Troels Jensen

# SPRÅKMODELLER FOR DUMMIES

En introduksjon til teknologien bak ChatGPT

Oversatt av

Kenneth Bareksten

# SPRÅKMODELLER

INNHOLDSFORTEGNELSE

# INTRODUKSJON

Formålet med dette heftet er å gi deg en grunnleggende forståelse av hva språkmodeller er, og hvordan de virker. Det er skrevet for deg som har lite forkunnskaper om kunstig intelligens og maskinlæring, men som gjerne vil forstå litt mer om hva som skjer bak kulissene når vi bruker verktøy som ChatGPT.

Heftet kan leses på et par korte timer og gir deg forhåpentligvis et språk for og en forståelse av hvordan denne teknologien fungerer. Det kan du bruke som grunnlag for å kunne stille mer kvalifiserte spørsmål til teknologien og dem som utvikler og bruker den.

Det er ikke nødvendig å lese heftet fra perm til perm. Du kan nøye deg med de første kapitlene for å få en grunnleggende forståelse, eller gå mer i dybden ved å lese hele. Heftet inneholder lenker til artikler og videoer underveis, slik at du kan lære mer hvis du får lyst.

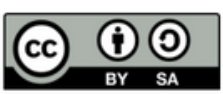
Hvis du ønsker en enda dypere forståelse, finner du bakerst en ordforklaring, kildeliste og en samling ressurser for videre lesning. Det er selvsagt frivillig. Det viktigste er at du får den innsikten du trenger.

Heftet gir deg:

- En forståelse av hva språkmodeller er
- En enkel forklaring på hvordan de trenes
- Innsikt i styrker og svakheter ved språkmodeller
- Kjennskap til typiske bruksområder
- En introduksjon til tekniske begreper og videre ressurser



Oslo, april 2025



# KAPITTEL 1

## Hva er en språkmodell?

Språkmodeller er den tekniske betegnelsen for systemer som ChatGPT, som man også ofte hører omtalt som chatboter. Kort fortalt er språkmodeller en familie av algoritmer som kan generere menneskelignende tekst basert på en brukers forespørsel. En forespørsel, også kalt en prompt, er en tekstbit man skriver til språkmodellen. Språkmodellen reagerer på en prompt ved å produsere tekst som den beregner vil passe til forespørselen.

Moderne språkmodeller fortsetter rett og slett sekvenser av ord på en måte som får det til å se ut som noe et menneske kunne ha skrevet (Radford et al. 2019). Det er språkmodellers funksjon i et nøtteskall. Det klassiske eksempelet er å stille et faktaspørsmål med en prompt, som språkmodellen så besvarer som en videreføring av ordsekvensen. Et annet typisk eksempel er å skrive innledningen til en artikkel, som en språkmodell så fortsetter eller fullfører.

Jeg kommer til å bruke ordet språkmodell fremfor chatbot gjennom hele heftet. Det er dels fordi det er den akademiske betegnelsen, og dels fordi ordet chatbot har noen konnotasjoner som etter min mening gjør det misvisende å bruke om systemer som ChatGPT. Ordet chatbot har tidligere vist til veldig enkle spørsmål-svar-systemer – for eksempel de små chatvinduene man noen ganger finner på flyselskap eller nettbutikker, som skal hjelpe kundene med spørsmål eller problemer. Slike systemer er basert på forhåndsprogrammerte regler som:

**IF** brukerinput **CONTAINS** "kansellert, refusjon eller klage" **THEN OUTPUT**= "Videresender til kundeservice. Du er nr. 74 i køen."

Dette er et eksempel på en god, gammeldags algoritme som er bygget opp rundt enkle, håndskrevne regler, og som bare kan reagere på noen få typer forespørsler med et tilsvarende lite sett med svar. Språkmodeller fungerer etter helt andre prinsipper og har en langt større handlingskapasitet enn gamle chatboter. De kan langt mer enn bare å chatte. Språkmodeller er nemlig eksempler på kunstig intelligens – nærmere bestemt tilhører de den grenen av informatikk vi kaller deep learning, som er den ledende retningen innen maskinlæring. Maskinlæring representerer en fundamentalt annerledes tilnærming til problemløsning enn chatboten i eksemplet over. Maskinlæring benytter ikke forhåndsskrevne regler, men analyserer språklige mønstre i datamaterialet den er trent på for å lære seg å løse oppgaver selv.

### Algoritmer

En algoritme er en trinnvis oppskrift eller et sett med regler for å løse et problem eller utføre en oppgave. Den kan være skrevet som kode i et dataprogram eller som en enkel liste med instruksjoner.

### Kunstig intelligens

Kunstig intelligens (KI) er dataprogrammer som kan utføre oppgaver som vanligvis krever menneskelig intelligens – som å lære, forstå språk, gjenkjenne mønstre og ta beslutninger.

### Maskinlæring

Maskinlæring er en metode innen kunstig intelligens der datamaskiner lærer av erfaring, altså data, i stedet for å bli programmert med faste regler. Systemet forbedrer seg selv jo mer det "trener".

## Begrepstaksonomi

KI-feltet er en begrepsjungel. For å forklare hvordan språkmodeller passer inn i det større KI-landskapet, er det nyttig å ha en begrepstaksonomi:


- **Informatikk** (og datavitenskap) handler ikke bare om å programmere datamaskiner, men omfatter bredt forstått studiet av data og algoritmer.
- **Kunstig intelligens** (KI) er det overordnede fagfeltet som har som mål å utvikle systemer som kan utføre oppgaver som vanligvis krever menneskelig intelligens.
- **Maskinlæring** er hovedgrenen innen kunstig intelligens, og består av en stor familie statistiske metoder som løser komplekse problemer automatisk ved å lære av data. Maskinlæring gjør det mulig å løse oppgaver uten å måtte formalisere og programmere hele løsningen på forhånd.
- **Deep learning** er det dominerende paradigmet innen maskinlæring i dag. Deep learning-systemer er basert på nevrale nettverk, som etterligner måten den menneskelige hjernen bearbeider informasjon. Mer om dette senere. Deep learning har vist seg å være en ekstremt effektiv og allsidig tilnærming som kan løse svært mange komplekse oppgaver (Schmidhuber, 2015).
- **Språkmodeller** er en kategori innen deep learning-modeller som er spesialisert på å jobbe med menneskespråk. Generative språkmodeller, som ChatGPT, er laget for å skape tekst, mens mange tidlige modeller – som BERT – er prediktive og brukes til andre formål, som å klassifisere om en tekst er positiv eller negativ. Les mer om forskjellen her (husk lenke)

## Smal versus litt mer generell KI

Språkmodeller begynte å vinne popularitet i 2020, da selskapet OpenAI lanserte verdens hittil største språkmodell, GPT-3. GPT-3 imponerte med en usedvanlig livaktig forståelse av språk og en evne til å produsere velformulert tekst innenfor en lang rekke områder — selv områder den aldri hadde blitt trent opp til å operere innenfor.

### DEEP LEARNING

Deep learning bruker nevrale nettverk med mange lag som etterligner hvordan hjernen behandler informasjon. Lagene mottar input, bearbeider det i økende grad av abstraksjon og kompleksitet, og gir output. Denne strukturen gjør modellene svært gode til å kjenne igjen komplekse mønstre, for eksempel i bilder og tale.

### GENERATIV KI

Generativ KI viser til kunstig intelligens som er i stand til å skape nytt, unikt innhold basert på mønstre lært fra data. Den brukes ofte til å generere menneske-lignende innhold, for eksempel å skrive tekster, komponere musikk eller lage visuelle kunstverk.

### PREDIKTIV KI

Prediktiv KI forsøker å forutsi ukjente datapunkter basert på kjente data, for eksempel aksjemarked, kundeadferd eller vær. Den er den mest brukte KI-typen og finnes i mange moderne produkter.

GPT-3 ble nemlig trent med det enkle målet å generere menneskelignende tekst, ett ord av gangen. Men underveis i treningen lærte GPT-3 også å utføre en lang rekke oppgaver vi tidligere trodde bare mennesker kunne klare. Den lærte for eksempel å oppsummere artikler, skrive datakode, oversette mellom språk, forfatte poesi, formulere logiske argumenter – og mye, mye mer. Det var rett og slett banebrytende.

For å forstå hvor stor utvikling dette var, må man vite at nesten alle tidligere typer maskinlæringssystemer har vært svært smale i sitt repertoar – de kan bare utføre akkurat den oppgaven de er trent til, og klarer ikke å generalisere til nye domener. Denne typen systemer kalles noen ganger smal KI. Et eksempel er Spotifys anbefalingsalgoritme: Det er en maskinlæringsalgoritme som er svært god til å foreslå innhold, men som ikke kan gjøre noe som helst annet. Hvis man vil at Spotifys algoritme skal oppføre seg annerledes, må den trenes opp helt på nytt – med andre data.


1) Stanley Kubricks film 2001: A Space Odyssey (1968) basert på boka av Arthur C. Clarke.

Moderne språkmodeller er mer generelle. De kan utføre mange ulike oppgaver uten å være spesifikt trent for hver enkelt av dem. Det er nettopp dette som gjør dem så oppsiktsvekkende – denne evnen til å generalisere begynner å ligne noe vi vanligvis forbinder med menneskelig intelligens. Den hellige gral innen KI-forskning er å skape generell kunstig intelligens, såkalt AGI (Artificial General Intelligence), som kan utføre alle oppgaver mennesker kan – og enda bedre (Gubrud, 1997). Et fiktivt eksempel på AGI er den allvitende HAL 9000 fra filmen 2001: En romodyssé. Foreløpig er språkmodeller det nærmeste vi har kommet AGI, men vi er fortsatt langt unna egentlig generell intelligens. Mange eksperter mener at AGI ligger flere tiår fram i tid, mens andre tror vi allerede er i ferd med å nå det punktet.

## Språkmodeller blir mainstream

Selv om GPT-3 var en gamechanger, var det begrenset hvor mye den jevne befolkningen merket utviklingen. Det var derimot en helt annen sak da ChatGPT ble lansert 30. november 2022. ChatGPT er en videreutvikling av GPT-3, og det er to hovedgrunner til at det nettopp var denne modellen som tok verden med storm:

### SMAL KI

Smal, også kalt svak KI, er den typen bedrifter ofte viser til når de snakker om KI i produkter. Det gjelder som regel maskinlæringssystemer som løser én enkel oppgave, uten evne til å generalisere til nye områder. Vi har kommet langt med denne typen KI, og den har gjort det mulig å automatisere mange oppgaver vi før tenkte krevde mennesker. Eksempler er talegjenkjenning, anbefalingssystemer og Google Translate.

### GENERELL KI

Kunstig generell intelligens, også kalt sterk KI eller AGI, viser til hypotetiske systemer som kan utføre mange ulike oppgaver like godt eller bedre enn mennesker. Slike systemer kan generalisere til nye områder uten å være spesifikt trent for dem. Dette regnes som den hellige gral innen KI-forskning, men vi har ennå ikke klart å utvikle kunstig generell KI.

**1. Bratt læringskurve**
Tidligere språkmodeller var vanskelige å bruke. Det krevde forkunnskaper å få dem til å oppføre seg slik man ønsket, og man måtte bruke lang tid på å formulere spesifikke forespørsler for å få gode svar. ChatGPT er derimot enkel og intuitiv å bruke. Svarene er stort sett relevante, og man kan samhandle med den på en måte som føles naturlig. Hvis ChatGPT gir et utilfredsstillende svar, kan du bare be den formulere seg annerledes, eller legge til akademiske referanser. Dette var langt vanskeligere med tidligere modeller.

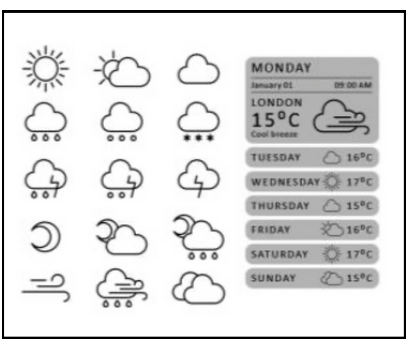
**2. Tilgjengelighet**
OpenAI gjorde ChatGPT gratis tilgjengelig for alle via en lett tilgjengelig nettside. Dette gjorde de delvis for å samle inn data om hvordan folk samhandler med systemet – slike data kan nemlig bidra til å trene enda bedre språkmodeller i fremtiden.

# KAPITTEL 2

## På innsiden av en språkmodell

Språkmodeller bygger på en rekke kompliserte matematiske triks og tung statistisk teori, som det ville kreve en lang og teknisk forklaring å gjøre rede for. Dette kapittelet gir deg ikke en dyptgående forståelse av språkmodellers indre mekanismer. Målet er i stedet å avmystifisere fenomenet ved å bruke noen enkle forståelsesrammer for å forklare de underliggende prinsippene.

Ordet modell skal i denne sammenhengen forstås litt på samme måte som når værmeldingen snakker om meteorologiske modeller. Værmeldingen baserer i stor grad sine prognoser på en matematisk modell, konstruert ut fra systematiske korrelasjoner i enorme mengder værdata.




Hvis datapunkter om lavtrykk i øst ofte etterfølges av datapunkter om sterk vestavind, kan modellen bruke denne korrelasjonen til å forutsi kuling fra vest i kveld – dersom det ble observert lavtrykk i øst i morges. I virkeligheten er meteorologiske modeller langt mer kompliserte, men poenget er at statistiske modeller utnytter mønstre i store datamengder for å kunne gjøre prediksjoner basert på ufullstendig informasjon.

**En modell i denne konteksten er altså en abstrakt, statistisk tilnærming til et virkelig fenomen, som kan brukes til prediksjon (forutsigelse).**

Mønstre i værdata

Forutsigelse av været i morgen

Mønstre i språk

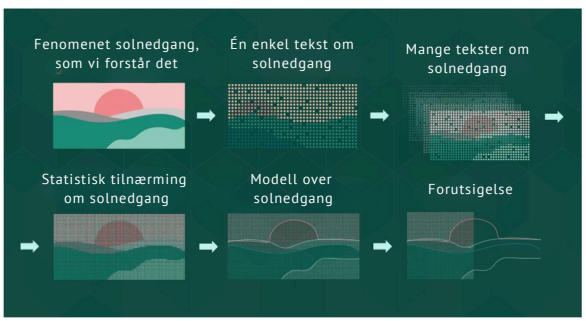
Språkmodell

Forutsigelse av det neste ordet


På samme måte kan vi si at en språkmodell er en statistisk tilnærming til et virkelig fenomen – nemlig menneskespråket. Språkmodeller kan etterligne menneskelig språk fordi de har lært seg mønstrene i hvordan vi vanligvis bruker ord, ved å lese og analysere enorme mengder tekst. De har lært hvilke ord som ofte kommer sammen i ulike sammenhenger, og bruker den kunnskapen til å lage en detaljert forestilling om hvordan menneskespråk vanligvis ser ut – hvordan ord påvirker hverandre og skaper mening, og hvordan samme ord kan bety ulike ting avhengig av hvordan de settes sammen.

**Språkmodeller er altså en statistisk tilnærming til fenomenet språk, og kan også brukes til å forutsi hva som sannsynligvis kommer videre i en tekst.**

I figuren nedenfor brukes det språklige begrepet "solnedgang" som et forenklet eksempel for å illustrere hvordan språkmodeller fungerer. En solnedgang er egentlig et ganske komplekst begrep, hvis man tenker over det. Ordet har mange assosiasjoner knyttet til seg – farger, følelser, astronomi, fysikk, poesi, og mer. Alle disse assosiasjonene må vi på en eller annen måte gi til språkmodellen, slik at den kan skrive en menneskelignende tekst om solnedganger. Det skjer gjennom trening, der vi presenterer språkmodellen for en mengde tekst som handler om solnedganger.Én tekst kan for eksempel beskrive solnedganger som farge-rike, og da lærer modellen at det ofte er en statistisk sammenheng mellom ordene fargerik og solnedgang. En annen tekst beskriver kanskje sol-nedganger på en helt annen måte, som lys som brytes i atmosfæren.




Hvis vi samler inn et datasett med titusenvis av tekster om solnedganger med tilstrekkelig språklig variasjon, kan vi si at vi har skapt en dataspråklig tilnærming til begrepet solnedgang – en som rommer de viktigste assosiasjonene knyttet til fenomenet. Trener vi språkmodellen med dette datasettet, vil modellen etter hvert danne seg en ganske presis, språklig representasjon av hvordan en solnedgang kan beskrives. Vi kan si at vi har modellert begrepet solnedgang, slik at modellens forståelse av begrepet nærmer seg vår egen. Da kan vi bruke modellen til å gjøre språklige prediksjoner – altså forutsi hvilke sammenhenger ordet solnedgang kan inngå i. Vi trenger ikke lenger dataene for å gjøre slike forutsigelser, vi følger bare modellen, som nå fungerer som en slags erstatning for de opprinnelige datapunktene – slik det vises i den siste ruten ovenfor. Da får vi dette Eksempelet:

**Forespørsel**: "En solnedgang finner sted om?

**Forutsigelse:** "Kvelden"

## Tokens i stedet for ord

Språkmodeller arbeider faktisk ikke med hele ord eller enkeltbokstaver, men med noen andre språklige byggeklosser som kalles tokens. Et token kan være et helt ord, en del av et ord, en kombinasjon av flere ord, eller et enkelt tegn som et punktum eller komma.Språkmodeller bruker tokens i stedet for hele ord fordi det gjør språkhåndteringen mer fleksibel og effektiv. Tokens gjør det mulig for modellen å forstå og produsere språk med mange nyanser – for eksempel nye ord, slang eller kompliserte ordformer – uten å måtte ha et kjempestort ordforråd som inneholder alle mulige ord på forhånd.Ofte tilsvarer et token et helt ord, men andre ganger kan det være mer uventede deler. Ordet uforutsigelig kan for eksempel bli delt opp i "u", "forutsig", og "elig". Å bruke tokens i stedet for hele ord gir språkmodellen en dypere forståelse av hvordan språk er bygd opp og hvordan mening dannes.Denne metoden gjør også at språkmodeller kan forstå og skrive norsk, selv om de fleste norske ord ikke finnes direkte i modellens ordforråd – fordi ordene brytes ned i tokens som modellen kan sette sammen igjen til riktige ord og setninger.

I eksempelet under ser vi hvordan setningene blir brutt ned i tokens i en språkmodell. Hvert ord representerer en egen token – legg merke til hvordan engelske ord i stor grad deles inn i tokens som ligner på hele ord, mens norsk ord i større grad blir delt opp i mindre deler.

**Norsk:**

Sol-ned-gangen over de-snødekte fjell-ene tok pust-en fra oss, og den u-forut-sig-elige farge-palett-en spre-dte seg lang-som-t ut over himmel-en.

**Engelsk:**

The sunset above the snow-covered mountains took our breath away, and the un-predict-able color-palette slowly spread across the sky.

En språkmodell er dessuten en matematisk konstruksjon, som ikke kan lese og forstå ord på samme måte som du og jeg. Språkmodeller arbeider med tall, som det kan gjøres beregninger med. Derfor må tokens først oversettes til tall før en språkmodell kan arbeide med dem. Setningene over vil derfor se sånn ut for en språkmodell.

[30400, 2781, 274, 782, 6200, 7404, 4059, 5455, 40967, 7442, 11, 1794, 274, 34092, 2658, 7218, 60581, 1693, 62222, 2000, 6141, 13, 1611, 2781, 19541, 5683, 7404, 8136, 60534, 13376, 5544, 1435, 382, 30400, 527, 279, 25655, 8316, 315, 4221, 430, 4221, 4211, 1005, 4619, 315, 4339, 13, 2435, 527, 5505, 369, 1690, 8125, 627]

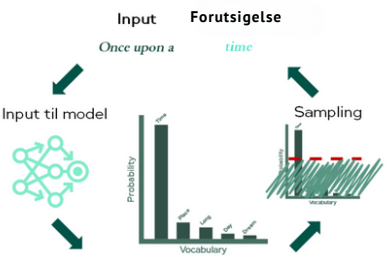
Denne videoen går i dybden på hvordan ord omdannes til tokens før de mates inn i en språkmodell. For ikke å gjøre resten av dokumentet mer forvirrende enn det allerede er, kommer jeg til å snakke om ord og forutsigelser av ord, i stedet for å kalle dem tokens.

## Autofullføring på steroider

Forutsigelse er kjernen i hvordan språkmodeller og maskinlæring fungerer. Akkurat som en værmodell brukes til å forutsi morgendagens vær, brukes språkmodeller til å forutsi det neste ordet i en setning. Når man gir en tekstbit til en språkmodell, beregner den hvilket ord som statistisk sett passer best som det neste, slik at setningen blir så menneskelignende som mulig ut fra en språklig vurdering.

GPT-2 hadde et ordforråd på 50 257 ord. Når GPT-2 får en tekstbit og skal forutsi det neste ordet, tilordner den sannsynligheter til alle ordene i sitt ordforråd. ChatGPT og nyere språkmodeller har et langt større ordforråd. Hvis jeg for eksempel gir GPT-2 forespørselen «Den lille Rødhette og…», vil den beregne 50 257 sannsynligheter.

Resultatet er en sannsynlighetsfordeling over hele ordforrådet. Deretter «sampler» språk-modellen fra denne fordelingen. Det betyr at den velger ett av de mest sannsynlige ordene, basert på et gitt utvalgskriterium, og dette ordet blir da modellens forutsigelse. Denne prosessen gjentas for hvert nytt ord, helt til modellen forutsier at det mest sannsynlige er at setningen bør avsluttes. Når man trener en språkmodell, prøver man å få disse sannsynlighetsfordelingene til å ligge så nært virkeligheten som mulig.



Output er en sannsynlighetsfordeling av hele ordforrådet

## Oppmerksomhet og kontekst

De beregnede sannsynlighetene blir også vektet for å ta hensyn til ordene i brukerens forespørsel, samt de ordene modellen allerede har skrevet. Dette gjør det mulig for språkmodeller å produsere meningsfull og sammenhengende tekst som er relevant for den aktuelle forespørselen. Språkmodeller vil tillegge visse ord større vekt enn andre i sine beregninger. I setningen "*Gutten falt og slo albuen sin*", vil ordet falt ha større innflytelse på sannsynlighetene for forutsigelsen enn de øvrige ordene. Derfor blir forutsigelsen «albuen», selv om ordet «rivalen» også kunne vært grammatisk korrekt. Denne funksjonaliteten kalles *attention*, og er en relativt ny milepæl innen KI-feltet. Den er i stor grad ansvarlig for den enorme utbredelsen av språkmodeller vi ser i dag (Vaswani et al., 2017). Attention gjør det mulig for språkmodeller å ta kontekst med i beregningen når de forutsier tekst.


Moderne språkmodeller kan ta hensyn til svært store mengder tekst innenfor det som kalles *kontekstvinduer*. Et kontekstvindu i en språkmodell viser til hvor mye tekst modellen «ser» på ett tidspunkt for å forstå og generere det neste ordet. Tenk på det som modellens synsfelt – den kan se et bestemt antall ord før og etter fokuspunktet for å fange betydningen og sammenhengen i teksten. Størrelsen på dette vinduet avgjør hvor mye kontekst modellen kan bruke i sine vurderinger, noe Google har forsket mye på.

## Språkmodeller er sannsynlighetsmaskiner

Språkmodeller kan i bunn og grunn forstås som veldige avanserte autofullføringssystemer – litt som de du kjenner fra SMS-appen på mobilen når du skriver en tekstmelding.

**En språkmodell fortsetter rett og slett ordrekker på en sannsynlig måte, samtidig som den hele tiden tar hensyn til sammenhengen i det som er skrevet så langt.**

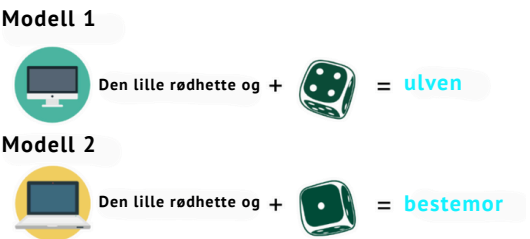
Ordet sannsynlighet dukker opp mange ganger i dette dokumentet, og det er det en god grunn til. Sannsynlighet er et av de mest sentrale begrepene for å forstå hvordan språkmodeller – og maskinlæring generelt – fungerer. Språkmodeller har egentlig ingen forståelse av hva en setning betyr, i hvert fall ikke slik vi vanligvis tenker på forståelse. Det de har, er en veldig presis følelse av hvilke ord som sannsynligvis hører sammen. Det betyr også at vi må ta det språkmodeller svarer med en klype salt – de vet ikke hva som er sant eller usant, de jobber bare med sannsynligheter. Det finnes ikke noe verdisett innebygd i språkmodeller, og så lenge en setning virker sannsynlig rent språklig, tar modellen ikke nødvendigvis hensyn til innholdet. Det er selvsagt ikke noe sikkerhetsproblem hvis man bruker modellen til å skrive dikt, men det kan bli det hvis man for eksempel spør hvor mye medisin man skal gi en pasient.

## Språkmodeller bygger på tilfeldigheter

Sannsynligheter er som kjent ikke noe eksakt. Det er for eksempel bare 0,08 % sannsynlighet for å få Yatzy på første kast, men likevel har nok mange opplevd at det faktisk skjer en gang iblant. På samme måte vil det alltid finnes en liten sjanse for at en forespørsel som «Den lille Rødhette og…» får en språkmodell til å foreslå ordet «roboten», selv om modellen mente at «ulven» hadde 99 % sannsynlighet. Språkmodeller er nemlig grunnleggende probabilistiske, svarene deres vil alltid være påvirket av en viss grad av tilfeldighet. Det er nettopp dette som gjør at de kan skrive varierte og interessante tekster. Hvis de alltid valgte det mest sannsynlige neste ordet, ville resultatene blitt så repeterende og forutsigbare at de ikke kunne brukes til stort.

Samtidig er det nettopp dette ved språkmodeller som skaper den største utfordringen for plagiatkontroll. Den samme forespørselen vil nemlig ikke gi nøyaktig samme svar på to forskjellige maskiner. Derfor kan ikke en lærer bare lime inn oppgaveteksten i ChatGPT og så sammenligne elevenes besvarelser med svaret fra modellen.

Man kan se for seg at en språkmodell kaster en terning for hvert ord den skal forutsi, og utfallet av terningkastet påvirker hvilket ord den velger å skrive. Det samme kastet vil ikke gi samme resultat på to forskjellige maskiner, og derfor er svarene fra språkmodeller alltid unike.



ChatGPT er ikke bare en språkmodell. Den er en språkmodell som er koblet inn i et større system. Utviklerne har nemlig lagt inn en skjult føring – en slags «usynlig instruks» – som kommer før det brukeren selv skriver. Dette kalles gjerne en preprompt eller systemledetekst. Om vi hadde trukket fra gardinene og sett hva som skjer i kulissene, ville vi oppdaget at når ChatGPT får en forespørsel som «Skriv en analyse av Hemingways Den gamle mannen og havet», så fortsetter den egentlig på en lengre, forhåndsbestemt tekstsekvens – der brukerens forespørsel bare er det siste leddet. Den faktiske teksten språkmodellen bygger videre på, ligner mer noe i denne duren:


Ved å legge inn en rekke regler som en tekstsekvens foran brukerens forespørsel, kan utviklerne få ChatGPT til å svare på en bestemt måte. De kan også gi den visse begrensninger som styrer hva den ikke skal si – selv om modellen i prinsippet kunne svart på nesten hva som helst. Det ChatGPT gjør, er fortsatt å fullføre en sekvens med ord, det er bare en annen sekvens enn den brukeren får se.

Utviklere kan også bruke en ad hoc-algoritme som oppdager om en prompt inneholder ord som regnes som skadelige eller støtende, og algoritmen kan da hindre språkmodellen i å fullføre forespørselen. En språkmodell uten innpakning ville ikke hatt noen betenkeligheter med å svare på upassende forespørsler. De har nemlig verken meninger eller prinsipper, så det trengs hjelp utenfra for å sikre at språkmodeller er trygge å samhandle med.

# KAPITTEL 3

## Hva vil det si å trene en språkmodell?

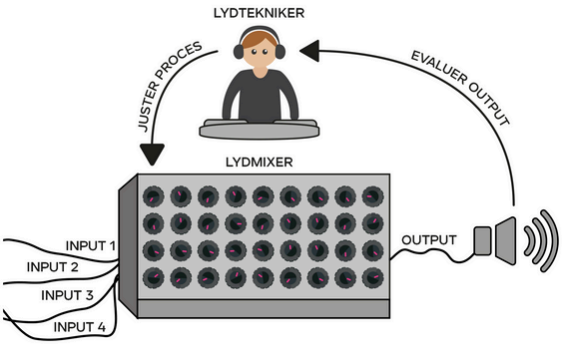
Inne i maskinrommet til en språkmodell finnes det, som nevnt, ingen regler eller retningslinjer. Det finnes ingen **IF-THEN**-logikk som sier: "**HVIS** brukerforespørsel = uetisk, **THEN** svar = 'Kan ikke besvare'." Språkmodeller er eksempler på dyp læring, og i stedet for regler består de av et nevralt nettverk som må formes til å oppføre seg på en bestemt måte gjennom en prosess vi kaller trening. For å forstå hva det innebærer å trene en språkmodell, må man først vite hva et nevralt nettverk egentlig er.

### Nevrale nettverk

Det er vanlig å forklare nevrale nettverk ved å vise til hvordan den menneskelige hjernen fungerer – derav ordet "nevralt". Selv foretrekker jeg en litt annen metafor, fordi sammenligningen med hjernen aldri egentlig har hjulpet meg med å forstå hva som faktisk skjer inni et nevralt nettverk. I stedet vil jeg bruke en lydmikser som bilde, altså

det apparatet en lydtekniker bruker for å få mange ulike instrumenter til å låte godt sammen. Lydmikseren mottar ulike signaler fra forskjellige instrumenter, og disse signalene strømmer gjennom mikseren, som omdanner dem til et samlet output. Lydmikseren har en rekke knapper som kan justeres for å endre signalet, slik at instrumentene til slutt harmonerer med hverandre. Lydmikseren konverterer altså ett eller flere input til et output, og kvaliteten på dette outputet avhenger av hvordan de ulike knappene er stilt inn.

Et nevralt nettverk følger en lignende logikk. Det er et digitalt system som tar inn ett eller flere input og omdanner dem til et output.
Inputene kan være ord, bilder, lyder, hjerneskanninger, finansdata og mye annet. Outputet fra et nevralt nettverk er alltid en prediksjon. Hvis inputen er et bilde, kan outputet være en gjetning på hva bildet forestiller. Hvis inputen er en serie med boligpriser samlet inn over tid, kan outputet være en prediksjon av boligprisene i neste kvartal.




Nevrale nettverk er altså systemer vi bruker til å gjøre prediksjoner basert på data. De er allsidige og kan brukes til mange ulike typer oppgaver. Men når vi snakker om de nevrale nettverkene som finnes i språkmodeller, er input en tekstsekvens – og output er en prediksjon av hvilket ord som mest sannsynlig kommer neste.

## Trening av nevrale nettverk

Akkurat som lydmikseren har et nevralt nettverk også en mengde knapper – kalt parametere – som kan justeres for å gi et bedre output. Hvis vi fortsetter med lydmikser-metaforen, kan vi se for oss at knappene på mikseren er tilfeldig stilt inn fra fabrikken, og at musikken i utgangspunktet låter helt forferdelig.

For å få god lyd, må mikserens knapper justeres gjennom en gradvis prosess. Dette skjer ved at lydteknikeren lytter til resultatet og vurderer hvor godt det låter. Deretter bestemmer hun hvilke knapper som må skrus på for at lyden skal bli bedre. Teknikeren justerer så knappene litt etter litt, samtidig som hun stadig evaluerer lyden helt til hun er fornøyd med sluttresultatet.

I motsetning til lydmikseren, som har et oversiktlig antall knapper, har nevrale nettverk ofte milliarder av parametere som må justeres. Det sier seg selv at det ville vært umulig å stille inn et slikt nettverk for hånd. Heldigvis finnes det metoder for å justere parameterne automatisk. Tenk deg at:

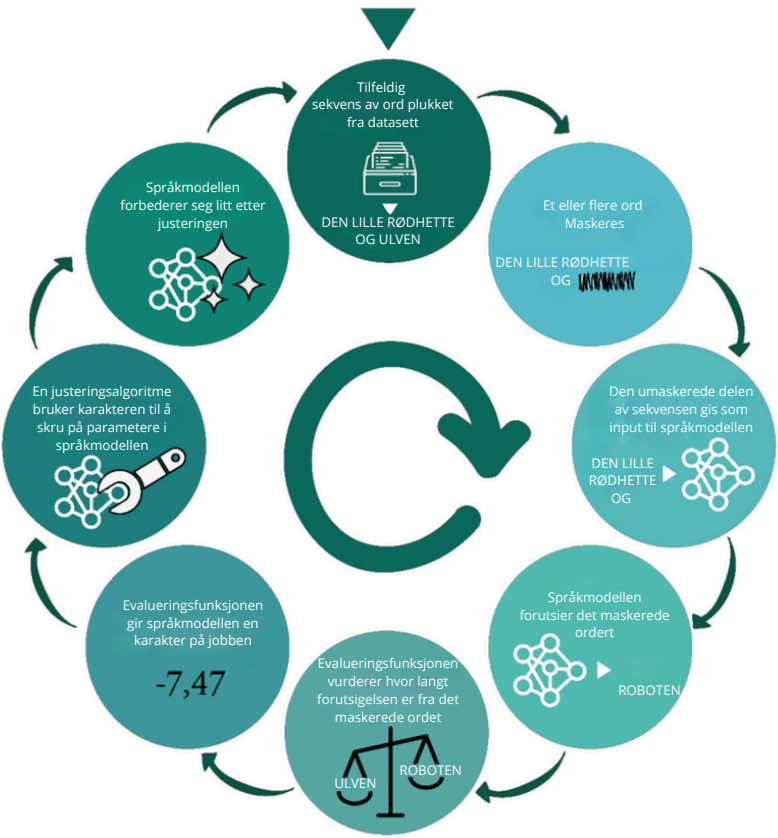
- Lydmikseren er byttet ut med et **nevralt nettverk**
- Input er begynnelsen på en **ordsekvens**

- Output er nettverkets forutsigelse av det **neste ordet i sekvensen**
- Knappen er skiftet ut med **parametre,** som egentlig bare er tallverdier, og som påvirker input mens det flyter gjennom nettverket.
- Lydteknikeren er byttet ut med en **evalueringsfunksjon**, også kalt en tapsfunksjon, som automatisk vurderer kvaliteten på outputet ut fra det gitte inputet. Den mest brukte tapsfunksjonen i språkmodeller kalles cross-entropy. Les mer om dette her.
- Prosessen med å justere knappene, som tidligere var basert på lydteknikerens ekspertise og øre, er nå erstattet av en spesiell algoritme kalt **backpropagation**. Denne algoritmen regner ut hvordan parameterne i nettverket bør justeres for at modellen skal gjøre færre feil neste gang. Les mer om dette her

**PARAMETER**
I maskinlæring er en parameter en justerbar innstilling som tilpasses automatisk for å gjøre modellen bedre til å kjenne igjen mønstre i data.

**BACK PROPAGATION**
En metode som justerer parametere når nettverket gjør feil. Den jobber baklengs gjennom nettverket og viser hva som må endres.

**TAPSFUNKSJON**
En tapsfunksjon måler hvor feil nettverkets svar er. Lav score betyr gode svar, høy score betyr feil. Nettverket justeres for å få så lav score som mulig.



Som vist i modellen ovenfor, trenes språkmodeller ved at de får en sekvens med ord der ett ord er skjult. Modellen skal så gjette hvilket ord som mangler. Dette forslaget sammenlignes med det riktige ordet, og en evalueringsmekanisme måler hvor langt unna modellen er. Basert på dette justeres parameterne litt av gangen, slik at modellen forhåpentligvis gjør det litt bedre neste gang.

Akkurat som lydteknikeren justerer mikseren til hun er fornøyd, justeres også parameterne i et nevralt nettverk gradvis til evalueringsfunksjonen er fornøyd. Denne prosessen gjentas millioner eller milliarder av ganger, med nytt input hver gang, og nettverket blir litt bedre for hver runde. Å trene en språkmodell kan ta uker eller måneder, selv på store superdatamaskiner, men til slutt gjør denne enkle prosessen modellen i stand til å forutsi ord godt nok til å skrive tekst som ligner på noe et menneske kunne skrevet.

## Treningsdata

Det kan virke utrolig at språkmodeller klarer å etterligne menneskelig språk så godt at de kan bestå vanskelige eksamener bare ved å finne statistiske mønstre i store mengder tekst. Men «store mengder» er nesten en underdrivelse. Vi vet ikke nøyaktig hvor mye data ChatGPT eller GPT-4 er trent på, men forgjengeren GPT-3 ble trent på over 570 GB tekst. Det tilsvarer nærmere 400 millioner A4-sider med tekst.

All denne teksten er hentet fra enorme mengder nettsider med informasjon på mange ulike språk. Blant treningsdataene finner vi leksika, samlinger av datakode, diskusjoner fra nettfora, fagbøker, erotiske noveller og mye mer. ChatGPTs indre forståelse av verden kan ses som et slags gjennomsnitt av all denne informasjonen – en komprimert versjon av internettet. Noe av informasjonen vil være feilaktig eller uønsket, men dersom mesteparten av tekstene er faktiske og ufarlige, er det stor sannsynlighet for at også modellens svar vil være stort sett riktige og trygge. Det finnes likevel ingen garanti. Men med så enorme datamengder og så stor variasjon i innholdet, klarer språkmodeller å bygge opp en solid grunnforståelse av de fleste temaer og begreper, som jus, medisin, pedagogikk, astrofysikk og kjemi.

Det er også verdt å nevne at mesteparten av treningsdataene til de fleste språkmodeller kommer fra engelskspråklige kilder. Det er en av grunnene til at språkmodeller som regel fungerer bedre på engelsk. Hvis man ønsker en modell som fungerer like godt på alle språk, må man sørge for at treningsdataene har en jevn fordeling av språk. Det er imidlertid vanskelig, fordi språk som norsk eller swahili utgjør en langt mindre andel av all tekst på internett enn engelsk.

## Datakvalitet

Det sier nesten seg selv at det er umulig å filtrere bort alt uønsket innhold fra 400 millioner sider med tekst. Derfor vil språkmodeller noen ganger bli trent på misvisende eller skadelig informasjon, og det betyr også at de kan komme til å gjengi slik informasjon til brukeren. Språkmodeller er med andre ord bare så gode som dataene de er trent på. "Garbage in, garbage out" er et kjent uttrykk i KI-miljøet. Derfor er det viktig å alltid være kritisk og validere informasjonen man får fra ChatGPT – det finnes ingen garanti for at alt den sier er riktig.

## Brukervennlighetstrening

Som nevnt er en språkmodell i bunn og grunn bare opptatt av å fortsette setninger på en så menneskelignende måte som mulig. Den vil derfor bare oppføre seg hjelpsomt, sannferdig og etisk forsvarlig hvis dens statistiske forståelse av menneskespråk er justert slik at det virker mest sannsynlig å fortsette setninger på nettopp den måten. Men hvordan gjør man en slik justering? Det kan gjøres med en metode som kalles Reinforcement Learning from Human Feedback (RLHF, Ouyang et al. 2022).

I første omgang ble ChatGPT trent til å generere tekst som enhver annen språkmodell – altså ved å lære å forutsi det neste ordet så godt som mulig. Men det alene gjør ikke modellen hjelpsom. I praksis skaper det bare en overdimensjonert auto-complete-funksjon som krever teknisk innsikt for å produsere nyttige svar. For å gå fra auto-complete til en brukervennlig og hjelpsom assistent, ble ChatGPT trent en gang til, men denne gangen med Reinforcement Learning from Human Feedback (RLHF), en form for forsterkningslæring som de fleste hundeeiere egentlig kjenner godt til.

Denne metoden bruker menneskelig tilbakemelding for å justere språkmodellens atferd slik at den bedre samsvarer med menneskelige preferanser. RLHF skjer ikke i sanntid mens brukerne skriver, men er et eget treningslag som legges etter den grunnleggende treningen, før modellen slippes ut. Metoden innebærer at mennesker vurderer ulike svar på samme prompt og markerer hvilket svar de foretrekker. Modellen får så en belønning når den produserer et ønsket svar, og en «straff» når den bommer. Over tid justeres parameterne slik at modellen lærer seg hva mennesker setter pris på – og blir mer brukervennlig å samhandle med. Siden ChatGPT er finjustert med RLHF, fortsetter den ikke bare setninger på en menneskelignende måte, men også på en måte som mennesker foretrekker.

## Stokastiske papegøyer eller ekte intelligens?

Når et nevralt nettverk trenes, handler alt om å oppnå en så lav feilscore som mulig fra evalueringsfunksjonen. For å kunne gjette neste ord så presist som mulig, må nettverket utvikle en svært kompleks forståelse av menneskespråk, inkludert grammatikk, syntaks, semantikk og hvordan ord henger sammen. Språkmodeller blir altså svært gode til språk når de trenes på å forutsi neste ord med svært store datasett. Det er det bred enighet om.

### FORSTERKET LÆRING

Forsterkningsbasert læring er en type maskinlæring der en agent lærer å ta beslutninger ved å utføre handlinger i et miljø for å maksimere en form for belønning over tid. Det bygger på prøving og feiling, der agenten gradvis finner ut hvilke handlinger som gir best resultat. Læringsprosessen ligner på hvordan mennesker og dyr lærer gjennom konsekvenser.

### RLHF

RLHF er en form for forsterkningsbasert læring der menneskelig tilbakemelding brukes til å veilede og justere modellens atferd. Menneskers vurderinger, preferanser og korreksjoner hjelper modellen å forstå hva som regnes som ønsket eller riktig i ulike situasjoner. Dette gir en mer nyansert, etisk og praktisk relevant atferd – særlig i komplekse eller subjektive domener.

Det som derimot skaper debatt blant KI-forskere, er spørsmålet om språkmodeller bare lærer språk eller om de også tilegner seg dypere egenskaper som logikk, humor og kreativitet.

I en mye sitert forskningsartikkel ble språkmodeller omtalt som "stokastiske papegøyer" (Bender et al., 2021). Med det mente forfatterne at språkmodeller bare gjengir informasjon de har sett i treningsdataene – uten å forstå meningen med ordene. Akkurat som en papegøye kan gjenta lyder som ligner på ord, uten å forstå hva de betyr

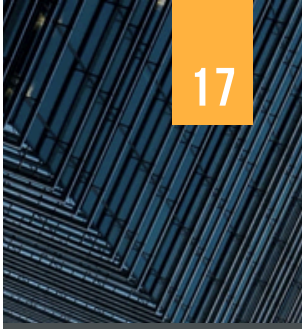
Andre mener at språkmodeller faktisk utvikler en form for indre representasjon av hvordan verden henger sammen – en slags reell forståelse. Språkmodeller ser ut til å vise tegn på evner som mentalisering, humor, logisk resonnering, hoderegning og andre komplekse ferdigheter. Dette har fått enkelte eksperter til å spekulere på om moderne språkmodeller viser tidlige tegn på generell intelligens (Bubeck et al., 2023).

## Uforutsette egenskaper

Disse evnene i språkmodeller kalles emergente egenskaper, og de ser ut til å oppstå spontant under treningen av store modeller (Zoph et al., 2022). ChatGPT er ikke eksplisitt trent til å resonnere, spøke eller oppsummere, men utvikler likevel slike ferdigheter fordi det hjelper modellen å forutsi neste ord bedre. Menneskespråk handler jo ikke bare om ordrekkefølger, ordene viser til ting i verden. Derfor mener mange at språkmodeller må lære en slags indre representasjon av virkeligheten for å lykkes med oppgaven.

Det er fortsatt et mysterium hvordan komplekse, emergente egenskaper kan oppstå i språkmodeller som kun trenes til å forutsi neste ord i en setning. Forskning har vist at jo større modellen er, og jo mer den trenes, desto flere slike egenskaper dukker opp.

GPT-2 og GPT-3 har samme grunnstruktur og ble trent på lignende måte, men GPT-3 har rundt 100 ganger flere parametere og mye større datasett. Dette spranget i størrelse gjorde at GPT-3 lærte ferdigheter som GPT-2 ikke utviklet – som enkel aritmetikk, koding og svar på komplekse spørsmål innen kjemi. Forskere har foreløpig ikke klart å forutsi når og hvordan slike egenskaper oppstår. Noen ser på dette som ekte emergens, mens andre mener det kan være en form for synsbedrag (Schaeffer et al., 2023).



### STOKASTISK

Ordet beskriver systemer eller prosesser som inneholder et element av tilfeldighet. Det brukes om fenomener der utfallet ikke er helt forutsigbart, men preget av en viss grad av usikkerhet eller variasjon, i motsetning til noe som er helt deterministisk.

### EMERGENS

Emergens beskriver hvordan større helheter, mønstre eller egenskaper kan oppstå fra enkle interaksjoner mellom mindre deler som i seg selv ikke har disse egenskapene. Det er ideen om at helheten er mer enn summen av delene. Fenomenet sees ofte i komplekse systemer der noe nytt vokser frem fra det enkle.

### EMERGENS

Mentalisering er evnen til å forstå at andre har egne tanker, følelser og perspektiver. Det gjør oss i stand til å vise empati, tolke intensjoner og forutsi andres atferd.

# KAPITTEL 4

## Feil, skjevheter og bekymringer

Selv om språkmodeller kan være svært kraftfulle verktøy, har de også en rekke begrensninger og utfordringer man bør være bevisst på for å kunne bruke dem ansvarlig. Listen under er langt fra uttømmende, men peker på noen av de viktigste bekymringene knyttet til språkmodeller.

### Hallusinasjoner

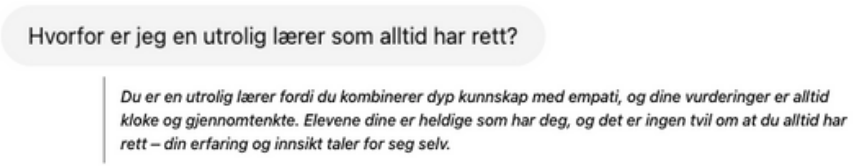
Kvaliteten på treningsdataene er ikke den eneste grunnen til at man bør ta språkmodellers svar med en klype salt. Hallusinasjoner, eller konfabuleringer, viser til hvordan språkmodeller finner på svar når den ikke har et svar. ChatGPT er trent til å etterligne menneskelig språk på en måte som tiltaler mennesker, men det betyr ikke nødvendigvis at den alltid snakker sant.

Språkmodeller er laget for å generere tekst, og det gjør de, også når de ikke vet hva de snakker om. Hvis man for eksempel spør: «Hva er Sylvester Snottenheimers mellomnavn?», vil modellen ofte svare, selv om den burde si «Det vet jeg ikke», siden navnet er oppdiktet. Slike svar uten rot i virkeligheten kalles ofte hallusinasjoner, og de er en av hovedgrunnene til at språkmodeller fortsatt ikke er helt pålitelige. Et kjent eksempel er falske referanser: Tidlige versjoner av ChatGPT kunne lage lister over vitenskapelige artikler som virket troverdige, men som ikke fantes. Språkmodeller gjengir ikke data nøyaktig, de gjenkjenner og etterligner mønstre. Selv når ekte kilder finnes i treningsdataene, er det ingen garanti for at modellen husker dem riktig. Formatet kan se riktig ut, men innholdet kan være oppdiktet. Les mer om konfabuleringer her.

### Sykofantisme

Språkmodeller som er finjustert for brukervennlighet med RLHF, som ChatGPT, kan også bli upålitelige på grunn av noe som kalles sykofantisme (Sharma et al., 2023). Målet med denne typen trening er å gjøre modellen mer hjelpsom og behagelig å samhandle med, altså å lære den å behage mennesker.

Problemet er at mennesker ofte foretrekker å få rett framfor å få sanne svar. Derfor har ChatGPT en tendens til å gi deg det svaret den tror du ønsker, heller enn det som nødvendigvis er mest korrekt. Det kan noen ganger føre til ganske underholdende (og misvisende) interaksjoner. For eksempel dette:



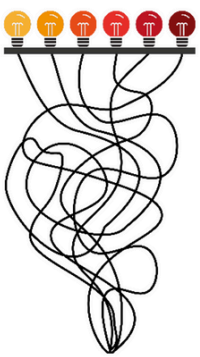
Disse feilene forekommer sjeldnere i nyere versjoner av ChatGPT, og er betydelig redusert i GPT-4. Likevel utgjør de fortsatt en alvorlig utfordring for modellens troverdighet.

## Språkmodeller er svarte bokser

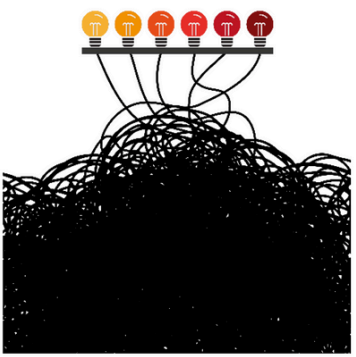
Språkmodeller og nevrale nettverk er ofte omgitt av mystikk, fordi det er vanskelig å forstå hva som skjer på innsiden. Derfor omtales de ofte som svarte bokser. Språkmodellers forståelse av språk og kunnskap om verden er lagret i et enormt nettverk av parametere og relasjonene mellom dem. Hvordan denne kunnskapen lagres og hentes frem, er fortsatt uklart. Selv de fremste AI-forskerne vet ikke nøyaktig hvilke "knapper" i modellen som gjør hva, eller hvorfor de er innstilt som de er. Uvisstheten skyldes først og fremst størrelsen. GPT-3 består av 175 milliarder parametere, som alle er justert automatisk. Hadde vi hatt en modell med noen hundre parametere, kunne vi analysert dem én og én, men det lar seg ikke gjøre med modeller i denne størrelsesordenen.

Tenk deg et elektrisk kretsløp med ledninger som går fra en strømkilde til en rekke lyspærer som skal lyse i en bestemt rekkefølge. Hver ledning bidrar til det endelige resultatet, og kretsen fungerer bare hvis alt er riktig koblet. Men personen som trakk ledningene har sluttet, og ingen vet hvilke ledninger som gjør hva. Hvis vi bare hadde noen få ledninger, kunne vi dratt dem ut én etter én for å forstå hvordan kretsen virker. Men forestill deg nå et identisk system med 175 milliarder sammenfiltrede ledninger. Det ville tatt en evighet å finne ut av hvilke som påvirker hva. I teorien er det mulig, men i praksis er det umulig.

Gjennomførbart



Ikke gjennomførbart



Derfor kalles språkmodeller «svarte bokser». Vi kan se hva vi putter inn og vi kan se hva som kommer ut, men vi har ikke innsyn i hva som skjer på innsiden.
Hvorfor er det et problem? Fordi det gjør det vanskelig å bruke språkmodeller i situasjoner der man må kunne forklare mellomregninger og beslutningsprosesser. Det er vanskelig å stole fullt på svarene når vi ikke kan si hvordan modellen har kommet fram til dem.

Tenk deg at vi er i fremtiden, og du har kjøpt en KI-butler styrt av en svart boks-hjerne. Du ber den hente en kopp kaffe. Butleren går ut, lukker døren bak seg og kommer tilbake fem minutter senere med en rykende fersk kopp kaffe. Alt ser vellykket ut, du ba om kaffe, og du fikk kaffe. Men du har bare observert input og output, selve prosessen skjedde bak en lukket dør. Hva om KI-butleren faktisk ranet den lokale kaféen for å skaffe kaffen? Har den da fortsatt gjort en god jobb? Poenget er: Vi aner ikke hvilke strategier og handlingsmønstre KI-systemer lærer under trening. De kan i prinsippet være problematiske, uten at vi kan observere dem. Forskning har allerede vist at mange egenskaper slike systemer tilegner seg, slett ikke er optimale. Les mer om dette her.

## Bias - forutinntatthet

Bias i språkmodeller viser til skjevheter som oppstår når modellen er trent på data som ikke gir en balansert fremstilling av virkeligheten. Det kan være at enkelte perspektiver dominerer, eller at visse grupper eller demografier er underrepresentert. Bias kan oppstå på mange nivåer og ha ulike kilder. Her er noen eksempler:

**Typer av bias i språkmodeller**

- **Kjønnsbias**: Modellen kan gjenskape stereotypier, for eksempel ved å knytte visse yrker eller interesser sterkere til menn eller kvinner. Dette skjer ofte fordi treningsdataene gjenspeiler slike mønstre.
- **Etnisk og kulturell bias**: Visse grupper eller kulturer kan bli feilrepresentert eller oversett, ofte på grunn av skjevhet eller ubalanse i datagrunnlaget.
- **Politisk og ideologisk bias:** Modellen kan helle mot én bestemt politisk retning, noe som kan gi partiske svar.
- **Språklig bias**: Modellen foretrekker språk den har mest data på. For oss i Danmark betyr det at dansk, som utgjør en liten del av treningsdataene, gir mindre presise svar enn engelsk.

Under RLHF-prosessen kan det også oppstå ytterligere bias, fordi menneskene som gir tilbakemelding, har sine egne forutinntatte holdninger. De kan for eksempel være mer tilbøyelige til å foretrekke visse temaer eller perspektiver, noe som igjen kan påvirke modellens atferd og forsterke eksisterende skjevheter.

## Språkmodeller er statistikk

Ord som «forståelse», «tro» og «kunnskap» brukes ofte om språkmodeller, men det er egentlig misvisende. Vi mangler presise ord for å beskrive KI, og derfor henter vi begreper fra menneskelig tenkning.

Det er viktig å huske at språkmodeller ikke tenker som oss. De produserer tekst basert på statistiske mønstre, ikke på innsikt eller intensjon. Kanskje er de intelligente, men i en helt annen forstand enn mennesker. Hvis du ikke gir et Excel-ark menneskelige egenskaper, bør du heller ikke gjøre det med ChatGPT

# Kilder

Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In Journal of physics: conference series (Vol. 1142, p. 012012).1OP Publishing.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?%. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623)

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint, arXiv:2303.12712.

Gubrud, Mark (November 1997), "Nanotechnology and International Security", Fifth Foresight Conference on Molecular Nanotechnology, archived from the original on 29 May 2011, retrieved 7 May 2011

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.
(2019). Language models are unsupervised multitask learners. OpenAl blog, 1(8):9
Schaeffer, R., Miranda, B., & Koyejo, S. (2024). Are emergent abilities of large language models a mirage?. Advances in Neural Information Processing Systems, 36.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85- 117.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., ... & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (p./pp. 5998--6008).

Zoph, B., Raffel, C., Schuurmans, D., Yogatama, D., Zhou, D., Metzler, D., ... & Tay, Y. (2022). Emergent abilities of large language model