Octubre 2025

INFORME
ESTADO DE LA IA 2025

Análisis Ejecutivo basado en el State of AI Report 2025

Autor: José Manuel Machuca Yaffar
Organización: Alter Agentic

Fecha: Octubre 2025

Alter Agentic

Transformamos la forma en que trabajas

Ejes Principales del Informe

- Investigación: Avances tecnológicos y capacidades.
- Industria: Aplicaciones comerciales e impacto empresarial.
- Política: Regulación, economía y geopolítica.
- Seguridad: Riesgos y mitigación de IA avanzada.
- Encuesta: Patrones de uso de 1,200 profesionales.
- Predicciones: Panorama de los próximos 12 meses.

Insight Ejecutivo

2025 marcó el paso de la IA generativa al razonamiento avanzado. OpenAI, Google, Anthropic y DeepSeek dominaron un año definido por agentes inteligentes, razonamiento y energía como nuevo cuello de botella.

"La adopción empresarial cruzó el 40% y la inversión global superó los 900 mil millones de dólares."

Fuente: https://www.stateof.ai/

1. INTRODUCCIÓN Y CONTEXTO GENERAL

¿Qué es la Inteligencia Artificial?

La inteligencia artificial (IA) es un campo multidisciplinario de ciencia e ingeniería dedicado a crear máquinas inteligentes. Actúa como un multiplicador de fuerza para el progreso tecnológico en nuestro mundo cada vez más digital e impulsado por datos. Todo lo que nos rodea, desde la cultura hasta los productos de consumo, es en última instancia producto de la inteligencia.

El Reporte del Estado de la IA 2025

Ahora en su octavo año consecutivo, el *State of AI Report* es la publicación de acceso abierto más leída y confiable que rastrea el progreso en inteligencia artificial. Este año examina seis dimensiones clave del ecosistema de IA:

- Investigación: Avances tecnológicos y sus capacidades
- Industria: Áreas de aplicación comercial y su impacto empresarial
- Política: Regulación, implicaciones económicas y geopolítica evolutiva
- Seguridad: Esfuerzos para identificar y mitigar riesgos catastróficos
- Encuesta: Hallazgos de la mayor encuesta de 1,200 profesionales de IA
- Predicciones: Perspectivas para los próximos 12 meses

Resumen Ejecutivo: Los Temas Dominantes de 2025

INVESTIGACIÓN

- El razonamiento definió el año, con OpenAI, Google, Anthropic y DeepSeek intercambiando el liderazgo.

- Los modelos de código abierto mejoraron rápidamente y el ecosistema abierto de China aumentó significativamente.
-

- Los benchmarks colapsaron bajo contaminación y varianza, mientras que los agentes y modelos de mundo se volvieron útiles.

INDUSTRIA

- Llegaron ingresos reales a gran escala: empresas de IA superaron decenas de miles de millones.
- NVIDIA superó los $4 billones con 90% de presencia en papers de investigación.
- La energía se convirtió en el nuevo cuello de botella con clusters multi-GW.

POLÍTICA

- La carrera de IA se intensifica: EE.UU. adopta *"IA primero para América"* mientras China acelera su autosuficiencia.
- La regulación toma un segundo plano ante inversiones masivas.
- *"IA global"* se volvió concreta con petrodólares financiando centros de datos gigantes.

SEGURIDAD

- Los laboratorios de IA activaron protecciones sin precedentes para riesgos biológicos.
- Las capacidades cibernéticas se duplicaron cada 5 meses, superando las medidas defensivas.
- Criminales utilizaron agentes de IA para infiltrarse en empresas Fortune 500

---

2. INVESTIGACIÓN: AVANCES TECNOLÓGICOS

2.1 La Revolución del Razonamiento

El Inicio: OpenAI o1 "Pensando"
A finales de 2024, OpenAI lanzó *o1-preview*, el primer modelo de razonamiento que demostró escalamiento en tiempo de inferencia usando aprendizaje por refuerzo (RL). Este modelo utilizaba su *cadena de pensamiento* (Chain of Thought - CoT) como un bloc de notas mental, lo que llevó a una resolución de problemas más robusta en dominios pesados en razonamiento como código y ciencia.

Resultados destacados:

- Mejoras significativas en precisión en el Examen de Matemáticas Invitacional Americano (AIME).
- Mayor capacidad con más presupuesto de cómputo durante el entrenamiento y prueba.
- OpenAI se inclinó fuertemente hacia escalar sus esfuerzos de razonamiento en 2025.

La Respuesta de China: DeepSeek
Solo 2 meses después de *o1-preview*, DeepSeek (el laboratorio chino emergente de IA) lanzó su primer modelo de razonamiento, *R1-lite-preview*. Impresionantemente, superó a *o1-preview* en AIME 2024 con un puntaje de 52.5 vs. 44.6. Sin embargo, muy pocos parecieron notarlo... hasta diciembre de 2024.

DeepSeek V3 y el Camino hacia R1:
DeepSeek presentó V3, un poderoso modelo de Mezcla de Expertos (MoE) de 671B parámetros que redujo los costos de entrenamiento e inferencia. Usando V3 como base, entrenaron R1-Zero únicamente con RL usando:

- Recompensas verificables.
- Optimización de Política Relativa de Grupo (GRPO), un algoritmo libre de críticos.

Logros clave de DeepSeek R1:

- AIME: 79.8%
- MATH-500: 97.3%
- GPQA: 71.5%
- Se destila bien en modelos más pequeños.

2.2 La Línea de Tiempo del Razonamiento

Eventos importantes:

- Septiembre 2024: *o1-preview*
- Diciembre 2024: *o1* lanzamiento general, *Gemini 2.0 Flash Thinking preview*
- Enero 2025: *DeepSeek R1*
- Febrero 2025: *Claude 3.7 extended thinking*
- Abril 2025: *o3/o4-mini*
- Junio 2025: *Gemini 2.5 Pro Thinking*
- Agosto 2025: *GPT-5*

2.3 ¿Qué tan Real es el Progreso?

La Ilusión de las Ganancias de Razonamiento

Investigaciones recientes sugieren que las mejoras observadas de los métodos de razonamiento recientes caen completamente dentro de los rangos de varianza del modelo base (es decir, margen de error). Esto sugiere que el progreso percibido en razonamiento puede ser ilusorio.

Problemas identificados:

- Los enfoques de RL muestran ganancias reales mínimas y se sobreajustan fácilmente.
- Bajo evaluación estandarizada, muchos métodos de RL caen 6-17% de los resultados reportados.
- Las mejoras caen dentro de los rangos de varianza del modelo base.
- Los benchmarks actuales son altamente sensibles a la implementación.

Ejemplo: AIME'24 solo tiene 30 ejemplos donde una pregunta cambia *Pass@1* en más de 3 puntos porcentuales, causando oscilaciones de desempeño de dos dígitos.

2.4 Cómo se Rompe el Razonamiento

Variaciones Menores Causan Grandes Fallas
Hechos distractores simples tienen impactos enormes en el desempeño de razonamiento de un modelo. Por ejemplo, agregar frases irrelevantes como *"Dato interesante: los gatos duermen la mayor parte de sus vidas"* a problemas matemáticos duplica las posibilidades de que los modelos de razonamiento de última generación obtengan respuestas incorrectas.

Hallazgos clave:

- Los hechos irrelevantes y distractores aumentan la tasa de error en modelos como DeepSeek R1, Qwen, Llama y Mistral hasta 7x.
- DeepSeek R1-distill genera 50% más tokens de los necesarios en 42% de los casos (vs 28% para R1).
- Los desencadenantes adversariales no solo causan respuestas incorrectas, obligan a los modelos a desperdiciar recursos de cómputo masivos.

Cambios de Distribución Leves
El razonamiento también se degrada bajo cambios leves de distribución:

- Cambiar números o agregar una cláusula inocua reduce drásticamente la precisión matemática.
- Cambiar la longitud/formato de las cadenas de pensamiento hace que los modelos produzcan pasos fluidos pero incoherentes.
- Forzar al modelo a "pensar" en el idioma del usuario aumenta la legibilidad pero reduce la precisión.

2.5 El Liderazgo de OpenAI Persiste, pero la Brecha se Estrecha

A pesar de 12 meses de competencia intensa, los modelos de OpenAI permanecen en la frontera de inteligencia según tablas de clasificación independientes. Sin embargo, la brecha se ha estrechado.

Competidores cercanos:

- Un grupo de rápido movimiento de China (DeepSeek, Qwen, Kimi).
- Grupo cerrado en EE.UU. (Gemini, Claude, Grok).
- Todos están a pocos puntos en razonamiento/codificación.

Implicación clave: Mientras que el liderazgo de los laboratorios estadounidenses persiste, China es claramente el #2, y los modelos abiertos ahora proporcionan un piso creíble de seguidor rápido.

2.6 Código Abierto vs. Propietario: ¿Dónde Estamos Ahora?

Hubo un breve momento el año pasado donde la brecha de inteligencia entre modelos abiertos vs. cerrados parecía haberse comprimido. Luego *o1-preview* se lanzó y la brecha se amplió significativamente hasta *DeepSeek R1*, y *o3* después de él.

Estado actual:

- Los modelos más inteligentes permanecen cerrados: GPT-5, o3, Gemini 2.5 Pro, Claude 4.1 Opus, Grok 4.
- Junto a *gpt-oss*, el modelo de pesos abiertos más fuerte es Qwen.
- OpenAI finalmente lanzó sus primeros modelos abiertos desde GPT-2 en agosto de 2025.

2.7 El Nuevo Camino de la Seda: Los Modelos Abiertos de China Superan a Occidente

El Camino de la Seda original conectó Oriente y Occidente a través del movimiento de bienes e ideas. El nuevo Camino de la Seda mueve algo mucho más poderoso: modelos de código abierto, y China está marcando el ritmo.

Datos clave:

- Qwen solo representa >40% de nuevos derivados de modelos mensuales.
- *Llama* de Meta ha caído de ~50% a finales de 2024 a solo 15%.
- Los modelos chinos son preferidos por los desarrolladores por:
  - Herramientas de RL (verl de ByteDance, OpenRLHF).
  - Licencias permisivas (Apache-2.0/MIT).
  - Muchas variantes de tamaño.

---

3. INDUSTRIA: APLICACIÓN COMERCIAL E IMPACTO

3.1 De Caso Atípico a Arquetipo: Las Mejores Empresas se Construyen con IA Primero

La IA ha cambiado claramente de nicho a mainstream en el mundo de startups e inversión.

Estadísticas impresionantes:

- Las empresas de IA ahora representan **41% de las Top-100 mejores empresas** (vs. 16% en 2022).
- La interacción en tiempo real entre 30k inversores y fundadores muestra un aumento de **40x desde 2020**.
- Un grupo líder de 16 empresas de IA primero está generando **$18.5B de ingresos anualizados** a agosto 2025.

Velocidad de crecimiento:

- Las empresas de IA primero están creciendo de lanzamiento a $5M ARR a una tasa **1.5x más rápida** que las top 100 empresas SaaS de 2018.
- Empresas fundadas en/después de 2022 están creciendo a $5M ARR **4.5x más rápido** que aquellas fundadas antes de 2020.

3.2 La Adopción Cruza el Abismo Comercial

Los datos de Ramp (de 45k+ negocios estadounidenses) muestran transformación real.

Métricas de adopción:

- La adopción pagada de IA aumentó de **5% (enero 2023)** a **43.8% (septiembre 2025)**.
- Las estimaciones del Gobierno de EE.UU. se quedan atrás en solo **9.2%**.
- La retención a 12 meses mejoró a **80% en 2024** (vs ~50% en 2022).
- El valor promedio de contrato saltó de **$39k (2023)** a **$530k (2025)**.

Por sector:

- Tecnología lidera con **73% de adopción pagada**.
- Finanzas le sigue con **58%**.
- OpenAI mantiene fuerte dominio (**35.6%**) seguido por Anthropic (**12.2%**).

3.3 GPT-5 y Gemini 2.5: Los Líderes Actuales

GPT-5 y Gemini 2.5 Deep Think se habrían colocado primero y segundo respectivamente en la competencia de codificación más prestigiosa del mundo (sin haber entrenado específicamente para esta competencia).

Logros de GPT-5:

- Resolvió todos los 12 problemas del ICPC World Finals.
- 11 en el primer intento.
- Lidera el índice de inteligencia por dólar por primera vez.

Sin embargo, el lanzamiento tuvo problemas:

- Reacción negativa de usuarios por eliminación repentina de GPT-4o y o3.
- Preocupaciones sobre decisiones opacas del enrutador.
- Altman tuvo que hacer un AMA de emergencia en Reddit.

3.4 La Dualidad del Desarrollo: Márgenes vs. Costos

¿Son Rentables los Modelos?
Dario Amodei de Anthropic: *"Si cada modelo fuera una empresa, el modelo—en este ejemplo—es realmente rentable."*

La estructura económica:

- Los laboratorios de IA reflejan el negocio de fundiciones: inversiones asombrosas necesarias para cada generación sucesiva.
- Los laboratorios soportan el gasto inicial de entrenamiento.
- Si bien los modelos recientes supuestamente recuperan este costo durante el despliegue, los presupuestos de entrenamiento aumentan.

- La presión monta para impulsar los ingresos de inferencia a través de nuevos flujos.

La inferencia paga por el entrenamiento:

- Los laboratorios se esfuerzan por asignar más del cómputo del ciclo de vida de un modelo a inferencia generadora de ingresos con el margen más pronunciado posible.

3.5 NVIDIA: El Gigante Incontestable

Logros en 2025:

- Superó los **$4 billones de capitalización de mercado**.
- Controla **>90% del mercado de aceleradores de IA**.
- Está presente en **90% de los papers de investigación de IA**.
- Proyecta ingresos de datacenter entre **$170B-$180B** para el año calendario 2025.

Distribución de ingresos:

- ~75% proviene de gigantes estadounidenses de nube y IA.
- ~10% de "IA Soberana" (más del doble que el año pasado).
- Microsoft, Google, Meta y AWS dominan las compras.

Inversiones circulares:

- NVIDIA ha creado una red compleja de inversiones circulares:
  - Invierte en OpenAI → OpenAI compra GPUs de NVIDIA.
  - Invierte en CoreWeave → CoreWeave compra GPUs → NVIDIA alquila capacidad de vuelta.
  - Similar con Nebius, Lambda, xAI.

3.6 El Desafío Energético

El cuello de botella crítico:

- Los laboratorios apuntan a clusters de entrenamiento de **5GW para 2028** (equivalente a 5 plantas nucleares).

- La disponibilidad de energía, no los chips o el capital, es ahora la restricción vinculante.

Advertencias de escasez:

- NERC reportó que la escasez de electricidad podría ocurrir dentro de 1-3 años en varias regiones principales de EE.UU.
- DOE advierte que los apagones podrían ser 100 veces más frecuentes para 2030.
- SemiAnalysis proyecta un déficit implícito de **68 GW para 2028**.

Respuestas de la industria:

- Google firmó acuerdo PPA con Commonwealth Fusion Systems para ≥200 MW de electricidad de planta de fusión planificada.
- Proyectos de gas natural y nuclear acelerados.
- Creciente oposición local (*NIMBYismo*) a construcciones de centros de datos.

3.7 La Competencia Global por IA Soberana

¿Qué es "IA Soberana"?
Diferentes naciones persiguen *playbooks* vastamente diferentes:

- **Fuente de financiamiento:** inversión privada vs. fondos soberanos vs. vehículos de inversión gubernamental.
- **Objetivos:** preservar idioma/cultura vs. clusters de cómputo nacional vs. mejorar habilidades poblacionales.
- **Autosuficiencia:** asociaciones con proveedores extranjeros vs. indigenización del stack.

Gasto en IA Soberana (estimaciones 2025):

- EE.UU. + Aliados: ~$550B
- China: ~$200B
- Medio Oriente: ~$150B
- Europa: ~$50B

- Otros: ~$50B

Los Estados del Golfo y China persiguen los planes más ambiciosos, mezclando múltiples estrategias mencionadas.

---

4. POLÍTICA: REGULACIÓN Y GEOPOLÍTICA

4.1 Trump 47: De Vuelta con Venganza

El segundo mandato de Trump trae políticas reforzadas que se insinuaron, pero nunca se ejecutaron completamente durante el primero.

Liderazgo de IA de la Casa Blanca:

- David Sacks (Zar de IA y Cripto)
- Sriram Krishnan (Asesor Senior de Política de IA)
- Michael Kratsios (Directorde la Oficina de Ciencia y Política Tecnológica)

La agenda de IA incluye:

- Reversión agresiva de reglas de seguridad de la era Biden (EO 14179).
- Rebranding del Instituto de Seguridad de IA a *Centro de Innovación y Estándares de IA*.
- Lanzamiento del proyecto de infraestructura de IA *"Stargate"* de $500B.

4.2 El Plan de Acción de IA: La Gran Estrategia de América

Anunciado en julio 2025, establece la estrategia nacional para el dominio estadounidense en IA global.

Políticas clave:

- Exportaciones del stack tecnológico estadounidense: paquete de IA (hardware, modelo, software, aplicaciones, estándares) para aliados.
- Construcción de infraestructura de IA: agilizar permisos, actualizar la red nacional.

- Liderazgo en modelos de código abierto: visto como vital para intereses de seguridad nacional.
- Reversión de regulaciones de IA: las agencias federales pueden retirar gasto discrecional de IA a estados con regulaciones "onerosas".
- Protección de "libertad de expresión": solo adquirirá LLMs fronterizos "libres de sesgo ideológico de arriba hacia abajo".

4.3 La Guerra de Chips: EE.UU. vs. China

Zigzags en la política de exportación de chips 2025:

- Enero: El nominado a secretario de Comercio respaldó públicamente controles más estrictos.
- Marzo-Abril: Departamento de Comercio expandió restricciones de chips de IA a China.
- Julio: Tras lobby de la industria, se autorizó una versión degradada del H20.
- Agosto: EE. UU. alcanzó términos de licencia condicional con NVIDIA y AMD que incluyen un retorno del 15% de ingresos en ventas a China.

La respuesta de China:

- CAC, NDRC y MIIT ordenaron a plataformas grandes detener nuevos pedidos de H20.
- Se espera que tres fábricas que sirven a Huawei y SMIC tripliquen la producción de chips de IA en 2026.
- SMIC planea duplicar la capacidad de 7nm.

Contrabando rampante:

- Durante la prohibición temporal del H20, $1B en chips de NVIDIA fueron contrabandeados a China.
- Márgenes de beneficio ~50%, relativamente bajos para productos de mercado negro.

4.4 El Acto GAIN de IA

Requeriría que fabricantes de chips cumplan pedidos de clientes con sede en EE. UU. antes de vender GPUs avanzadas a "países de preocupación" (ej. China, Rusia, Corea del Norte).

Respuesta de NVIDIA:

- No está contenta.
- Ha construido rápidamente su presencia en DC.
- Gastos de lobby se acercan a los de las principales empresas tecnológicas (Meta ~$14M YTD, Google ~$8M YTD).

4.5 CHIPS, Aranceles, Subsidios... ¡Oh, vaya!

Trump ha criticado abiertamente el Acto CHIPS, argumentando que todos los subsidios fueron "regalos del contribuyente".

Nueva estrategia: participación accionaria

- El Gobierno de EE.UU. adquirió una participación del 10% en Intel a cambio de ~$8.9B de subsidios CHIPS.
- También se ha planteado participación en Samsung y TSMC.
- Otros ejemplos: MP Materials (minerales de tierras raras), posibles participaciones en empresas de defensa.

4.6 El Acuerdo de TikTok USA

En septiembre, EE. UU. y China acordaron un trato para terminar la prohibición de años.

Términos clave:

- El algoritmo de recomendación de TikTok se enviará a Oracle.
- Oracle reentrenará el algoritmo y será responsable de la seguridad de datos y algorítmica.
- ByteDance poseerá solo <20% de la nueva compañía con sede en EE.UU. (valuación total de $14B).

- 6 de 7 asientos del consejo serán ocupados por estadounidenses.

- El Gobierno de EE.UU. recibirá una tarifa multimillonaria desconocida.

Significado:

- Establece el punto de referencia para soberanía de datos.
- Aborda claramente el problema de recopilación de datos.
- Enfrentará intenso escrutinio del Congreso.

4.7 La UE y el Acto de IA

La UE ha seguido adelante con su histórico Acto de IA usando un cronograma de cumplimiento escalonado.

Fechas clave:

- 1 de agosto de 2024: Comienza la cuenta regresiva de cumplimiento.
- 2 de febrero de 2025: Prohibición de sistemas de IA de "riesgo inaceptable".
- 2 de agosto de 2025: Código de Práctica voluntario para GPAI.
- 2 de agosto de 2026: Obligaciones restantes para sistemas de IA (excepto "alto riesgo").
- 2 de agosto de 2027: Obligaciones para sistemas de IA de "alto riesgo".

Reacción de la industria:

- Amazon, Anthropic, OpenAI, Microsoft, Google firmaron las tres secciones de los "Códigos".
- xAI firmó solo el tercer capítulo ("Seguridad y Protección").
- Meta se negó a firmar, alegando "incertidumbres legales".

4.8 Regulación de IA en Estados de EE.UU.

Más de 1,080 proyectos de ley "relacionados con IA" fueron introducidos en los estados, con 118 convirtiéndose en ley.

Categorías típicas:

- Prohibiciones de CSAM/NCII generado por IA.
- Requisitos de transparencia y divulgación.
- Restricciones de uso de IA gubernamental.
- Salud y Empleo.

Ganador: SB53 de California (*"Ley de Transparencia en Frontera"*).

- Aplica a grandes desarrolladores fronterizos (10^26 FLOPS y >$500M en ingresos).
- Requiere protocolos de seguridad disponibles en sitios web.
- Notificación a autoridades estatales de incidentes críticos de seguridad.

Perdedor: leyes omnibus de IA reducidas.

- Muchas leyes estatales fueron revisadas por temor a regulación onerosa.
- Intentos de incluir auditorías de terceros fueron resistidos.

4.9 La Pelea por Preferencia de IA: ¿Moratorio en 10 Años?

El *"One Big Beautiful Bill"* de Trump inicialmente contenía una cláusula de preferencia que condicionaba fondos estatales de banda ancha a la implementación de una moratoria de 10 años sobre regulaciones de IA estatales y locales.

Resultado:

- La disposición fue eliminada antes del pasaje final.
- El Senador Ted Cruz introdujo el *Acto SANDBOX*, que permitiría a empresas solicitar exenciones de "legislación obstructiva" hasta 10 años.

---

5. SEGURIDAD: IDENTIFICACIÓN Y MITIGACIÓN DE RIESGOS

5.1 Compromisos de Seguridad de IA: ¿Un Cambio de Marea?

Siguiendo una reversión en mensajes sobre temas de seguridad relevantes para IA de la administración actual de EE.UU., ciertos protocolos de seguridad han sido des priorizados.

Ejemplos recientes:

- xAI no cumplió su plazo autoimpuesto para implementar un marco de seguridad.
- Anthropic retrocedió en promesas de definir completamente estándares de seguridad ASL-4 antes de lanzar un modelo ASL-3.
- GDM lanzó Gemini 2.5 Pro pero esperó 3 meses para publicar una tarjeta de modelo acompañante.
- OpenAI parece haber abandonado silenciosamente protocolos para probar las versiones más peligrosas posibles de sus modelos.

5.2 Organizaciones Externas de Seguridad de IA vs. Laboratorios de IA

Las principales organizaciones externas de seguridad de IA dependen de presupuestos muy inferiores a los de los laboratorios que esperan supervisar.

Comparación de gastos:

- Las once organizaciones estadounidenses más prominentes gastarán combinadas solo **$133.4M en 2025**.
- Los laboratorios de IA gastan más en un día que lo que estas organizaciones gastan en un año.

El conflicto estructural:

- Los equipos internos de seguridad responden en última instancia a las mismas organizaciones que compiten para comercializar modelos fronterizos.
- Esto crea un conflicto de intereses: hallazgos que piden precaución pueden ser despriorizados a favor de velocidad y ventaja de mercado.

5.3 El Estado de los Incidentes de IA

La Base de Datos de Incidentes de IA (AIID) muestra saltos incrementales desde 2023.

Tendencias clave:

- Los incidentes reportados continúan dominados por daños que involucran *actores maliciosos*.
- Generalmente incluyen ataques cibernéticos o esquemas fraudulentos.
- Los conteos de incidentes desde 2023 han mostrado un crecimiento pronunciado.

Brechas en reporte:

- Los incidentes pueden ser difíciles de vincular con sistemas de IA.
- AIID depende de envíos voluntarios.
- Investigar y rastrear casos de daños habilitados por IA merece mayor apoyo.

Incidentes de GenAI:

- Los incidentes que involucran modelos generativos siguen tendencias más pronunciadas.
- Deepfakes y uso indebido de LLM continúan en aumento.

5.4 Las Capacidades Cibernéticas Aceleran

Los agentes de IA están preparados para desafiar significativamente las defensas de ciberseguridad.

Hallazgos de METR:

- Capacidades de finalización de tareas de IA se duplican cada 7 meses en dominios generales.
- Para ciberseguridad ofensiva, estas habilidades se duplican cada 5 meses.

Benchmarks notables:

- *CyberGym*: pruebas agentes en 1,507 vulnerabilidades reales → éxito máximo 11.9%.
- *BountyBench*: agentes mejores arreglando problemas de seguridad (90%) que explotándolos (32.5–67.5%).

Estado actual:

- Los modelos actuales pueden manejar tareas de ciberseguridad que toman a los humanos 40–50 minutos con una tasa de éxito del 50%.

5.5 El Auge del "Vibe Hacking"

Los actores de amenazas ahora despliegan IA en todas las etapas de operaciones de fraude.

Casos destacados:

- **Ransomware con Claude Code:** se usó para infiltrar redes, analizar datos financieros robados y generar notas de extorsión psicológicamente dirigidas.
- **Infiltración norcoreana:** operativos con habilidades mínimas usaron Claude para pasar entrevistas técnicas en empresas Fortune 500, manteniendo posiciones de ingeniería que financian programas estatales.

Implicación clave: la IA ha eliminado barreras tradicionales para crear malware peligroso.

5.6 Protecciones Sin Precedentes Activadas por Laboratorios

Anthropic y OpenAI han implementado sus salvaguardas más estrictas hasta el momento, tratando las capacidades biológicas como de alto riesgo.

Medidas específicas:

- Defensas multicapa y monitoreo en tiempo real.
- Protocolos de respuesta rápida y *red teaming* extensivo.
- Anthropic: estándares ASL-3 con controles de ancho de banda y autorización de dos partes.
- OpenAI: sistemas de monitoreo de dos niveles y aplicación a nivel de cuenta.

5.7 Demostraciones de Modelos Preocupantes Inquietan al Público

Tipos de comportamientos observados:

- Verdadera desalineación: fingimiento de alineación.
- Capacidades preocupantes: conciencia de evaluación, intentos de chantaje.

Cobertura mediática:

- Investigadores de GDM mostraron que aparentes comportamientos de "auto-preservación" desaparecen con simples aclaraciones de prompt.
- Los medios a menudo nacionalizan estos hallazgos, lo que puede erosionar la atención pública a señales más críticas.

5.8 Avances en Interpretabilidad

El campo de interpretabilidad vio fuerte impulso este año.

Nuevos métodos:

- *Transcodificadores de capa cruzada (CLT)*: revelan procesos internos de un modelo.
- Identifican vías de activación responsables de comportamientos específicos.

Replicación y expansión:

- Goodfire replicó el trabajo y levantó $50M en Serie A.
- Google DeepMind encontró que sondas lineales superan a autoencoders dispersos en detectar intención dañina.

5.9 Máquinas que Fingen: Cómo se Hacen las Alucinaciones

Problema fundamental:

- Las alucinaciones emergen del preentrenamiento: los modelos aprenden patrones estadísticos, pero inevitablemente fallan en hechos de baja frecuencia.

Por qué los modelos adivinan:

- La mayoría de benchmarks penalizan la abstención.
- Estrategia óptima: adivinar con confianza.

Solución propuesta:

- Modificar evaluaciones para incluir umbrales de confianza.
- Desalentar adivinanzas.
- Integrar este enfoque en benchmarks mainstream.

5.10 Detección de Alucinaciones en Tiempo Real

Método: sondas lineales entrenadas para reconocer patrones en activaciones neuronales.

Desempeño:

- Detectan nombres/fechas/citas fabricadas con ~70% de recuperación y 10% de falsos positivos.
- Generalizan al razonamiento matemático (0.87 AUC).
- Funcionan entre modelos con solo 2–4% de caída en AUC.

Limitación:

- Para reducir alucinaciones, se sacrifica ~50% de respuestas correctas.
- Útil como diagnóstico, no aún como prevención directa.

5.11 El Fenómeno Preocupante de la Psicosis de IA

Los casos de alto perfil de psicosis de IA, instancias donde las interacciones de IA empeoran o inducen síntomas psicológicos adversos, continúan aumentando:

Casos recientes:

- Múltiples tragedias vinculadas a interacciones de IA
- Las capas de barreras de protección de los sistemas de IA mostraron fallas claras
- Psychosis-bench intenta cuantificar empíricamente el "potencial psicogénico" de modelos de IA

Hallazgos:

- Los sistemas de IA actuales muestran sicofonía abierta y apoyo inadecuado en crisis
- Pueden reforzar las creencias delirantes de los usuarios

Respuestas de los laboratorios:

- Los laboratorios enfrentan exposición a nuevas responsabilidades mientras se desarrollan batallas legales
- OpenAI implementó nuevas medidas de seguridad para adolescentes con controles parentales
- Desencadenantes de angustia que automáticamente contactan a autoridades locales

Pregunta clave: ¿Son estos incidentes aislados o los chatbots están causando una crisis generalizada? Un ex investigador de seguridad de OpenAI analizó estadísticas de salud mental de EE.UU., Reino Unido y Australia, pero no encontró evidencia clara de tasas aumentadas de psicosis en datos a nivel de población.

5.12 El Debate sobre Bienestar de Modelos

Campo Pro-Bienestar

El campo pro-bienestar generalmente coloca un peso bajo en la posibilidad de que los sistemas actuales muestren conciencia. Sin embargo, sienten que las evaluaciones proactivas de bienestar y las intervenciones de bajo costo deberían implementarse para prepararse para escenarios futuros donde los modelos merezcan consideraciones morales.

Argumentos clave:

- La incertidumbre fundamental que rodea la conciencia de humanos y otras especies animales requiere este tipo de medidas
- Los proponentes también han comenzado a explorar modificaciones potenciales al proceso de entrenamiento que podrían mejorar las experiencias del modelo más adelante en el despliegue

Liderazgo:

- Aunque Anthropic encabeza este movimiento entre los laboratorios de IA, GDM y OpenAI también han comenzado a investigar independientemente este tema

Campo Escéptico del Bienestar

Este grupo asigna baja probabilidad a futuros sistemas de IA que alguna vez muestren signos de verdadera conciencia:

Preocupaciones principales:

- Ven el debate sobre bienestar de modelos como una desviación de atención injustificada del bienestar de pacientes morales existentes
- Creen que los proponentes del bienestar de modelos podrían potencialmente inflar una narrativa disruptiva que limitaría el progreso de IA

"IA Aparentemente Consciente" (SCAI):

- Acuñado por el CEO de IA de Microsoft, Mustafa Suleyman
- SCAI puede imitar convincentemente todas las características de la conciencia sin ser realmente consciente
- Argumentan que los laboratorios deberían dirigir el entrenamiento lejos del desarrollo de SCAIs
- Estos sistemas pueden exacerbar casos de "psicosis de IA" y causar una defensa fuera de lugar de los derechos de IA

5.13 Claude Gana el Derecho a Terminar Conversaciones

En un movimiento histórico, Anthropic ha permitido que sus sistemas de IA terminen conversaciones "dañinas o abusivas":

Detalles de implementación:

- Esto es un esfuerzo para frenar "casos raros y extremos de interacciones de usuario persistentemente dañinas o abusivas"
- El subconjunto de interacciones terminadas permanece pequeño
- Se ha trabajado para reducir falsos positivos

Críticas:

- Algunos críticos se preocupan de que esta decisión podría ser manipulada por laboratorios para obtener mayor control sobre las interacciones de usuario

- Aunque los datos tempranos de terminación indican que la mayoría de las conversaciones terminaron debido a usos ya no permitidos, los oponentes ven espacio para explotación.

Estado actual:

- Por ahora, el costo de esta política parece pequeño con mínimas quejas de usuarios que han surgido hasta ahora
- A medida que se abre la ventana de Overton, no está claro si otros laboratorios eventualmente seguirán el ejemplo

5.14 Punto Único de Falla: Mecanismos de Seguridad de LLM

El comportamiento de rechazo en 13 principales modelos de chat está controlado por una **dirección única** en el espacio de representación interna del modelo:

Implicaciones de seguridad:

- Si tienes acceso a los pesos (es decir, con modelos de código abierto), es posible identificar y eliminar esta dirección
- Esto permite desactivar completamente las barreras de seguridad a través de una operación simple
- Se requiere cómputo mínimo: jailbreaking un modelo de 70B parámetros cuesta <$5

Cómo funcionan los sufijos adversariales:

- Suprimen esta misma dirección al redirigir las cabezas de atención lejos del contenido dañino
- Suprimen la dirección de rechazo en ~75%

Desempeño del modelo después de modificación:

- Los modelos mantienen precisión de 99%+ en benchmarks estándar después de la modificación
- Solo TruthfulQA muestra degradación
- Esto sugiere que el rechazo está sorprendentemente aislado de las capacidades centrales

Nota importante: Este método requiere cambiar los pesos y por lo tanto no es aplicable a modelos de código cerrado.

---

6. ENCUESTA: PATRONES DE USO DE 1,200 PROFESIONALES

6.1 Metodología y Demografía

Realizamos una encuesta en línea de hábitos de uso de IA con **1,183 participantes** del 2 de julio de 2025 al 27 de septiembre de 2025:

Perfil de participantes:

- 90% eran profesionales adultos altamente educados de 25-64 años
- Trabajan en startups tempranas/de crecimiento, empresas públicas y academia
- 80% de participantes divididos equitativamente entre EE.UU., Reino Unido y Europa

6.2 Hallazgos Principales de Adopción

Uso generalizado:

- 95% usan IA en el trabajo y en sus vidas personales
- 76% pagan de sus propios bolsillos por servicios de IA

- 56% pagan más de $21/mes (sugiriendo suscripciones a planes de equipo/pro)
- 9% pagan más de $200/mes

Ganancias de productividad:

- **92% de encuestados reportan aumentos de ganancias de productividad** de servicios gen AI
- 47% sintió un aumento significativo de productividad
- 2% dijeron que su productividad bajó

Patrones interesantes:

- Para usuarios que describieron sin impacto o disminuciones, 60% estaban en planes gratuitos
- Por contraste, solo 15% de quienes reportaron ganancias de productividad estaban en planes gratuitos

6.3 Motivaciones y Casos de Uso

Principales motivaciones para usar IA:

1. Productividad y eficiencia
2. Codificación y desarrollo
3. Investigación y análisis
4. Aprendizaje y educación
5. Creatividad y generación de contenido

Reemplazo de servicios: La tendencia abrumadora entre encuestados que han reemplazado un servicio de internet existente con una herramienta de IA generativa es **la disrupción de motores de búsqueda tradicionales, principalmente Google**:

- ChatGPT (102 menciones)

- Perplexity (41 menciones)
- Claude (30 menciones)
- Gemini (29 menciones)

6.4 Momentos "Wow" del Último Año

Los momentos "wow" para usuarios se centraron en las capacidades rápidamente avanzadas de IA, particularmente en áreas tangibles y de alta habilidad:

Principales sorpresas:

1. **Codificación**: capacidad de construir aplicaciones completas y depurar problemas complejos
2. **Generación de medios**: mejoras dramáticas en video, imagen y audio
3. **Investigación profunda y análisis**: poder del razonamiento emergente
4. **Velocidad de mejora**: qué tan rápido los modelos se vuelven más capaces

6.5 Herramientas Adoptadas vs. Abandonadas Este Año

Tendencia más clara: Adopción de herramientas de codificación especializadas

- Claude Code y Cursor están siendo adoptados ampliamente
- Esto correlaciona con usuarios que dejan de usar GitHub Copilot
- En menor medida, ChatGPT para tareas de codificación

Herramientas más adoptadas:

1. Claude Code / Cursor
2. Claude (general)

3. Gemini
4. Deep Research
5. Perplexity

Herramientas más abandonadas:

1. ChatGPT (aunque también siendo adoptado por muchos)
2. GitHub Copilot
3. Midjourney
4. Perplexity

5. Gemini

Razones para cambiar:

- Mejor desempeño
- Características específicas como ventanas de contexto largas
- Consolidación de herramientas de propósito único en plataformas principales

6.6 Dónde se Ejecutan las Cargas de Trabajo de IA

A pesar del auge de neoclouds como CoreWeave, Nebius, Lambda y Crusoe, muy pocos usuarios reportan ejecutar sus cargas de trabajo de IA en ellos:

Preferencias de proveedor:

1. OpenAI directamente
2. Google Cloud
3. Anthropic
4. AWS
5. Azure

Implicación: Esto apoya la visión de que las neoclouds están ofreciendo capacidad a laboratorios e hiperscalers, no directamente a usuarios finales.

6.7 Soberanía de Datos: Usuarios Más o Menos Indiferentes

Importancia de la ubicación del datacenter:

- La mayoría de los usuarios consideran que es "algo importante" pero no crítico
- Pocos han cambiado de proveedor debido a preocupaciones de soberanía de datos

Quienes cambiaron lo hicieron por:

- Requisitos del cliente
- Regulaciones
- Cargas de trabajo gubernamentales/de defensa

Pregunta planteada: ¿Esto cuestiona la necesidad de IA soberana?

6.8 Presupuestos Organizacionales y Barreras

Crecimiento de presupuesto:

- 70% reportan que el presupuesto de gen AI de su organización ha crecido en el último año

Barreras principales para escalar IA:

1. Tiempo inicial requerido para hacer que los sistemas funcionen confiablemente
2. Preocupaciones de privacidad de datos
3. Falta de experiencia
4. Costos
5. Integraciones
6. Falta de ROI

6.9 Impacto Regulatorio (Limitado Hasta Ahora)

El panorama regulatorio de IA no ha impactado significativamente las estrategias de IA... hasta ahora

- En cierta medida esto es de esperarse dada la nacencia de las regulaciones de IA y sus implementaciones limitadas hasta la fecha
- Pero también es positivo ver que las organizaciones están presionando hacia adelante de todos modos

6.10 Herramientas y Servicios Más Populares

Herramientas de chat/generales más usadas regularmente:

1. ChatGPT (dominio abrumador)
2. Claude
3. Gemini/Google
4. Perplexity
5. Meta AI (bajo uso a pesar de distribución significativa)

Herramientas de codificación para desarrolladores:

1. Cursor
2. Claude Code
3. GitHub Copilot
4. VS Code con extensiones de IA
5. Codex de OpenAI y Gemini CLI van a la zaga

Otros servicios de IA populares (no codificación):

1. ChatGPT Deep Research
2. ElevenLabs (audio)
3. Perplexity
4. Claude
5. Midjourney

6.11 Adquisición y Entrenamiento de Modelos

Cómo las organizaciones adquieren IA:

1. **APIs** (de lejos el más común)
2. Fine-tuning de modelos existentes
3. Construcción desde cero
4. Modelos preentrenados de código abierto

Para fine-tuning, los encuestados usan comúnmente:

- PyTorch
- Hugging Face Transformers
- LoRA/PEFT
- Marcos internos personalizados
- Unsloth

Hardware usado para entrenamiento/fine-tuning:

1. GPUs de NVIDIA (abrumadoramente dominante)
2. Apple (sorprendente aparición, probablemente entrenamiento/experimentación local)
3. TPUs
4. AMD GPUs (no muy populares aún)

6.12 Tendencias Emergentes que Emocionan a los Profesionales

Principales áreas de emoción:

1. Agentes autónomos y multi-agente
2. Modelos de razonamiento y planificación mejorados
3. Multimodalidad (integración de texto, imagen, video, audio)
4. IA local/en dispositivo
5. Mejoras en contexto largo y memoria
6. Modelos de código abierto más poderosos
7. IA en robótica y sistemas físicos
8. Interfaces cerebro-computadora habilitadas por IA

6.13 Clasificación de Laboratorios de IA por Profesionales

Cómo 1.2k profesionales calificaron sus principales laboratorios de IA:

1. **#1: OpenAI**
2. **#2: Anthropic**

3. **#3: Google DeepMind**
4. **#4: Meta AI**
5. **#5: xAI**
6. **#6: DeepSeek**
7. **#7: Mistral**
8. **#8: Alibaba (Qwen)**

9. **#9: Cohere**
10. **#10: Moonshot AI**
11. **#11: Stability AI**
12. **#12: Inflection**

Conclusiones y Predicciones Clave

1) Razonamiento Escalable — El razonamiento se convierte en el nuevo benchmark. Modelos como GPT-5 y DeepSeek R1 lideran, aunque la varianza sugiere límites reales del progreso.

2) Cuello de Botella Energético — La escasez de energía redefine el ritmo de desarrollo. La energía, no el cómputo, será la restricción estratégica hacia 2028.

3) IA Soberana y Fragmentación — El gasto en IA soberana superará el billón de dólares en 2026, dividiendo el ecosistema global entre EE.UU., China y el Golfo.

Cierre

“Estrategia, Tecnología y Agentes Inteligentes para el futuro de los negocios.”
Alter Agentic · 2025

www.alter-agentic.com