# Detectores de IA en la academia.

¿Funcionan? ¿Se puede confiar en ellos?
¿O son otro problema disfrazado de solución?

**Oscar Obando Chaves** | Senior AI Counsel | CAIO
oscarobando.ai
Marzo 2026

# No es lo mismo detectar plagio que detectar IA

**Hay una confusión generalizada entre dos tecnologías que resuelven problemas completamente distintos. Un detector de plagio, como el módulo clásico de Turnitin, compara el texto contra una base de datos de millones de documentos existentes buscando coincidencias textuales. Si copiaste un párrafo de Wikipedia, lo encuentra. Un detector de IA hace algo diferente. Analiza patrones estadísticos internos del texto sin compararlo contra ninguna fuente externa.**

Esto significa que un ensayo generado por ChatGPT, que es técnicamente original y no existe en ninguna base de datos, pasa limpio por un detector de plagio. El detector de IA intenta identificarlo por sus "huellas estadísticas" midiendo dos cosas. La perplejidad, que es qué tan predecible es cada palabra en la secuencia (la IA tiende a elegir palabras probables, los humanos hacen elecciones más inesperadas). Y la explosividad (burstiness), que mide la variación en la complejidad de las oraciones. Los humanos alternan entre oraciones cortas y largas. La IA produce estructuras más uniformes.

Hay una tercera vía en desarrollo. El marcado de agua (watermarking), donde el propio proveedor de IA incrusta una señal estadística oculta durante la generación del texto. Google DeepMind ya ha marcado más de 10.000 millones de piezas de contenido con SynthID desde 2023. El problema es que requiere cooperación del proveedor, no funciona con modelos de código abierto, y se puede eliminar parafraseando. OpenAI evaluó implementar watermarking pero lo pausó por temor a que los usuarios migraran a competidores sin marca de agua.

*Fuentes: GPTZero (perplexity/burstiness), Nature oct 2024 (SynthID), JISC 2025 (OpenAI watermarking), Brookings 2024.*

# Los números que las empresas no ponen en su marketing

**Turnitin declara menos del 1% de falsos positivos y 98% de precisión. GPTZero dice 99,3%. Originality.ai dice 99%. Winston AI proclama 99,98%. Los estudios independientes cuentan otra historia.**

Weber-Wulff et al. (2023, International Journal for Educational Integrity) evaluaron 14 detectores con investigadores de HTW Berlin, Uppsala y otras seis instituciones. Ninguno superó el 80% de precisión en muestras diversas. Hadra, Cambridge y Mesbah (febrero 2026, misma revista) evaluaron 192 textos y encontraron que Originality.ai alcanzó 0,69 de precisión global y Turnitin 0,61. El benchmark RAID (Dugan et al., ACL 2024, Universidad de Pensilvania), con 6,2 millones de generaciones, reveló que ZeroGPT se estabilizó en 16,9% de falsos positivos sin poder reducirlo más.

Perkins et al. (2024, International Journal of Educational Technology in Higher Education) midieron la precisión base de seis detectores principales en 39,5%. Después de que los estudiantes aplicaran ediciones manuales básicas (sinónimos, cambios de longitud de oraciones), la precisión cayó a 22,1%. Estas eran ediciones que cualquier estudiante puede hacer en veinte minutos.

El sesgo contra hablantes no nativos de inglés está documentado por Stanford. Liang et al. (2023, revista Patterns) evaluaron siete detectores en 91 ensayos TOEFL de estudiantes internacionales. La tasa promedio de falsos positivos fue del 61,22%. Más de la mitad de los ensayos fueron incorrectamente clasificados como generados por IA. Un detector marcó casi el 98% de los ensayos TOEFL como artificiales. La razón es que los hablantes no nativos usan vocabulario más simple y estructuras más predecibles, exactamente la señal que los detectores interpretan como IA.

*Fuentes: Weber-Wulff et al. 2023, Hadra et al. 2026, RAID/ACL 2024, Perkins et al. 2024, Liang et al./Stanford 2023.*

# Comparativa de los principales detectores

| Detector | Precisión declarada | Precisión independiente | Español | Precio | Punto débil |
|---|---|---|---|---|---|
| Turnitin AI | 98%, <1% FP | 0,61 global (Hadra 2026). ~30% con parafraseo | Sí | Solo institucional | Cae con parafraseo. 15% de miss rate admitido |
| GPTZero | 99,3% | 52% (Scribbr 2024). 42% post-retrotraducción | Sí | Gratis (10k pal) / $10/mes | Vulnerable a parafraseo y retrotraducción |
| Originality.ai | 99% | 0,69 global (Hadra 2026). 10% FP en español | Sí | $14,95/mes | Agresivo. Más FP en escritura formal |
| Copyleaks | >99%, 0,2% FP | ~50% con parafraseo. Inferior en español | Sí | $9,99/mes | Ausente del benchmark RAID |
| ZeroGPT | >98% | 35-74%. 16,9% FP mínimo (RAID) | Sí | Gratis / $9,99/mes | Mayor tasa de FP de todos los principales |
| Winston AI | 99,98% | ~71% (RAID). Hasta 90% en GPT-4 puro | Sí | $12/mes | Débil con modelos open-source |

OpenAI lanzó su propio detector en enero de 2023 y lo retiró seis meses después. Identificaba correctamente solo el 26% del texto de IA y marcaba incorrectamente el 9% del texto humano. Que la empresa que posiblemente sabe más sobre cómo funcionan los modelos de lenguaje haya logrado 26% de precisión es un indicador devastador sobre la viabilidad de la detección.

*Fuentes: RAID/ACL 2024, Scribbr 2024, Hadra et al. 2026, CorpIdentIA 2025, OpenAI ene-jul 2023.*

# Lo que rompe a los detectores

**Krishna et al. (NeurIPS 2023) crearon DIPPER, un modelo de parafraseo de 11.000 millones de parámetros. DIPPER redujo la precisión de DetectGPT del 70,3% al 4,6%. Evadió exitosamente GPTZero, DetectGPT, marcas de agua y el clasificador de OpenAI. Anderson et al. (2023) encontraron que el parafraseo elevó la puntuación de "escrito por humano" del 0,02% al 99,52%. Evasión completa.**

La retrotraducción (traducir a otro idioma y de vuelta) también funciona. El estudio ESPERANTO (Ayoobi et al., 2024) evaluó nueve detectores con 720.000 textos retrotraducidos. GPTZero cayó del 97% al 42% en noticias y del 65% al 9% en reseñas. Solo Pangram Labs mantuvo más del 96% después de retrotraducción.

Existe un ecosistema creciente de sistemas diseñados para hacer indetectable el texto de IA. Undetectable.ai tiene más de 21 millones de usuarios. StealthWriter, WriteHuman, HIX Bypass y otros ofrecen lo mismo. Turnitin lanzó en agosto 2025 una detección específica de "humanizadores de IA" reconociendo que constituyen una nueva categoría de trampa. Un estudio de Northumbria University (Blommerde, 2025) evaluó seis de estas plataformas y encontró resultados dispares. Algunas fueron detectadas. Otras permanecieron invisibles.

Sadasivan et al. (ICLR 2024, Universidad de Maryland) proporcionaron la prueba teórica más importante. La detección fiable se vuelve matemáticamente imposible cuando las distribuciones de texto humano y de IA se solapan más del 11%. A medida que los modelos mejoran en modelar el lenguaje humano, el mejor detector posible se acerca al rendimiento de un clasificador aleatorio. 50%.

*Fuentes: Krishna et al./NeurIPS 2023, ESPERANTO/Ayoobi et al. 2024, Sadasivan et al./ICLR 2024, Blommerde/Northumbria 2025.*

# Las universidades ya están tomando partido

**Más de 40 universidades de primer nivel han prohibido o desactivado los detectores de IA. MIT, Yale, Northwestern, NYU, Georgetown, UC Berkeley, UCLA, Boston University, Syracuse, University of Toronto, University of British Columbia, entre otras. Vanderbilt University desactivó la detección de IA de Turnitin en agosto de 2023 citando la escala de falsos positivos, el sesgo contra hablantes no nativos, y la duda fundamental de que la detección de IA sea posible. UT Austin prohibió la compra de detectores de IA por completo.**

Los casos de acusaciones falsas son reales. En Texas A&M; (mayo 2023), un profesor acusó a toda su clase de usar ChatGPT después de enviar los trabajos al propio ChatGPT y preguntarle si los había escrito. Los estudiantes recibieron calificaciones incompletas y diplomas retenidos. La universidad confirmó después que ningún estudiante reprobó. En Yale (febrero 2025), un estudiante de EMBA, hablante nativo de francés, fue suspendido por un año después de que GPTZero marcara su examen. Presentó demanda federal alegando discriminación. En University of Houston (2025), una estudiante que escribió su trabajo con un tutor del centro de escritura de la universidad fue marcada como 100% IA por Turnitin.

Ninguna universidad identificada en esta investigación utiliza los puntajes de detección como prueba única de deshonestidad académica. Todas incluyen advertencias. University of Kentucky indica que la escritura marcada "no puede verificarse contra otra evidencia". Johns Hopkins reconoce "un amplio rango en la precisión". MIT Sloan publicó una guía titulada "Los detectores de IA no funcionan. Esto es lo que hay que hacer en su lugar."

*Fuentes: Vanderbilt 2023, PLEASE.edu tracker, Rolling Stone 2023, Crowell & Moring 2025, CalMatters/GradPilot 2025-2026.*

# ¿Y América Latina?

**El 65% de las instituciones latinoamericanas aún carecen de políticas oficiales sobre IA. Mientras 40+ universidades del norte global ya prohibieron los detectores por poco fiables, la región está importando estas tecnologías sin el debate que las acompaña.**

Chile lidera el Índice Latinoamericano de IA 2024 con una política nacional proyectada a 2031. México tiene el Observatorio Interinstitucional de IA en la Educación Superior (OIIAES) con participación de UNAM, UAM, La Salle y Anáhuac, pero las universidades mexicanas aún no han desarrollado marcos regulatorios específicos para la detección de IA. Colombia está adoptando marcos tipo "semáforo" que clasifican actividades de IA como permitidas, restringidas o prohibidas. En Brasil, la Universidad Federal de Goiás ha pasado de prohibir la IA a requerir citación explícita de su uso.

Un dato de alerta. En la Universidad de la República (Uruguay), el porcentaje de estudiantes con calificación "excelente" saltó del 11,9% al 48,3% en un año tras la adopción de IA. En España, el 75% del profesorado ha encontrado plagio con IA generativa (Sánchez-Vera et al., Frontiers in Education, 2025). El primer estudio específico en español (CorpIdentIA, 2025) encontró que Originality.ai alcanza 100% de precisión en textos de IA pura en español pero con 10% de falsos positivos en textos humanos, y que ningún detector es fiable para textos híbridos en nuestro idioma.

UNESCO publicó en 2023 la primera guía global sobre IA generativa en educación pero no recomienda detectores específicos. El AI Act de la Unión Europea clasifica los sistemas de IA usados para evaluar resultados de aprendizaje y detectar comportamiento de estudiantes en exámenes como "alto riesgo", lo que podría obligar a los detectores a cumplir estándares estrictos de precisión y transparencia que hoy no cumplen.

*Fuentes: Sociedad 3.0 feb 2026, Infobae ago 2025, UNIVIM, CorpIdentIA 2025, UNESCO 2023/2025, AI Act/EU 2024.*

***"Nuestros resultados apuntan a la imposibilidad de la detección de texto generado por IA en escenarios prácticos."***

— Soheil Feizi, profesor de Ciencias de la Computación, Universidad de Maryland. Coautor de "Can AI-Generated Text be Reliably Detected?" (ICLR 2024)

**oscarobando.ai**

linkedin.com/in/oscar-obando-chaves-237495133

Sígueme en LinkedIn para análisis semanal sobre el impacto real de la IA