LAS MENTIRAS DE LA IA

Fuentes de documentación por capítulo

Canal Gustavo Entrala · 2026

CAPÍTULO INTRO — El abogado de Nevada

Afirmación: *Un abogado presenta ante un juez federal sentencias inventadas por ChatGPT; el juez le impone una sanción creativa obligándole a dar charlas en facultades de derecho.*

Este caso ocurrió en Nevada en 2025 (Uprise case, juez Hardy). El caso seminal de 2023 (Mata v. Avianca) en Nueva York sentó el precedente con multas de $5.000 a los abogados.

Caso Nevada — juez con sanción creativa (LawNext, sep. 2025)
https://www.lawnext.com/2025/09/nevada-judge-takes-creative-and-unusual-approach-to-combat-ai-generated-fictitious-citations.html

Caso original Mata v. Avianca (LegalDive, may. 2023)
https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/

Sanción del juez Castel y detalles (ABA, jun. 2023)
https://www.americanbar.org/groups/litigation/resources/litigation-news/2023/use-chatgpt-research-bogus-cases-sanctions/

Base de datos global de alucinaciones en casos legales (Damien Charlotin)
https://www.damiencharlotin.com/hallucinations/ *+1.031 casos documentados*

Epidemia en el ámbito jurídico — 300+ casos (Jones Walker)
https://www.joneswalker.com/en/insights/blogs/ai-law-blog/from-enhancement-to-dependency-what-the-epidemic-of-ai-failures-in-law-means-for.html

---

CAPÍTULO 1 — Los números son impresionantes, si sabes leerlos

GPT-5 alucina un 26% menos que GPT-4o en contextos médicos y científicos

⚠️ *El anuncio oficial de OpenAI cita '~45% menos errores factuales con búsqueda web activa'. El 26% corresponde a una métrica específica sin búsqueda, confirmada por el artículo científico de PMC. Conviene matizarlo en el vídeo.*

Anuncio oficial OpenAI — Introducing GPT-5 https://openai.com/index/introducing-gpt-5/ *Fuente primaria*

GPT-5 System Card (PDF oficial) https://cdn.openai.com/gpt-5-system-card.pdf *Métricas detalladas de alucinaciones*

Artículo revisado por pares — PMC (reducción 26%) https://pmc.ncbi.nlm.nih.gov/articles/PMC12701941/

GPT-5.2 mejora otro 30% en tareas de resumen con documentos

Anuncio oficial OpenAI — Introducing GPT-5.2 https://openai.com/index/introducing-gpt-5-2/ *Fuente primaria — cita exacta del 30%*

Los laboratorios destacan tasas de 0,5–1%, pero solo en la tarea más favorable

⚠️ *Esos porcentajes corresponden al benchmark Vectara (summarización de documento propio). En preguntas abiertas las tasas son mucho más altas.*

Vectara Leaderboard — tasas por modelo 2025
https://www.aboutchromebooks.com/ai-hallucination-rates-across-different-models/ *Gemini 0,7% / GPT-4o 1,5%*

Análisis con contexto metodológico completo (Suprmind)
https://suprmind.ai/hub/ai-hallucination-rates-and-benchmarks/

CAPÍTULO 2 — ¿Qué es exactamente una alucinación?

Afirmación: *Hay médicos que han recibido recomendaciones de medicamentos que no existen; periodistas que han publicado citas falsas; firmas de consultoría que han devuelto el dinero por estadísticas fabricadas.*

Epidemia de fallos de IA en el derecho y más allá (Jones Walker, 2025)
https://www.joneswalker.com/en/insights/blogs/ai-law-blog/from-enhancement-to-dependency-what-the-epidemic-of-ai-failures-in-law-means-for.html

Abogados y alucinaciones — estado global (Cronkite News, oct. 2025)
https://cronkitenews.azpbs.org/2025/10/28/lawyers-ai-hallucinations-chatgpt/

Afirmación: *(FUENTE indicada en el guión) AIMultiple Research 2026 — gráfico de alucinaciones por año*

AIMultiple — AI Hallucination Statistics https://aimultiple.com/ai/hallucination *Verificar disponibilidad*

Alternativa con datos equivalentes (Suprmind, 2026)
https://suprmind.ai/hub/insights/ai-hallucination-statistics-research-report-2026/

CAPÍTULO 3 — La mecánica del engaño

Afirmación: *Un LLM es un sistema de compresión con pérdidas (Rate-Distortion Theorem, arXiv 2025).*

Rate-Distortion Theorem for Membership Testing (arXiv 2025) https://arxiv.org/abs/2503.10724 *Paper citado en el propio guión*

Afirmación: *Los modelos aprenden que dar una respuesta —aunque sea incorrecta— tiene mejor recompensa que no dar ninguna.*

OpenAI — Why Language Models Hallucinate https://openai.com/index/why-language-models-hallucinate/
Explicación técnica oficial

CAPÍTULO 4 — Cómo han mejorado (y por qué no es suficiente)

Gemini con ventana de contexto gigante baja al 0,7% — el dato más bajo de cualquier benchmark en 2025

Vectara Leaderboard — Gemini 2.0 Flash lidera con 0,7%
https://www.aboutchromebooks.com/ai-hallucination-rates-across-different-models/ *Fuente principal*

FACTS Grounding Benchmark (Google DeepMind)
https://deepmind.google/discover/blog/facts-grounding-a-new-benchmark-for-evaluating-the-factuality-of-large-language-models/ *Benchmark de Google*

FACTS Grounding — resumen técnico (Emergent Mind)
https://www.emergentmind.com/topics/facts-grounding-benchmark

El modelo o3 de OpenAI alucinó un 33% en preguntas sobre personas reales — el doble que su predecesor o1

TechCrunch — OpenAI's new reasoning AI models hallucinate more (abr. 2025)
https://techcrunch.com/2025/04/18/openais-new-reasoning-ai-models-hallucinate-more/ *Fuente primaria de medios*

System Card oficial o3/o4-mini (PDF OpenAI)
https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf *Documento técnico oficial*

CAPÍTULO 5 — ¿Quién nos engaña más y cuánto?

Benchmarks mencionados en el guión con sus enlaces de referencia:

SimpleQA — preguntas verificadas a Gemini en 2025 (HuggingFace)
https://huggingface.co/datasets/google/simpleqa-verified/viewer/simpleqa_verified/eval?row=6 *Ya indicado en el guión*

Vectara Hallucination Leaderboard (GitHub) https://github.com/vectara/hallucination-leaderboard

HalluLens — cómo se equivocan los modelos (GitHub Microsoft) https://github.com/microsoft/HalluLens

FACTS Grounding (Google DeepMind — blog oficial)
https://deepmind.google/discover/blog/facts-grounding-a-new-benchmark-for-evaluating-the-factuality-of-large-language-models/

FUENTE del ranking visual (indicada en el guión):
Voronoi / Visual Capitalist — AI Hallucination Rankings 2025
https://www.visualcapitalist.com/ranked-ai-hallucination-rates-by-model-2025/

TABLA 2 — FUENTE indicada en el guión:
AIMultiple Research https://aimultiple.com/ai/hallucination

Alternativa con datos comparables — AI Hallucination Report 2026
https://www.allaboutai.com/resources/ai-statistics/ai-hallucinations/

CAPÍTULO 6 — ¿Tiene solución? (La respuesta honesta)

"Colapso del modelo" — los modelos se entrenan con sus propios errores pasados

Knowledge Collapse in LLMs (arXiv 2509.04796) — NeurIPS 2025 https://arxiv.org/html/2509.04796v1
Paper exacto citado en el guión

Yann LeCun: la arquitectura actual hace imposible resolver las alucinaciones; propone los "modelos del mundo"

⚠️ *LeCun fundó AMI Labs (Advanced Machine Intelligence) en París en marzo de 2026, recaudando $1.030 millones — la mayor ronda semilla europea de la historia. Alex LeBrun (CEO) viene de la startup de salud digital Nabla. Ambos son tus interlocutores en París.*

Lanzamiento AMI Labs — $1.03B para los modelos del mundo (The Next Web)
https://thenextweb.com/news/yann-lecun-ami-labs-world-models-billion

LeCun advierte del 'callejón sin salida' de los LLMs (Creati.ai)
https://creati.ai/ai-news/2026-01-26/ai-pioneer-yann-lecun-warns-tech-dead-end-llms/ *Argumento técnico sobre alucinaciones*

Citas de LeBrun sobre alucinaciones en medicina (BetaNews)
https://betanews.com/article/yann-lecun-ami-labs-raises-1-03b-world-models-ai/ *Relevante para contexto entrevista*

ZME Science — perfil de AMI Labs y visión de LeCun
https://www.zmescience.com/science/news-science/father-of-ai-yann-lecun-raises-1-billion-with-startup-betting-that-the-industry-is-wrong/

Afirmación: *(FUENTE indicada en el guión) Perplexity AI Statistics 2025-2026, Incremys*
Perplexity AI Stats 2025-2026 (Incremys) https://incremys.com/ai/perplexity-ai-statistics/ *Verificar disponibilidad*

CAPÍTULO 7 — Kit del usuario escéptico

Afirmación: *Hábito 1 — Pídele que diga 'no lo sé'. (FUENTE indicada en el guión: Anthropic Docs)*
Anthropic Docs — Reduce Hallucinations
https://docs.anthropic.com/en/docs/test-and-evaluate/strengthen-guardrails/reduce-hallucinations *Ya indicado en el guión*

RESUMEN — Fuentes clave sin enlace previo en el guión

Tabla de referencia rápida para verificación:

Afirmación	Enlace recomendado
Caso abogado Nevada (juez creativo)	https://www.lawnext.com/2025/09/nevada-judge-takes-creative-and-unusual-approach-to-combat-ai-generated-fictitious-citations.html
GPT-5 System Card oficial (PDF)	https://cdn.openai.com/gpt-5-system-card.pdf
GPT-5.2 mejora 30% (anuncio oficial)	https://openai.com/index/introducing-gpt-5-2/
o3 hallucina 33% PersonQA (TechCrunch)	https://techcrunch.com/2025/04/18/openais-new-reasoning-ai-models-hallucinate-more/
o3 System Card oficial (PDF OpenAI)	https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf
Gemini 0,7% Vectara (About Chromebooks)	https://www.aboutchromebooks.com/ai-hallucination-rates-across-different-models/
Knowledge Collapse arXiv 2509.04796	https://arxiv.org/html/2509.04796v1
AMI Labs / Yann LeCun (The Next Web)	https://thenextweb.com/news/yann-lecun-ami-labs-world-models-billion
Anthropic — Reduce Hallucinations	https://docs.anthropic.com/en/docs/test-and-evaluate/strengthen-guardrails/reduce-hallucinations
Por qué alucinan los LLM (OpenAI)	https://openai.com/index/why-language-models-hallucinate/
Base de datos legal global (Charlotin)	https://www.damiencharlotin.com/hallucinations/