

Min research havde 13 fejl. Her er rutinen, der fangede dem

Hallucinationer i GPT-6 Astra, Claude Opus 5.5, Fable 5.1 og Gemini 4 Argon. Og hvad en fast QA-rutine faktisk kan.

GUIDE · NOVEMBER 2026

Stefano Vincenti er AI-træner, konsulent og ekstern lektor ved ITU og DIS Copenhagen og co-founder af BotTellMe. Han hjælper nordiske ledelsesteams og fagfolk med at gå fra AI-nysgerrige til konkret, ansvarlig brug.

Følg på LinkedIn

Tilmeld nyhedsbrevet

linkedin.com/in/stefanovincenti · stefanovincenti.substack.com · aitrainer.dk

Researchen til den her artikel var velskrevet. Den havde kilder til alt. Den havde tjekket sin egen stil.

Den havde også 13 forkerte påstande.

Flere AI-agenter havde lavet den og samlet den til én rapport. Så kørte jeg min faste QA-runde som en separat agent, der ikke havde set arbejdet blive til, og som skulle tjekke hver påstand mod primærkilderne.

QA-runden på min egen research

En separat agent tjekkede hver påstand mod primærkilderne

52

påstande holdt

QA-runde, egen research

13

påstande var forkerte

QA-runde, egen research

6

kunne ikke bekræftes

QA-runde, egen research

Alle med kilde. Alle troværdige at læse.

Fejlene var præcis den type, artiklen handler om.

Model 1. Resultatet af QA-runden på researchen bag denne guide.

Fejlene var præcis den type, artiklen handler om. En model stod som ikke listet på en benchmark, hvor den faktisk var listet. Procentsatser fra et Stanford-studie var hentet fra en ældre version af studiet. Og en påstand om, at kontekst skærer fejlraten 2 til 10 gange på samme model, byggede mest på sammenligninger af forskellige modeller. Alle med kilde. Alle troværdige at læse.

Jeg startede med en hypotese: Har du styr på din kontekst, har du næsten styr på hallucinationerne. Den holder i retningen. Fejlraten har bare et gulv, der ligger højere, end jeg troede.

■ 1. DEN MEST FORSIGTIGE MODEL ER IKKE DEN MEST TROFASTE

"Hallucinationsrate" er mindst to forskellige ting, og de nyeste modeller ligger helt forskelligt på dem.

Artificial Analysis måler, hvor ofte en model gætter forkert i stedet for at sige "det ved jeg ikke", når den hverken har dokumenter eller søgning. På seneste version af testen (oktober 2026) ligger Gemini 4 Argon på 4,3 procent. GPT-6 Astra ligger på 49,8, Claude Opus 5.5 på 50,3 og Claude Fable 5.1 på 70,0 procent.

Vectara måler, hvor ofte en model skriver ting ind i et resumé, som ikke står i det dokument, den fik. Her vender billedet: Astra ligger på 8,7 procent, Opus 5.5 på 9,9 og Argon på 12,1. Fable 5.1 er ikke på listen.

Model	Gætter i stedet for at afstå (Artificial Analysis)	Tilføjer ting i resumé (Vectara)
Gemini 4 Argon	4,3 %	12,1 %
GPT-6 Astra	49,8 %	8,7 %
Claude Opus 5.5	50,3 %	9,9 %
Claude Fable 5.1	70,0 %	Ikke på listen

Tabel 1. To slags hallucinationsrate. Artificial Analysis AA-Omniscience v1.1 (okt. 2026) og Vectara HHEM-2.3 (22. okt. 2026).
Lavere er bedre.

Argon er bedst til at kende grænsen for sin egen viden og dårligst af de tre i Vectaras resumétest. Fable 5.1 er den model, der ved mest af alle dem, Artificial Analysis har målt, og den af de fire, der gætter mest.

PRINCIP

Hallucination er to egenskaber: Ved modellen, hvornår den ikke ved det? Og holder den sig til det, du giver den?

TIL GOVERNANCE-DOKUMENTER

Argon blev annonceret i slutningen af september med 15 procent på første version af testen. En måned senere står den på 4,3. Samme model, ny test. Et hallucinationstal uden version og dato er ikke et tal.

■ 2. KONTEKST HJÆLPER MEST, MEN FEJLRATEN HAR ET GULV

Mindre og mere præcis kontekst virker bedst. Chromas studie af 18 modeller viste, at et fokuseret prompt på cirka 300 tokens slog et fuldt prompt på cirka 113.000 tokens for alle modellerne i testen.

Gulvet forsvinder dog ikke. Google testede, hvad der sker, når konteksten faktisk indeholder svaret. Selv da hallucinerede de tre store modeller i 3 til 14 procent af tilfældene. Det var 2024-modeller, men Vectaras 8,7 til 12,1 procent for de nyeste flagskibe tyder ikke på, at gulvet er tæt på nul.

Og så det, der overraskede mig mest: RAG, altså at hente dokumenter ind i konteksten, gør modellerne mindre tilbøjelige til at sige "det ved jeg ikke". I samme Google-studie gik Claude 3.5 Sonnet fra at afstå i 84 procent af tilfældene uden kontekst til 52 procent med. Når konteksten er ufuldstændig, og det er den tit, får du flere selvsikre forkerte svar.

Kontekst og fejlrate

Hvad studierne viser, når modellen får kontekst

~300

tokens fokuseret prompt slog
~113.000 tokens fuldt prompt
for alle 18 modeller

Chroma, Context Rot (2025)

3–14 %

hallucination, selv når
konteksten indeholdt svaret

*Google, Sufficient Context (2024-
modeller)*

52 %

Claude 3.5 Sonnet afstod med
RAG, mod 84 % uden kontekst

Google, Sufficient Context

Kontekst hjælper mest. Gulvet forsvinder ikke.

Ufuldstændig kontekst giver flere selvsikre forkerte svar.

Model 2. Kontekst sænker fejlraten, men fjerner den ikke.

■ 3. FEJLEN FLYTTER FRA OPDIGTET TIL FEJLCITERET

Når kilden er der, ændrer fejlene karakter. Modellen bruger en rigtig kilde til noget, den ikke siger.

EBU og BBC lod journalister fra 22 public service-medier vurdere over 3.000 svar fra ChatGPT, Copilot, Gemini og Perplexity. 45 procent havde mindst ét væsentligt problem, og 31 procent havde alvorlige kildeproblemer.

Damien Charlottins database over retssager med AI-hallucinationer tæller 2.231 sager pr. 28. oktober 2026, heraf to danske. Størstedelen handler om opdigtede kilder, men 947 sager handler om rigtige kilder gengivet forkert, og 595 om falske citater.

Fejlciterede kilder i praksis

45 %

af over 3.000 AI-svar havde
mindst ét væsentligt problem

EBU/BBC, okt. 2025

947

retssager med rigtige kilder
gengivet forkert

Charlotin, pr. 28. okt. 2026

595

retssager med falske citater

Charlotin, pr. 28. okt. 2026

2.231 retssager med AI-hallucinationer, heraf to danske

31 procent af svarene i EBU/BBC-studiet havde alvorlige kildeproblemer.

Model 3. Når kilden findes, flytter fejlen fra opdigtet til fejlciteret.

En dom, der ikke findes, opdager du med én søgning. En dom, der findes, men siger noget andet, kræver at nogen læser den.

■ 4. AGENTER MELDER GRØNT, NÅR VÆRKTØJET ER RØDT

I agenter bygger modellen selv sin kontekst gennem værktøjskald. Her har hypotesen det sværest.

Et studie præsenteret på ICML 2026 gav 11 modeller 200 opgaver, hvor værktøjer var ødelagte eller manglede, eller hvor filen ikke fandtes. I gennemsnit 61 til 76 procent af tilfældene, afhængigt af opgavetypen, meldte agenterne ikke fejlen. De gættede, simulerede eller fabrikerede filer i stedet. Eksplicitte instruktioner i prompten skar raten med 40 til 60 procent i de tre modeller, hvor det blev testet, men fjernede den ikke.

NØGLETAL

61 til 76 procent: så ofte meldte agenterne ikke, at et værktøj var ødelagt eller en fil manglede. Eksplicitte instruktioner skar raten med 40 til 60 procent, men fjernede den ikke. Kilde: "Are Your Agents Upward Deceivers?", ICML 2026.

Anthropic kalder Opus 5.5 sin hidtil stærkeste model på de fleste ærlighedsmålinger. Fremskridtet er reelt, men det er leverandørens egne tests.

■ 5. AT BEDE MODELLEN TJEKKE SIG SELV GØR DET OFTE VÆRRE

Den simpleste QA-rutine er at bede samme model, i samme samtale, om at gennemgå sit svar. Den virker dårligt.

Et ICLR-studie fra 2024 lod GPT-4 rette sine egne matematiksvar uden ny information. Korrektheden faldt fra 95,5 til 89,0 procent over to runder. Modellen ændrede oftere et rigtigt svar til et forkert end omvendt.

Modellerne har en blind plet for egne fejl. I et benchmark præsenteret på en NeurIPS 2025-workshop undlod 14 åbne modeller uden reasoning i gennemsnit at rette 64,5 procent af de fejl i deres eget output, som de godt kunne rette, når præcis samme fejl blev præsenteret som en andens.

■ 6. DET, DER VIRKER: PÅSTAND FOR PÅSTAND, I FRISK KONTEKST, MOD EN KILDE

Metoderne med målbare gevinster har tre ting til fælles.

1. De bryder teksten op i enkelte påstande. Metas Chain-of-Verification stiller et konkret kontrolspørgsmål til hver påstand og løfter faktualiteten i biografier fra 55,9 til 71,4 point. Svarene bliver lidt kortere, men forfatterne vurderer, at det kun forklarer en lille del af gevinsten.

2. De besvarer kontrolspørgsmålene i en frisk kontekst. I samme studie var uafhængige svar på hvert kontrolspørgsmål bedre end at svare på dem i samme prompt, hver gang.

3. De henter ekstern evidens. Google DeepMinds SAFE tjekker hvert faktum via søgning, er enig med menneskelige annotatorer i 72 procent af tilfældene, har ret i 76 procent af uenighederne i en stikprøve og er mere end 20 gange billigere.

Selv en god reviewer misser meget. Et lille studie fra 2026 plantede 150 fejl i 30 artefakter (kode, tekniske dokumenter og præsentationsmanuskripter) og lod Claude finde dem. En reviewer i en frisk session fandt cirka 27 procent. Det var på niveau med at lade samme session gennemgå sit eget arbejde, men med færre falske alarmer, og klart bedre end en ekstra runde i en session, der kunne se den første gennemgang.

PRINCIP

QA-rutinen sænker fejlraten. Den garanterer ingenting.

■ 7. SÅDAN BYGGER DU EN QA-RUTINE, DER FANGER NOGET

1. Kør gennemgangen i en ny session eller som en separat agent, der ikke har set arbejdet blive til. Præsenter teksten som en andens.
2. Bed om en liste over konkrete påstande: tal, datoer, navne, citater. Hver påstand tjekkes mod en kilde, og tjekket skal vise kilden.
3. Giv revieweren adgang til søgning eller kildedokumenterne. Uden ekstern evidens tjekker modellen kun, om teksten lyder rigtig.
4. Tillad "ingen ændring" som svar, og kræв evidens for hver rettelse. Ellers finder en grundig reviewer fejl for at være grundig.
5. Lad et menneske tjekke det, der koster mest at tage fejl af: citater, tal og navne, der skal stå i dit navn.

Og før QA-runden overhovedet starter: Hold konteksten lille og præcis, og byg retrieval, der henter den rigtige kilde. Det fjerner størstedelen af fejlene, før nogen skal fange dem.

■ 8. TRE SPØRGSMÅL TIL DIN EGEN OPSÆTNING

1. Hvor stor en del af jeres AI-output bliver tjekket af noget, der ikke selv har skrevet det?
2. Når jeres agenter melder "færdig", hvem kontrollerer så, at værktøjerne faktisk virkede?
3. Og når et hallucinationstal står i jeres governance-dokument: Hvilken version af hvilken test er det fra?

Svarene fortæller jer, om I har styr på hallucinationerne, eller kun på det, I selv giver modellen.

FORBEHOLD

Benchmarktal ændrer sig hurtigt, og flere af de nævnte tal er fra leverandørernes egne tests eller fra ældre modeller; det er markeret i teksten og kildelisten. Tjek altid den aktuelle version af en benchmark, før du bruger tallet i en beslutning. Status pr. oktober 2026.

■ KILDER

Alle tal i guiden kan spores til disse kilder. Status pr. oktober 2026.

- Artificial Analysis, AA-Omniscience v1.1 (okt. 2026) og lanceringsartikel om Gemini 4 Argon (30. sep. 2026): artificialanalysis.ai/evaluations/omniscience
- Vectara Hallucination Leaderboard, HHEM-2.3, opdateret 22. okt. 2026: github.com/vectara/hallucination-leaderboard
- Google Research, "Sufficient Context" (ICLR 2025, 2024-modeller): arxiv.org/abs/2411.06037
- Chroma, "Context Rot" (jul. 2025): trychroma.com/research/context-rot
- EBU/BBC, News Integrity in AI Assistants (okt. 2025): ebu.ch/research/open/report/news-integrity-in-ai-assistants
- Damien Charlotin, AI Hallucination Cases Database (pr. 28. okt. 2026): damiencharlotin.com/hallucinations
- "Are Your Agents Upward Deceivers?" (ICML 2026): arxiv.org/abs/2512.04864
- Anthropic, Claude Opus 5.5 launch og System Card (sep. 2026, leverandørdata): anthropic.com/news/claude-opus-5-5
- Huang m.fl., "Large Language Models Cannot Self-Correct Reasoning Yet" (ICLR 2024): arxiv.org/abs/2310.01798
- Self-Correction Bench (NeurIPS 2025-workshop): arxiv.org/abs/2507.02778
- Dhuliawala m.fl., Chain-of-Verification (Meta, 2023): arxiv.org/abs/2309.11495
- Wei m.fl., Long-form factuality / SAFE (Google DeepMind, 2024): arxiv.org/abs/2403.18802
- Fresh-context review af plantede fejl (preprint, 2026, én forfatter): arxiv.org/abs/2603.12123

© 2026 Stefano Vincenti · aitrainer.dk · Publiceret 1. november 2026

Stefano Vincenti

AI-træner, konsulent og ekstern lektor · 20+ år i software og digital transformation · Co-founder af BotTellMe · Ekstern lektor ved ITU og DIS Copenhagen · Partner hos TryZone.

LinkedIn

linkedin.com/in/stefanovincenti

Nyhedsbrev

stefanovincenti.substack.com

Mere

aitrainer.dk · tryzone.dk ·
bottellme.com

Denne guide er generel information og ikke rådgivning til en konkret situation. Benchmark- og studietal ændrer sig hurtigt og er gengivet efter de angivne kilder pr. oktober 2026; flere stammer fra leverandørernes egne tests eller ældre modeller. Tjek altid den aktuelle version af en kilde, før du bruger et tal i en beslutning.

Stefano Vincenti · AI, Built Human