Hallucination

hallucination · safety · concept

Source: https://policywindow.org/wiki/hallucination

Generated 2026-05-30T22:09:07 UTC

Summary

Confidently-asserted but factually incorrect output produced by an AI model — including fabricated citations, invented people or events, and confabulated numerical values — that the model cannot reliably distinguish from correct output at generation time.

At a glance

Used by

4 instrument(s)

Related concepts

retrieval-augmented-generation, model-card, training-data-attribution, scalable-oversight

Primary source

Ji, Z., et al. (2023), 'Survey of Hallucination in Natural Language Generation,' ACM Computing Surveys 55(12): 1-38.

Details

Hallucination, in the foundation-model-output sense, was named by Ji et al. (2023, 'Survey of Hallucination in Natural Language Generation') and has become the canonical term for LLM factual error. The phenomenon decomposes into intrinsic hallucination (output contradicts available context) and extrinsic hallucination (output asserts facts that aren't grounded in context). NIST AI RMF GenAI Profile (NIST AI 600-1) names 'Confabulation' as a primary risk category, capturing the same phenomenon under a different label (NIST's choice signals a preference against anthropomorphic framing).

Governance relevance touches four surfaces. (a) Liability — when an AI-mediated legal brief contains hallucinated citations (Mata v. Avianca, 2023, S.D.N.Y.), who bears responsibility: the lawyer, the AI provider, or the AI deployer? EU AI Act Art. 13 transparency requirements + Art. 86 right-to-explanation are the closest binding frame. (b) Disclosure — should providers disclose hallucination rates as part of model-card disclosures (EU AIA Art. 53)? Industry practice is partial. (c) Redress — when hallucinated output causes harm (defamation via fabricated facts, financial loss via wrong numbers), redress mechanisms are unclear. EU AIA Art. 85 + OECD Principle 1.5 (accountability) frame the obligation; operationalisation is inconsistent. (d) Sectoral safety — hallucination in healthcare (medical-misinformation), criminal-justice (false-positive risk scores), and education (factual errors as authoritative output) drives most sectoral guidance. NIST AI 600-1 explicitly treats confabulation as a primary risk; UK AISI evaluations include factuality probes; Brazil PL 2338/2023 includes accuracy obligations.

Methodologically, hallucination cannot be eliminated by current architectures (Xu et al. 2024, 'Hallucination is Inevitable'). Mitigation is via retrieval-augmented generation, confidence calibration, and post-hoc verification — not architectural fixes.

How to cite this article

APA

Policy Window. (n.d.). Hallucination [Wiki article — Concept]. https://policywindow.org/wiki/hallucination

CHICAGO

Policy Window. n.d.. "Hallucination." Wiki article (Concept). https://policywindow.org/wiki/hallucination.

HARVARD

Policy Window (n.d.) 'Hallucination', Wiki article — Concept, available at: https://policywindow.org/wiki/hallucination.

OSCOLA

Policy Window, 'Hallucination' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/hallucination> accessed [date].

BIBTEX

```
@misc{policywindow-hallucination,
title  = {Hallucination},
author = {Policy Window},
year   = {n.d.},
howpublished = {hallucination — safety},
url    = {https://policywindow.org/wiki/hallucination},
note   = {Primary source: https://arxiv.org/abs/2202.03629}
}
```