COMPAISS EVIDENCE BASE

Hallucination Rates in Contemporary AI Systems (2023-2025)

Verified sources only: peer-reviewed studies, official benchmarks, and institutional research reports | Prepared for institutional evaluation | All sources verified April 2026

50-83%

Adversarial hallucination rate across 6 major models in clinical vignette study (Mount Sinai / Nature Medicine, 2025)

17-34%

Hallucination rate in specialized legal RAG tools marketed as 'hallucination-free' (Stanford HAI / RegLab, 2024)

33-51%

Hallucination rate for OpenAI o3 reasoning models on open-domain factual benchmarks (SimpleQA / PersonQA, 2025)

The central finding across all evidence reviewed: Hallucination is not a temporary limitation that improving models will eliminate. It is a persistent, structural property of inference-first AI systems that changes shape as models improve. On tightly constrained, grounded tasks, rates decline. On complex reasoning, open-domain questioning, and adversarial inputs, rates remain high - and in some cases increase as models become more capable.

VERIFIED EVIDENCE BASE - STUDY BY STUDY

Omar et al. (2025) - Adversarial hallucination attacks in clinical decision support

CLINICAL / NATURE MEDICINE

MODELS TESTED

GPT-4o, Distilled-DeepSeek-Llama, Phi-4, Gemma-2-27B-it, Qwen-2.5-72B, Llama-3.3-70B (6 models, 5,400 outputs)

DEFINITION AND METHOD

300 physician-validated vignettes each containing one fabricated clinical detail (lab value, sign, or condition). Hallucination = model elaborates on fabricated detail as real.

REPORTED RATES

Overall default: 65.9%
DeepSeek-Llama: 80-83%
GPT-4o (best case): 50-53%
With mitigation prompt: 44.2%
GPT-4o + mitigation: 23%

Peer-reviewed in Communications Medicine (Nature Publishing Group), PMCID: PMC12318031 | pmc.ncbi.nlm.nih.gov/articles/PMC12318031 | Source verified April 2026

Magesh et al. (2024) - Hallucination in specialized legal AI RAG tools

LEGAL / STANFORD HAI

MODELS TESTED

Lexis+ AI (LexisNexis), Westlaw AI-Assisted Research (Thomson Reuters), Ask Practical Law AI (Thomson Reuters), GPT-4

DEFINITION AND METHOD

200+ manually constructed open-ended legal queries. Hallucination = incorrect legal statement or misgrounded citation that does not support the claim made.

REPORTED RATES

Westlaw AI-Assisted Research: >34%
Lexis+ AI: >17%
Ask Practical Law AI: >17%
GPT-4 general-purpose (prior study): 58-82%

Stanford RegLab and HAI | hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries | Source verified April 2026
Key finding: even purpose-built legal RAG tools with 'hallucination-free' marketing claims hallucinated on more than 1 in 6 queries. RAG reduces errors but does not eliminate them.

Graffius (2026) - Are AI Hallucinations Getting Better or Worse?

CROSS-MODEL / BENCHMARK SYNTHESIS

MODELS TESTED	DEFINITION AND METHOD	REPORTED RATES
OpenAI o3, o1; Gemini-2.0-Flash; multiple vendors. Benchmarks: Vectara HHEM, SimpleQA, PersonQA, domain-specific evaluations.	Synthesis of 2024-2025 benchmark data. Two divergent patterns identified: grounded summarization tasks (rates declining) vs. complex open-domain reasoning (rates increasing).	**Grounded summarization, top models: 0.7-1.5%** **OpenAI o3 on PersonQA / SimpleQA: 33-51%** **OpenAI o1 (prior generation): ~16%** **Broad task evaluations 2025: 3-20%+**

DOI: 10.13140/RG.2.2.33179.53285 | scottgraffius.com/blog/files/ai-hallucinations-2026.html | Source verified April 2026

Key finding: newer reasoning-focused models hallucinate more on complex factual tasks than earlier models. Stronger reasoning ability does not translate to stronger factual reliability.

Chelli et al. (2024) - Hallucination rates in AI-generated academic citations

CLINICAL / JMIR

MODELS TESTED	DEFINITION AND METHOD	REPORTED RATES
GPT-3.5 (ChatGPT), GPT-4, Bard/Gemini	Generated reference not matching any real paper metadata. Systematic review of AI-generated citations for medical literature.	**GPT-3.5: 39.6%** **GPT-4: 28.6%** **Bard/Gemini: 91.4%**

Journal of Medical Internet Research, 2024, e53164 | jmir.org/2024/1/e53164

Vectara HHEM Leaderboard (2025) - Grounded summarization benchmark

BENCHMARK

MODELS TESTED	DEFINITION AND METHOD	REPORTED RATES
Gemini-2.0-Flash, GPT-4o, GPT-o1, DeepSeek-R1, DeepSeek-V3, and others	Hughes Hallucination Evaluation Model (HHEM): faithfulness of model output to a provided source document. Tightly constrained grounded summarization - best-case conditions.	**Gemini-2.0-Flash: ~0.7%** **GPT-o1 / GPT-4o: ~0.9-1.9%** **DeepSeek-R1: 14.3%** **DeepSeek-V3: 3.5%**

Vectara Hallucination Leaderboard | github.com/vectara/hallucination-leaderboard

Note: these are best-case rates on a constrained summarization task - not representative of open-domain or complex reasoning performance.

WHAT THE EVIDENCE ESTABLISHES

The research record across 2023-2025 establishes three findings that are directly relevant to institutional AI governance.

First, hallucination rates are task-dependent, not model-dependent.

The same model family that achieves below 1% on a constrained document summarization task can produce errors on 33-51% of open-domain factual queries. Vectara's leaderboard shows top models approaching 0.7% on grounded summarization. OpenAI's own system card for o3 documents 33-51% error rates on SimpleQA and PersonQA. These are not contradictory findings - they reflect the same underlying architecture operating under different conditions. Institutions evaluating AI on vendor-supplied benchmarks are typically seeing best-case constrained performance, not performance on the kinds of complex, policy-specific questions their users will actually ask.

Second, RAG reduces hallucination but does not eliminate it.

Stanford HAI's independent evaluation of the leading legal AI platforms - tools marketed as 'hallucination-free' by LexisNexis and Thomson Reuters - found hallucination rates of 17-34% in real-world legal query conditions. These are purpose-built RAG systems with curated legal databases. The finding is not that RAG fails - it reduces errors substantially compared to general-purpose models - but that it introduces no architectural guarantee. Hallucination rates depend on retrieval quality, query complexity, and the presence of false premises, all of which vary in real institutional environments.

Third, more capable models can hallucinate more on certain tasks.

The Graffius (2026) synthesis documents a counterintuitive pattern: OpenAI's o3 series, optimized for complex reasoning, hallucinates at 33-51% on open-domain factual recall - more than double the rate of the earlier o1 series at approximately 16%. The Mount Sinai clinical study found that prompt mitigation reduced GPT-4o's adversarial hallucination rate from 53% to 23% - a significant improvement, but still a 23% failure rate under the best available mitigation strategy. Hallucination is not a problem of model quality that will be solved by the next generation. It is a structural property of inference-first architectures that scales with model capability in complex reasoning contexts.

GOVERNANCE IMPLICATION

For regulated institutions, the relevant question is not whether current AI systems hallucinate at acceptable rates on controlled benchmarks. Several do. The question is whether the conditions under which acceptable rates are achieved can be reliably maintained in institutional deployment - and what the consequences are when they are not.

The Stanford HAI legal study is the clearest illustration of this gap. Platforms that claimed to be 'hallucination-free' based on narrow citation accuracy metrics produced incorrect legal information on more than one in six real-world queries. A university, hospital, or government agency that relies on vendor benchmark claims to satisfy its governance obligations is relying on best-case performance figures that may not hold under actual conditions of institutional use.

An architecture that conditions the existence of inference on pre-execution authorization does not compete with these systems on hallucination rate. It removes the structural conditions under which hallucination occurs for unauthorized or out-of-scope queries - not by achieving a lower rate, but by preventing inference from running at all when authoritative sources do not exist to support it.

VERIFIED REFERENCES

1. Omar M. et al. (2025). Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Communications Medicine (Nature). PMCID: PMC12318031. pmc.ncbi.nlm.nih.gov/articles/PMC12318031
2. Magesh V. et al. (2024). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Stanford RegLab / HAI. hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries
3. Graffius S.M. (2026). Are AI Hallucinations Getting Better or Worse? We Analyzed the Data. DOI: 10.13140/RG.2.2.33179.53285. scottgraffius.com/blog/files/ai-hallucinations-2026.html
4. Chelli M. et al. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews. Journal of Medical Internet Research, e53164. jmir.org/2024/1/e53164
5. Vectara (2025). Hughes Hallucination Evaluation Model (HHEM) Leaderboard. github.com/vectara/hallucination-leaderboard
6. OpenAI (2025). o3 and o4-mini System Card. Documents 33-51% hallucination rates on PersonQA and SimpleQA benchmarks. Referenced and synthesised in Graffius (2026).