección de plagio académico, la evaluación de competencias profesionales, la identificación de delegación indebida de funciones o la determinación del grado de asistencia algorítmica empleada).

En cualquier supuesto, el uso de IA en la producción de un texto no debiera implicar necesariamente consecuencias negativas, siempre que el contenido haya sido objeto de una intervención humana suficiente que preserve su integridad y validez jurídica, académica o administrativa, según el caso.

### 02. Dimensiones del marco de validación



## NIVEL 1: Detección algorítmica automatizada (uso instrumental)

La detección algorítmica automatizada comprende el uso de herramientas existentes destinadas a identificar posibles indicios de contenido generado mediante inteligencia artificial, tales como GPTZero, ZeroGPT, SEO GPT, GPTKit, Sapling, YouScan, Phrasly, Sidekicker, GoWinston, Quillbot, My Detector y Smodin.

Estas herramientas se emplean exclusivamente con carácter instrumental y preliminar, como mecanismos de alerta temprana orientados a identificar señales que puedan justificar un análisis posterior más profundo, sin que sus resultados posean valor concluyente por sí mismos.

Desde el punto de vista técnico, estas soluciones operan a partir del análisis estadístico y lingüístico de los textos, evaluando métricas de previsibilidad léxica, regularidad sintáctica y patrones probabilísticos de secuenciación del lenguaje que suelen asociarse a producciones generadas por modelos de lenguaje.

Dicho análisis se basa en la comparación del texto evaluado con modelos clasificatorios entrenados sobre grandes volúmenes de escritura humana y contenidos sintéticos, lo que permite asignar probabilidades o indicadores de riesgo respecto de una posible generación automatizada.

En consecuencia, los resultados obtenidos deben interpretarse estrictamente como estimaciones probabilísticas y no como determinaciones fácticas sobre la autoría del contenido.

De esta manera, la utilización de herramientas de detección algorítmica no permite afirmar, con certeza, la existencia de autoría algorítmica ni descartar la intervención humana en el proceso de producción del texto, ya sea en forma de edición, corrección, reformulación o coautoría asistida. La naturaleza híbrida de los procesos actuales de escritura, así como las limitaciones técnicas conocidas de estos sistemas, impiden atribuirles valor probatorio autónomo.

Por estas razones, los resultados obtenidos mediante detección algorítmica automatizada deben tener un peso decisorio bajo dentro del esquema general de evaluación.

Su función debería limitarse a orientar y priorizar instancias de revisión humana cualitativa, enmarcadas en criterios institucionales claros, transparentes y explicables. Cualquier decisión con efectos jurídicos, académicos o administrativos debe fundarse necesariamente en un análisis integral que considere el contexto, la intencionalidad, la responsabilidad y la intervención humana verificable, elementos que no son inferibles de manera automatizada.

---

<sup>1</sup>Ver más en: <https://www.scribbr.com/ai-tools/how-do-ai-detectors-work/>

<sup>2</sup>Ver más en: <https://gptzero.me/news/how-ai-detectors-work/>

Concebir esta dimensión como un insumo auxiliar y no como un mecanismo de atribución definitiva resulta clave para evitar enfoques presuntivos basados exclusivamente en herramientas algorítmicas, y para asegurar un uso responsable de la inteligencia artificial compatible con los principios de debido proceso, razonabilidad y gobernanza ética.

## **Estado del arte de la detección algorítmica automatizada**

Existen diversos estudios empíricos recientes que examinan, desde enfoques metodológicos variados, la capacidad de los sistemas automáticos disponibles en el mercado para identificar contenido generado por modelos de lenguaje.

Algunos de estos trabajos abordan, desde perspectivas empíricas distintas pero complementarias, el problema de la detección de texto generado por inteligencia artificial, analizando tanto la distinción entre texto humano y texto generado por IA como la persistencia de huellas estilísticas en muestras breves, así como el desempeño de las herramientas de detección frente a intervenciones habituales tales como correcciones de estilo, reformulaciones automáticas o ediciones humanas parciales.

En conjunto, permiten delinear un panorama consistente sobre el estado actual de la detección automatizada y ponen de manifiesto una brecha significativa entre la alta viabilidad técnica observada en entornos controlados y la menor robustez práctica que presentan las mismas herramientas en contextos reales de uso. Esta brecha resulta determinante para evaluar su idoneidad como insumo en procesos de evaluación, control o toma de decisiones institucionales.

## **Clasificación binaria en entornos controlados**

El trabajo “Detection of AI-generated Text: A Multifaceted Approach to Binary and Multiclass Classification”<sup>3</sup> describe pruebas diseñadas para evaluar la capacidad de sistemas automáticos para distinguir entre texto humano y texto generado por IA, y en menor medida, de identificar el modelo generador. El núcleo del experimento se concentra en la clasificación binaria, donde se observa un desempeño muy elevado cuando se utilizan arquitecturas híbridas que combinan modelos neuronales, embeddings semánticos y rasgos estilométricos.

El resultado más relevante consiste en la constatación de que, en condiciones controladas y con textos no intervenidos, la detección de artificialidad alcanza niveles de precisión casi perfectos. Sin embargo, el propio estudio evidencia que esta solidez disminuye cuando se avanza hacia escenarios más finos o menos controlados.

---

<sup>3</sup>Detección de texto generada por IA: un enfoque multifacético para la clasificación binaria y multiclase. Harika Abburi, Deloitte & Touche Assurance and Enterprise Risk Services India Private, Limited, India Deloitte & Touche LLP, EE. UU. Sanmitra Bhattacharya, Edward Bowen, Nirmala Pudota (2025). <https://arxiv.org/html/2505.11550v1>

## Persistencia de huellas estilísticas

El segundo paper, “Stylometry recognizes human and LLM-generated texts in short samples”, describe pruebas de naturaleza distinta. Aquí no se prioriza la construcción de sistemas complejos, sino la evaluación de si los rasgos estilísticos clásicos del lenguaje permiten distinguir textos humanos de textos generados por modelos de lenguaje, incluso cuando las muestras son breves.

El estudio demuestra que los LLMs tienden a producir textos con una mayor regularidad gramatical, una menor variabilidad estilística y patrones lingüísticos más homogéneos, lo que deja huellas detectables mediante técnicas estilométricas tradicionales.

Este enfoque resulta particularmente relevante porque muestra que la detección no depende exclusivamente de arquitecturas profundas opacas, sino que puede apoyarse en señales lingüísticas interpretables. No obstante, el propio diseño del experimento, centrado en textos enciclopédicos y dominios relativamente homogéneos, limita la extrapolación directa de los resultados a otros contextos discursivos.

## Degradación de la detección ante intervención humana

El estudio, “AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI...”, introduce un cambio decisivo de perspectiva al trasladar el problema al terreno operativo.

A diferencia de los trabajos anteriores, se pregunta qué ocurre cuando textos generados por IA circulan, se corrigen, se reformulan o son parcialmente editados por personas.

Las pruebas muestran que si bien algunos detectores mantienen un desempeño relativamente estable frente a textos generados directamente por IA, la mayoría de las herramientas pierde eficacia de manera significativa ante intervenciones simples como la corrección de estilo, la paráfrasis automatizada o una edición humana moderada.

Este hallazgo es central porque evidencia que la detectabilidad no es una propiedad fija del texto, sino una condición frágil que se degrada rápidamente en escenarios realistas.

---

<sup>4</sup>Stylometry recognizes human and LLM-generated texts in short samples - arXiv, fecha de acceso: enero 12, 2026, <https://arxiv.org/html/2507.00838v2>

<sup>5</sup>Muhammad Abid Mali, Amjad Islam Amjad, “AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by ChatGPT, Perplexity, and Gemini?”, Journal of Applied Learning & Teaching Vol.8 No.1 (2025). [https://www.researchgate.net/publication/388103693\\_AI\\_vs\\_AI\\_How\\_effective\\_are\\_Turnitin\\_ZeroGPT\\_GPTZero\\_and\\_Writer\\_AI\\_in\\_detecting\\_text\\_generated\\_by\\_ChatGPT\\_Perplexity\\_and\\_Gemini#:~:text=adversarial%20techniques,detection%20tools](https://www.researchgate.net/publication/388103693_AI_vs_AI_How_effective_are_Turnitin_ZeroGPT_GPTZero_and_Writer_AI_in_detecting_text_generated_by_ChatGPT_Perplexity_and_Gemini#:~:text=adversarial%20techniques,detection%20tools)

## Evaluación exhaustiva con modelos de vanguardia

De la misma manera, en el reciente estudio del Becker Friedman Institute (BFI), titulado "Artificial Writing and Automated Detection", los investigadores Brian Jabarian y Alex Imas realizaron una evaluación exhaustiva de los detectores más utilizados en el mercado para determinar si realmente son capaces de distinguir la autoría humana de la sintética en el actual ecosistema de modelos como GPT-4.1, Claude 4 y Gemini 2.0.

Lo que distingue a este último trabajo es su rigor metodológico y la escala de su corpus. Los autores no se limitaron a pruebas superficiales; emplearon un corpus de 1,992 pasajes de texto que abarca seis géneros cotidianos (noticias, blogs, reseñas de consumidores, novelas, reseñas de restaurantes y currículos).

Los textos verificados de autoría humana se contrastaron con textos generados por IA utilizando cuatro modelos de lenguaje (LLM) de vanguardia: GPT-4.1, Claude Opus 4, Claude Sonnet 4, Gemini 2.0 Flash. También se examinó la efectividad de los 'humanizadores' de IA (Stealth GPT) para evadir potencialmente a los detectores. Este diseño permitió analizar variables críticas como la longitud del texto y la resistencia de los detectores ante herramientas de "humanización" diseñadas específicamente para evadir algoritmos de detección.

Los resultados revelan una marcada jerarquía de rendimiento. El detector Pangram emergió como el líder indiscutible, logrando tasas de error cercanas a cero y demostrando una robustez excepcional incluso en textos de menos de 50 palabras, un área donde la mayoría de los competidores suelen fallar. En un segundo escalón se ubicaron OriginalityAI y GPT Zero, que si bien son efectivos en pasajes largos, pierden precisión ante textos breves o manipulados. Por el contrario, el modelo de código abierto RoBERTa demostró ser inadecuado para entornos de alta responsabilidad, al no detectar más de la mitad de los contenidos generados por IA.

El documento advierte sobre la naturaleza de la "carrera armamentista" en este campo. La competencia entre los modelos de lenguaje que buscan sonar más humanos y los detectores que buscan identificarlos es constante y dinámica. Los autores concluyen que, si bien existen herramientas potentes, la detección no es una solución mágica. La implementación de estas tecnologías requiere un equilibrio ético y práctico que permita mitigar el fraude sin sofocar el uso legítimo y productivo de la IA como herramienta de asistencia en la redacción moderna.

---

<sup>6</sup>Jabarian, B., e Imas, A. (2025). Artificial writing and automated detection (Working Paper No. 2025-116). Becker Friedman Institute for Economics at the University of Chicago. [https://bfi.uchicago.edu/wp-content/uploads/2025/09/BFI\\_WP\\_2025-116.pdf](https://bfi.uchicago.edu/wp-content/uploads/2025/09/BFI_WP_2025-116.pdf)

## Conclusiones del estado del arte

La lectura integrada de las investigaciones relevadas permite extraer una conclusión convergente: la detección de texto generado por IA es técnicamente viable en condiciones controladas y sobre textos no intervenidos, pero su confiabilidad disminuye cuando el texto atraviesa procesos normales de uso, revisión y adaptación. Tanto los enfoques basados en deep learning como los estilométricos dependen, en mayor o menor medida, de regularidades que se diluyen cuando el contenido es reescrito o mediado por humanos.

Desde una perspectiva institucional, académica o jurídica, estos resultados son particularmente relevantes: los estudios muestran que los detectores pueden funcionar como herramientas auxiliares o indicativas, pero no como mecanismos concluyentes para establecer autoría, intencionalidad o uso indebido de IA.

La facilidad con la que la detección se debilita frente a intervenciones humanas o de otra IA, sumada a su opacidad inherente, pone en jaque su uso como prueba autónoma o decisiva, especialmente en contextos donde rigen principios de debido proceso, presunción de buena fe o evaluación contextual del contenido.

En conjunto, los trabajos relevados no desacreditan la detección de texto generado por IA, pero sí delimitan con claridad su alcance: se trata de una capacidad técnica útil, pero intrínsecamente contingente, dependiente del contexto, del tipo de texto y del grado de intervención humana.

Por ello, cualquier marco normativo, académico o institucional que incorpore estas herramientas debería partir de esa premisa y evitar atribuirles un grado de certeza que, a la luz de la evidencia empírica disponible, hoy no pueden garantizar.

Eje de análisis	Conclusión general	Implicancia práctica / institucional
Detectabilidad del texto generado por IA	La detección de texto generado por IA es técnicamente viable en condiciones controladas y sobre textos no intervenidos.	Los detectores pueden ser útiles como herramientas indicativas en contextos acotados, pero no garantizan resultados concluyentes.
Huellas estilísticas	Los textos generados por IA presentan patrones estilísticos relativamente homogéneos que pueden ser identificados, incluso en muestras breves.	La estilometría aporta señales interpretables, pero dependientes del dominio y del tipo de texto analizado.

<sup>6</sup>Jabarian, B., e Imas, A. (2025). Artificial writing and automated detection (Working Paper No. 2025-116). Becker Friedman Institute for Economics at the University of Chicago. [https://bfi.uchicago.edu/wp-content/uploads/2025/09/BFI\\_WP\\_2025-116.pdf](https://bfi.uchicago.edu/wp-content/uploads/2025/09/BFI_WP_2025-116.pdf)

<b>Robustez frente a intervenciones</b>	<p>La eficacia de la detección disminuye significativamente ante correcciones de estilo, paráfrasis automatizadas o ediciones humanas parciales.</p>	<p>Intervenciones habituales pueden invalidar o debilitar los resultados de los detectores, limitando su fiabilidad operativa.</p>
<b>Uso de herramientas comerciales</b>	<p>El desempeño de los detectores comerciales es heterogéneo y altamente sensible al tipo de texto y a las modificaciones aplicadas.</p>	<p>No resulta adecuado basar decisiones académicas o institucionales únicamente en resultados automáticos de detección.</p>
<b>Valor probatorio</b>	<p>Los resultados de detección no alcanzan, por sí solos, un estándar probatorio fuerte.</p>	<p>La detección automática debe integrarse con evaluaciones humanas, contextuales y procedimentales.</p>
<b>Idoneidad institucional</b>	<p>La detección de IA es una herramienta auxiliar, no determinante.</p>	<p>Su uso requiere marcos claros, criterios de prudencia y resguardo del debido proceso.</p>

## Evaluación de herramientas realizada por IALAB

De cara al objetivo general del estudio, orientado a obtener insights exploratorios iniciales, contrastar y consolidar líneas investigativas ya identificadas en la literatura reciente, se optó por el diseño deliberado de un conjunto de datos acotado, controlado y analíticamente segmentado para evaluar la performance de algunas de las herramientas de detección disponibles en el mercado. Este enfoque metodológico responde a una lógica exploratoria y comparativa, más que a la pretensión de producir inferencias estadísticas concluyentes.

El experimento desarrollado por IALAB se estructuró en dos fases metodológicamente equivalentes, cada una basada en el análisis de 100 inputs distribuidos de manera equilibrada entre: textos íntegramente producidos por autores humanos; textos íntegramente generados por sistemas de inteligencia artificial; y textos híbridos, resultantes de la interacción humano-IA.

Esta decisión permitió asegurar la comparabilidad de los resultados y evaluar el comportamiento de las herramientas de detección, tanto en escenarios "puros" como en contextos híbridos que son, en la práctica, los más frecuentes en ámbitos académicos, jurídicos e institucionales.

La diferencia central entre ambas fases no residió en la cantidad ni en la tipología de los insumos, sino en su tratamiento previo. Mientras que en la primera etapa los textos fueron analizados en su estado original, en la segunda los mismos documentos fueron sometidos a procesos de humanización estilística mediante herramientas basadas en IA para luego ser nuevamente evaluados por los detectores.

Esta decisión metodológica permitió explorar, de manera controlada, cómo las intervenciones humanas mediadas por IA inciden en la detectabilidad algorítmica y, en particular, en las zonas grises de autoría que desafían los enfoques binarios tradicionales.

## Primera fase: detección sobre textos originales (100 inputs)

La base inicial de pruebas estuvo compuesta por 33 textos 100% humanos, 31 textos 100% IA y 36 textos de autoría mixta (IA + intervención humana). Esta distribución permite observar el desempeño de los detectores frente a casos extremos y también su comportamiento ante escenarios intermedios, característicos del uso real de herramientas de IA.

**Textos 100% humanos.** Los resultados evidenciaron una problemática particularmente relevante: el 51,52% fue correctamente identificado como humano, mientras que el 48,48% restante fue clasificado erróneamente como generado por IA. Esta tasa de falsos positivos resulta extraordinariamente elevada y pone de manifiesto el riesgo concreto de atribuir incorrectamente autoría automatizada a producciones humanas legítimas.

**Textos 100% IA.** En contraste, los textos íntegramente generados por IA fueron identificados como tales en el 100% de los casos, sin registrarse falsos negativos. Este resultado confirma que las herramientas de detección tienden a mostrar un buen desempeño cuando se enfrentan a textos completamente automatizados y no intervenidos.

**Textos híbridos.** El análisis de los textos mixtos revela un comportamiento significativamente menos estable. En este grupo, aproximadamente un 61,11% de los documentos fue clasificado como humano y el 38,89% como IA, lo que demuestra que la presencia de edición, reformulación o contextualización humana interfiere de manera sustantiva con los criterios estadísticos de detección (en términos binarios), y complejiza una identificación consistente de la autoría.

En términos agregados, la primera fase arrojó un 60 % de clasificaciones correctas y un 40 % de errores, siendo el error predominante la atribución de IA a textos humanos (falsos positivos).

Tipo de documento	Detectado como humano	Detectado como IA	Observación principal
100 % humanos (33)	17 (51,52%)	16 (48,48%)	Falsos positivos elevados
100 % IA (31)	0 (0 %)	31 (100 %)	Sin falsos negativos
Híbridos (36)	22 (61,11%)	14 (38,89%)	Alta inestabilidad
<b>Total</b>	<b>≈ 39</b>	<b>≈ 61</b>	<b>Error concentrado en humanos</b>

## Segunda fase: textos humanizados y revalidados (67 inputs)

En la segunda fase del experimento se incorporaron exclusivamente documentos generados íntegramente por inteligencia artificial y documentos de carácter híbrido.

Ambos conjuntos fueron sometidos a procesos de humanización estilística y posteriormente re-analizados mediante las herramientas de detección. La nueva distribución estuvo compuesta por 31 documentos originalmente generados por IA y 36 documentos híbridos.

Los resultados evidencian un impacto particularmente significativo de la humanización en los textos originalmente producidos en un 100 % por IA. Tras el tratamiento estilístico, sólo el 35,48 % de estos documentos fue identificado como generado por IA, mientras que el 64,52 % restante fue clasificado como de autoría humana.

Este comportamiento pone de manifiesto una reducción sustancial de la capacidad de detección frente a transformaciones estilísticas relativamente superficiales.

En el caso de los documentos híbridos, el 72,22 % fue clasificado como humano y apenas el 27,78 % como generado por IA, lo que refuerza la hipótesis de que, una vez intervenido el estilo, las herramientas de detección tienden a perder la capacidad de discriminar entre escenarios de colaboración humano-máquina y producciones de autoría predominantemente algorítmica.

Tipo de documento	Detectado como humano	Detectado como IA	Observación principal
100 % IA humanizada (31)	20 (64,52 %)	11 (35,48 %)	Colapso de detección
Híbridos humanizados (36)	26 (72,22 %)	10 (27,78 %)	Predominio del "estilo humano"
<b>Total</b>	<b>67</b>	<b>33</b>	<b>El detector deja de distinguir autoría</b>

## Análisis comparativo e interpretación

La comparación entre ambas fases permite extraer las siguientes conclusiones:

**01. Colapso de la detección tras humanización.** Se verifica una caída abrupta en la detección de IA pura, que pasa del 100% en la primera fase al 35,48% tras la humanización (reducción del 64,52%).

**02. Persistencia de falsos positivos.** Los textos 100% humanos mantienen una tasa inaceptable de clasificación errónea (48,48%), lo que da cuenta de que el problema de falsos positivos no es marginal sino estructural.

**03. Incapacidad de discriminar autoría híbrida.** El comportamiento frente a los textos mixtos confirma que los detectores no evalúan procesos de producción ni niveles de intervención humana o de IA.

**04. Predominio del estilo sobre la autoría.** Los resultados demuestran que una intervención estilística relativamente simple puede hacer que textos íntegramente generados por IA sean clasificados como humanos.

En conjunto, los resultados muestran que la principal debilidad de estas herramientas reside en su incapacidad para discriminar adecuadamente la intervención humana, especialmente en contextos híbridos, que constituyen el escenario predominante en la práctica institucional, académica y profesional contemporánea.

Estos hallazgos refuerzan la necesidad de un enfoque metodológico multinivel que no dependa exclusivamente de la detección algorítmica automatizada, sino que integre evidencia técnica objetiva y, fundamentalmente, la identificación de huellas humanas irreductibles como criterio decisorio principal.

## NIVEL 2: Trazabilidad técnica - metadatos y marcas de agua

### Alcance del análisis técnico y principio probatorio

El Nivel 2 del marco metodológico se orienta al análisis de evidencia técnica directa asociada al documento bajo examen, en la medida en que dicha evidencia exista, sea accesible y resulte técnicamente verificable.

Su función no es establecer autoría de manera concluyente ni determinar, por sí sola, el grado de intervención algorítmica en la producción del texto, sino aportar elementos instrumentales que permitan reforzar, debilitar o mantener abierta una hipótesis previamente formulada a partir de otras dimensiones del análisis.

Este encuadre responde tanto a limitaciones técnicas inherentes (derivadas de la naturaleza fácilmente modificable de los metadatos y de su pérdida frecuente en operaciones habituales de edición y conversión) como a exigencias normativas vinculadas al debido proceso, la razonabilidad probatoria y la necesidad de evitar inferencias falaces.

En este sentido, el análisis de metadatos, historiales de versión, marcas técnicas o posibles mecanismos de watermarking se rige por un principio rector ampliamente reconocido en ámbitos científicos y jurídicos: la ausencia de evidencia no constituye evidencia de ausencia. Dicho de otro modo, la falta de trazas técnicas no permite inferir que no hubo intervención algorítmica, del mismo modo que su presencia no prueba autoría íntegramente automatizada.

El Nivel 2 ocupa, por tanto, una posición intermedia dentro del esquema de evaluación. A diferencia del Nivel 1, basado en señales probabilísticas de bajo peso decisorio, la trazabilidad técnica puede aportar evidencia objetiva cuando se encuentra disponible y preservada. Y a diferencia del Nivel 3, su peso decisorio es medio y condicionado, y su valor interpretativo depende de la integridad de la evidencia, de la posibilidad de verificación independiente y de su articulación coherente con el contexto institucional y procedimental del documento analizado.

## **Metadatos documentales: naturaleza, categorías y persistencia**

A los fines del presente marco, los metadatos documentales se entienden como el conjunto de datos estructurales, temporales y procedimentales que describen las condiciones de creación, edición, transformación y almacenamiento de un documento digital, con independencia de su contenido semántico.

Estos metadatos constituyen artefactos técnicos derivados del funcionamiento de los sistemas informáticos utilizados en el proceso de producción documental, y no declaraciones intencionales de autoría o responsabilidad.

Desde una perspectiva forense, resulta fundamental distinguir entre los metadatos como registros instrumentales generados por el sistema y las afirmaciones normativas o contextuales que puedan inferirse a partir de ellos.

Los metadatos no describen directamente quién “es” el autor en sentido jurídico o institucional, sino qué operaciones fueron registradas por determinadas herramientas, en determinados momentos y bajo configuraciones específicas. En consecuencia, su valor analítico depende tanto de su contenido como de su persistencia, coherencia interna y compatibilidad con el contexto procedimental en el que el documento fue producido.

En términos analíticos, los metadatos relevantes para el Nivel 2 pueden agruparse en cuatro categorías principales, cuya presencia y fiabilidad varían significativamente según el formato del documento y las operaciones a las que haya sido sometido.

## **Metadatos de autoría y atribución**

Incluyen campos como autor declarado, usuario creador, identificadores de cuenta, firmas electrónicas o referencias explícitas al software generador.

Estos datos pueden ser fácilmente editables por el usuario o regenerados automáticamente por la aplicación, por lo que su valor probatorio aislado es bajo. No obstante, cuando se presentan de manera consistente a lo largo de múltiples artefactos o versiones, pueden aportar indicios contextuales compatibles con determinados flujos de trabajo humano.

## **Metadatos temporales**

Comprenden fechas y horas de creación, modificación, guardado, exportación o conversión del archivo. Estos registros permiten realizar análisis de coherencia temporal, evaluando si los intervalos observados resultan razonablemente compatibles con procesos humanos situados o si, por el contrario, sugieren automatización intensiva o generación masiva.

Sin embargo, debe enfatizarse que los timestamps internos son, en muchos formatos, técnicamente modificables y sensibles a operaciones rutinarias del sistema operativo, por lo que requieren siempre corroboración cruzada.

## **Metadatos de herramientas y entorno técnico**

Identifican el software, la versión, la plataforma o el sistema utilizado para crear o editar el documento. La aparición de determinadas herramientas puede resultar compatible con el uso de sistemas de generación o asistencia algorítmica, pero nunca constituye, por sí sola, evidencia suficiente de tal intervención.

Su interpretación exige considerar prácticas habituales de la organización, configuraciones por defecto y procesos de conversión intermedios.

## **Metadatos de versión y trazabilidad**

Incluyen historiales de cambios, identificadores persistentes, registros de edición colaborativa y otros mecanismos que permiten reconstruir parcialmente la genealogía del documento.

Esta categoría es la que, potencialmente, puede ofrecer el mayor valor analítico dentro del Nivel 2, en la medida en que dichos identificadores sobrevivan a copias, guardados sucesivos o redistribución de archivos. No obstante, su preservación es altamente dependiente del formato y se ve frecuentemente degradada o destruida por operaciones comunes como el copiado de contenido, la exportación o la conversión entre tipos de archivo.

## **Limitaciones del nivel**

Un aspecto central para la correcta interpretación de los metadatos documentales es reconocer que su persistencia no es uniforme. Distintos formatos priorizan distintos objetivos (interoperabilidad, colaboración, preservación visual o archivística) y, en consecuencia, conservan o descartan información de trazabilidad de manera diferencial.

Así, la ausencia de determinados metadatos puede ser el resultado esperable de una operación legítima y habitual, y no necesariamente de una intención de ocultamiento o de la inexistencia de un proceso previo.

Por ello, el análisis de metadatos en el Nivel 2 no se orienta a la búsqueda de marcadores universales ni a la aplicación de reglas binarias, sino a la evaluación contextual de qué tipo de información técnica es razonable esperar que sobreviva dadas las características del documento, su formato, su circulación y las prácticas normales del entorno institucional en el que fue generado. Solo sobre esa base resulta metodológicamente defendible extraer conclusiones prudentes acerca del alcance y los límites de la trazabilidad técnica disponible.

## NIVEL 3: Identificación de huellas humanas irreductibles

### **Rendición de cuentas, juicio situado y legitimidad decisoria en contextos de escritura asistida por IA**

El Nivel 3 tiene por finalidad evaluar la existencia de intervención humana sustantiva en la producción de un texto, con independencia de que se hayan utilizado sistemas de inteligencia artificial como herramientas de apoyo.

La evaluación no persigue la detección de contenido generado por IA ni la determinación de autoría en sentido instrumental, sino la identificación de operaciones cognitivas, decisorias y éticas que no emergen naturalmente de procesos de optimización estadística y que, por su propia naturaleza, requieren validación humana, asunción de responsabilidad epistémica y contextualización situada.

Este enfoque se apoya en la premisa desarrollada por Porayska-Pomsta y Rajendran, quienes sostienen que la diferencia clave entre la toma de decisiones humana y la de IA radica en que las decisiones humanas implican flexibilidad individual, juicios relevantes al contexto, empatía, así como juicios morales complejos, ausentes en la inteligencia artificial.

Los autores argumentan que la rendición de cuentas constituye un atributo irreductiblemente humano, dado que los sistemas de IA carecen de agencia moral y, por tanto, no pueden ser considerados sujetos de responsabilidad por las consecuencias éticas o institucionales de una decisión.

Desde esta perspectiva, la diferencia entre la toma de decisiones humana y la automatizada no radica únicamente en el grado de complejidad del procesamiento, sino en la capacidad humana de ejercer juicio contextual, deliberación moral, empatía, flexibilidad frente a la excepción y rendición de cuentas. Mientras la IA opera sobre patrones históricos y objetivos previamente definidos, el ser humano puede interrogar el propio objetivo, ponderar valores en tensión, reconocer circunstancias singulares y asumir conscientemente las consecuencias de la decisión adoptada.

### **Enfoque metodológico**

El enfoque adoptado es funcional, cualitativo y gradual, y se basa en el análisis del texto entendido como un acto decisorio o comunicacional, y no como mero artefacto lingüístico. La evaluación se realiza a partir del análisis cualitativo de ocho dimensiones, cada una de las cuales permite identificar operaciones que requieren intervención humana sustantiva y que resultan centrales para la atribución de autoría y responsabilidad.

Dimensiones de análisis:

## a. Juicio evaluativo explícito

Se analiza si el texto formula valoraciones, juicios o ponderaciones y si estas se encuentran debidamente fundamentadas y contextualizadas. La presencia de evaluaciones situadas constituye una huella de juicio humano, y reconoce la capacidad humana de comprender excepciones y circunstancias que pueden no ser capturadas por los sistemas de IA.

### Indicadores:

Formulación de valoraciones  
fundamentadas

Ponderación entre opciones  
con criterios explícitos

Evaluación contextualizada  
(no genérica)

## b. Toma de posición frente a alternativas

Se evalúa si el texto identifica opciones posibles y adopta una decisión fundada entre ellas, especialmente en contextos donde existe margen de discrecionalidad. La adopción de una posición argumentada refleja agencia decisoria humana y excede la mera selección algorítmica entre alternativas predefinidas.

### Indicadores:

Identificación explícita  
de alternativas

Elección fundamentada  
entre opciones

Argumentación de la  
posición adoptada

## c. Asunción de responsabilidad

Esta dimensión examina si el texto reconoce explícitamente las consecuencias derivadas del contenido (jurídicas, institucionales o prácticas) y si dichas consecuencias son asumidas por quien lo emite.

### Indicadores:

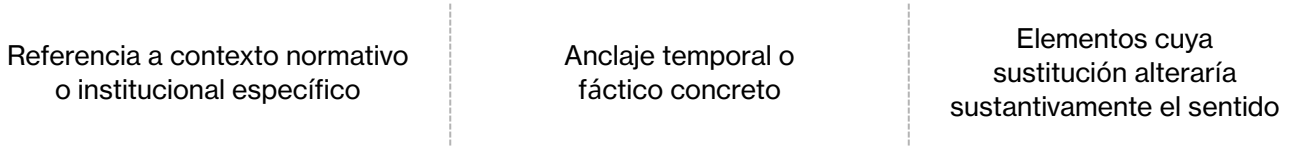
Reconocimiento explícito  
de consecuencias

Asunción de efectos jurídicos,  
institucionales o prácticos

### d. Contextualización situada

Se analiza la inserción del texto en un contexto normativo, institucional, temporal o fáctico específico, cuyo intercambio alteraría su sentido o alcance. La contextualización no intercambiable opera como anclaje de significado y de responsabilidad, y revela comprensión situada del entorno decisorio.

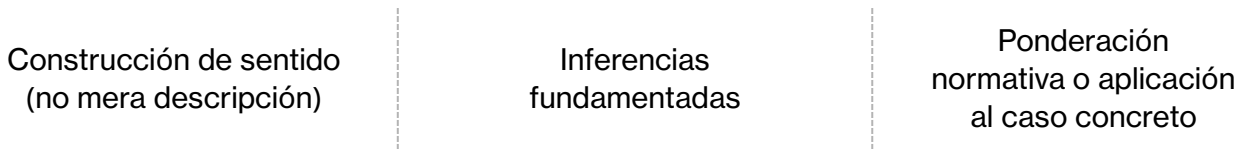
**Indicadores:**



### e. Grado de interpretación

Se distingue entre textos meramente descriptivos y aquellos que construyen sentido, realizan inferencias, ponderan normas y fuentes de información, o traducen información en criterios aplicables al caso concreto. Esta capacidad interpretativa refleja una operación hermenéutica típicamente humana, no siempre presente en los sistemas de IA.

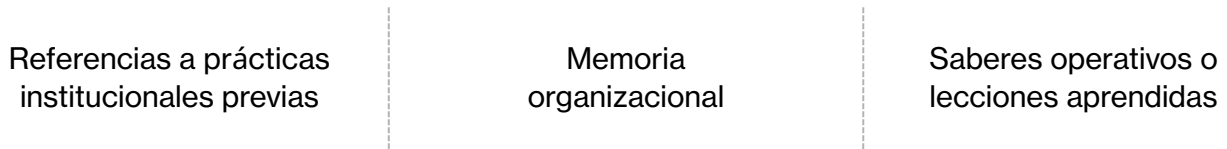
**Indicadores:**



### f. Experiencia humana o institucional reconocible

Se evalúa la presencia de referencias a prácticas reiteradas, memoria organizacional, aprendizajes previos o saberes operativos no explícitos en normas formales. Esta dimensión introduce un componente reflexivo y temporal que refuerza la legitimidad del contenido producido.

**Indicadores:**



## g. Ética aplicada al caso concreto

Se verifica si el texto pondera valores, riesgos o impactos de manera situada, atendiendo a derechos, proporcionalidad, razonabilidad e impactos diferenciales, y no mediante referencias éticas abstractas o genéricas. Esta dimensión se corresponde directamente con la afirmación de que la IA carece de brújula moral propia.

### Indicadores:

Ponderación de valores  
en tensión

Análisis de impactos  
diferenciales

Aplicación situada de  
principios éticos  
(no abstractos)

## h. Creatividad funcional orientada a un fin

Finalmente, se analiza si el texto propone soluciones, enfoques o cursos de acción instrumentales, viables y adaptados al contexto institucional y a un objetivo real.

### Indicadores:

Propuestas de solución  
adaptadas al contexto

Cursos de acción  
viables e instrumentales

Innovación funcional  
orientada a objetivos  
específicos

## Criterio de valoración

La valoración final no depende de la presencia aislada de alguno de estos elementos, sino de su concurrencia, articulación y coherencia interna. A partir de ello, el texto puede clasificarse según el siguiente gradiente:

## Gradiente de intervención humana

<b>Nivel Bajo</b>	Ausencia total o muy significativa de huellas humanas irreductibles. El texto podría ser íntegramente generado por IA sin intervención sustantiva.
<b>Nivel Medio</b>	Presencia parcial de huellas humanas, pero con limitada articulación o profundidad. Indicios de intervención humana pero no sustantiva.
<b>Nivel Alto</b>	Concurrencia consistente de múltiples dimensiones, con articulación coherente. Evidencia clara de intervención humana sustantiva con alto peso decisorio dentro del marco de evaluación.

El Nivel 3 constituye el criterio decisorio principal del marco metodológico. Se articula con los restantes pero, finalmente, prevalece sobre los resultados de detección automatizada (Nivel 1) y la evidencia técnica disponible (Nivel 2), para operar como salvaguarda fundamental contra falsos positivos y como mecanismo de validación de la legitimidad institucional, académica o jurídica del documento frente a una realidad de omnipresencia de la inteligencia artificial en la mayoría del software que se utiliza.

Con el propósito de operacionalizar la evaluación del Nivel 3 en contextos aplicados, se incluye como el Anexo I una matriz de valoración de huellas humanas irreductibles. Esta herramienta proporciona criterios orientativos para el análisis de las ocho dimensiones y debe entenderse como un instrumento de apoyo metodológico que complementa (pero no reemplaza) el ejercicio de juicio experto contextualizado requerido para una evaluación integral de la autoría.

## C. Reflexiones finales

La detección de inteligencia artificial en la producción textual no debe ser entendida como un indicador negativo ni como un mecanismo automático de deslegitimación. Interpretarla de ese modo implica trasladar acríticamente al plano institucional una lógica binaria “humano versus artificial” que ya no resulta adecuada para describir los procesos contemporáneos de producción de conocimiento, redacción técnica y toma de decisiones.

El marco propuesto adopta una premisa deliberadamente más exigente: en un escenario de escrituras híbridas, asistidas y colaborativas, el foco no debe colocarse en el origen instrumental del texto, sino en la existencia de intervención humana sustantiva suficiente para atribuir responsabilidad, validez y legitimidad.

En este esquema, la detección algorítmica cumple un rol meramente preliminar y la trazabilidad técnica aporta contexto cuando existe, pero es la identificación de huellas humanas irreductibles la que define, en última instancia, el peso decisorio del análisis.

La evidencia empírica relevada en la literatura académica reciente y en las evaluaciones realizadas por IALAB confirma que los detectores actuales presentan limitaciones estructurales severas, particularmente en textos técnicos, jurídicos y administrativos, donde patrones lingüísticos genuinamente humanos suelen ser clasificados erróneamente como sintéticos. La tasa de falsos positivos documentada (que alcanzó el 48,48% en textos 100% humanos en las pruebas de IALAB) evidencia que su utilización como prueba autónoma o como base de decisiones sancionatorias resulta técnicamente infundada e institucionalmente riesgosa.

Por ello, el marco metodológico propuesto opera bajo una jerarquía inversa: la presencia consistente de huellas humanas irreductibles (Nivel 3) prevalece sobre cualquier resultado de detección automatizada (Nivel 1) y de trazabilidad técnica (Nivel 2), incluso cuando estas últimas arrojen resultados positivos.

En definitiva, la pregunta relevante no es si un texto fue escrito por una inteligencia artificial, sino si el contenido expresa una voluntad responsable, una interpretación situada y una decisión atribuible a un humano. Solo desde este desplazamiento conceptual es posible construir criterios de validación que sean técnica, jurídica y éticamente defendibles en aquellos ámbitos donde la palabra escrita sigue siendo, y debe seguir siendo, un acto de responsabilidad humana.

El marco propuesto reconoce la creciente incorporación de la IA generativa en las prácticas institucionales, académicas y profesionales contemporáneas, y propone criterios de evaluación que permitan diferenciar entre el uso instrumental de IA y la delegación acrítica e indebida de autoría y responsabilidad decisoria.

Esta distinción resulta clave para preservar la integridad de los procesos que requieren rendición de cuentas humana, sin inhibir la innovación tecnológica ni sancionar injustamente a quienes utilizan estas herramientas de manera apropiada y transparente (de acuerdo con los estándares exigidos en cada caso).

La aplicación efectiva de este marco requiere, finalmente, formación específica de los evaluadores en dos dimensiones complementarias: primero, en la comprensión profunda de las limitaciones estructurales de la detección automatizada y los riesgos asociados a su uso acrítico; segundo, en el desarrollo de competencias para el análisis cualitativo de huellas humanas irreductibles mediante la aplicación de las ocho dimensiones propuestas en el Nivel 3.

Solo mediante esta profesionalización del análisis será posible evitar los falsos positivos que penalizan injustamente la autoría humana, y también los falsos negativos que permiten la circulación de contenido sin validación o intervención humana suficiente en contextos donde esta resulta indispensable.

Finalmente, resulta indispensable el desarrollo de protocolos institucionales que fijen reglas claras y realistas acerca de las posibilidades y términos de uso de la IA; el grado de transparencia exigible en cada caso; y las consecuencias institucionales, académicas o jurídicas en juego.

## D. Alcances, limitaciones y líneas de trabajo futuro

El presente trabajo constituye un primer paso hacia la construcción de un enfoque prudente y no reduccionista para la identificación y evaluación de textos generados o asistidos por inteligencia artificial generativa.

Lejos de proponer soluciones definitivas o criterios automáticos de atribución de autoría, el marco desarrollado busca ordenar el análisis en contextos reales de uso, caracterizados por prácticas híbridas de producción textual y por exigencias institucionales de responsabilidad, validez y debido proceso.

Los principales alcances, limitaciones y posibles extensiones del estudio pueden sintetizarse del siguiente modo:

### **Alcance del marco propuesto**

El modelo se orienta a la evaluación de textos en contextos académicos, administrativos y jurídicos, donde la autoría y la responsabilidad conservan relevancia normativa. No pretende ser un sistema universal de detección ni un sustituto del juicio humano experto, sino una herramienta metodológica de apoyo para evaluaciones situadas y proporcionales.

### **Dependencia del contexto de aplicación**

El valor interpretativo de cada nivel del marco (detección algorítmica, trazabilidad técnica e identificación de huellas humanas) depende fuertemente del tipo de documento, del entorno institucional y del objetivo de la evaluación. Futuros trabajos podrían profundizar la adaptación del esquema a dominios específicos, tales como evaluación académica, procedimientos administrativos sancionatorios, producción de resoluciones judiciales o entornos de auditoría corporativa.

### **Limitaciones empíricas del estudio experimental**

Las evaluaciones realizadas por IALAB se basan en un conjunto de datos acotado y controlado, diseñado con fines exploratorios y comparativos. Si bien los resultados permiten ilustrar patrones relevantes (como la elevada tasa de falsos positivos -48,48% en textos 100% humanos- y la fragilidad de la detección frente a intervenciones de humanización estilística), no habilitan inferencias estadísticas generalizables a gran escala.

Estudios futuros podrían:

- Ampliar el corpus analizado en términos cuantitativos
- Diversificar géneros discursivos (sentencias judiciales, artículos científicos, informes técnicos)
- Explorar variaciones lingüísticas, culturales y disciplinares
- Incorporar textos en múltiples idiomas para evaluar la robustez transcultural.

## **Evolución tecnológica y obsolescencia de señales**

El marco reconoce que tanto los modelos generativos como las herramientas de detección evolucionan de manera acelerada y dinámica. En consecuencia, las señales técnicas y estilísticas aquí consideradas no deben entenderse como estables ni definitivas, sino como indicadores históricamente situados en el estado del arte de 2025 y comienzos de 2026.

Una línea futura relevante consiste en:

- Evaluar periódicamente la vigencia de estos indicadores frente a nuevos modelos (GPT-5, Claude 5, Gemini 3.0, etc.)
- Analizar el impacto de técnicas avanzadas de humanización y estrategias de evasión de detectores
- Actualizar las dimensiones del Nivel 3 si los modelos generativos desarrollan capacidades de emulación de huellas humanas más sofisticadas.

## **Necesidad de operacionalización institucional**

Si bien el trabajo describe criterios analíticos claros, su implementación efectiva requiere:

- Protocolos institucionales específicos que establezcan procedimientos de aplicación.
- Capacitación sistemática de evaluadores en las dimensiones formativas propuestas.
- Definición de estándares procedimentales para garantizar consistencia y trazabilidad.
- Mecanismos de revisión y apelación cuando las evaluaciones tengan consecuencias sancionatorias.

Futuros desarrollos podrían traducir el marco en guías operativas, matrices de decisión detalladas o instrumentos de apoyo técnico para organismos educativos, judiciales o administrativos.

## Necesidad de operacionalización institucional

El estudio se concentra en la asistencia algorítmica, sin contrastar sistemáticamente con otras formas de apoyo humano en la producción textual, tales como tutorías académicas, corrección editorial profesional, coautoría colaborativa o asistencia técnica especializada.

Investigaciones futuras podrían:

- Comparar el impacto de distintos tipos de asistencia sobre la autoría, la responsabilidad y la calidad del contenido producido.
- Analizar las implicaciones éticas y normativas de tratar de manera diferenciada la asistencia humana y la algorítmica.

## E. Cierre

En conjunto, este trabajo no propone "resolver" el problema de la detección de escritura generada por IA, sino un marco inicial para **enmarcarlo conceptual y metodológicamente**. La evidencia disponible sugiere que insistir en enfoques binarios o exclusivamente técnicos conduce a errores sistemáticos, riesgos institucionales y decisiones difícilmente defendibles desde perspectivas jurídicas, éticas o epistemológicas.

En cambio, avanzar hacia evaluaciones basadas en la identificación de intervención humana sustantiva, juicio situado y responsabilidad explícita ofrece una vía más sólida y compatible con los principios de debido proceso, razonabilidad probatoria y gobernanza ética que rigen los ámbitos donde la palabra escrita conserva efectos jurídicos, académicos o administrativos reales.

Desde esta perspectiva, **el desafío futuro no reside en perfeccionar indefinidamente la detección automática** (frente a lo cual siempre aparecerán herramientas y técnicas diseñadas para evadir esos mecanismos) **sino en diseñar marcos de gobernanza del uso de IA que preserven la centralidad de la decisión y la supervisión humanas**, incluso en un ecosistema de herramientas cada vez más sofisticadas y de fronteras cada vez más difusas entre producción humana y asistencia algorítmica.

# Anexo I - Matriz de evaluación - Nivel 3: Huellas humanas irreductibles

## Instrucciones de uso

Para cada dimensión, marque el nivel que mejor describe la presencia de huellas humanas en el texto analizado. La evaluación final se basa en la concurrencia y articulación de múltiples dimensiones, no en la suma mecánica de puntos.

Dimensión	Ausente (0)	Presente parcialmente (1)	Presente consistentemente (2)	Observaciones
<b>01</b> Juicio evaluativo explícito	<input type="checkbox"/> No formula valoraciones o son genéricas sin fundamentación	<input type="checkbox"/> Formula algunas valoraciones pero con fundamentación limitada o descontextualizada	<input type="checkbox"/> Valoraciones fundamentadas, contextualizadas y con criterios explícitos	
<b>02</b> Toma de posición frente a alternativas	<input type="checkbox"/> No identifica alternativas o no adopta una posición clara	<input type="checkbox"/> Identifica alternativas pero la posición adoptada carece de argumentación sólida	<input type="checkbox"/> Identifica alternativas y adopta posición clara con argumentación fundamentada	
<b>03</b> Asunción de responsabilidad	<input type="checkbox"/> No reconoce consecuencias ni asume responsabilidad	<input type="checkbox"/> Reconoce consecuencias de manera general pero sin asunción explícita de responsabilidad	<input type="checkbox"/> Reconoce explícitamente consecuencias y asume responsabilidad por ellas	
<b>04</b> Contextualización situada	<input type="checkbox"/> Contexto ausente o intercambiable sin afectar el sentido	<input type="checkbox"/> Menciona contexto pero de manera superficial o genérica	<input type="checkbox"/> Inserción específica en contexto normativo, institucional o fáctico no intercambiable	
<b>05</b> Grado de interpretación	<input type="checkbox"/> Texto meramente descriptivo o paráfrasis	<input type="checkbox"/> Alguna interpretación pero limitada o poco fundamentada	<input type="checkbox"/> Construcción de sentido, inferencias fundamentadas, ponderación normativa aplicada	

<p><b>06</b> Experiencia humana o institucional</p>	<input type="checkbox"/> No referencia prácticas, memoria o aprendizajes previos	<input type="checkbox"/> Referencias superficiales o genéricas a experiencia	<input type="checkbox"/> Referencias concretas a prácticas reiteradas, memoria organizacional o saberes operativos	
<p><b>07</b> Ética aplicada al caso</p>	<input type="checkbox"/> Sin consideraciones éticas o solo referencias abstractas	<input type="checkbox"/> Consideraciones éticas genéricas sin aplicación situada	<input type="checkbox"/> Ponderación situada de valores, riesgos, impactos diferenciales y principios aplicados	
<p><b>08</b> Creatividad funcional orientada a un fin</p>	<input type="checkbox"/> No propone soluciones o son genéricas no adaptadas	<input type="checkbox"/> Propone soluciones pero con limitada adaptación al contexto	<input type="checkbox"/> Soluciones viables, instrumentales y adaptadas al contexto institucional específico	

## Criterio de valoración final

Evalúe la articulación y coherencia interna del conjunto de dimensiones:

Clasificación	Criterio	Peso decisorio
<p><b>Nivel Bajo</b></p>	<p>0-5 puntos totales, con mayoría de dimensiones ausentes o presentes solo parcialmente. Escasa articulación entre dimensiones.</p>	<p>El texto presenta indicios insuficientes de intervención humana sustantiva. Prevalecen señales compatibles con generación algorítmica.</p>
<p><b>Nivel Medio</b></p>	<p>6-10 puntos totales, con presencia parcial en varias dimensiones o presencia consistente en algunas pero limitada articulación.</p>	<p>El texto presenta indicios moderados de intervención humana. Se requiere análisis complementario de Niveles 1 y 2 para evaluación integral.</p>
<p><b>Nivel Alto</b></p>	<p>11-16 puntos totales, con presencia consistente en la mayoría de dimensiones y fuerte articulación y coherencia interna.</p>	<p>El texto presenta huellas humanas irreductibles claras. Prevalece sobre resultados de Niveles 1 y 2. Evidencia de autoría humana sustantiva.</p>

## Observaciones metodológicas importantes

### 01. La evaluación no es mecánica

- La suma de puntos es orientativa, no determinante.
- La calidad de la articulación entre dimensiones es más relevante que el puntaje total.
- Algunas dimensiones pueden tener mayor peso según el contexto y tipo de documento.

### 02. Contextualización por tipo de documento

Tipo de documento	Dimensiones especialmente relevantes
<b>Resoluciones administrativas</b>	2 (Toma de posición), 3 (Responsabilidad), 4 (Contextualización), 5 (Interpretación)
<b>Dictámenes jurídicos</b>	2 (Toma de posición), 5 (Interpretación), 7 (Ética aplicada)
<b>Informes técnicos</b>	1 (Juicio evaluativo), 4 (Contextualización), 6 (Experiencia institucional), 8 (Creatividad funcional)
<b>Documentos académicos</b>	1 (Juicio evaluativo), 2 (Toma de posición), 5 (Interpretación), 8 (Creatividad funcional)

### 03. Principio de prevalencia jerárquica:

- Nivel 3 (evidencia de autoría humana sustantiva) prevalece sobre resultados positivos en Nivel 1 (detección algorítmica) y en Nivel 2 (trazabilidad técnica).
- La evaluación final debe ser integral, razonable y contextual.



**IALAB**