

# **“Fejl eller feature?” paneldebat**

**MHS #1: "Hvad skal balancen være mellem hallucination og kreativitet?"**

**TH #1: “Brug den rigtige teknologi (herunder rigtige type AI) til at løse de rigtige problemer”**

**MHS #2: “Kan vi leve med hallucinationer”**

**TH #2: “Du kan ikke bruge test til at sige at en sprogmodel er pålidelig”**

**MHS #3: “At AI tager fejl er ikke et tilstrækkelig argument for at man ikke kan bruge AI”**

**TH #3: “Sprogmodeller kan ikke ræsonnere logisk”**

**MHS #4: “Det kræver noget at bruge Generativ AI”**

**TH #4: “AI og Mennesker begår begge fejl, men på forskellig vis”**

**MHS #5: “Hvad skal vi være bange for ved Generativ AI”**

**TH #5: “Hvem skal tage risikoen og ansvaret for de fejl teknologien begår og hvordan sikrer vi, at det sker på en ansvarlig måde?”**

**MHS #1:** "Hvad skal  
balancen være mellem  
hallucination og kreativitet?"

# Hvad er en Hallucination?

- Et udsagn der er **falsk** 😬
- Et udsagn der involverer mennesker, begivenheder, osv., **der ikke findes** i den virkelige verden 😬
- Matematisk formalisering 😬
- Et udsagn der **ikke følger** af træningsmaterialet 🥰
- Vi må ikke kalde det hallucinationer [1] 😂

# Hallucinationer kan ikke undgås?

- A la **uendelige løkker er uundgåelige og uafgørlige i universelle sprog (Turing)**
- Bevis antager at LLM er deterministisk plus har eksternt while loop [2,3] 😂
- A la **de reelle tal er ikke tællelige (Cantor)**
- Ground truth functions er overtællelige [4] 😂
- A la **Lags (Post, Wang) [6]** tæt på ved at udnytte **autoregressive decoding**.
- Men lidt snyd at udnytte bounded input til at simulere skipping tokens 😂

# Årsager til Hallucination

- Næste token baseret på **sandsynlighed**, fejl akkumulerer over sekvens af tokens ✓
- **Fejlagtig træningsdata** ✓
- **Begrænset træningsdata**, men giver altid et svar i stedet for “ved ikke” ✓
- **Approksimering** af hensyn til beregningsmæssig kompleksitet ✓
- RLHF fejler pga **ukorrekt human feedback**. ✓
- Med temperature > 0 (ikke-deterministisk) **samples mindre sandsynlige svar** ✓

# Kreativitet vs. hallucination

- **Approksimering + temperature** er væsentlige aspekter
- LLM har utrolig evne til at abstrahere irrelevante detaljer
- Man føler LLM har **eget bidrag** til viden
- Måske er **temperature > 0 nødvendig** (kan styres i API)
- Måske kan vi **leve med hallucinationer** ligesom halting problem og uendelige løkker eller de ebber ud
- Der er en hel del **vendor og bruger kan gøre** (og har gjort!) som modvirker hallucinationer (→ **se senere**)

**TH #1:** “Brug den rigtige teknologi (herunder rigtige type AI) til at løse de rigtige problemer”

# Alternativer

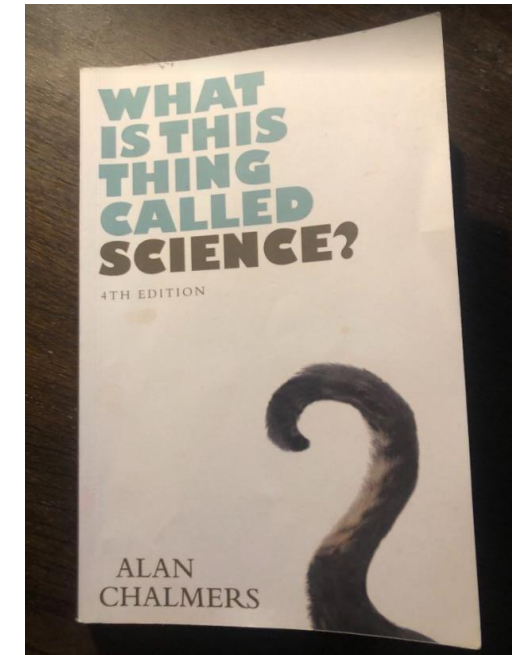
- Brug den **forkerte** teknologi til at løse de **forkerte** problemer
- Brug den **forkerte** teknologi til at løse de rigtige problemer
- Brug den rigtige teknologi til at løse de **forkerte** problemer

# Mening

- Brug algoritme, hvis den findes.
- Brug skak program til skak, det ser dybere end mennesker\*
- Brug X, hvis Y
- Flere eksempler!

\* Interessant fordi det viser i nogle tilfælde at:

- AI ikke bare kan **erstatte** menneske, men
- AI kan **gøre det bedre** end menneske



**MHS #2: “Kan vi leve med  
hallucinationer”**

# Fang Hallucinationerne

- Kode generering (Eksempel: POC)
- Fejlsøgning i kode (Eksempel: PyGame)
- Draft (Eksempel: Ansøgning)
- Review (Eksempel: Ansøgning)
- Search (indexing)
- Litteratursøgning (Eksempel: dage->1 min, ram 95%)
- Anden fejlsøgning (Eksempel: bilbatteri)
- 4 øjne → 2 øjne
- Kort sagt: Human in the loop

# Lev med Hallucinationerne

- Forvent der er fejl
- Du kan ikke bare stikke et svar ind som Appendix A.
- Lære spansk
- Læse artikler sammen

# Eliminer Hallucinationerne

- Bedre modeller
  - Statistiske modeller -> Neurale, måske ebber det ud
  - API-kald (Eksempel: Wolfram primtest)
  - Beregninger (Eksempel: Python call out)
  - Flere parametre, lag, kontekst
  - Track og eksponer score
  - Deterministiske svar(temperature close to 0)
- Mere træning
  - Mere data
  - Bedre data
  - Specifikke modeller\*
- Bedre prompting
  - COT 😊 [5]
  - Søg multiple svar & score (Ek
- RAG + systeminstruks
- Browsing
- RLHF
- Angiv sources

\* Nogle eksempler har vist at generelle modeller koder bedre end specifikke

**TH #2:** “Du kan ikke bruge test  
til at sige at en sprogmodel er  
pålidelig”



# Forudsigelighed af svar

- Hvis man spørger om det samme to gange kan man få forskelligt svar
  - Temperature  $> 0$
  - Mere data
  - Ny model
- Fair scenarie
  - Man har trænet med alle de relevante scenarier
  - Men benytter deterministisk model
  - Man stiller kun modellen spørgsmål vedrørende scenarierne
  - Er man så dårligere stillet end ved klassisk test af et program?
  - I begge tilfælde kan man ramme nyt grænsetilfælde
- Unfair scenarie
  - Vilkaarlige spørgsmål til generel model
  - Vi snakker om et program der giver feedback på alle problemer i verden
  - Det er ikke nødvendigvis et problem med modellen, men med scope
  - "Et klassisk testet program kan også give fejl"

**MHS #3:** “At AI tager fejl er ikke et tilstrækkeligt argument for at man ikke kan bruge AI”

## Før og Nu

- Der var penge i at lave AI
- Nu er der penge i at kalde AI for snake oil.

## Min kæphest

- "AI fejler så det kan ikke bruges".
- "Mennesker fejler også men kan bruges, og derfor kan du ikke ræsonnere på den måde"

*Saying that we can live with hallucinations because humans fail too is Erasmus Montanus logic.*

## En sten kan ikke flyve, Morlille kan ikke flyve, ergo er Morlille en Sten

$\sim \text{KanFlyve}(\text{MorLille})$

$\text{Ax. ErSten}(x) \Rightarrow \sim \text{KanFlyve}(x)$

$\text{ErSten}(\text{MorLille}) \Rightarrow \sim \text{KanFlyve}(\text{MorLille})$

$\text{ErSten}(\text{MorLille})$  ←

$\text{Ay.}(\text{Ax. S}(x) \Rightarrow \text{I}(x)) \ \& \ \text{I}(y) \Rightarrow \text{S}(y)$  ←

S = sten

I = ikke flyve

Hvis alle medlemmer af en population S har en egenskab I, og noget har egenskaben I, så er det med i populationen S

Validated by ChatGPT  
Ukorrekt i 1. ordens prædikatkalkyle  
 $A \Rightarrow B$  forvirret med  $B \Rightarrow A$

## Konklusion

- Kvanticer benefit, effort og risk (probability times impact) i forhold til human scenario

*GenAI and humans fail different and for different reasons – there is no scientific grounds for comparing it or arguing like this. Comparing the hit rate between two systems (a human vs GenAI) makes little sense when they work so differently.*

Nogle gange ræsonnerer ChatGPT bedre end mennesker

## At AI tager fejl er ikke et tilstrækkelig argument for at man ikke kan bruge AI

Ex.  $\text{ErMenneske}(x) \ \& \ \text{KanFejle}(x) \ \& \ \text{KanBruges}(x)$

$\text{ErMenneske}(M) \ \& \ \text{KanFejle}(M) \ \& \ \text{KanBruges}(M)$

$\text{KanFejle}(M) \ \& \ \text{KanBruges}(M)$

Ex.  $\text{KanFejle}(x) \ \& \ \text{KanBruges}(x)$

$\sim \sim \text{Ex. KanFejle}(x) \ \& \ \text{KanBruges}(x)$

$\sim \text{Ax.} \sim (\text{KanFejle}(x) \ \& \ \text{KanBruges}(x))$

$\sim \text{Ax.} \sim \text{KanFejle}(x) \vee \sim \text{KanBruges}(x)$

$\sim \text{Ax. KanFejle}(x) \Rightarrow \sim \text{KanBruges}(x)$  ←

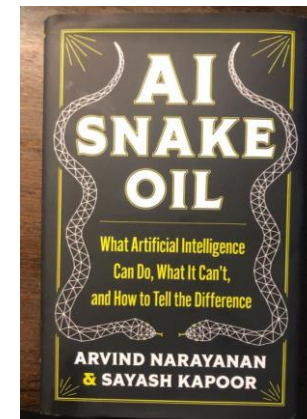
$(\text{Ex. M}(x) \ \& \ \text{F}(x) \ \& \ \text{B}(x)) \Rightarrow \sim \text{Ax. F}(x) \Rightarrow \sim \text{B}(x)$  ←

M = menneske

F = KanFejle

B = KanBruges

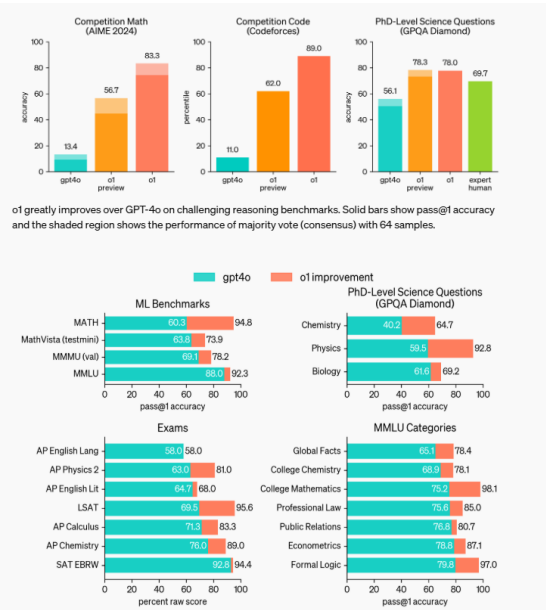
Validated by ChatGPT  
Korrekt i 1. ordens prædikatkalkyle



**TH #3:** “Sprogmodeller kan  
ikke ræsonnere logisk”

# Eksempler hvor de bliver bedre

- Checke primtal
  - Prompting: Heine regel
  - Delegering til APIs (Wolfram)
  - Kodeeksekvering (Python)
  - De blev hurtigt meget bedre
- River and boat
  - Prompting: fokus på score
- 1 kg fjer og 1 kg jern
- Patching af eksempler vs generel ændring
  - Library af algoritmer med deskriptiv tekst
  - Prompting, COT, inference scaling, ...
- Sprogmodeller kan heller ikke kode!



## River and boat (4o)

I have a boat that can hold one human and one animal. Create an optimized plan for getting me and my dog, Luna, across the river.

If you have multiple answers let me know what they are.

If the probability score is low for your answer, let me know.

Here is an optimized plan for getting you and your dog, Luna, across the river, along with alternate approaches and probability assessments.

### Optimized Plan: Single Crossing

1. You and Luna cross the river together in the boat.

• Result: Both you and Luna are on the far side.

• Probability of Success: High (100%). The plan requires only one trip, and the boat's capacity matches the requirements.

This plan is the most efficient and simple. There is no need for additional steps since there are no conflicts or constraints.

### Alternative Plan: Multiple Trips (Unnecessary but Valid)

If for some reason, a single trip is not possible or you wish to experiment:

1. You cross the river alone, leaving Luna behind.

2. You return to the starting side with the boat.

3. You and Luna cross the river together.

• Result: Both you and Luna are on the far side.

• Probability of Success: Medium to High (~95%), as unnecessary additional trips slightly increase the chance of delays, confusion, or errors.

This plan adds extra steps but still accomplishes the goal.

## Mor Lille-logik

### Analyzing for a Counterexample

To determine validity, we attempt to disprove the formula by finding an interpretation where it is false. A formula of the form  $P \implies Q$  is false only if  $P$  is true and  $Q$  is false.

In this case:

- $P = \forall x (S(x) \implies I(x)) \wedge I(y)$
- $Q = S(y)$

Thus, the formula is false if:

1.  $\forall x (S(x) \implies I(x))$  is true,
2.  $I(y)$  is true,
3.  $S(y)$  is false.

### Constructing a Counterexample

- Let the domain  $D$  be a set with at least one element.
  - Define predicates  $S$  and  $I$  as follows:
    - $S(x)$ : True for some  $x$  but not for all  $x$ .
    - $I(x)$ : True for all  $x$  where  $S(x)$  is true.
- Specifically:
- $S(x) \implies I(x)$ : This can hold since  $I(x)$  is true whenever  $S(x)$  is true.
  - Choose some  $y$  such that  $I(y)$  is true but  $S(y)$  is false.

Here:

- $\forall x (S(x) \implies I(x))$  holds true.
- $I(y)$  holds true for some  $y$ .
- However,  $S(y)$  is false.






In this scenario:

- The premises  $\forall x (S(x) \implies I(x)) \wedge I(y)$  are true.
- The conclusion  $S(y)$  is false.

This constitutes a counterexample, proving that the formula is not valid in first-order logic.

**MHS #4: “Det kræver noget  
at bruge Generativ AI”**

## Scenarier

- a) patient -> GenAI som læge  læge vil være human in the loop
- b) læge -> GenAI som læge  læge er allerede human in the loop
- 
- a) Junior dev -> Gen AI som Senior dev 
- b) Senior dev -> Gen AI som Senior dev 
- b) Senior dev, but not in this tech -> Gen AI som Senior in this tech 

Prompt engineering

Validering af svar

**TH #4:** “AI og Mennesker begår  
begge fejl, men på forskellig vis”



## Udvikler fejl

- Manglende library til kode (React eksempel)
- Cross Origin Resource Sharing issue ved API-kald
- Manglende rettigheder (PLSQL eksempel)
- Genintroducere fejl der tidligere er løst.
- håndsoprækning hvem har set det i et projekt

## Conjecture

- For enhver fejl AI har begået,  
er der et menneske der har begået samme fejl,  
hvor andre troede at der ikke var nogen fejl.

## Eksempler

- **Falske referencer** (Eksempel: litteratursøgning/gymnasiestil)
- **Forkert sagsbehandling** (Eksempel: sanktionering af ledige/førtidspension)
- Osv

Undgå fraud kunne være en væsentlig benefit.

**MHS #5:** “Hvad skal vi være  
bange for ved Generativ AI”

## Risks

- Agents
  - Claude der styrer laptop
- Træning med AI genereret indhold
  - Scheeming
- Rettigheder (Eksempel: StackOverflow)
- Deep fake
- Kriminelle med modeller
  - FraudGPT, WormGPT

## Detaljer

- Om vi kalder hallucinationer for hallucinationer
- Om vi benytter antropomorf sprogbrug
- Om vi siger han/hun eller den/det

## Trends

- Agents
- Scaling
- Inference scaling
- Reasoning model o1
- O1 replication

**TH #5:** “Hvem skal tage risikoen og ansvaret for de fejl teknologien begår og hvordan sikrer vi, at det sker på en ansvarlig måde?”

## **Generelt**

- Er det ikke det samme som med andre teknologier

## **Kodning**

- Den der koder med LLM er selv ansvarlig for koden

[1] **ChatGPT is bullshit**, Ethics and Information Technology, <https://doi.org/10.1007/s10676-024-09775-5>

[2] **Memory Augmented Large Language Models are Computationally Universal**,  
<https://arxiv.org/abs/2301.04589>

[3] **LLMs Will Always Hallucinate, and We Need to Live with this**, presented at NLP Summit 2024,  
<https://arxiv.org/abs/2409.05746>

[4] **Hallucination is Inevitable: An Innate Limitation of Large Language Models**,  
<https://arxiv.org/pdf/2401.11817>

[5] **The Unreasonable Effectiveness of Eccentric Automatic Prompts**, <https://arxiv.org/abs/2402.10949>

[6] **Autoregressive Large Language Models are Computationally Universal**,