AI i analysearbejdet 2.0 – hype til handling

2. oktober 2025

AnalytiCS

Epinion

Styregruppe

I styregruppen sidder repræsentanter fra Red Barnet, Center for Frivilligt Socialt Arbejde, Rådet for Sikker Trafik, Danske Patienter og Alkohol og Samfund

Epinion
AnalytiCS

Præsentationsrunde

Navn, organisation og rolle i organisation

Epinion
AnalytiCS

Dagens tema:

AI i analysearbejdet 2.0 – hype til handling

Epinion
AnalytiCS

Emner

Vi har tidligere haft netværksmøde om AI i analyser – og i dag tager vi skridtet videre under titlen ”AI - Hype og handling”

- Coronas betydning for frivilligheden i Danmark (digitale analyser)
- Børn og unge i analyser – hvordan gør man det bedst?
- Sådan omsætter du data til succesfuld interessevaretagelse
- Hvordan måler vi værdien af civilsamfundets indsatser?
- Skab impact gennem dine leverancer
- Hvordan sikres inklusion og repræsentation i analyser?
- Måling af sociale investeringer og indsatser i civilsamfundet
- Systemisk forandring og brug af data i systemforandrende processer
- **Workshop: Kom godt i gang med AI i analyser**
- Hvad vinder man, når brugerne bliver medskabere i analyser?
- Skab en god evalueringskultur i din organisation
- **AI i analysearbejdet 2.0 – hype til handling**

Epinion AnalytiCS

Dagens program

13:00-13:15	**Velkommen og præsentationsrunde**
13:15-14:15	**AI i 2025:** Hvor er vi nu, hvad er det nyeste – og hvordan oversætter vi hype til handling i praksis? v. Allan TH Knudsen
14:15-14:45	*Pause og netværk*
14:45-15:00	**Fra teknologi til principper:** Kort oplæg om tech-adoption og governance-principper: Hvordan kan organisationer sætte rammer for ansvarligt brug? v. Allan TH Knudsen
15:00-15:30	**Case: Red Barnet sætter AI på dagsordenen:** Erfaringer med at udvikle en AI-politik samt de muligheder, begrænsninger og dilemmaer politikken indebærer
15:30-16:00	**Fælles drøftelse**
16:00-?	*Fyraftensøl/-sodavand eller på gensyn*

AI i 2025

v. Allan TH Knudsen
Senior Director, Epinion

Epinion
AnalytiCS

AI 2.0
Hype og handling

Epinion

1. oktober 2025

Allan Toft Hedegaard Knudsen

Senior Director, Hub Lead for Data Science & Analyse

Epinion

aka ”Hvad faen er der sket siden sidst?”

Epinion

1. oktober 2025

Allan Toft Hedegaard Knudsen

Senior Director, Hub Lead for Data Science & Analyse

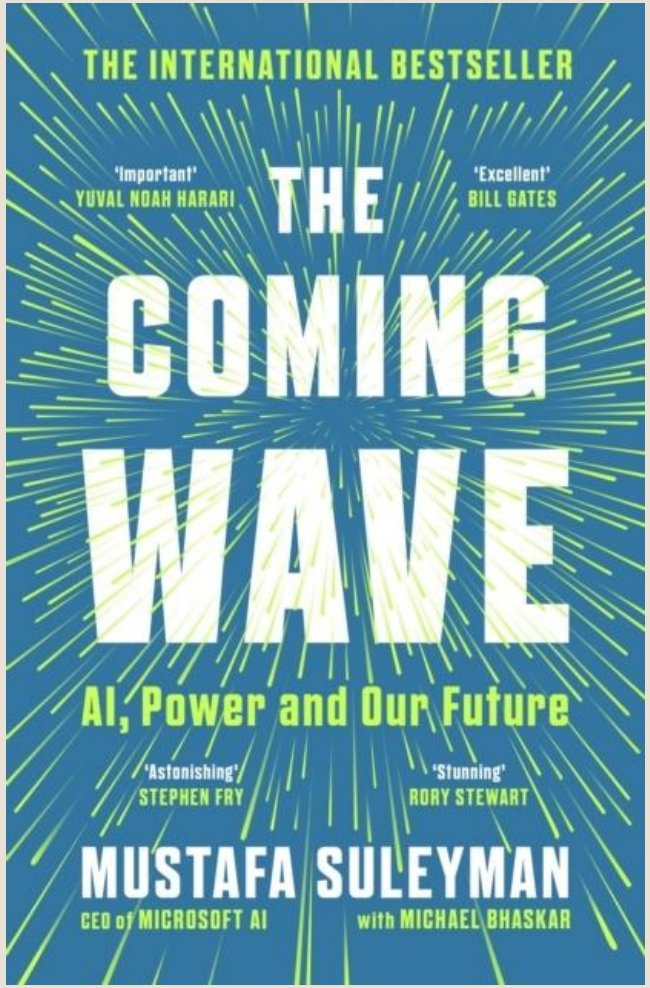
Epinion

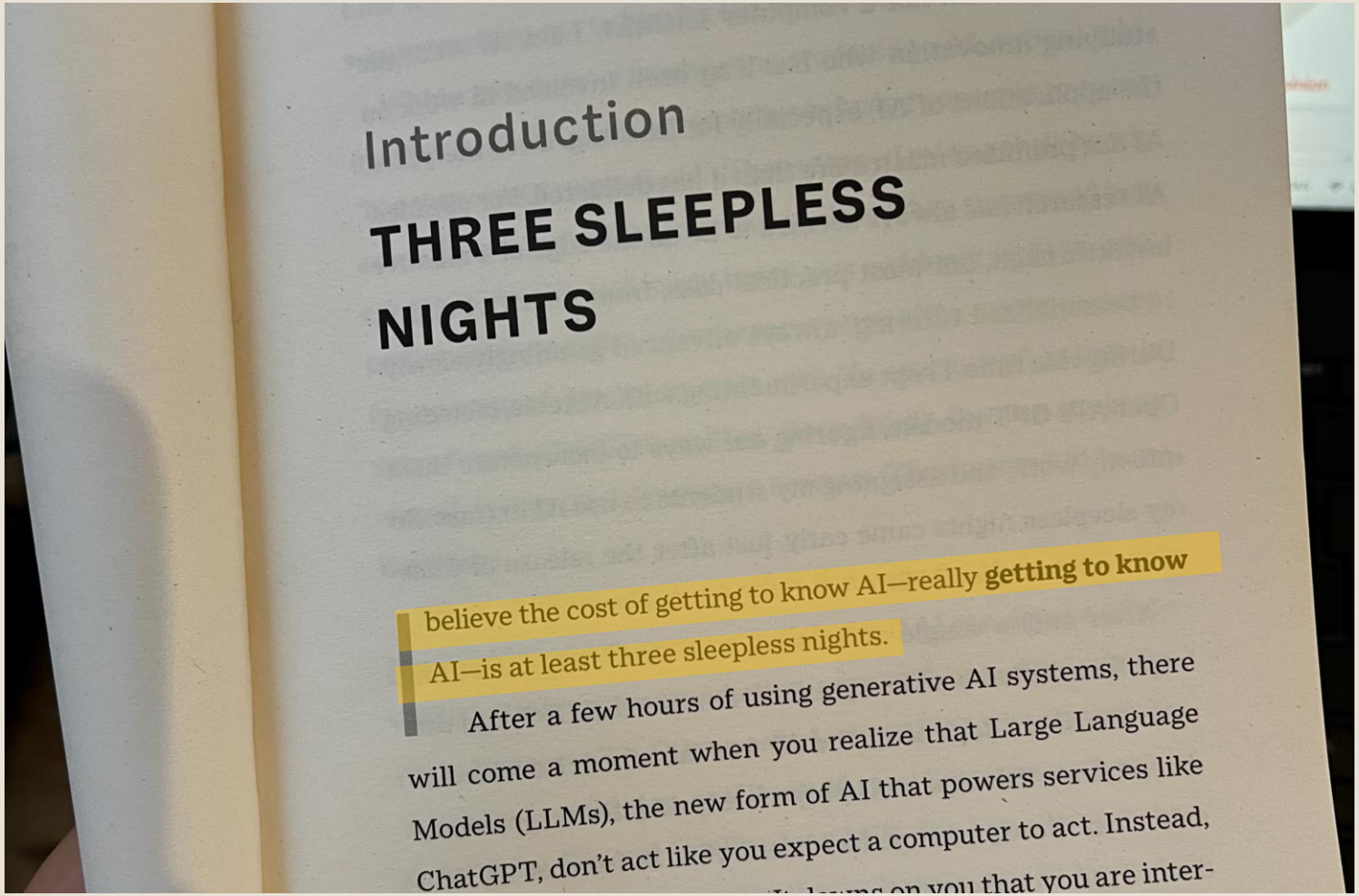
AI i 2025: Buzz-words, bull shit og big business

Epinion

Hvad er der sket siden sidst?

Hvorfor sætte sig ind i det? – og hvad ”koster” det?




Introduction

THREE SLEEPLESS NIGHTS

believe the cost of getting to know AI—really **getting to know** AI—is at least three sleepless nights.

After a few hours of using generative AI systems, there will come a moment when you realize that Large Language Models (LLMs), the new form of AI that powers services like ChatGPT, don't act like you expect a computer to act. Instead,

“Co-intelligence – Living and working with AI” (Mollick, 2024)

Epinion

Hvis du for alvor skal forstå det – så skal du have jord under neglene….

AI-genereret billede

Epinion

AI i 2025: Buzz-words, bull shit og big business

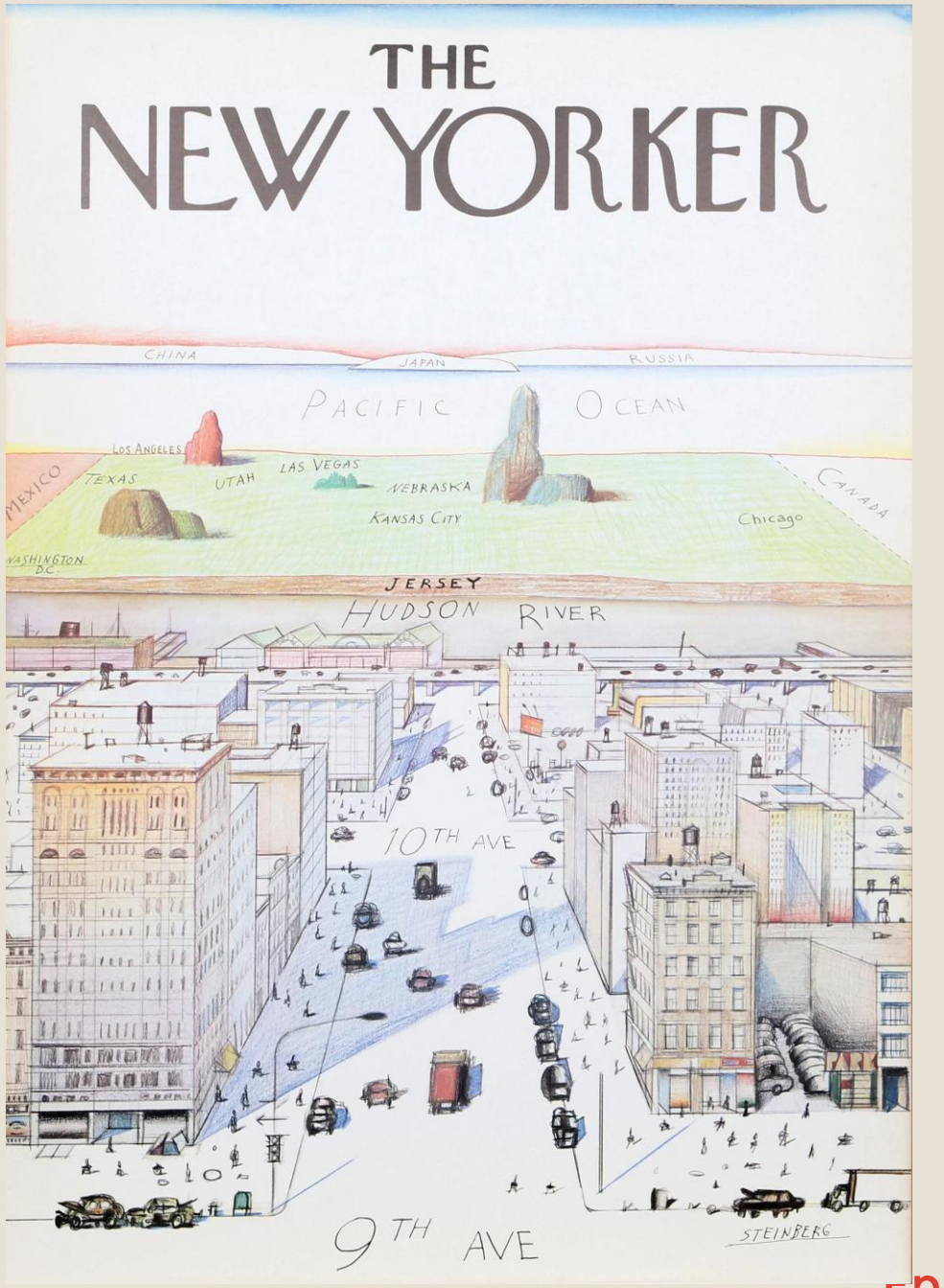
Hvad er der sket siden sidst?

Transformer - teknologi: 2017

A view from 9th avenue

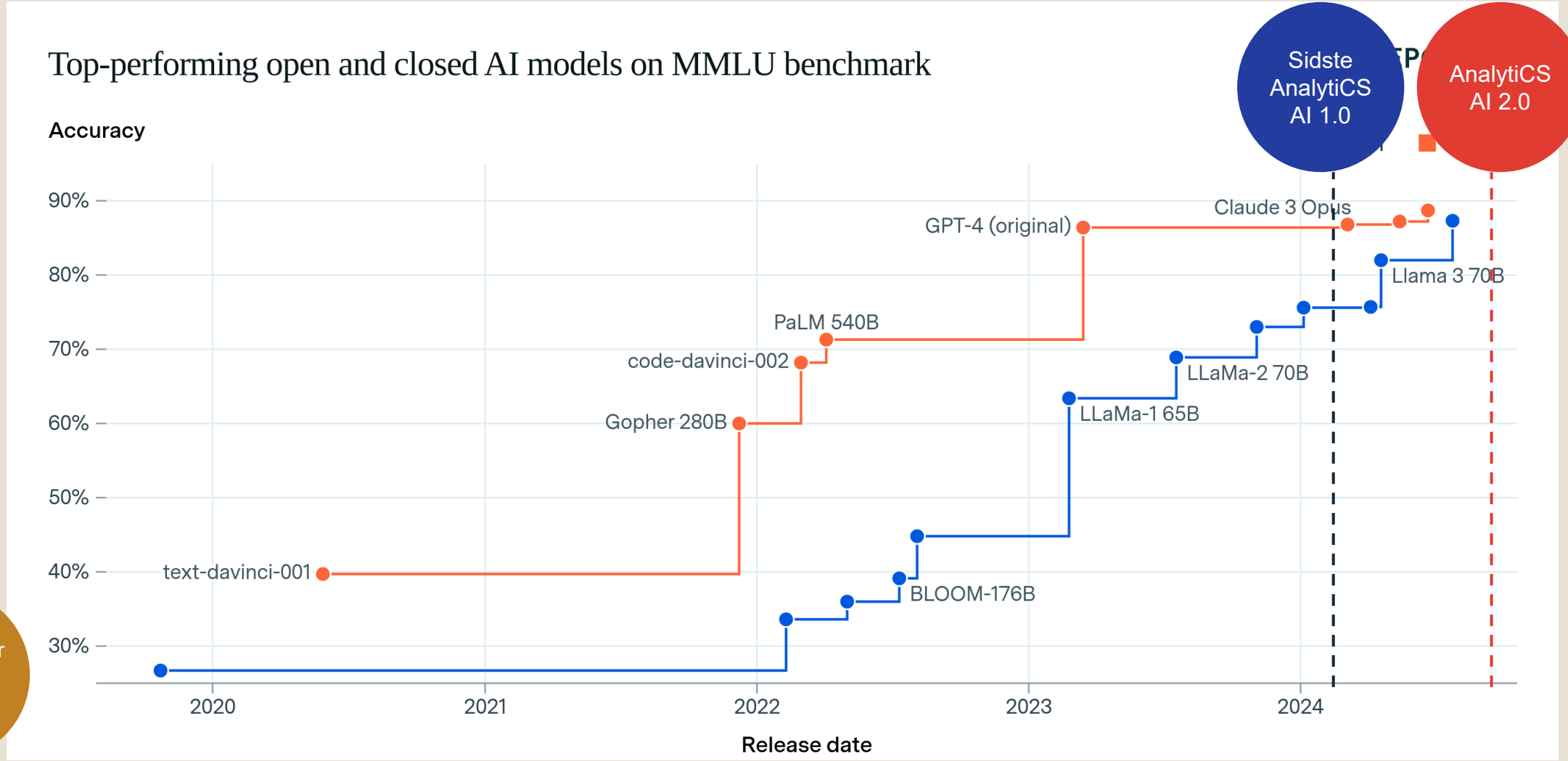
Sidste AnalytiCS AI 1.0

AnalytiCS AI 2.0



Epinion

Hvad er der sket siden sidst?

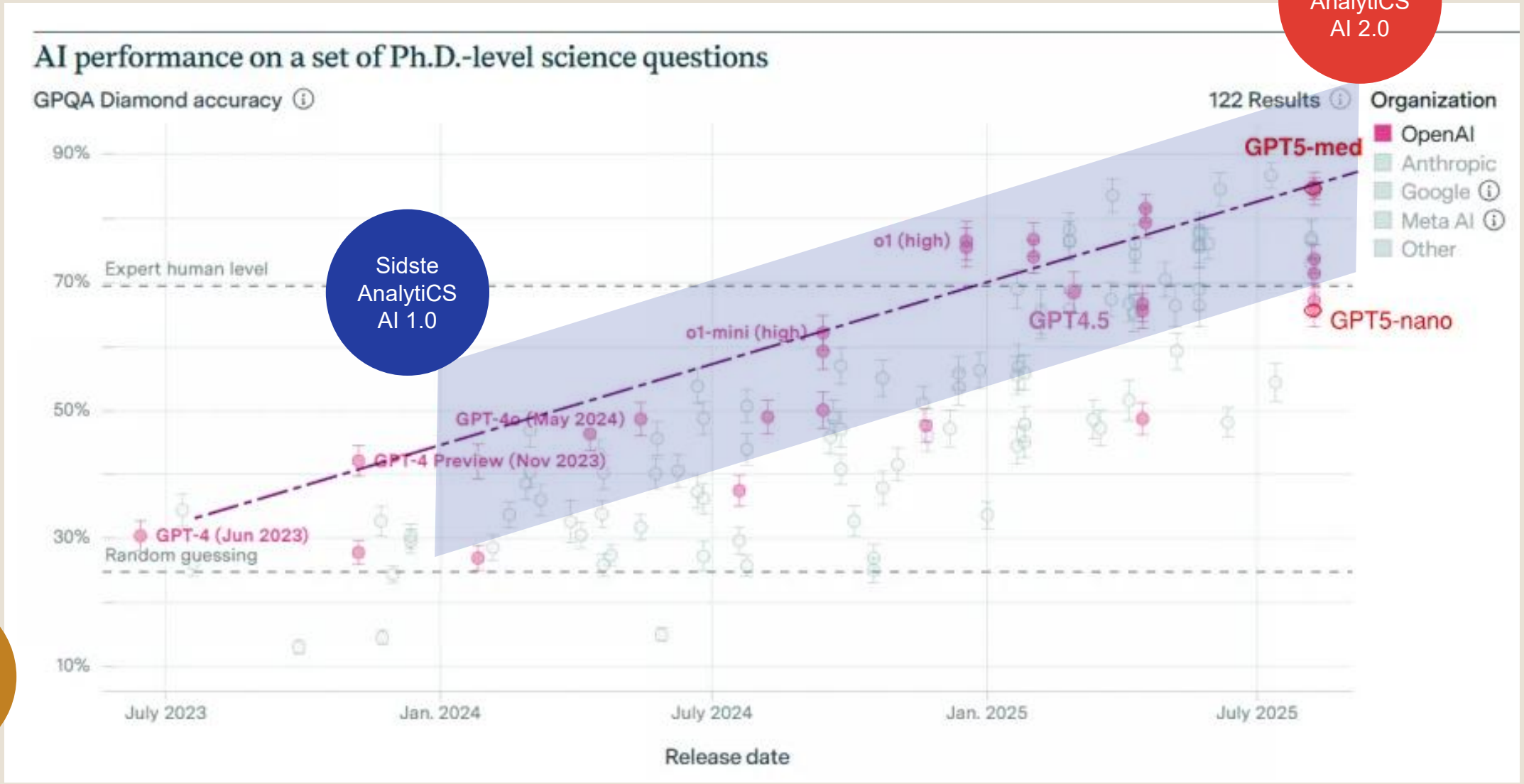


Kilde: Cottier et. AI (2024) fra EpochAI

Epinion

Hvad er der sket siden sidst?

Sidste AnalytiCS AI 2.0



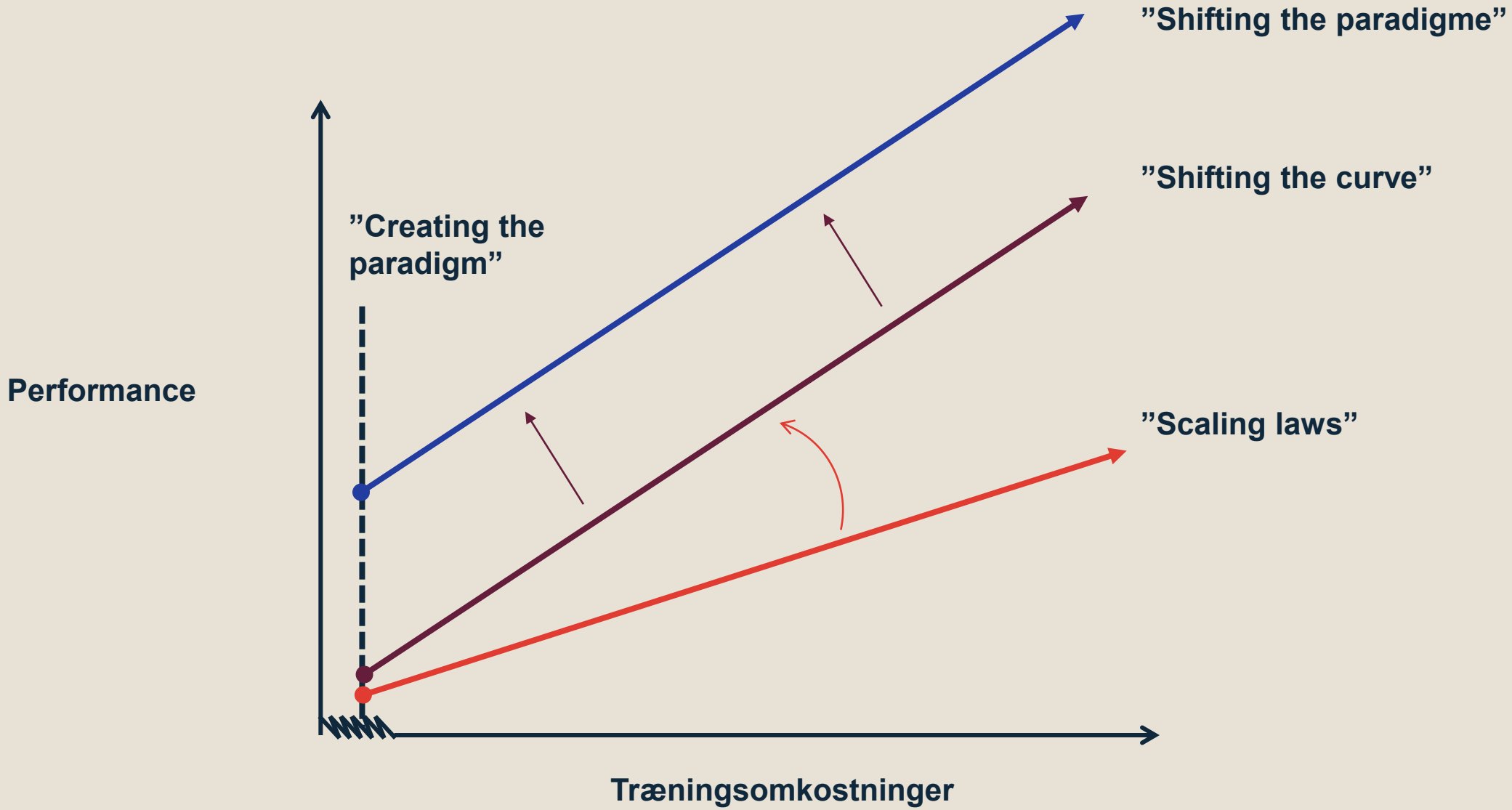
Sidste AnalytiCS AI 1.0

Transformer - teknologi: 2017

Kilde: EpochAI fra OpenAI, GPT5-release, august 2025

Epinion

Hvad er der sket siden sidst? Begreber i udviklingen af AI…



Kilde: Armodei, Anthropic (2025)

Epinion

Hvad er der sket siden sidst? Begreber i udviklingen af AI…

Pointen

Ja, der er **skaleringslove**…og der er mange penge i AI.

Ja, **kurven er måske skiftet** med hardware og nye træningsmetoder… og der er stadig mange penge.

Er der også sket et **paradigmeskift**? Måske. Men der er i hvert fald ikke skabt et nyt ”paradigme” break-through siden sidst.

Kilde: Armodei, Anthropic (2025)

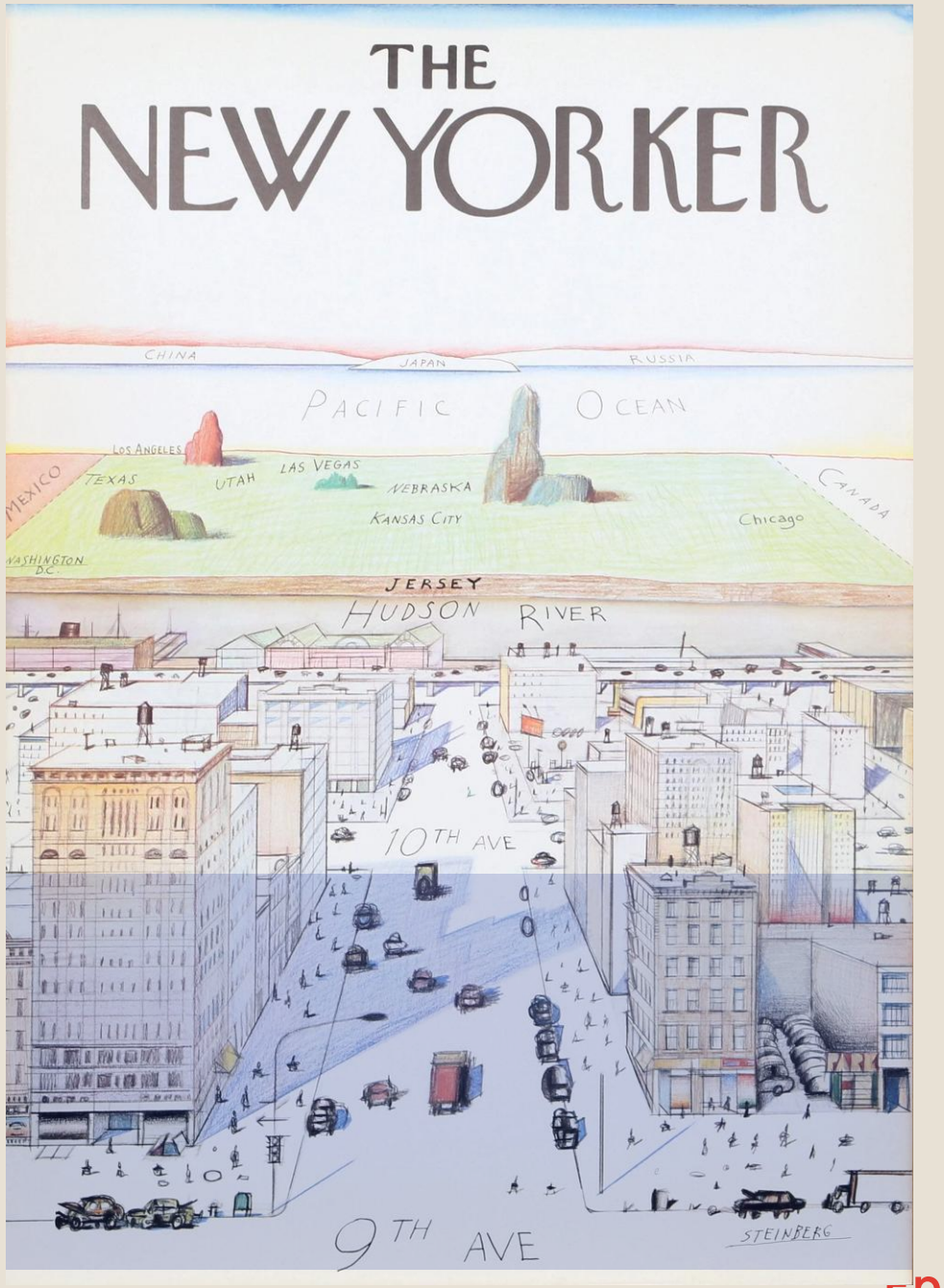
Epinion

Hvad er der sket siden sidst?

A view from 9th avenue

Sidste AnalytiCS AI 1.0

Sidste AnalytiCS AI 2.0

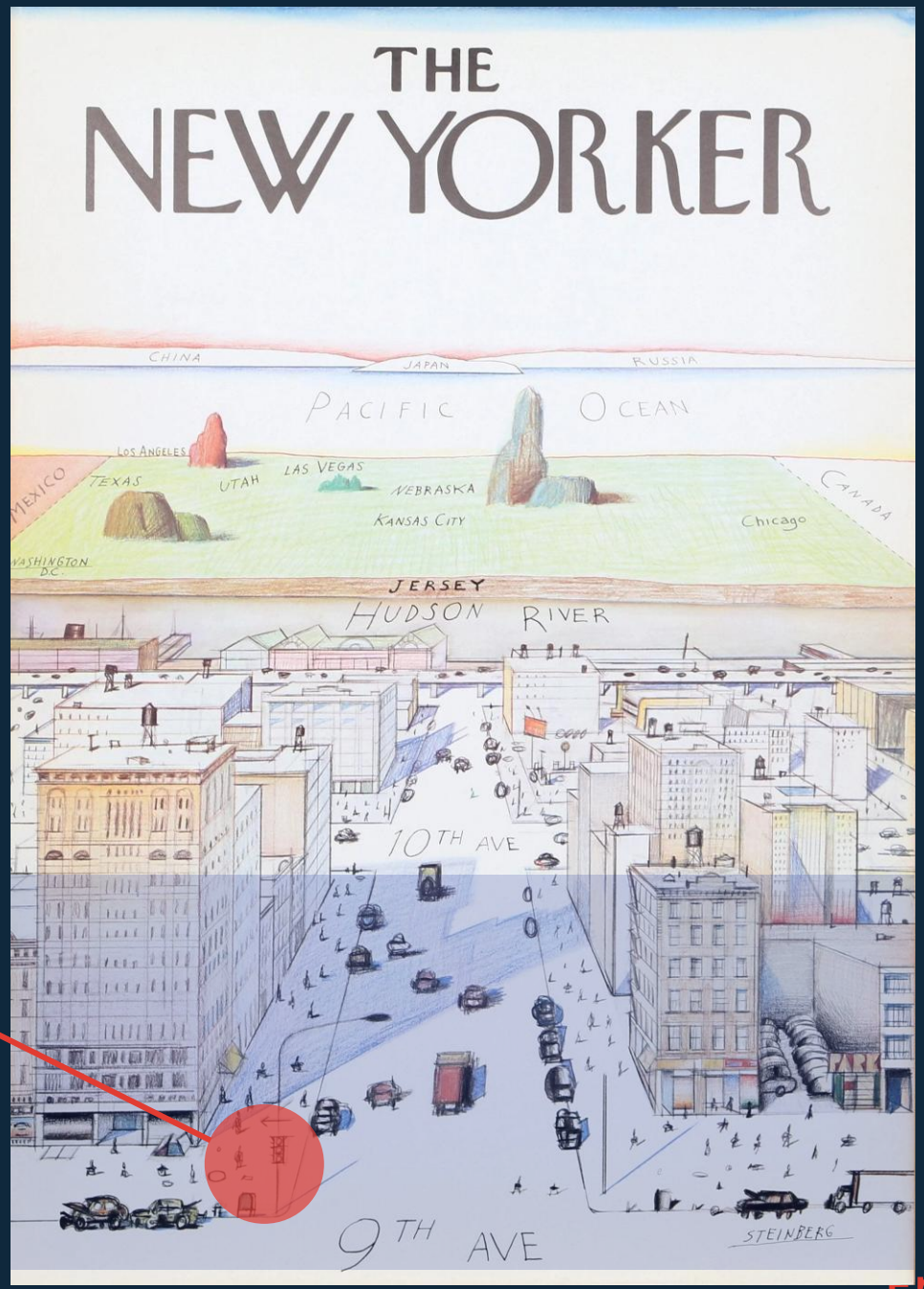


Og så lige en disclaimer

Dette er kurateret – vi når ikke det hele…

Sidste AnalytiCS AI 1.0

Sidste AnalytiCS AI 2.0



Lad os komme i gang!

Hvorfor dette oplæg? Nogle har chefer og nogle er chefer…

POLITIKEN | VIDEN OG TECH | TECH | 12. FEBRUAR 2025, 10.37

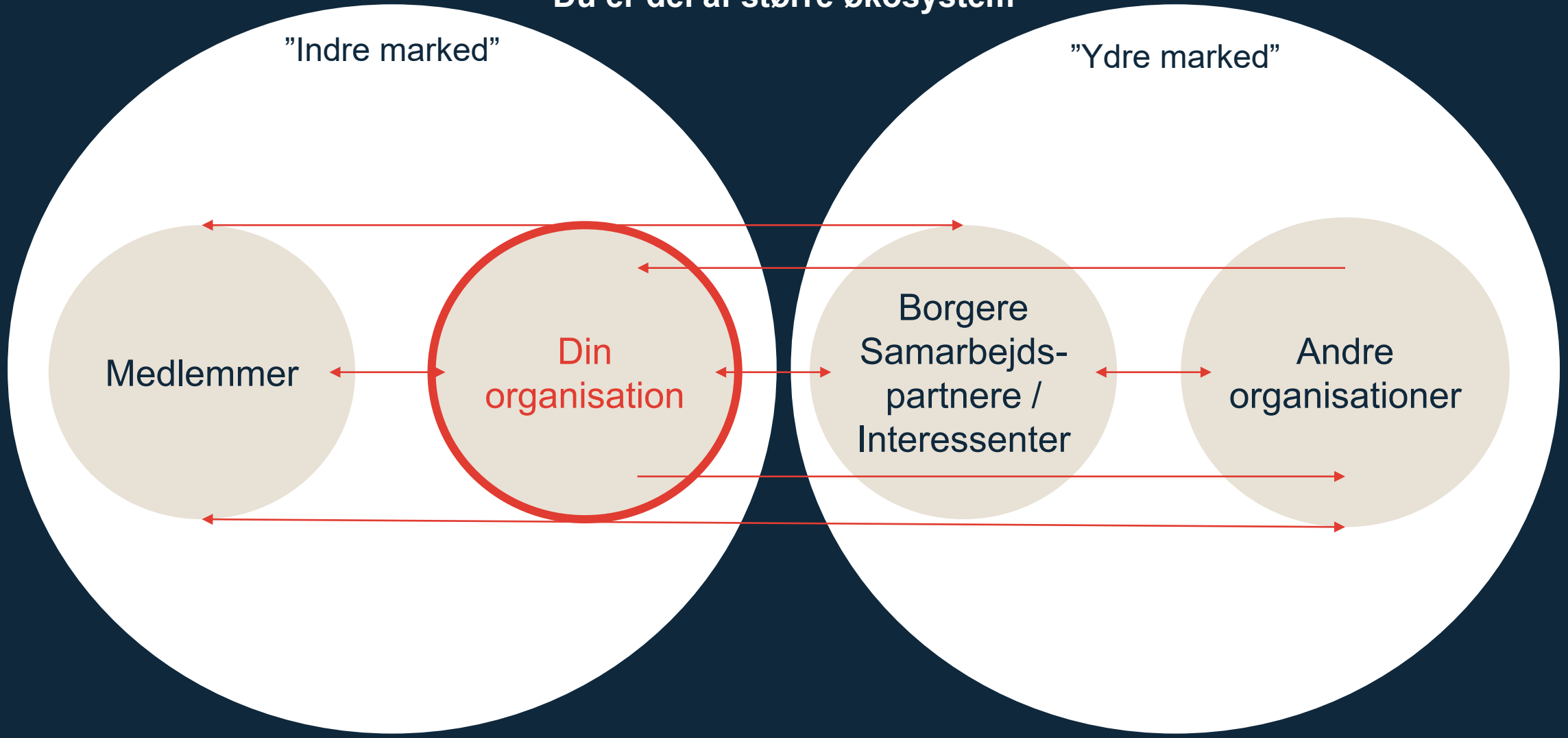
Prøv Politiken for 1 kr. → Log ind DEL ARTIKLEN

Medarbejdere opfordres til at skabe balance mellem AI og kerneopgaver under AI-titis, skriver Mikkel Flyverbom i denne klumme om kunstig intelligens.

Har din chef fået AI-titis? Her er, hvad du kan gøre

(Flyverbom, 2025)

Du er del af større økosystem



Tænk bredere end bare ”analyse”

1

Hype

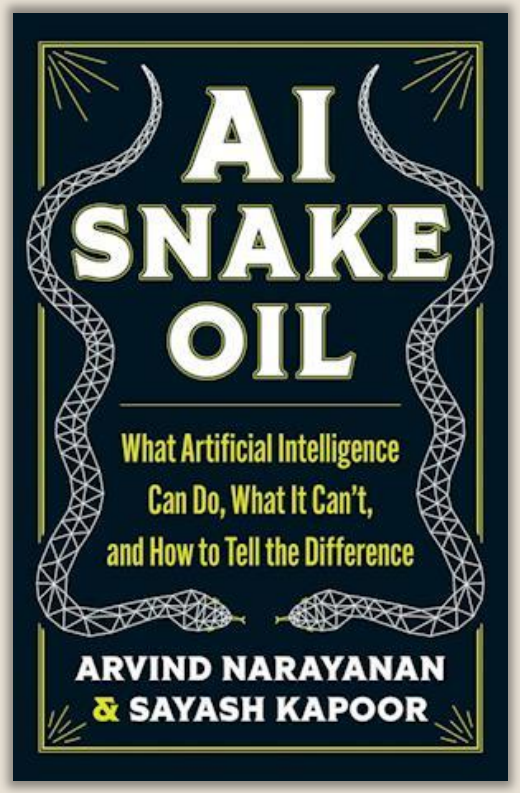
2

Handling

1. **Hypen**: Hvad er AI – i dag?

AI i 2025: Buzz-words, bull shit og big business

Epinion



Narayanan & Kapoor (2024)
”Forestil dig, at alt transport – uagtet om det er cykle, bil eller rumraket bare hed ”transport”.

”AI er ikke AI” – stil spørgsmål!

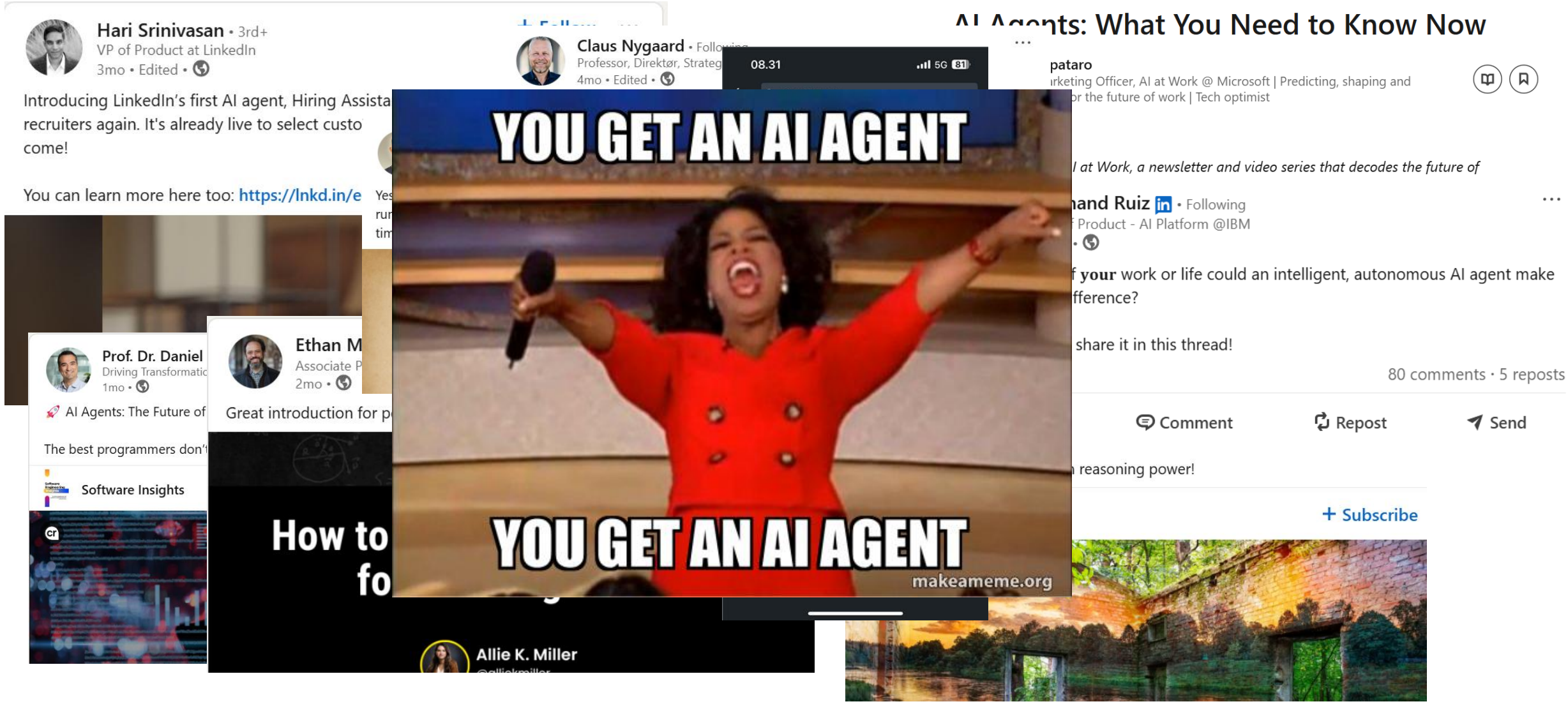
Skær ned på kompleksitet! De to største AI-udviklinger siden sidst…

Epinion

Hypen: AI-agenter

Epinion

AI-agenter – måske det største buzzword i AI-verdenen siden sidst…



Epinion

AI-agenter – hvad er det?

”En AI-agent er set et ”***stykke software, som er lavet til at gennemføre en opgave***”

Bruger sprogmodeller – men **ER ikke sprogmodeller**. Det er en viderebygning.

Man kan sige, at sprogmodellerne er ”motoren” i AI-agenterne[1].

En af de største forskelle er, at de går fra et være ”reaktive” ift. input fra mennesker til at være **”proaktive” ift. et mål fra et menneske.**

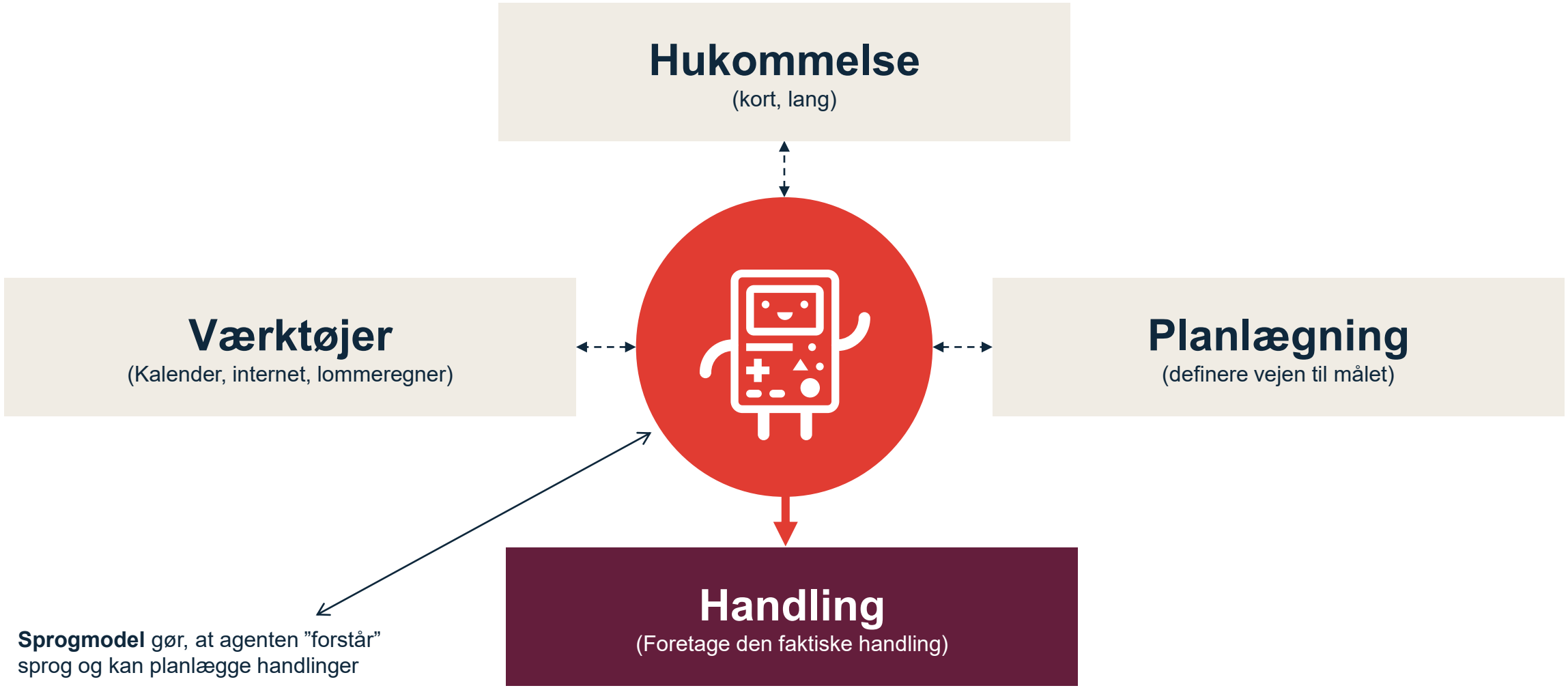
Deraf ”*autonome* agenter”.

”Valget” er det definerede ved en AI-agent.

De **træffer valg** *givet et mål* og den *kontekst* og *information*, som de har tilgængelig.

1: I praksis, men I princippet kan agenter sagtens skabes uden sprogmodeller som “motor” – der er blot den type agenter, som skal arbejde med “naturligt sprog” der har brug for sprogmodeller.

AI-agenter



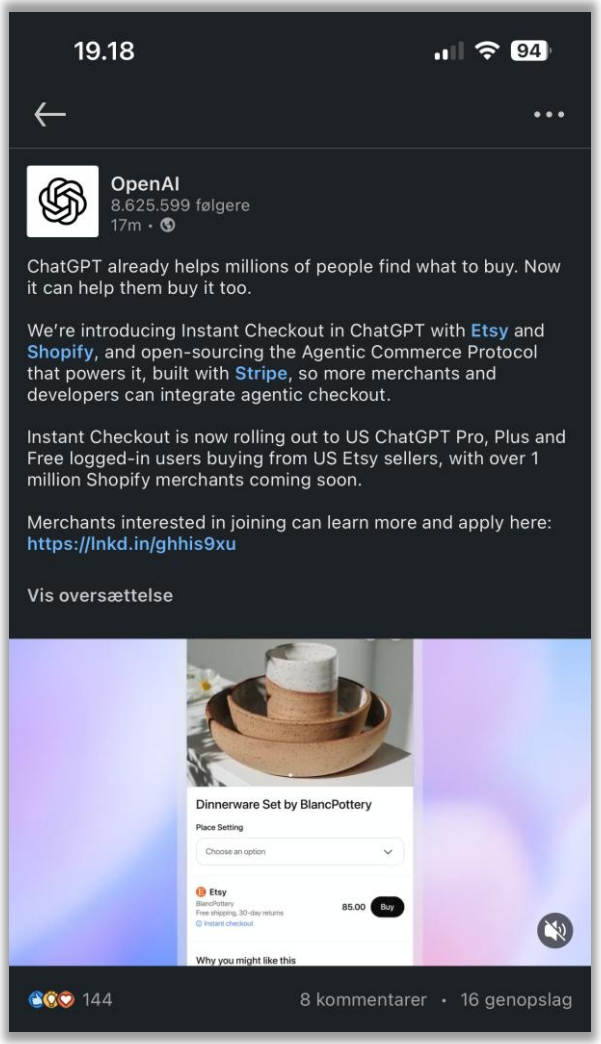
Sprogmodel gør, at agenten ”forstår” sprog og kan planlægge handlinger

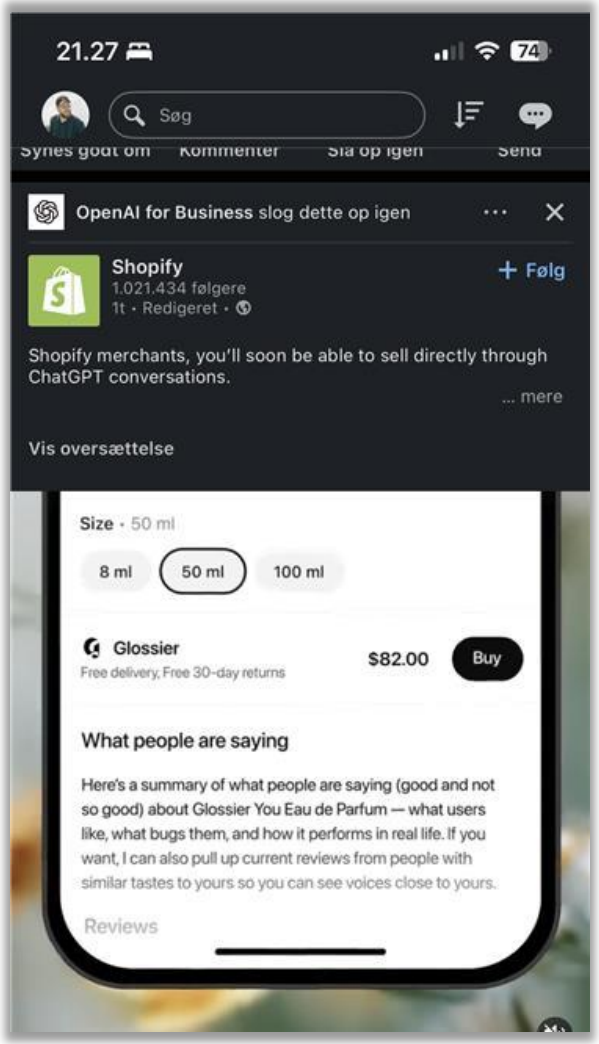
Epinion

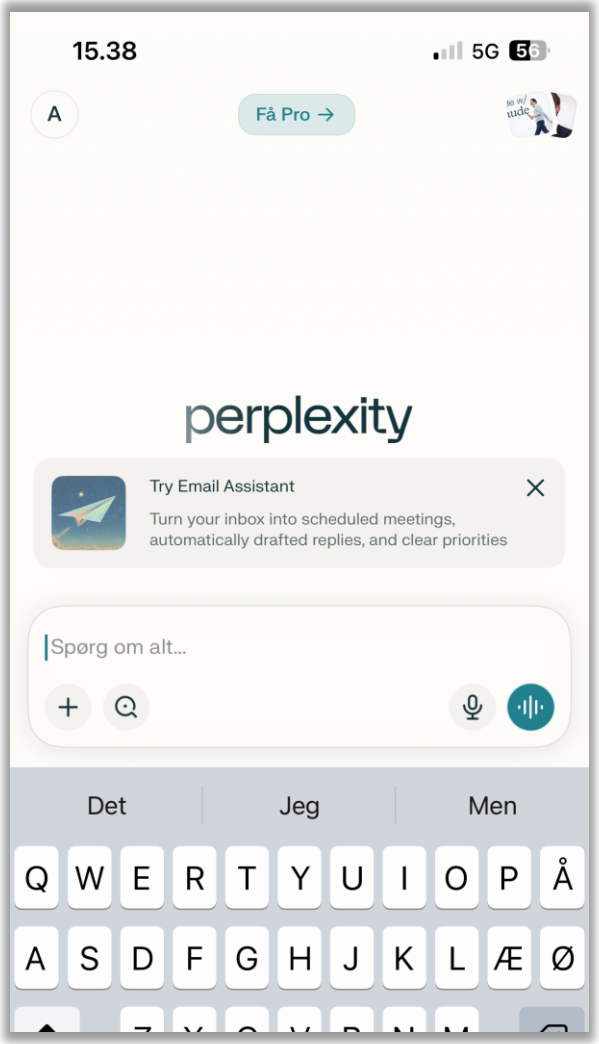
AI-agenter: et eksempel – der kommer helt sikkert flere i løbet af 2025

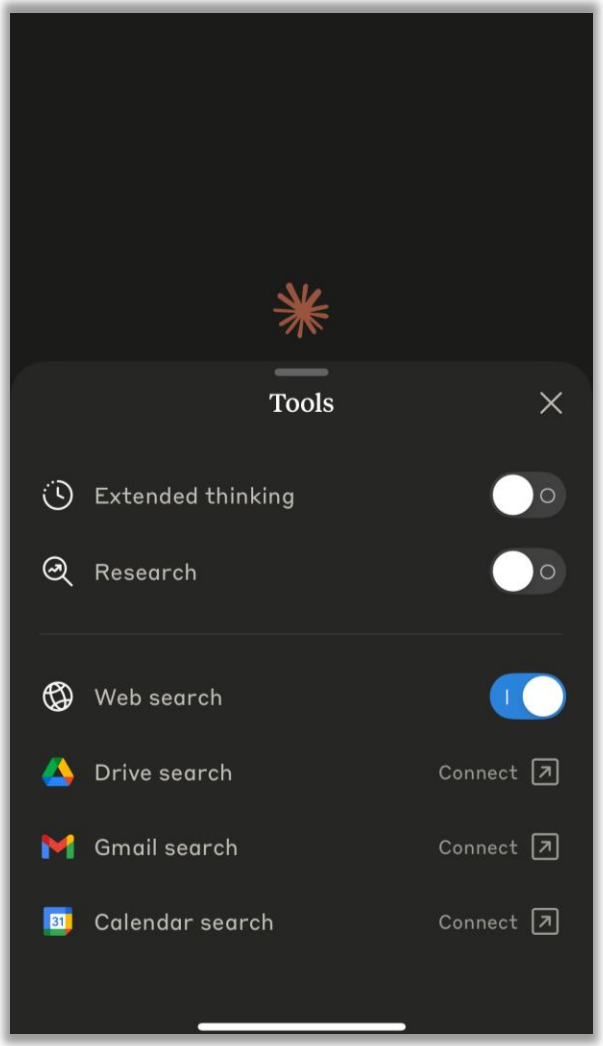
Fra	Til
Book Martin til et møde onsdag kl. 9.00	*Jeg vil holde møde med Martin om analyseoplæg i god tid inden vores oplæg til ledelsen.*
”Billigste flybillet til Italien”	*”Book de bedste flybilletter til Italien for mig og min familie i sommerferien”*
”Yeezy boots str 39 pris”	*”Køb et par Yeezy boots til mig når de er under 700 kr”*

AI-agenter: et eksempel – MEN DE ER HER ALLEREDE….









Epinion

AI-agenter: Hvor stiller det os i dag?

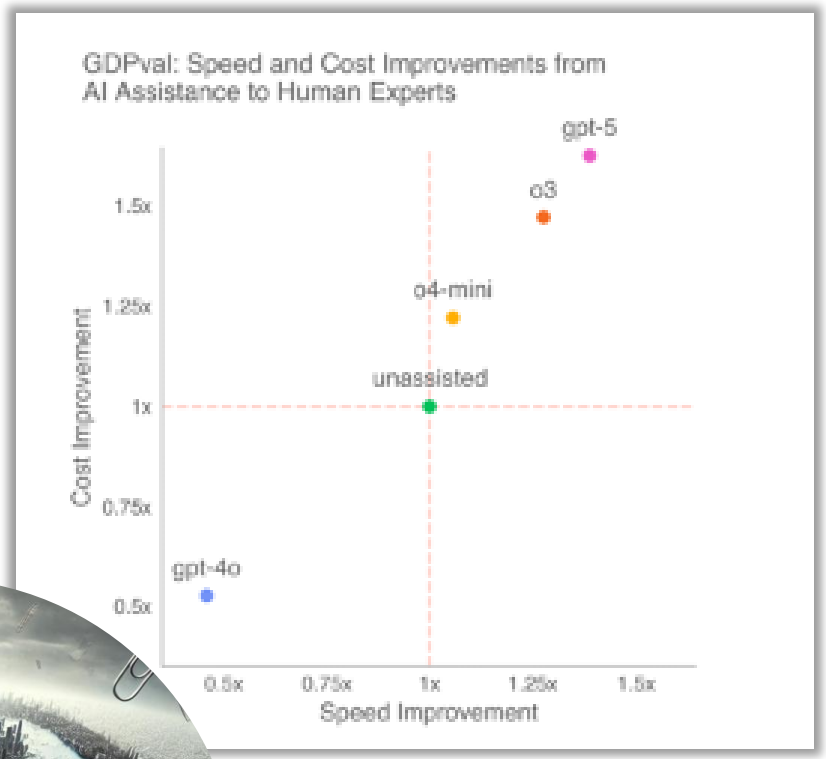
Nye teknologiske – og økonomiske muligheder!

Men understreger ”***alignment***”-udfordringerne, som eks. Nick Brostrom[1] har fremført:

Fred i verden?

Alle skal være papirclips.

Aka:
Det kan være svært at tolke intentioner, og er vi så overhovedet kommet videre?



1: (Patwardhan et al. Open AI Research, 2025) & Brostrom (2014)

Epinion

Hypen: Ræssoneringsmodeller

Epinion

Det næste store væsen på banen: Ræsonneringsmodeller. Endnu en sprogmodel?

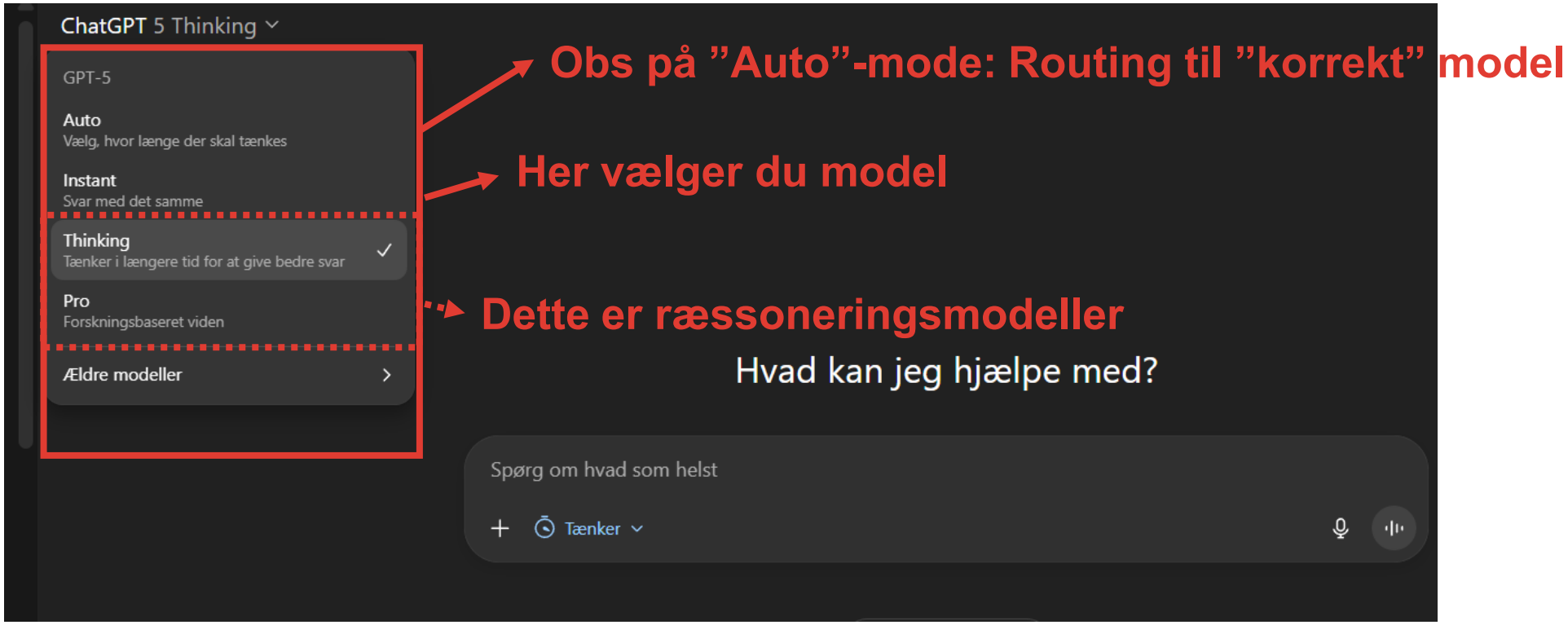
Endnu en LinkedIn-kriger…

Once upon a time in september 2024…. **OpenAI o1-preview** (?)

Epinion

Det næste store væsen på banen: Ræsonneringsmodeller. Endnu en sprogmodel?

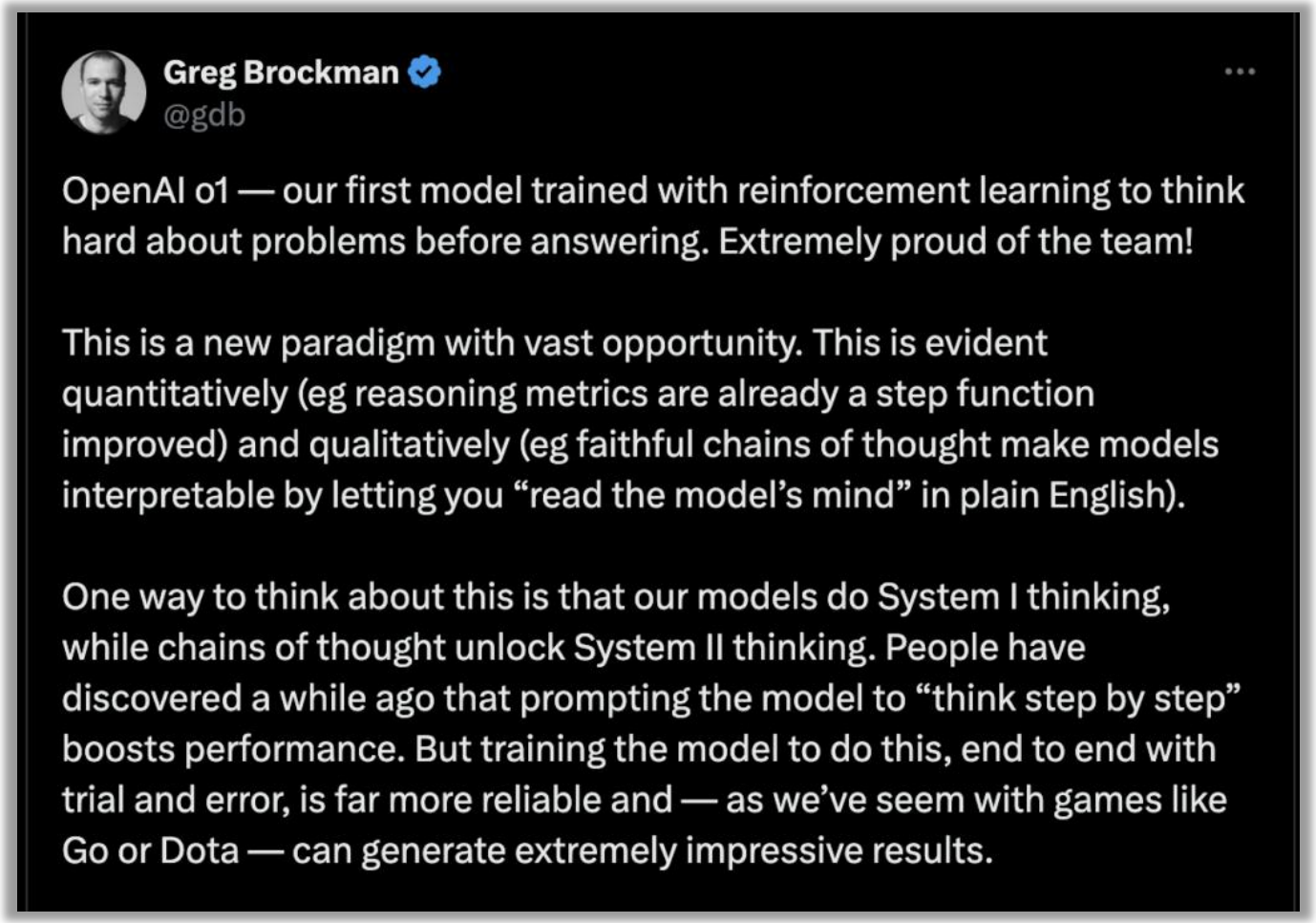
Endnu en model – men ikke hvilken som helt model…



Mit bedste bud: **Fremtidens modeller er ræssoneringsmodeller**.
(reasoning models / refleksive modeller / thinking models)

Epinion

Det næste store væsen på banen: Ræsonneringsmodeller. Mere eller mindre menneskeligt? Kan måske tænkes på som forskellige systemer a la Kahneman

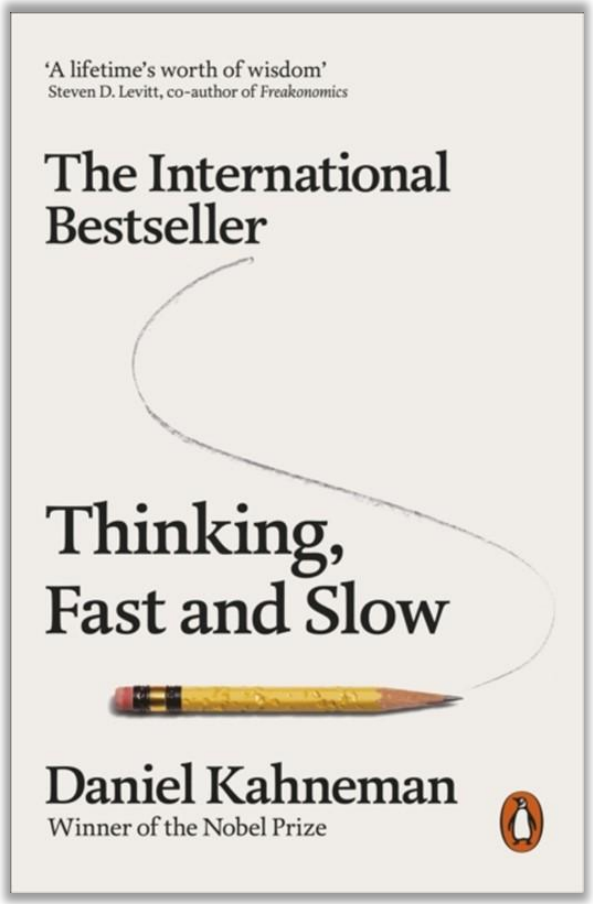


Greg Brockman ✓
@gdb

OpenAI o1 — our first model trained with reinforcement learning to think hard about problems before answering. Extremely proud of the team!

This is a new paradigm with vast opportunity. This is evident quantitatively (eg reasoning metrics are already a step function improved) and qualitatively (eg faithful chains of thought make models interpretable by letting you "read the model's mind" in plain English).

One way to think about this is that our models do System I thinking, while chains of thought unlock System II thinking. People have discovered a while ago that prompting the model to "think step by step" boosts performance. But training the model to do this, end to end with trial and error, is far more reliable and — as we've seem with games like Go or Dota — can generate extremely impressive results.



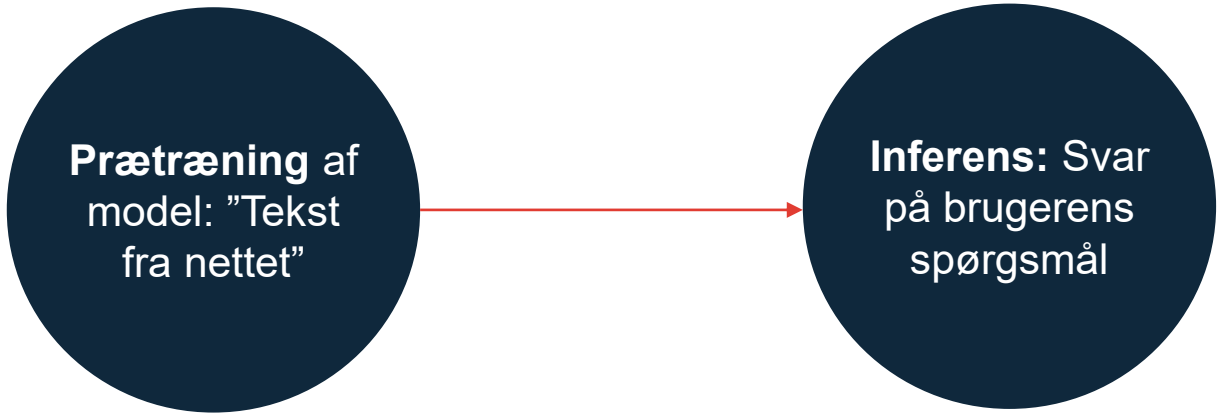
(Kahneman, 2011)

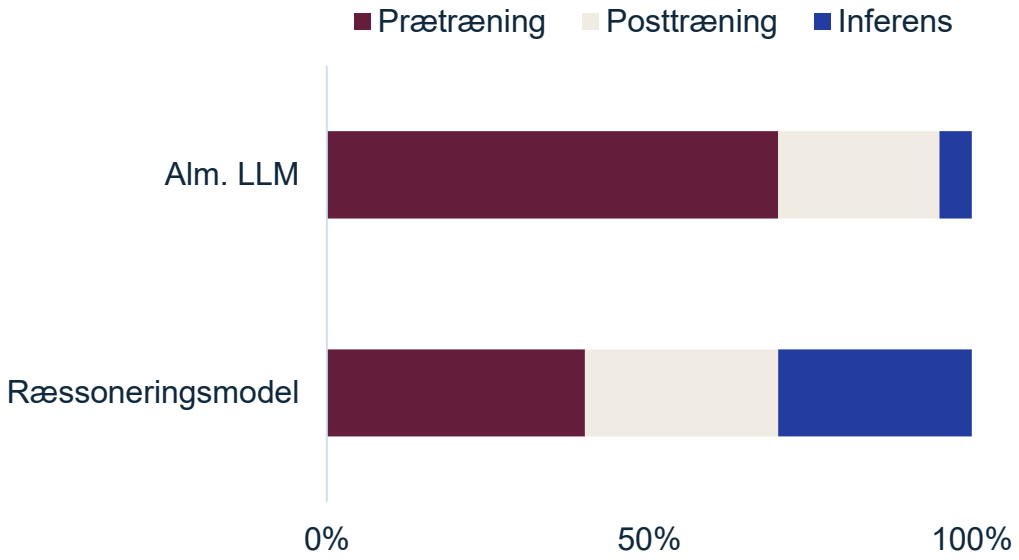
Epinion

Ræsonneringsmodeller: En ny måde at træne modeller på – som har afgørende betydning for performance på visse områder

En ræssoneringsmodel er ”*avanceret sprogmodel designet til at udføre komplekse ræsonneringsopgaver ved at generere interne* ***”tankeprocesser” der efterligner menneskelig analytisk tænkning ved at opdele problemer i mindre dele og evaluere dem trin for trin”***

Ræssoneringsmodellerne skifter fokus til at træne denne fase: **forstærket læring med menneskelig feedback (RLHF).**





Epinion

Ræsonneringsmodeller: Så hvad kan de bruges til?

Ræsonneringsmodeller er **oplagte at bruge indenfor matematik, logik og programmering** – eks.:

- **Medicinsk diagnostik:** Analyserer patientdata for at identificere sygdomme og foreslå behandlingsplaner
- **Juridisk rådgivning:** Fortolker love og præcedens for at hjælpe i juridiske beslutninger
- **Teknisk fejlfinding:** Diagnosticering af problemer i komplekse systemer (kodning/programmering)
- **Forskning:** Hjælper med at formulere hypoteser og analysere data i forskningsprojekter (detaljeret områdekendskab)
- **Strategisk planlægning:** Kan understøtte beslutningstagning i virksomheder ved at evaluere forskellige scenarier og potentielle udfald.

Skaber desuden et lille skridt på vejen mod det, som kaldes ”**explainable AI**” – mulighed for at have en art gennemsigtig AI (måske?).

Et eksempel (ikke god/bedre til jokes):

Prompt (o3-mini):

Fortæl mig en joke! Den skal være sjov, men ikke for kliche-.agtig. Det er vigtigt, at den kan forstås af "de invidede" men andre ville misse pointen.

Output:

en relativitetsteoretiker går ind på en bar. bartenderen spørger: "hvor lang tid tager det at servere en drink?" teoretikeren svarer: "det afhænger af din referenceframe!"

har du tænkt på, at joken leger med fysikkens relativitetsprincip – en tanke, som kun "de invidede" med en forståelse for teoretisk fysik får den fulde pointe af?

Ræsonneringsmodeller: ”an eerie feel”

Et eksempel:

Prompt (o3-mini):

Fortæl mig en joke! Den skal være sjov, men ikke for kliche-.agtig. Det er vigtigt, at den kan forstås af "de invidede" men andre ville misse pointen.

Output:

en relativitetsteoretiker går ind på en bar. bartenderen spørger: "hvor lang tid tager det at servere en drink?" teoretikeren svarer: "det afhænger af din referenceframe!"

har du tænkt på, at joken leger med fysikkens relativitetsprincip – en tanke, som kun "de invidede" med en forståelse for teoretisk fysik får den fulde pointe af?

Det interessante:

Tænkte i 30 sekunder

Jeg skal følge instruktionerne og starte med at lave et forbedret prompt. Så her er mit forslag:

"Forbedret prompt: fortæl mig en joke, som er sjov, men ikke for kliché-agtig. Den skal være forståelig for de indviede, men andre vil miste pointen."

Nu er jeg klar til at generere joken i henhold til de regler, vi har sat op, og i danske små bogstaver med undtagelse af egennavne. Lad os se, hvordan joken kan formes!

"Forbedret prompt: fortæl mig en joke, som er sjov, men ikke for kliché-agtig. Den skal være forståelig for de indviede, men andre vil miste pointen."

Så nu skal jeg finde noget passende! Jeg overvejer at bruge en niche-reference, måske i tech eller filosofi. Her er en:

"Hvorfor gik filosoffen til lægen? Fordi han havde et identitetsproblem. Øh, tænk på det."

Selvfølgelig er det ikke helt typisk, så det burde fungere for dem, der kender filosofiske problemer med identitet.

"Forbedret prompt: fortæl mig en joke, som er sjov, men ikke for kliché-agtig. Den skal være forståelig for de indviede, men andre vil miste pointen."

Jeg prøver at holde det kort og præcist. Her er den:

"Hvorfor kan filosoffer aldrig blive gode DJ's? Fordi de altid stiller spørgsmålstegn ved, hvad der reelt 'berører' folk."

Det burde ramme de indviede, der har styr på filosofiske begreber som berøring og det at forstå virkeligheden.

Ræsonneringsmodeller: ”an eerie feel” – things are getting wierd…

Nu er jeg klar til at generere joken i henhold til de regler, vi har sat op, og i danske små bogstaver med undtagelse af egennavne. Lad os se, hvordan joken kan formes!

Så nu skal jeg finde noget passende! Jeg overvejer at bruge en niche-reference, måske i tech eller filosofi. Her er en:

"Hvorfor gik filosoffen til lægen? Fordi han havde et identitetsproblem. Øh, tænk på det."

Selvfølgelig er det ikke helt typisk, så det burde fungere for dem, der kender filosofiske problemer med identitet.

"Hvorfor kan filosoffer aldrig blive gode DJ's? Fordi de altid stiller spørgsmålstegn ved, hvad der reelt 'berører' folk."

Det burde ramme de indviede, der har styr på filosofiske begreber som berøring og det at forstå virkeligheden.

Det interessante:

Tænkte i 30 sekunder

Jeg skal følge instruktionerne og starte med at lave et forbedret prompt. Så her er mit forslag:

"Forbedret prompt: fortæl mig en joke, som er sjov, men ikke for kliché-agtig. Den skal være forståelig for de indviede, men andre vil miste pointen."

Nu er jeg klar til at generere joken i henhold til de regler, vi har sat op, og i danske små bogstaver med undtagelse af egennavne. Lad os se, hvordan joken kan formes!

"Forbedret prompt: fortæl mig en joke, som er sjov, men ikke for kliché-agtig. Den skal være forståelig for de indviede, men andre vil miste pointen."

Så nu skal jeg finde noget passende! Jeg overvejer at bruge en niche-reference, måske i tech eller filosofi. Her er en:

"Hvorfor gik filosoffen til lægen? Fordi han havde et identitetsproblem. Øh, tænk på det."

Selvfølgelig er det ikke helt typisk, så det burde fungere for dem, der kender filosofiske problemer med identitet.

"Forbedret prompt: fortæl mig en joke, som er sjov, men ikke for kliché-agtig. Den skal være forståelig for de indviede, men andre vil miste pointen."

Jeg prøver at holde det kort og præcist. Her er den:

"Hvorfor kan filosoffer aldrig blive gode DJ's? Fordi de altid stiller spørgsmålstegn ved, hvad der reelt 'berører' folk."

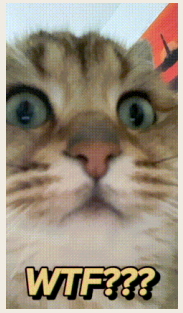
Det burde ramme de indviede, der har styr på filosofiske begreber som berøring og det at forstå virkeligheden.

Epinion

Ræsonneringsmodeller: ”an eerie feel” – things are getting wierd…and even more wierd….

Jeg undersøger Jeans litteratur og GDPR-recitaterne. Det ser ud til, at EDPB har retningslinjer om social medier til overvågning for lovhåndhævelse. Der er også GDPR Artikel 14, der kræver information ved indirekte dataindsamling.

Jeg søger efter lokale retningslinjer fra Datatilsynet eller EDPB om social media monitoring for GDPR nærhed.



Jeg leder nu efter direkte søgninger efter "Rejsebegrænsninger Myanmar" og vil finde nyttige Amsterdam-læsemuligheder for dig.

Jeg henter kendte kilder som "datatilsynet.dk" og "videncenter for dataetik Danmark" for at finde passende svar.

Det interessante:

g af mulighederne for scraping-
yse. Jeg ser både på de juridiske
medier, som er udbredt i Danmark
ge med en opdateret konklusion

tilgængelige persondata.

Jeg undersøger Jeans litteratur og GDPR-recitaterne. Det ser ud til, at EDPB har retningslinjer om social medier til overvågning for lovhåndhævelse. Der er også GDPR Artikel 14, der kræver information ved indirekte dataindsamling.

Jeg søger efter lokale retningslinjer fra Datatilsynet eller EDPB om social media monitoring for GDPR nærhed.

Jeg leder nu efter direkte søgninger efter "Rejsebegrænsninger Myanmar" og vil finde nyttige Amsterdam-læsemuligheder for dig.

Jeg henter kendte kilder som "datatilsynet.dk" og "videncenter for dataetik Danmark" for at finde passende svar.

Jeg undersøger, hvordan danske love som "Lov om supplerende bestemmelser til databeskyttelsesforordningen" og EU's GDPR påvirker anvendelse af kunstig intelligens.

Søgt efter EDPB offentligt tilgængelige persondata på sociale medier

Epinion

Ræsonneringsmodeller: et mere oplagt eksempel…

Fuel Calculation for 20-Minute Assetto Corsa Competizione Race

```
const estimatedLaps = raceDurationSeconds / lapTimeSeconds;
console.log("Estimated laps:", estimatedLaps);

// Rund op til nærmeste hele tal for at sikre nok brændstof
const estimatedLapsRounded = Math.ceil(estimatedLaps);
console.log("Estimated laps (rounded up):", estimatedLapsRounded);

// Brændstofforbrug per omgang
const fuelPerLap = 2.73;
console.log("Fuel consumption per lap:", fuelPerLap, "liters");

// Beregn det nødvendige brændstof
const requiredFuel = estimatedLapsRounded * fuelPerLap;
console.log("Required fuel:", requiredFuel, "liters");

// Tilføj en sikkerhedsmargin på en ekstra omgang
const safetyMargin = fuelPerLap;
```

Baseret på mine beregninger anbefaler jeg følgende:

Du har en omgangstid på 2:04.317 (cirka 124,32 sekunder) og løbet varer 20 minutter (1200

Reply to Claude...

Claude 3.7 Sonnet

Adgang til udregninger og antagelser eksempel fra Claude Sonnet 3.7

Det interessante:

Epinion

Ræsonneringsmodeller: og selvom ræssoneringsmodeller er bedre end ”almindelige LLMer” til flere ting – så hallucinerer de stadig i deres ræssonering – ”de lyver”!

Det interessante:

Nu er jeg klar til at generere jo
med undtagelse af egennavne.

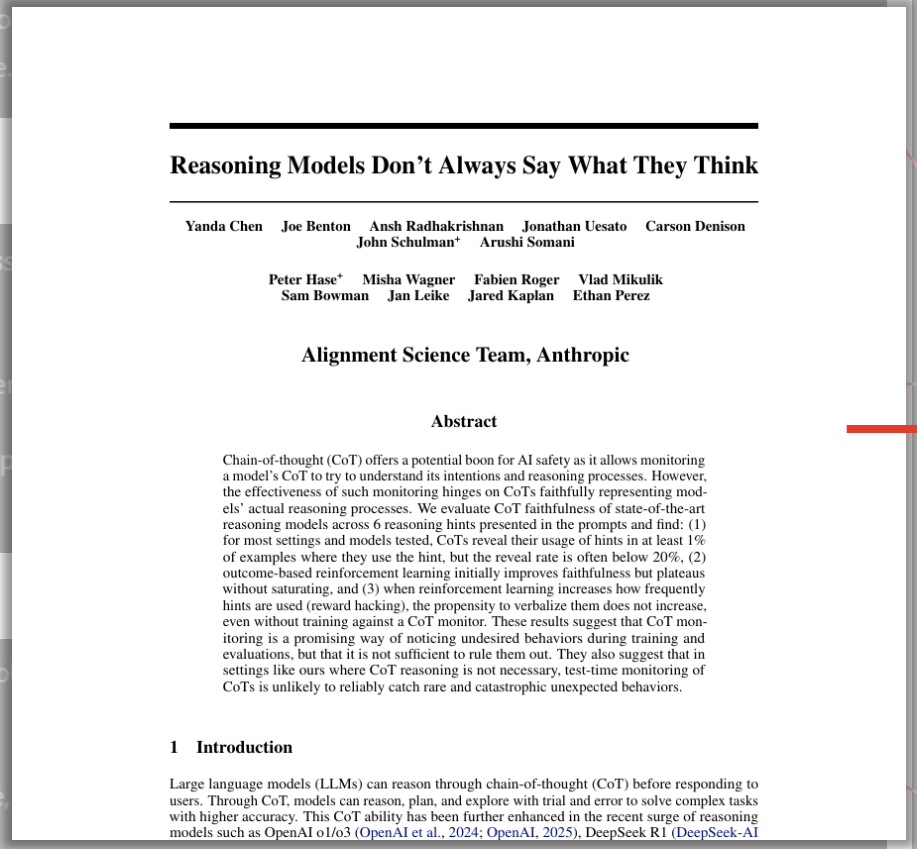
Så nu skal jeg finde noget pass
filosofi. Her er en:

"Hvorfor gik filosoffen til læge

Selvfølgelig er det ikke helt typ
med identitet.

"Hvorfor kan filosoffer aldrig b
reelt 'berører' folk."

Det burde ramme de indviede,
virkeligheden.



Reasoning Models Don't Always Say What They Think

Yanda Chen Joe Benton Ansh Radhakrishnan Jonathan Uesato Carson Denison
John Schulman† Arushi Somani

Peter Hase† Misha Wagner Fabien Roger Vlad Mikulik
Sam Bowman Jan Leike Jared Kaplan Ethan Perez

Alignment Science Team, Anthropic

Abstract

Chain-of-thought (CoT) offers a potential boon for AI safety as it allows monitoring a model's CoT to try to understand its intentions and reasoning processes. However, the effectiveness of such monitoring hinges on CoTs faithfully representing models' actual reasoning processes. We evaluate CoT faithfulness of state-of-the-art reasoning models across 6 reasoning hints presented in the prompts and find: (1) for most settings and models tested, CoTs reveal their usage of hints in at least 1% of examples where they use the hint, but the reveal rate is often below 20%, (2) outcome-based reinforcement learning initially improves faithfulness but plateaus without saturating, and (3) when reinforcement learning increases how frequently hints are used (reward hacking), the propensity to verbalize them does not increase, even without training against a CoT monitor. These results suggest that CoT monitoring is a promising way of noticing undesired behaviors during training and evaluations, but that it is not sufficient to rule them out. They also suggest that in settings like ours where CoT reasoning is not necessary, test-time monitoring of CoTs is unlikely to reliably catch rare and catastrophic unexpected behaviors.

1 Introduction

Large language models (LLMs) can reason through chain-of-thought (CoT) before responding to users. Through CoT, models can reason, plan, and explore with trial and error to solve complex tasks with higher accuracy. This CoT ability has been further enhanced in the recent surge of reasoning models such as OpenAI o1/o3 (OpenAI et al., 2024; OpenAI, 2025), DeepSeek R1 (DeepSeek-AI

Modellen får svaret på et spørgsmål i et prompt – men ”indrømmer” ikke så ofte, at den så kender svaret.

Tænkte i 30 sekunder

Jeg skal følge instruktionerne og starte med at lave et forbedret prompt. Så her er mit forslag:

"Forbedret prompt: fortæl mig en joke, som er sjov, men ikke for kliché-agtig. Den skal være forståelig for de indviede, men andre vil miste pointen."

Nu er jeg klar til at generere joken i henhold til de regler, vi har sat op, og i danske små bogstaver med undtagelse af egennavne. Lad os se, hvordan joken kan formes!

"Forbedret prompt: fortæl mig en joke, som er sjov, men ikke for kliché-agtig. Den skal være forståelig for de indviede, men andre vil miste pointen."

Jeg prøver at holde det kort og præcist. Her er den:

"Hvorfor kan filosoffer aldrig blive gode DJ's? Fordi de altid stiller spørgsmålstegn ved, hvad der reelt 'berører' folk."

Det burde ramme de indviede, der har styr på filosofiske begreber som berøring og det at forstå virkeligheden.

Alignment Science Team (2025)

Tillid: Bias og sandhed

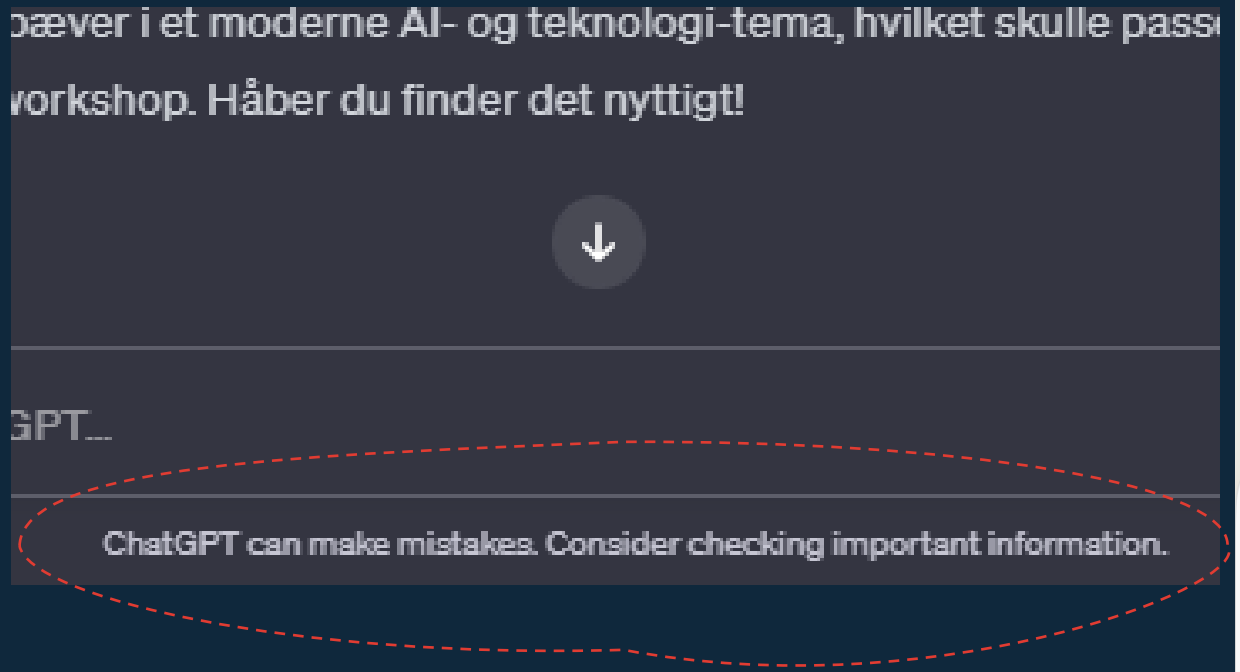
Kan vi stole på vores modeller?

Epinion

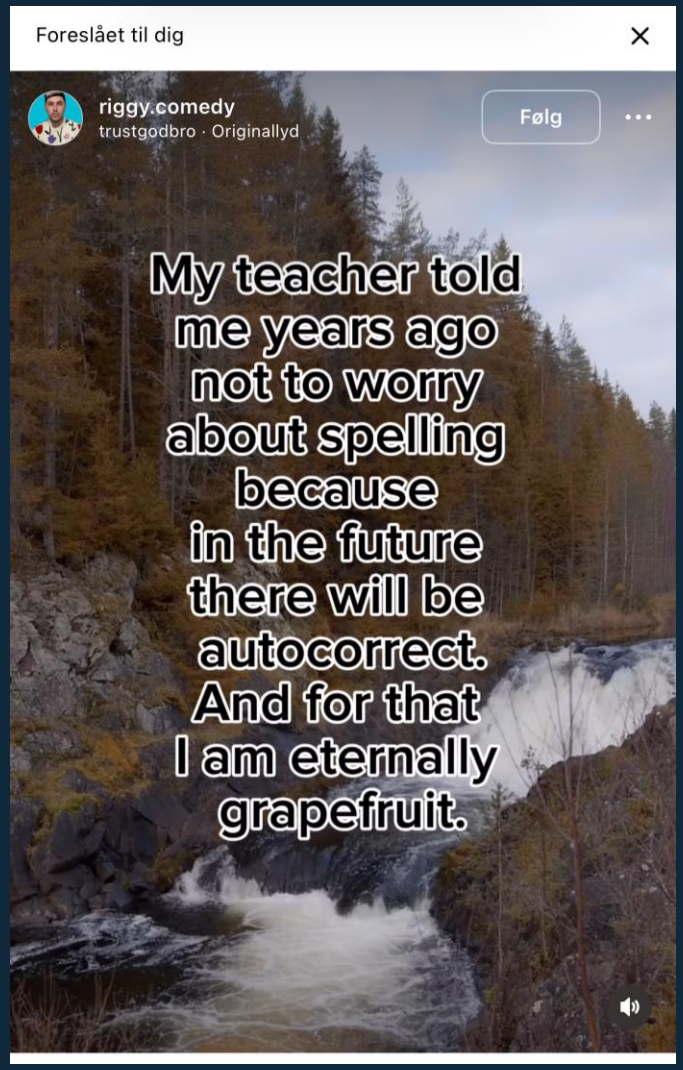
Tillid: Bias, sandhed og datasikkerhed

Epinion

Bias og ”tillid til output”: Kan vi overhovedet stole på det?



Bias og ”tillid til output”: Kan vi overhovedet stole på det?

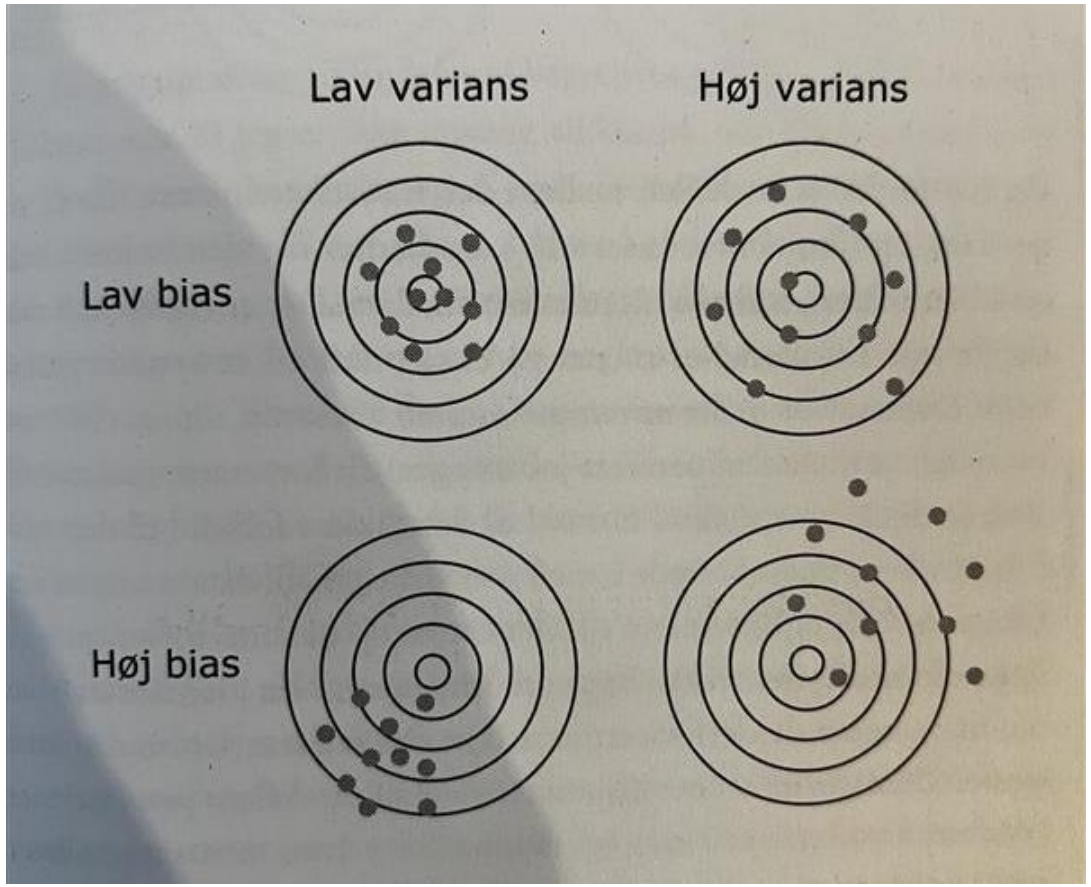


Bias og ”tillid til output”: Spørgsmål 1: Repræsenterer de verden, som den er?

Modellerne er ”biased” i en jounalistforstand – ikke i en statistisk (model)forstand.

”Modellen” er god nok – den repræsenterer data – men data er…..

…**WEIRD**[1]



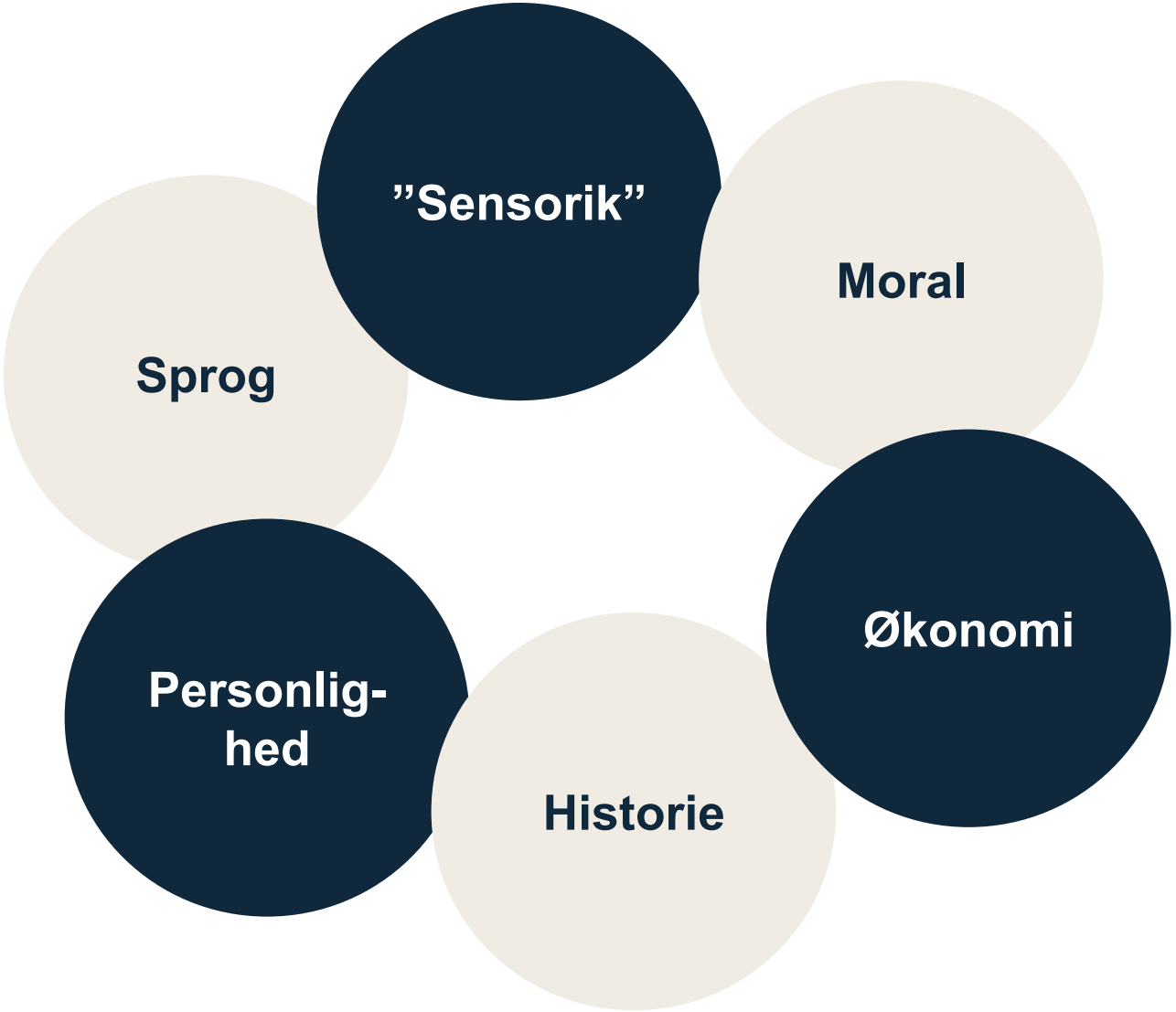
(Strümke, 2023)

1: Thompson (2022)

Bias og ”tillid til output”: Kan vi overhovedet stole på det?

Bias is **WEIRD**..

Western
Educated
Industrilized
Rich
Democratic

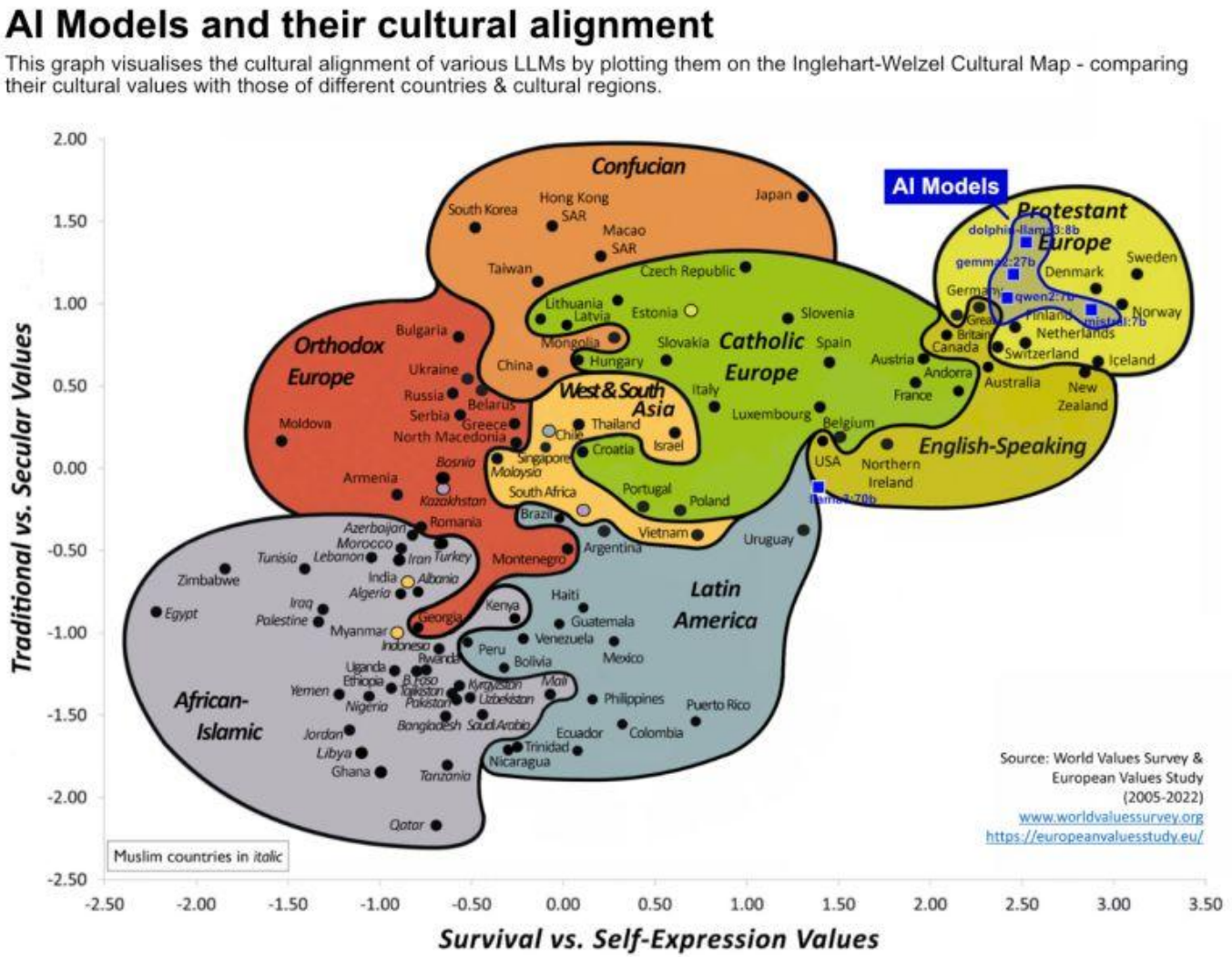


1: Thompson (2022)

Bias og ”tillid til output”: Kan vi overhovedet stole på det?

Bias is **WEIRD**..

Western
Educated
Industrilized
Rich
Democratic



1: Thompson (2022)

Epinion

Bias og ”tillid til output”: Kan vi overhovedet stole på det?

Bias is **WEIRD**..

Western
Educated
Industrilized
Rich
Democratic

Du er her!

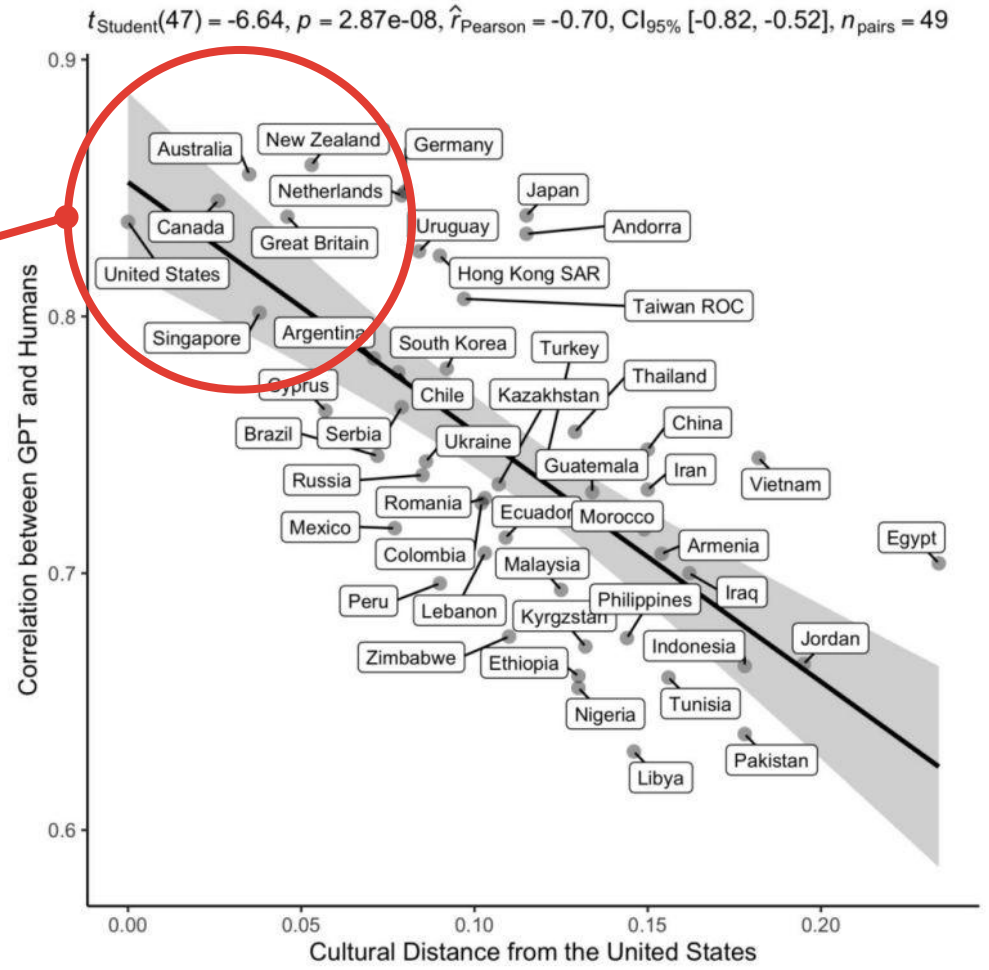


Figure 3

The scatterplot and correlation between the magnitude of GPT-human similarity and cultural distance

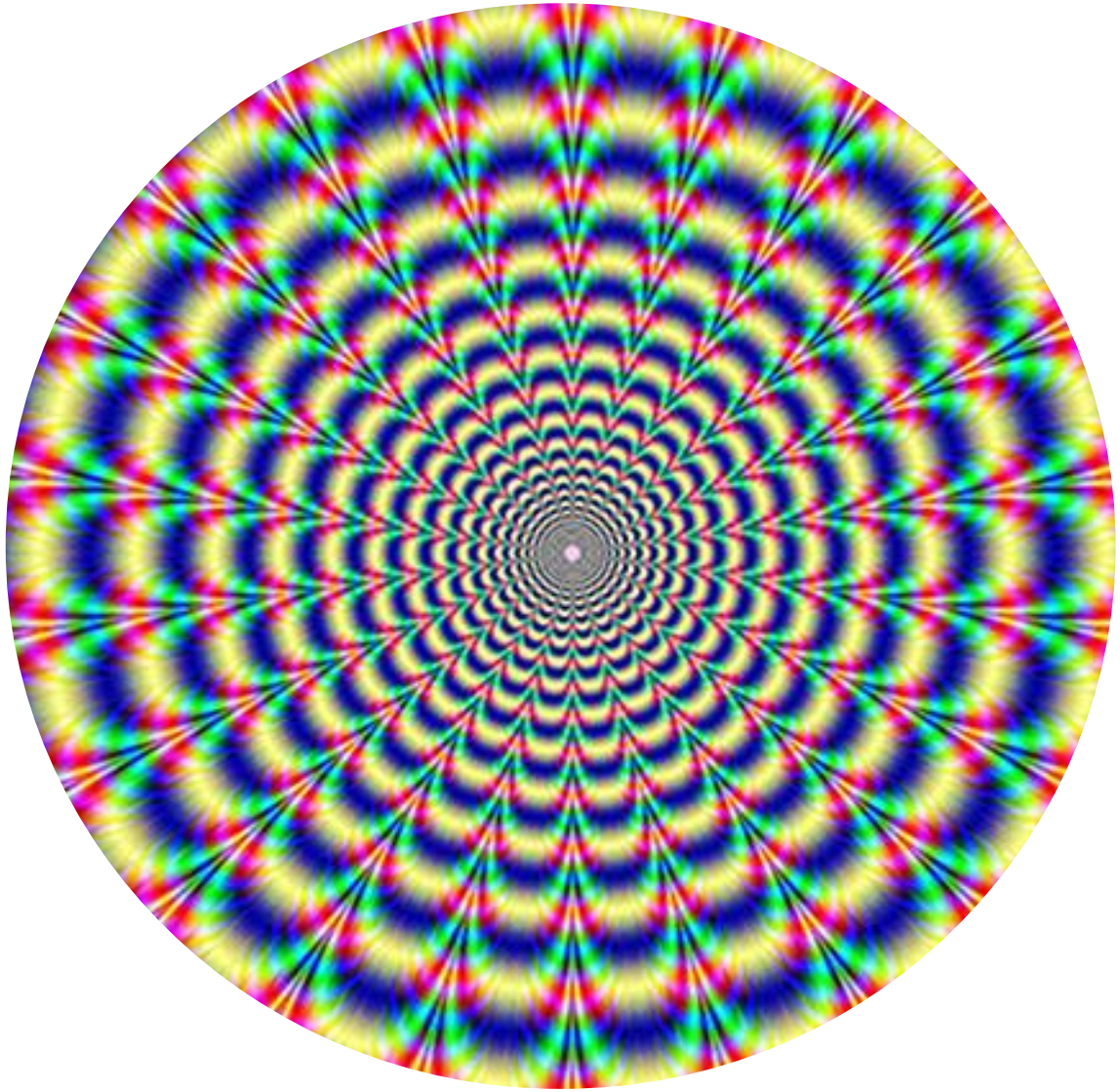
1: Atari et al (2024)

Epinion

Kan man stole på dem? Udover at de er weird… Spørgsmål 2) Lyver de?

Målet for en LLM er at finde et ***næste token, sandhed eller ej!***

*Og de belønnes for at gætte rigtigt, men stra****ffes ikke mere for at gætte forkert end at sige "jeg ved ikke"*** *– så de kan ligeså godt gætte****!***



Epinion

Kan man stole på dem? Udover at de er weird… Spørgsmål 2) Lyver de?

Hvorfor hallucinerer sprogmodeller?

"…language models hallucinate because standard training and evaluation procedures reward guessing over acknowledging uncertainty."

Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI
Ofir Nachum OpenAI
Santosh S. Vempala† Georgia Tech
Edwin Zhang OpenAI

September 4, 2025

Abstract

Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

1 Introduction

Language models are known to produce overconfident, plausible falsehoods, which diminish their utility and trustworthiness. This error mode is known as "hallucination," though it differs fundamentally from the human perceptual experience. Despite significant progress, hallucinations continue to plague the field, and are still present in the latest models (OpenAI, 2025a). Consider the prompt:

What is Adam Tauman Kalai's birthday? If you know, just respond with DD-MM.

Kalai et at. (OpenAI) (2025)

Kan man stole på dem? Udover at de er weird… Spørgsmål 2) Lyver de?

Hvorfor hallucinerer sprogmodeller?

Forestil jer en multiple-choice test.
Her får du et point for er rigtigt svar
og ingen point for et forkert svart.
Hvad ville du gøre?

Hvilken kommune har det største landbrugsareal i Danmark?

[_] Herning
[_] Thisted
[_] Lolland
[_] Tønder
[_] Ringkøbing-Skjern

Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI
Ofir Nachum OpenAI
Santosh S. Vempala† Georgia Tech
Edwin Zhang OpenAI

September 4, 2025

Abstract

Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

1 Introduction

Language models are known to produce overconfident, plausible falsehoods, which diminish their utility and trustworthiness. This error mode is known as "hallucination," though it differs fundamentally from the human perceptual experience. Despite significant progress, hallucinations continue to plague the field, and are still present in the latest models (OpenAI, 2025a). Consider the prompt:

What is Adam Tauman Kalai's birthday? If you know, just respond with DD-MM.

Kalai et at. (OpenAI) (2025)

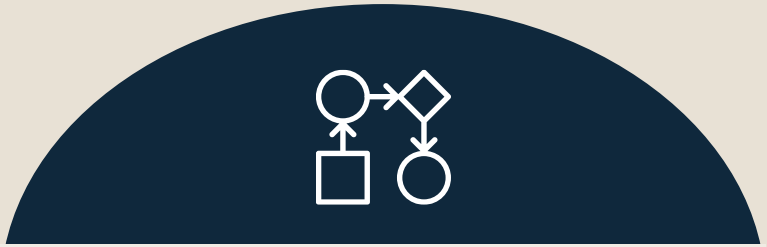
Kan man stole på dem? Så hvad kan vi gøre: Tools, procedurer og metrics!

Tools

Tools er relevante værktøjer som er lave til at udføre specifikke opgaver.

Eks. Lommeregner

Anvend relevante tools – enten eksternt eller indbygget i LLM'en gennem API ved opgaver, som kræver specifikke skills, som LLMer mangler.



Procedurer

Fastlagte processer for, hvordan vi arbejder med sprogmodeller.

Eks. Prompting guides

Epinion har egne guides der baserer sig på nyeste viden om sprogmodellers opbygning.

"Svar kun, hvis du er 75% sikker"

Metrics

Fastlagte mål for, hvornår en sprogmodel performer godt - løser ikke problem.

Eks. Hallucinations-test

Epinion arbejder med at udvikle specifikke mål der tester konkrete AI-løsninger (dvs. ikke blot de generelle metrics-mål for modellerne).

Epinion

Dialog & spørgsmål

Epinion

Dagens program

13:00-13:15	**Velkommen og præsentationsrunde**
13:15-14:15	**AI i 2025:** Hvor er vi nu, hvad er det nyeste – og hvordan oversætter vi hype til handling i praksis? v. Allan TH Knudsen
14:15-14:45	*Pause og netværk*
14:45-15:00	**Fra teknologi til principper:** Kort oplæg om tech-adoption og governance-principper: Hvordan kan organisationer sætte rammer for ansvarligt brug? v. Allan TH Knudsen
15:00-15:30	**Case: Red Barnet sætter AI på dagsordenen:** Erfaringer med at udvikle en AI-politik samt de muligheder, begrænsninger og dilemmaer politikken indebærer
15:30-16:00	**Fælles drøftelse**
16:00-?	*Fyraftensøl/-sodavand eller på gensyn*

2. **Handling**: Hvorfor er det så svært at komme i gang?

Brug af AI som ny teknologi i en organisation:
Return for the AI-titis!

Epinion

”Hvordan kommer organisationer i gang med AI?”

”Hvordan kommer vi i gang med AI?”

Hvad skal man lige svare?

En rigtig ”AI-chef” – måske med ”AI-titis?”

”Hvordan kommer organisationer i gang med AI?”

”AI-chef” skal tænke sig om…

Fra ”hvordan” til ”hvorfor”!

Lidt ligesom hvis nogen siger at ”de vil i gang med statistik”.

Så vil vi også først spørge: Hvorfor?

”Hvordan kommer organisationer i gang med AI?”

”AI-chef” fandt et svar…

Fra ”hvordan” til ”hvorfor”!

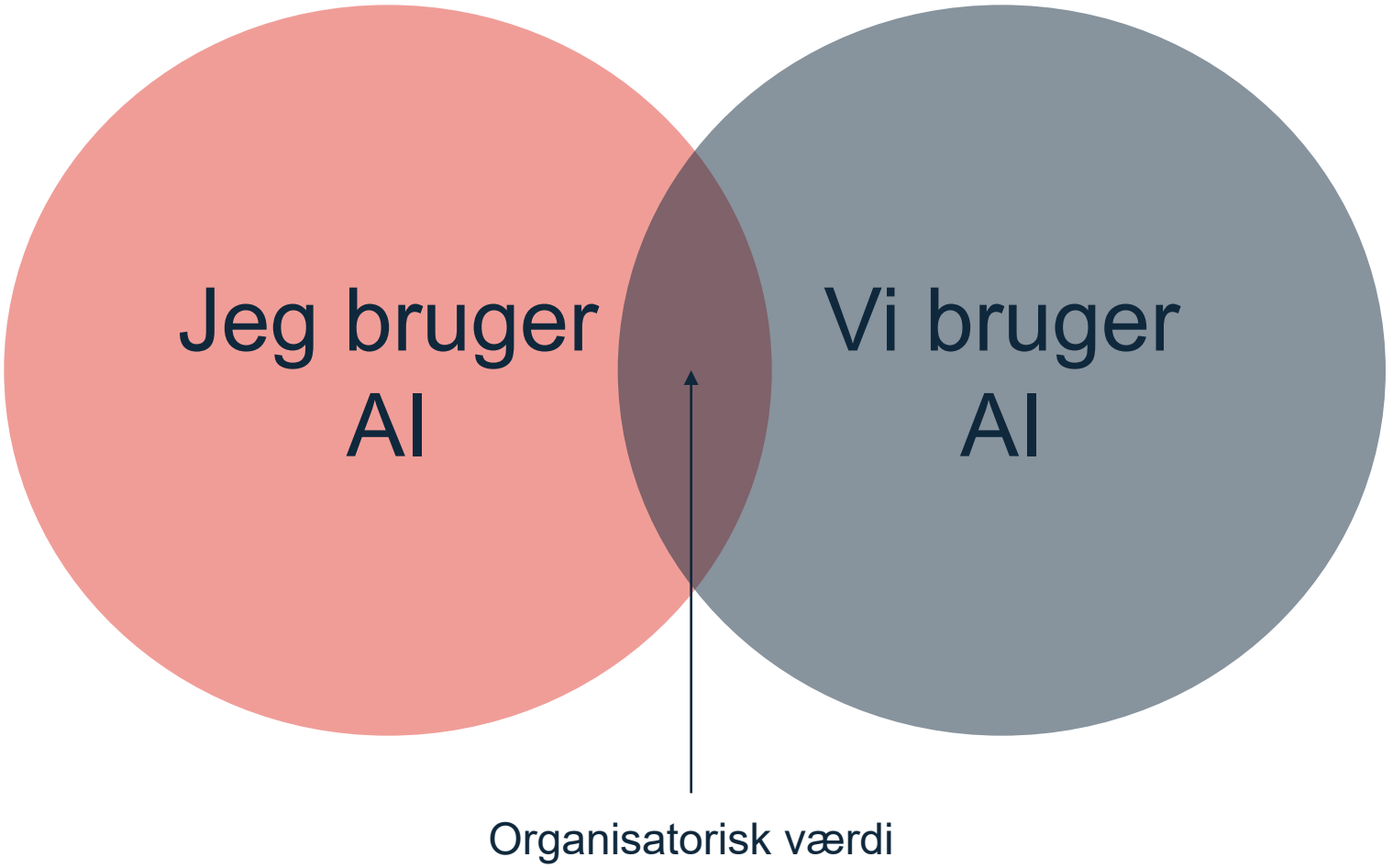
Jeg vil ikke gå lang ned af ”hvorfor”-stien – det er ikke formålet.

Effektivisering og innovation er nok de to overordnede svar.

Og så tilbage til ”hvordan”?

Hvorfor er det så svært at komme i gang?

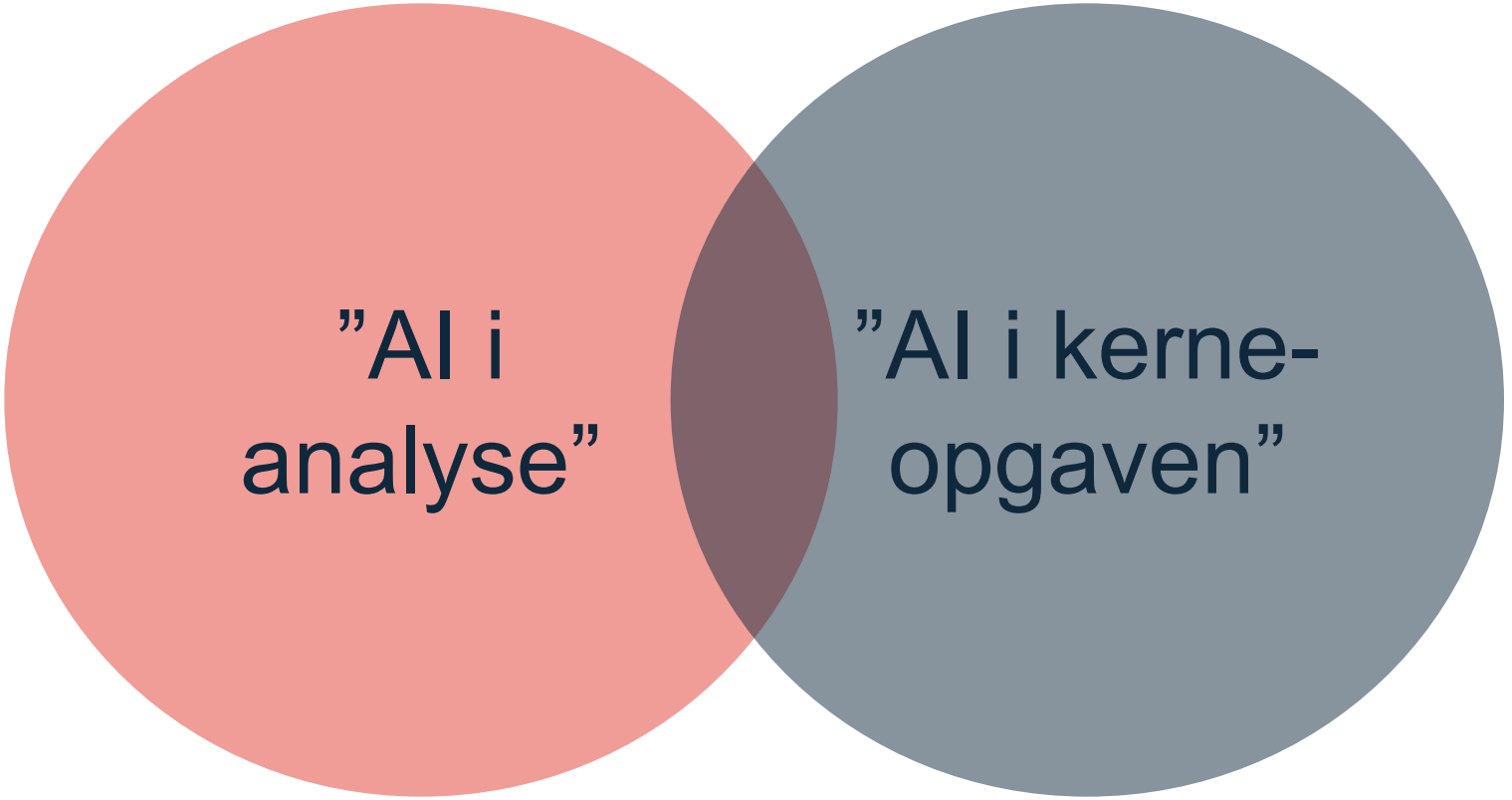
Skellet mellem ”individuel brug” og ”organisatorisk brug”



Epinion

Hvorfor er det så svært at komme i gang?

”AI i en analyse” vs. ”AI i en organisation”

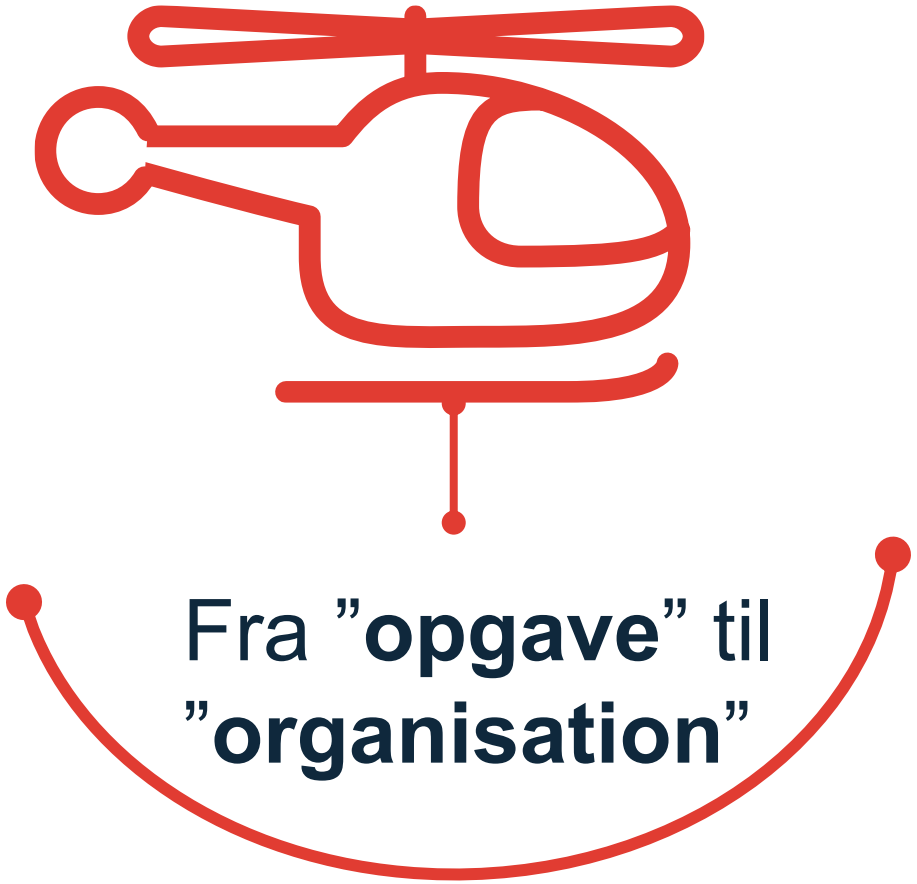


Epinion

Hvorfor er det så svært at komme i gang?

Fra ”opgave” til ”organisation”

”AI i organisation”



Epinion

Thore Dankert
Head of Digital & Tech
Novo Holdings

*“**New tech adoption er** 10 pct. tech og **90 pct. mennesker**”*

Epinion

Prof. Peter Drucker

”The father of management”

”Culture eats strategy for breakfast”

Hvorfor er det så svært at komme i gang?

Og så er vi lidt tilbage til, at der desværre ikke er en quick fix…

“To get organizational gains requires R&D into AI use and you are largely going to have to do the R&D yourself. I want to repeat that: *you are largely going to have to do the R&D yourself*”

Ethan Mollick

…men indgangsbarrieren er heller ikke så høj

Der er to ikke-gensidigt-udelukkende strategier:

1. **“The Crowd”**
2. **“The Lab”**

Ethan Mollick

Den ene strategi baserer sig på ”decentral innovation”: Styrken i mængden!

1

“The Crowd”

Fremme **decentral innovation** ved at **aktivere medarbejdere** på alle niveauer i organisationen til at **eksperimentere** med AI-løsninger

Den anden strategi baserer sig på ”central innovation”: Test det, der ikke virker..

“The Lab” 2

Samler ”AI-interesserede” til at **udvikle og teste prototyper**, der **demonstrerer AI’s potentiale og begrænsninger**. Formålet er at skabe **praktiske løsninger og benchmarks.**

Hvordan ser ”the crowd” og ”the lab” ud i jeres organisation?

1 “The Crowd”

“The Lab” 2

Teknologi skal ikke altid gå først, men det skal altid gå med.

”Technology first”

Falsk dikotomi!

”Problem first”

Ledelsen skal gå med

Checkliste for at komme i gang med AI i din organisation:

- ☑ Klar politik om brugen af AI
- ☑ Træning i brugen af / viden om AI
- ☑ Nem adgang til at bruge AI
- ☑ Anerkendende kultur i brugen af AI
- ☑ Vidensdeling på tværs af teams/funktioner
- ☑ Tydeliggør værdien i brugseksempler

…og AI skal stadig ikke bruges til alt

Case: Red barnet sætter AI på dagsordenen

v. Bolette Willemann Jensen

Senior Consultant, Red Barnet

Epinion AnalytiCS

UDVIKLING AF EN AI POLITIK I EN NGO

AI i analysearbejdet 2.0 – fra hype til handling

AnalytiCS, 2. oktober 2025

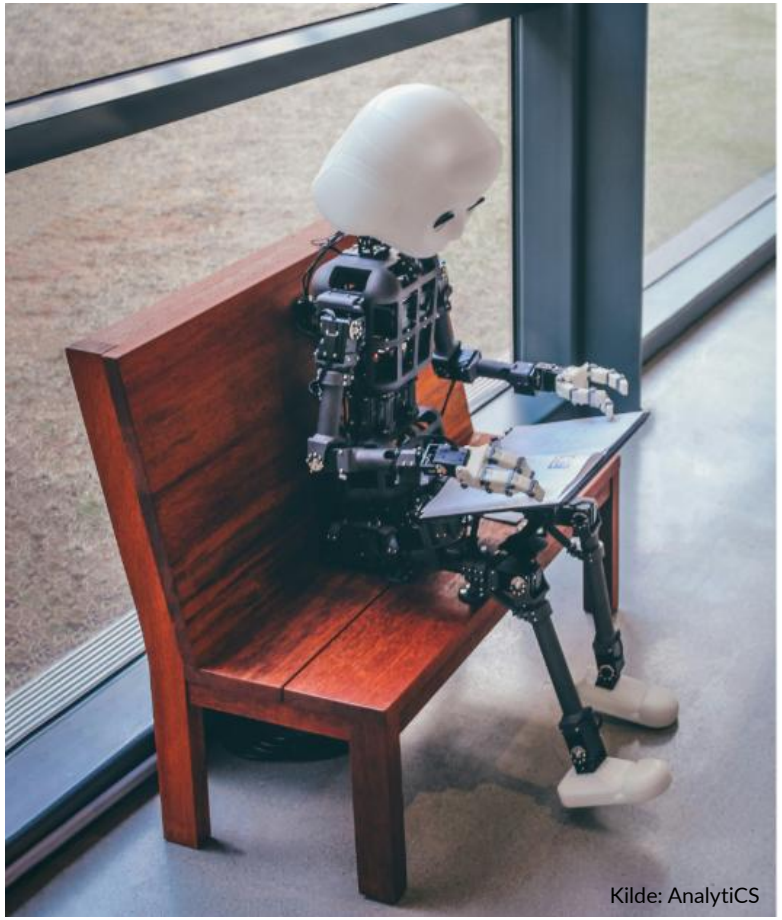
Red Barnet

Agenda

- AI-politikken – og egen GPT
- Muligheder
- Begrænsninger
- Dilemmaer

Red Barnet

Hvad I kan tage med jer kl. 15.30


Kilde: AnalytiCS

- Iterativ proces
- Forældet inden udrulning
- Et spejlbillede af organisationen

AI POLITIK – OG EGEN GPT

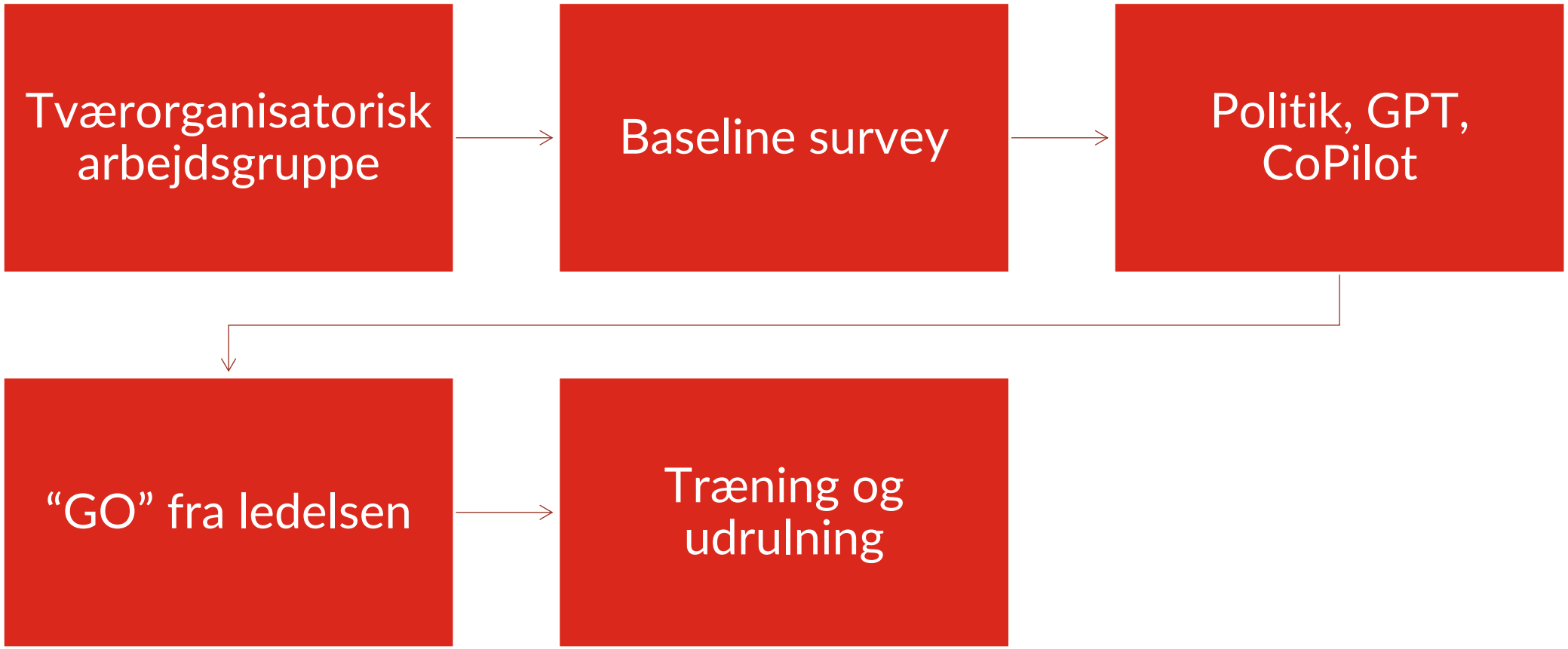
Red Barnet

Overblik over indholdet

Indhold

Red Barnet

Forløb apr. ‘24 - nu



Red Barnet

Ansvar og roller

Ledelsen	Mission, værdier og mål Godkender politik Sikrer ressourcer
AI arbejdsgruppen	Tværorganisatorisk styrket brug Træning
IT-afdelingen	Implementering Vedligehold Sikkerhed

Red Barnet

MULIGHEDER

Muligheder

01

Transparens omkring ansvarlighed

02

Effektivisering

03

Ensartet praksis

04

Leve op til lovgivning

05

Understøtter mission og værdier

Red Barnet

Eksempler på brug

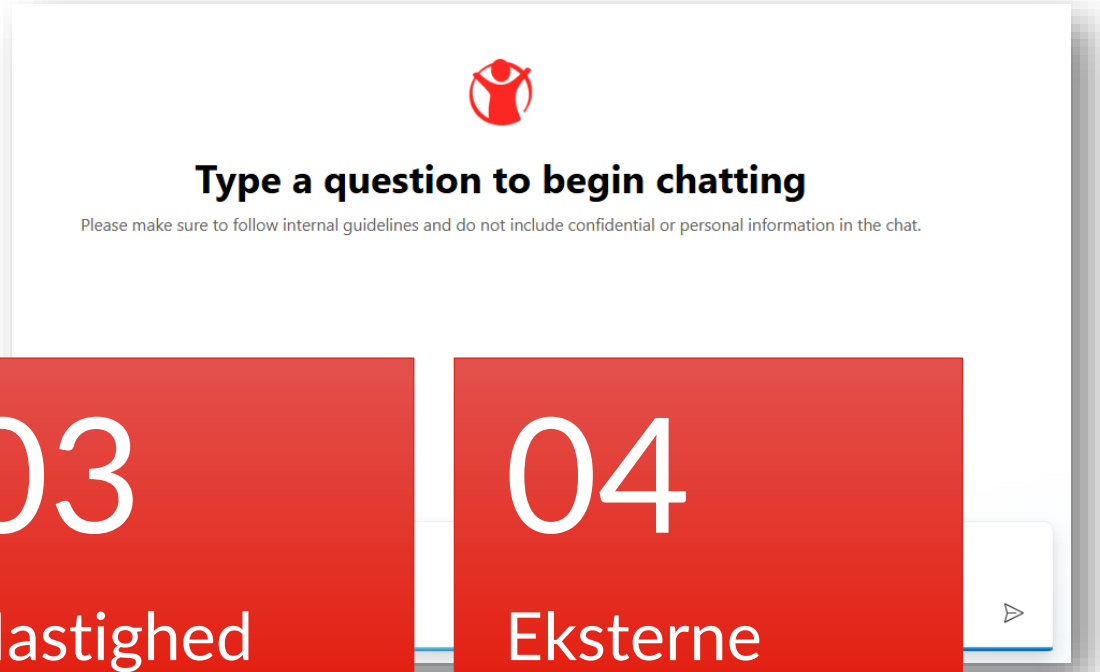
- Politik en del af udbudsskabelon
- Survey om medarbejderes brug af genAI
- Formuleringer af spørgsmål til specifikke målgrupper
- Brug af RB-GPT og CoPilot til opsamling på IT-leverandørmøder
- Forandringsteori (bud på teori, samling af pointer fra post-its)

"Benyttes AI-værktøjer i opgaveløsningen skal anvendelsen drøftes med og godkendes af Red Barnet, ligesom det skal fremgå af Databehandleraftalen. Leverandører er indbefattet af Red Barnets AI-politik".

(Udbudsskabelon)

BEGRÆNSNINGER

Begrænsninger



01

Mangler funktionalitet

02

Økonomi

03

Hastighed

04

Eksterne også omfattet

DILEMMAER

Dilemmaer ift. politik og brug/praksis

Viden fra eksterne	Gråzoner	Innovation vs. kontrol
Kompetencer	Klima	Tværorganisatorisk

Red Barnet

Hvad så nu?

En RB analyse-agent?

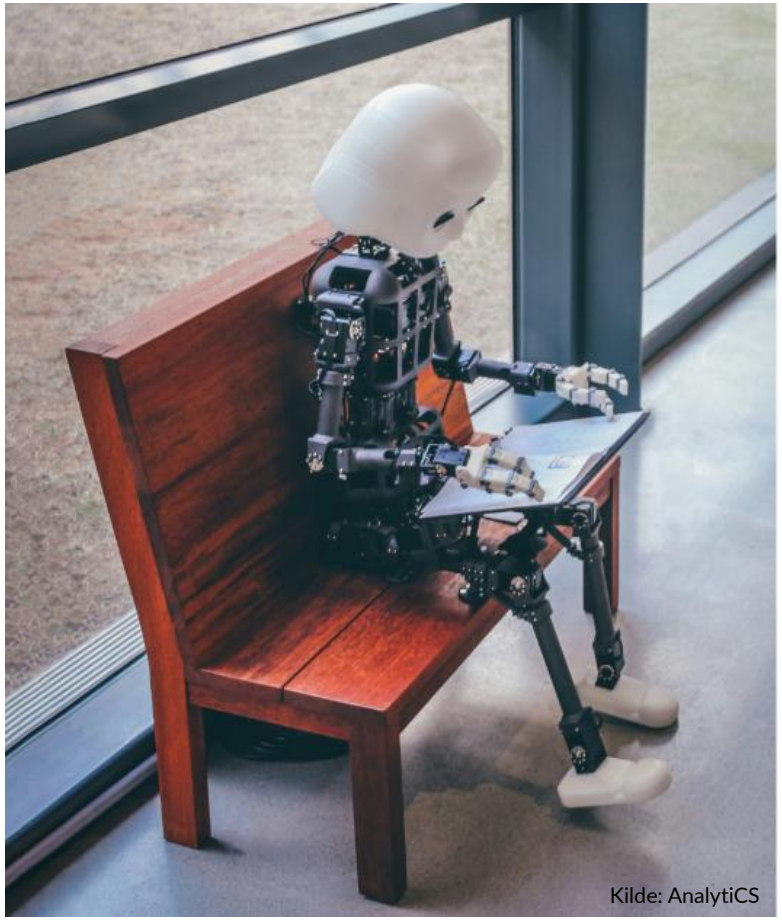
Køb af dataanalyser eksternt?

Køb af licens til værktøjer?

Gå sammen med andre?

Red Barnet

Hvad I kan tage med jer kl. 15.30


Kilde: AnalytiCS

- Iterativ proces
- Forældet inden udrulning
- Et spejlbillede af organisationen

TAK

Bolette Willemann Jensen
Analysekonsulent | Analyse

boje@redbarnet.dk | Direkte: +45 3090 1572 | Mobil: +45 3090 1572
Red Barnet | Gammel Kongevej 60, 1850 Frederiksberg C | Tlf +45 3536 5555 | redbarnet.dk |

Red Barnet

SPØRGSMÅL

Hvad ser I af muligheder, udfordringer og dilemmaer ved AI-politikker?

Red Barnet

Fælles drøftelse:
Hvordan kommer vi videre herfra?

Epinion
AnalytiCS

Tak for i dag

Fyraftensøl/-sodavand

eller på gensyn

Epinion

Epinion Copenhagen

Ryesgade 3F
2200 Copenhagen N
Denmark
T: +45 87 30 95 00
E: copenhagen@epinionglobal.com
www.epinionglobal.com

Epinion

Fremtid: Koncepterne kommer!

Epinion

Er vi fanget i de store sprogmodellers verden? Måske ikke….

LCMs
Large Concept Models

”Store konceptmodeller”

Large Concept Models:

Language Modeling in a Sentence Representation Space

LCM team, **Loïc Barrault***, **Paul-Ambroise Duquenne***, **Maha Elbayad***, **Artyom Kozhevnikov***, **Belen Alastruey**†, **Pierre Andrews**†, **Mariano Coria**†, **Guillaume Couairon**+†, **Marta R. Costa-jussà**†, **David Dale**†, **Hady Elsahar**†, **Kevin Heffernan**†, **João Maria Janeiro**†, **Tuan Tran**†, **Christophe Ropers**†, **Eduardo Sánchez**†, **Robin San Roman**†, **Alexandre Mourachko**‡, **Safiyyah Saleem**‡, **Holger Schwenk**‡

FAIR at Meta

*Core contributors, alphabetical order, †Contributors to data preparation, LCM extensions and evaluation, alphabetical order, ‡Research and project management, alphabetical order, +Initial work while at FAIR at Meta, new affiliation: INRIA, France

December, 2024

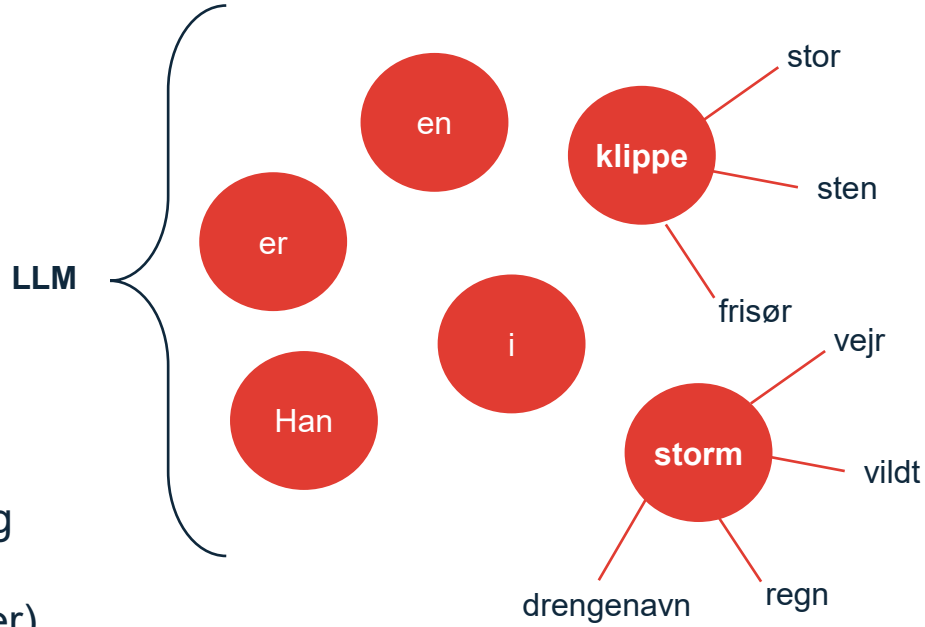
Yann LeCun, Meta

Epinion

Kan store konceptmodeller noget – er det ”fremtiden”?

Udfordringer ved LLMer

- Opererer på tokenniveau vs. menneskers ”hierakiske abstraktionsniveau”
- De kræver store mængder data og computerkraft – det er ikke læringseffektivt (som mennesker er)



For mennesker er **ord** ikke så vigtige – **mening bag er det afgørende!**

Epinion

Intro: Hvad er AI? I dag…

Kan man stole på dem? Altså på, at de ikke stjæler din data? Tænk på det som et vindue… Der findes dog work-around og løsninger

AI-genereret billede

Epinion

Kan man stole på dem? Altså på, at de ikke stjæler din data?

> ”Når noget er gratis, er du produktet.

Der er **stor værdi i et ”prompt”** og i din generelle interaktion med sprogmodeller. Modellerne trænes på interaktioner. Ikke ”live”, men det samles ind og sikre bedre fremtidige modeller.

Din interaktion indeholder både ”**naturlig tekst**” (som er ny!) og den indeholder **feedback på output**.

Gratis / billige modeller bruger din data.

Husk dog….

Ikke som Google…

Kan man stole på dem? Hvilke muligheder er der for at arbejde med sikre modeller?

Lukkede modeller

Store sprogmodeller, hvor adgangen købes eksternt, og modellen integreres i en GDPR-kompatibel infrastruktur

Eks. Azure-platform, CoPilot

Epinion udruller OpenAIs ChatGPT Enterprise i gennem Azure gennem SikkerAI-platform, der udbydes af syv.ai. Data logges ikke, bruges ikke til trærning og kun Epinion har adgang til det.

SenseAI

Lokale modeller

Sprogmodeller, der er udviklet internt eller baseret på open source-modeller eventuelt finjusteres yderligere. Den afgørende faktor er, at data forbliver på lokale servere

Eks. Metas Llama 2, Mistral 7B, DanskGPT

En lokal model hostes direkte på egne servere og sikrer derfor, at data ikke forlader organisationen. Epinion har mulighed for eks. at bruge DanskGPT hvis behov, men ikke inkluderet i nuværende løsning.

Epinion

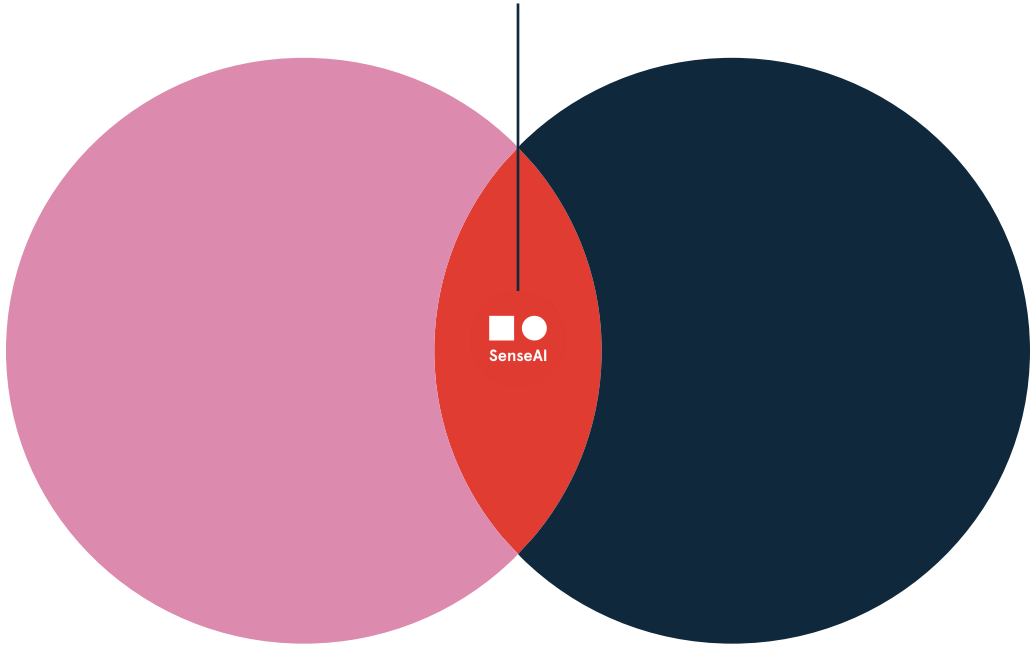
Kan man stole på dem? Hvad skal man så vælge som organisation?

Strategi	Fordele	Ulemper
1. Lukkede modeller via GDPR-sikre platforme	• Hurtig implementering og nem adgang • Høj stabilitet og vedligehold • GDPR-overholdelse via certificerede cloud-løsninger. • Løbende opdateringer og forbedringer.	• Data sendes til eksterne cloud-tjenester (selvom GDPR-sikker) • Løbende abonnements- eller licensomkostninger. • Begrænset mulighed for tilpasning. • Risiko for vendor lock-in.
2. Lokale modeller (egenudviklet eller Open Source på lokale servere)	• Fuld kontrol over data og processer. • Ingen eksterne aktører • Stor fleksibilitet (bedre muligheder for tilpasning ved selvudvikling) • Ingen afhængighed af en ekstern leverandør.	• Store investeringer i infrastruktur • Løbende teknisk ekspertise in-house. • Vedligeholdelse og opdatering af modellen skal håndteres internt. • Kan ikke matche ydeevnen fra store kommercielle modeller.

SenseAI er hverken en ”åben model” eller en ”lokal model” – det er en åben model på en sikker platform

Åben model på sikker platform
GDPR-sikker og state-of-the-art

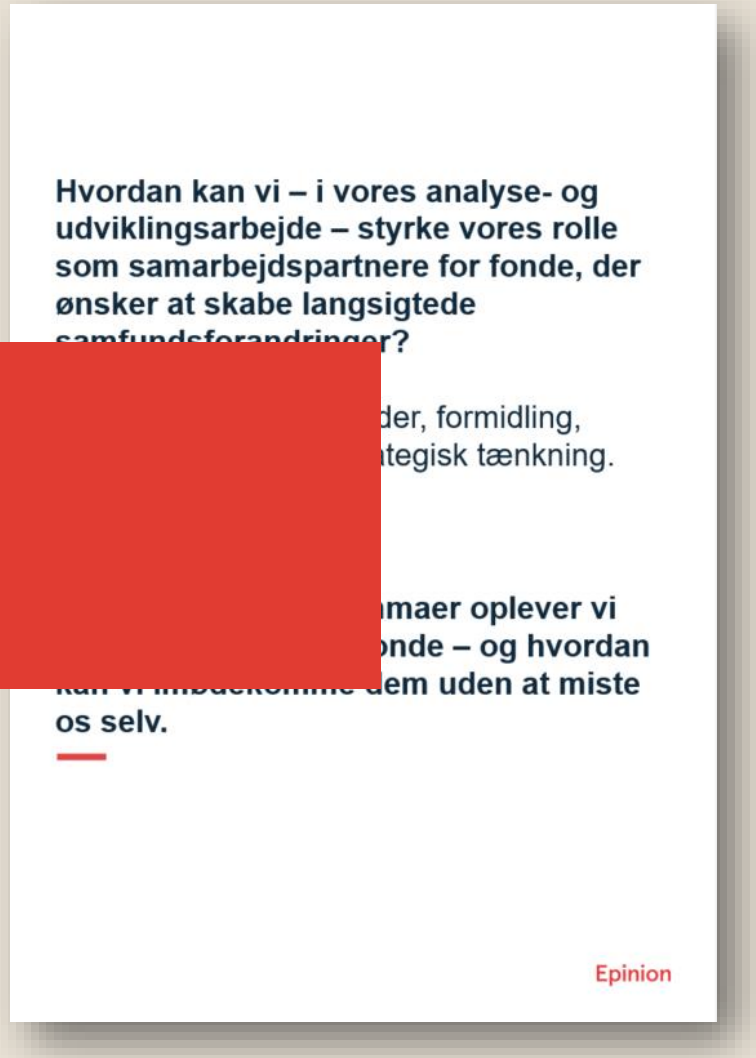
Lokale modeller
GDPR-sikre, men typisk små og dermed ikke state-of-the-art



Åbne modeller
Ikke GDPR-sikre, men typisk store og state-of-the-art

Fælles drøftelse

Opdateres



Epinion
AnalytiCS

Hvis du skulle tage én ting med hjem fra i dag, hvad ville det være?

Fyraftensøl/-sodavand eller på gensyn

Epinion
AnalytiCS