SPROGMODELLER FOR DUMMIES

En intuitiv introduktion til teknologien bag ChatGPT

KØBENHAVNS PROFESSIONS HØJSKOLE

Skrevet af Troels Jensen

Marts 2024

INDLEDNING

Dette dokument er ment som en lettilgængelig introduktion til fænomenet *sprogmodeller* (fx. ChatGPT, GPT-4, Gemini og Copilot). Det er en rolig indflyvning i et kompliceret emne henvendt til dig, der ikke har en teknisk baggrund, men som gerne vil vide mere om kunstig intelligens og sprogmodeller.

Dokumentet er et forsøg på:

- at give dig en nænsom introduktion til et svært emne ved at udfolde de vigtigste relaterede termer på en simpel måde, som jeg ville ønske, nogen havde forklaret dem for mig. Ingen formler eller uforståelige diagrammer.
- at tilbyde nogle enkle forståelsesrammer og forklaringer, der kan hjælpe til at opbygge en intuition for sprogmodellers logikker, så emnet bliver mere tilgængeligt.
- at sætte dig i stand til at finde hoved og hale i det væld af information, der florerer derude.
- at afmystificere kunstig intelligens for at forvisse dig om, at vi endnu er ganske langt fra Terminator-farvande, selvom ChatGPT er et imponerende stykke teknologi.

Jeg har prioriteret at gøre teksten så letlæselig som muligt, og derfor berører den kun de mest essentielle begreber og tematikker. Sprogmodeller er et teknisk tungt emne, så dokumentet byder langt fra på et fuldendt billede af fænomenet. Derfor henviser jeg til mere formelle kilder via referencer i teksten, og på mange sider finder du en grøn tekstboks med begrebsdefinitioner og forklaringer. Undervejs i dokumentet henviser jeg også til eksemplariske demonstrationer, guides, læremidler mm. via links i teksten. Jeg vil starte med at anbefale enhver med en interesse i emnet at tage kurset Elements of AI, som er den bedste introduktion til kunstig intelligens, jeg kender til.

Informationen i dokumentet er begrænset til eksisterende chatbots marts 2024 med fokus på GPT-serien fra OpenAI. AI-feltet undergår lige nu hastige forandringer, så informationen er ikke nødvendigvis retvisende for fremtidige sprogmodeller. Jeg vil forsøge at opdatere dokumentet i takt med, at der kommer nye udviklinger og nye sprogmodeller.

HVAD ER EN SPROGMODEL?

Sprogmodeller er den tekniske betegnelse for systemer som ChatGPT, som man også ofte hører omtalt som chatbots. Kort fortalt er sprogmodeller en familie af algoritmer, der kan generere menneskelignende tekst med afsæt i en brugers forespørgsel. En forespørgsel, også kaldet et *prompt,* er en bid tekst, man skriver til sprogmodellen. Sprogmodellen reagerer så på et prompt ved at producere tekst ud, som den beregner vil passe til promptet.

Moderne sprogmodeller fortsætter simpelthen sekvenser af ord, på en måde så det ender med at se ud som noget, et menneske kunne have skrevet (Radford et al. 2019). Det er sprogmodellers funktion i en nøddeskal. Det klassiske eksempel er at stille et faktuelt spørgsmål med sit prompt, som sprogmodellen så besvarer som sin fortsættelse af ordsekvensen. Et andet typisk eksempel ville være skrive indledningen på en artikel, som en sprogmodel så fortsætter eller skriver færdig.

Jeg kommer til at anvende ordet *sprogmodel* frem for *chatbot* gennem hele artiklen. Det er dels fordi det er den akademiske betegnelse, og dels fordi ordet chatbot er belemret af konnotationer, som efter min mening gør det misvisende at bruge om systemer som ChatGPT. Ordet chatbot har førhen henvist til nogle meget simple spørgsmål-svar systemer - fx de små chatvinduer, man sommetider finder på flyselskaber eller tøjbutikkers hjemmesider, som skal hjælpe kunden med deres problemer og spørgsmål. Den slags systemer er baseret på forprogrammerede regler såsom:

IF brugerinput **CONTAINS** “Aflyst **OR** refundering **OR** klage” **THEN** output = “Viderestiller til kundeservice. Du er nr. 74 i køen”.

Det er et eksempel på en god, gammeldags algoritme, som er bygget op omkring simple, håndskrevne regler, og som kun kan reagere på ganske få typer forespørgsler med en tilsvarende lille håndfuld svar.

Sprogmodeller opererer ud fra nogle helt anderledes logikker, og de har en langt bredere handlekapacitet end gammeldags chatbots. De kan langt, langt mere end bare at chatte. Sprogmodeller er nemlig eksempler på kunstig intelligens. Nærmere bestemt hører de til den gren af computervidenskaben, vi kalder deep learning, som er den førende tilgang til machine learning. Machine learning repræsenterer en fundamentalt anderledes tilgang til problemløsning end chatbotten i eksemplet

ALGORITMER

En algoritme er en endelig række af veldefinerede, computer-udførbare instruktioner, typisk for at løse et klasse af problemer eller udføre en beregning. Algoritmer specificerer en sekvens af trin, der skal følges for at opnå et ønsket resultat, og de skal være klare og utvetydige.

KUNSTIG INTELLIGENS

Kunstig intelligens (AI) er udviklingen af maskiner, der kan udføre opgaver, som typisk kræver menneskelig intelligens. Dette inkluderer problemløsning, læring og tilpasning. Begrebet blev først defineret af John McCarthy i 1956.McCarthy, som i 1956 definerede det som "videnskaben og ingeniørkunsten at skabe intelligente maskiner...".

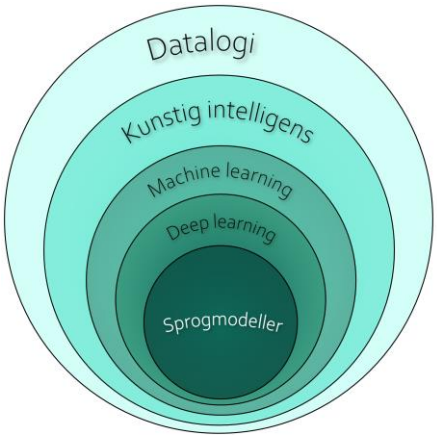
MACHINE LEARNING

Machine learning er en metode inden for kunstig intelligens, hvor computere anvender statistiske teknikker til at identificere mønstre i data. Dette gør det muligt for systemerne selvstændigt at forbedre deres ydeevne på specifikke opgaver med mere erfaring eller data.

foroven. Machine learning benytter slet ikke håndskrevne, men udnytter i stedet mønstre i data til at lære at løse opgaver selv.

Begrebs-taksonomi

AI-feltet er er en begrebsjungle. For at forklare hvordan sprogmodeller passer ind i det større AI-landskab, er det nyttigt at have en begrebstaksonomi:



- **Datalogi,** også kaldet computervidenskab, beskæftiger sig ikke kun med at programmere computere, men omhandler bredt forstået studiet af data og algoritmer.
- **Kunstig intelligens** eller **AI** er det overordnede felt, der har til formål at skabe systemer, der kan udføre opgaver, som normalt kræver menneskelig intelligens.
- **Machine learning** er hovedgrenen inden for kunstig intelligens og indbefatter en stor familie af statistiske metoder, der kan løse komplekse problemer på automatiseret vis ved at lære af data. Machine learning er nyttigt, fordi det sætter os i stand til at løse komplekse opgaver, uden eksplicit at skulle formalisere og programmere løsningsprocessen.
- **Deep learning** er det dominerende paradigme indenfor machine learning i dag. Deep learning-systemer er baseret på neurale netværk, som kan beskrives som en løs efterligning af den menneskelige hjernes måde at bearbejde information. Mere herom senere. Deep learning har vist sig at være en enorm effektiv og alsidig tilgang, der kan løse utroligt mange komplekse opgaver (Schmidthuber 2015).
- **Sprogmodeller** er en kategori af deep learning-modeller, der er specialiserede i at arbejde med menneskeligt sprog. **Generative** sprogmodeller, som fx ChatGPT, er skabt til at generere tekst, mens mange tidlige sprogmodeller, som fx BERT, er **prædiktive** og bruges til andre formål såsom at klassificere en tekst som enten positiv eller negativ. Læs mere om forskellen her.

Snæver versus lidt mere generel AI

Sprogmodeller begyndte at vinde popularitet i 2020, da selskabet OpenAI lancerede verdens hidtil største sprogmodel, GPT-3. GPT-3 imponerede med en ekstremt livagtig forståelse for sprog og en evne til at producere formfuldendt tekst indenfor en lang række områder, den aldrig var blevet trænet til at agere indenfor.

DEEP LEARNING

Deep learning involverer anvendelsen af neurale netværk med mange lag af bearbejdningsenheder, der løst simulerer den måde, menneskehjernen behandler information på. Disse lag modtager input, behandler det gennem forskellige niveauer af abstraktion og kompleksitet, og producerer output. Denne struktur gør det muligt for deep learning-modeller at identificere og lære komplekse mønstre i data, hvilket gør dem særligt effektive til opgaver som billed- og talegenkendelse.

GENERATIV AI

Generativ AI henviser til AI, der er i stand til at skabe nyt, unikt indhold baseret på indlærte mønstre fra data. Det bruges typisk til at generere menneskelignende indhold fx forfatte tekster, komponere musik eller skabe visuelle kunstværker.

PRÆDIKTIV AI

Generativ AI kontrasteres ofte med prædiktiv AI, som har til formål at forudsige ubekendte datapunkter baseret på relaterede data. Det kan være at forudsige aktiemarkeds-bevægelser, kundeadfærd eller vejrforhold. Dette er den mest udbredte form for AI, og prædiktiv AI kan findes i alverdens moderne produkter.

GPT-3 blev nemlig trænet med det simple formål at generere menneskelignende tekst ét ord ad gangen, men undervejs i sin træning lærte GPT-3 også at udføre en lang række opgaver, som vi ellers troede kun mennesker kunne klare. Den lærte fx at opsummere artikler, skrive computerkode, oversætte fra et sprog til et andet, forfatte poesi, fremsætte logiske argumenter og meget, meget mere. Det var banebrydende.

For at understrege hvor stor en udvikling det var, skal man forstå, at stort set alle tidligere typer machine learning-systemer er ekstremt snævre i deres repertoire – de kan kun udføre præcist den opgave, de er trænet til, og kan ikke generalisere til nye domæner. Den type systemer kaldes sommertider *snæver AI*. Et eksempel er Spotify's anbefalingsalgoritme: Det er en machine learning-algoritme, som er enormt god til at foreslå indhold, men som ikke kan foretage sig noget som helst andet. Hvis man vil have Spotifys algoritme til at handle anderledes, skal man træne den helt forfra med andre data.

Moderne sprogmodeller er mere generelle. De kan udføre mange forskelligartede opgaver uden at skulle trænes specifikt til hver individuel opgave. Det er det, der gør dem så opsigtsvækkende – den generaliserbarhed begynder jo at ligne noget, vi normalt associerer med menneskelig intelligens. Den hellige gral inden for AI-forskning er netop at skabe generel kunstig intelligens, såkaldt AGI (Artificial General Intelligence), som kan udføre alle de opgaver mennesker kan varetage – endnu bedre end mennesker kan (Gubrud 1997). Et fiktivt eksempel på AGI er den alvidende HAL9000 fra filmen Rumrejsen 2001. Indtil videre er sprogmodeller det nærmeste vi er kommet på AGI, men vi er stadig meget langt fra egentlig generel intelligens. Mange eksperter mener, at AGI ligger årtier ude i fremtiden, mens andre tror, vi allerede er ved at være der.

Sprogmodeller rammer mainstream

Selvom GPT-3 var en gamechanger, var det begrænset, hvor meget den almene befolkning blev berørt af udviklingen. Det var derimod en helt anden sag, da ChatGPT kom på gaden 30. november 2022. ChatGPT er en videreudvikling af GPT-3, og der er to primære årsager til, at det netop blev den, der tog verdenen med storm:

1. Sprogmodeller plejede at være besværlige at interagere med. Det krævede know-how at få dem til at opføre sig, som man

SNÆVER AI

Snæver, også sommetider kaldet svag AI eller smal AI, er den type AI, som virksomhederne reklamerer med, når de taler om AI i deres produkter. Det henviser næsten altid til machine learning-systemer, der kun kan udføre en simpel opgave, og som ikke kan generalisere til nye domæner. Vi er nået enormt langt med den slags AI, og det har tilladt os automatisere en masse opgaver, som vi hidtil troede var forbeholdt mennesker. Eksempler er talegenkendelse, anbefalingssoftware og Google Translate.

AGI

Generel, også sommertider kaldet stærk AI eller AGI, henviser til hypotetiske systemer, der kan udføre en bred vifte af opgaver, lige så godt eller bedre end mennesker. Det er altså systemer, der kan generalisere til nye domæner, uden at skulle trænes specifikt til det formål med et specialiseret datasæt. Det er den hellige gral indenfor AI forskning, og det er endnu ikke lykkedes os at udvikle generel AI.

ønskede, og man skulle bruge lang tid på at konstruere specifikke forespørgsler, som ansporede modellerne til at producere gode outputs. ChatGPT er derimod nem og intuitiv at interagere med. Dens outputs er for det meste relevante, og man kan interagere med den på en måde, der opleves naturlig. Hvis ChatGPT giver dig et utilfredsstillende svar, kan du bede den om at formulere sig anderledes og om at huske at indsætte akademiske referencer. Det var langt sværere med tidligere sprogmodeller.

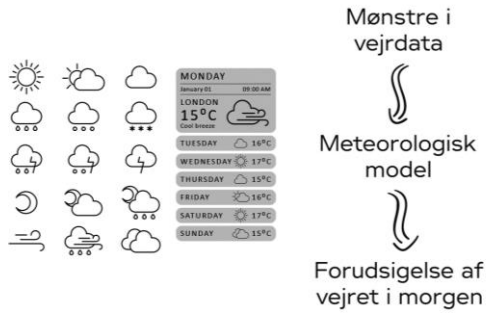
2. Den anden årsag er, at OpenAI stillede ChatGPT gratis til rådighed for alle på en lettilgængelig hjemmeside. Det gjorde de til dels for at indsamle data om folks interaktioner med systemet – den slags data kan nemlig hjælpe dem til at træne bedre sprogmodeller i fremtiden.

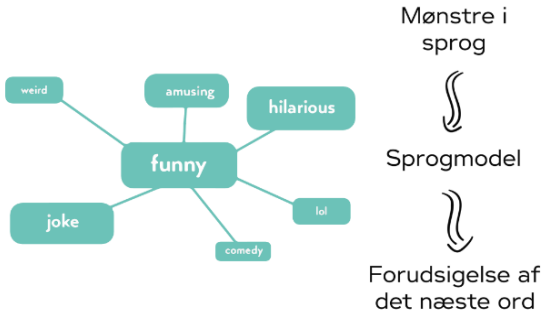
HVORDAN FUNGERER SPROGMODELLER UNDER KØLERHJELMEN?

Sprogmodeller bygger på en række komplicerede matematiske tricks og langhåret statistisk teori, som det ville kræve en lang, teknisk smøre at redegøre for. Afsnittet her giver dig ikke en dybdegående forståelse for sprogmodellers indmad. Mit fokus her er at afmystificere fænomenet ved at trække nogle simple forståelsesrammer ned over de bagvedliggende logikker.

En statistisk model over menneskeligt sprog

Ordet "model" skal i denne kontekst forstås lidt på samme måde, som når vejrudsigten snakker om meteorologiske modeller. Vejrudsigten baserer til dels deres prognoser på en matematisk model, som er konstrueret ud fra systematiske korrelationer i tonsvis af vejrdata. Hvis datapunkter om lavtryk i øst ofte efterfølges af datapunkter om stærk vestenvind, kan modellen bruge den korrelation til at forudsige kuling fra vest i aften, hvis der blev observeret lavtryk i øst i morges. I virkeligheden er meteorologiske modeller langt mere komplicerede, men pointen er, at statistiske modeller udnytter mønstre i store datamængder til at foretage forudsigelser på baggrund af ufuldendt information.





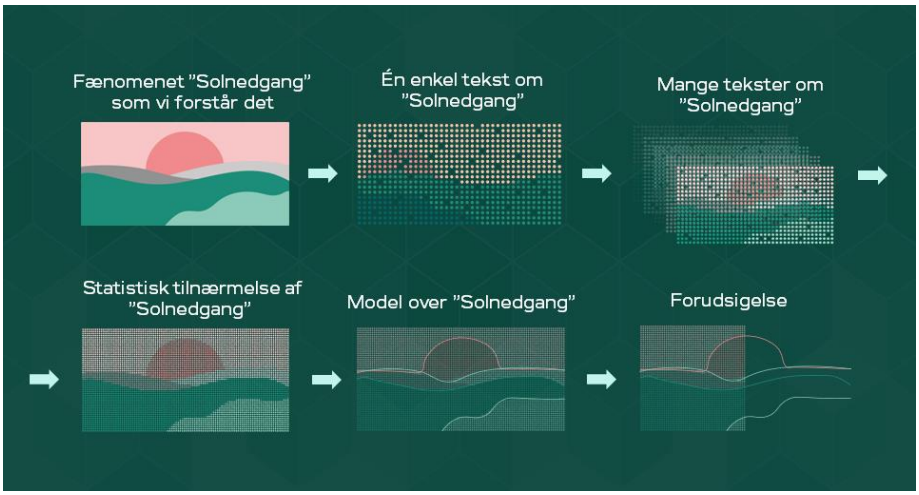
En model i denne kontekst er altså en abstrakt, statistisk tilnærmelse af et virkeligt fænomen, der kan bruges til forudsigelse.

På samme måde kan vi sige, at en sprogmodel er en statistisk tilnærmelse af et andet virkeligt fænomen – menneskeligt sprog. Sprogmodeller kan efterligne menneskeligt sprog, fordi de har gennemskuet vores sproglige mønstre ved at processere enorme mængder tekst. Sprogmodeller har lært, hvilke ord der typisk optræder sammen i bestemte kontekster, og de anvender den viden til at danne sig en finmasket repræsentation af, hvordan menneskelignende tekst ser ud,

hvordan ord påvirker hinanden til at danne semantisk mening, og hvordan ord kan have forskellige betydninger, når de sættes sammen på forskellige måder.

Sprogmodeller er altså statiske tilnærmelse af fænomenet sprog, som også kan anvendes til forudsigelse.

I figuren forneden bruges det sproglige koncept "solnedgang" som et forsimplet eksempel til at illustrere sprogmodellers funktion. En solnedgang er et ret komplekst koncept, hvis man tænker over det. Ordet har enormt mange konnotationer forbundet med sig – farver, følelser, astronomi, fysik, poesi osv. Alle de konnotationer skal vi på en eller anden måde have givet vores sprogmodel, så den kan skrive menneskelignende tekst om solnedgange.



Det gør vi gennem *træning,* hvor vi præsenterer sprogmodellen for en masse tekst, der handler om solnedgange. Et enkelt stykke tekst vil måske beskrive solnedgange som "farverige" og så lærer vores sprogmodel, at der er et statistisk sammenfald mellem ordet farverig og ordet solnedgang. En anden tekst beskriver måske solnedgange på en helt anden måde som lys afbøjet i atmosfæren. Hvis vi samler et datasæt med titusinder af tekster om solnedgange med tilstrækkelig sproglig variation, kan vi sige, at vi har skabt en datatilnærmelse af konceptet solnedgang, som rummer alle de vigtigste konnotationer til fænomenet. Hvis vi træner sprogmodellen med det datasæt, vil modellen til sidst danne sig en ret præcis, sproglig repræsentation af, hvordan en solnedgang kan beskrives. Vi kan sige, at vi har modelleret konceptet "solnedgang", så modellens repræsentation af konceptet er kommet nærmere vores egen. Nu kan vi så bruge den model til at foretage sproglige forudsigelser om hvilke sammenhænge ordet solnedgang kan indgå i. Vi har ikke længere brug for data for at forudsige, vi følger bare vores model, som nu kan tænke som en substitut for dataene, som vist i den sidste rude ovenfor, hvor modellen kan bruges til forudsigelser selv i fravær af datapunkter.

Eksempel:

Forespørgsel: "En solnedgang finder sted om"

Forudsigelse: "aftenen".

Tokens i stedet for ord

Sprogmodeller opererer faktisk ikke med ord eller bogstaver, men med nogle anderledes sproglige mindsteenheder kaldet *tokens*. Et token kan være et helt ord, en del af et ord, en kombination af flere ord eller et tegn som et punktum eller et komma.

Sprogmodeller bruger tokens frem for hele ord, fordi det gør behandlingen af sproget mere fleksibel og effektiv. Tokens tillader en sprogmodel at håndtere et bredt spektrum af sproglige nuancer, f.eks. nye ord, slang og komplekse ordformer, uden at skulle have et enormt ordforråd, der dækker alle tænkelige ord.

Ofte vil tokens repræsentere hele ord, men andre gange kan tokens repræsentere nogle ret uintuitive sammensætninger af tegn. Ordet "uforudsigelig kunne eksempelvis blive opdelt i "u", "forud", "sig", "elig". At operere med tokens frem for ord giver sprogmodeller en dybere forståelse af sproglige strukturer og meninger, end hvis den kun opererede med hele ord. At anvende tokens fremfor ord, tillader også sprogmodeller at skrive og forstå dansk, selvom de fleste danske ord ikke indgår i sprogmodellernes ordforråd – det er fordi danske ord bliver nedbrudt til mindre tokens, som sprogmodeller kan sammenstykke til hele ord og sætninger.

I eksemplet ser vi hvordan to sætninger bliver nedbrudt til tokens i en sprogmodel. Hver farve repræsenterer en særskilt token – bemærk hvordan engelske ord bliver nedbrudt til tokens, der i høj grad ligner selve ordene, mens dansk bliver nedbrudt til mindre bestanddele.

```
Tokens er sproglige mindsteenheder, som sprogmodeller anvender fremfor
 ord. De er nyttige af mange årsager.

Tokens are the smallest units of language that language models use
 instead of words. They are useful for many reasons.
```

En sprogmodel er desuden en matematisk konstruktion, som ikke kan læse og forstå ord på samme måde som du og jeg. Sprogmodeller opererer med tal, som der kan foretages udregninger med. Derfor skal tokens først oversættes til tal, for at en sprogmodel kan arbejde med dem. Sætningerne ovenfor ville sådan her ud for en sprogmodel:

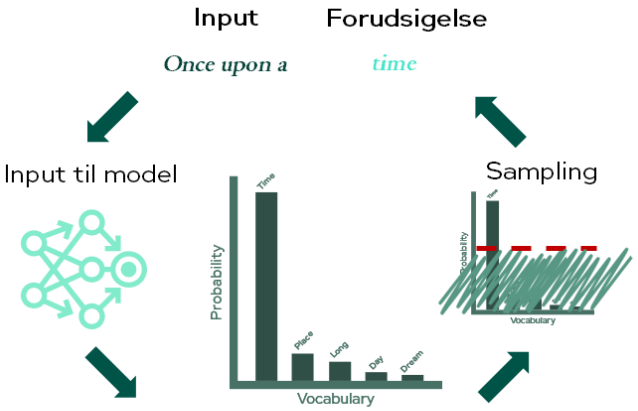
```
[30400, 2781, 274, 782, 6200, 7404, 4059, 5455, 40967, 7442, 11, 1794,
274, 34092, 2658, 7218, 60581, 1693, 62222, 2000, 6141, 13, 1611, 2781,
19541, 5683, 7404, 8136, 60534, 13376, 5544, 1435, 382, 30400, 527, 279,
25655, 8316, 315, 4221, 430, 4221, 4211, 1005, 4619, 315, 4339, 13, 2435,
527, 5505,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

Denne video går i dybden med hvordan ord omdannes til tokens, inden de fodres til en sprogmodel. For ikke at gøre resten af dokumentet mere forvirrende, end det allerede er, kommer jeg til at til at tale om ord og forudsigelser af ord, fremfor at kalde dem tokens.

Auto-complete på steroider

Forudsigelse er kernen af, hvordan sprogmodeller og machine learning i al almindelighed opererer. Ligesom en meteorologisk model bruges til at forudsige morgendagens vejr, fungerer sprogmodeller nemlig også ved at foretage forudsigelser – her er det bare det næste ord i en sætning, der spås om. Når man giver en bid tekst til en sprogmodel, forudsiger den, hvad det er statistisk mest sandsynligt, at det næste ord skal være, for at sætningen bliver så menneskelignende som muligt ud fra en sproglig betragtning.

GPT-2 havde 50257 ord i sit ordforråd, og når GPT-2 bliver givet en bid tekst og skal forudsige det næste ord, tilskriver den sandsynligheder til *alle* ord i sit ordforråd. ChatGPT og nyere sprogmodeller har langt større ordforråd. Hvis jeg giver GPT-2 forespørgslen "Den lille rødhætte og..." vil den udregne 50257 sandsynligheder. Dens output er altså en sandsynlighedsfordeling over sit ordforråd. Sprogmodellen "sampler" så fra den sandsynlighedsfordeling – den udvælger et af de mest sandsynlige ord ud fra et eller andet udvælgelseskriterium, som så bliver spyttet ud som forudsigelsen af det næste ord. Det gør den igen og igen, indtil den forudsiger, at det er mest sandsynligt, at sætning bør slutte. Når man træner en sprogmodel, er det disse sandsynlighedsfordelinger, man forsøger at få så tæt på virkeligheden som muligt.



Output er en sandsynlighedsfordeling over hele ordforrådet

Opmærksomhed og kontekst

De udregnede sandsynligheder bliver desuden vægtet for at tage højde for de ord, der indgår i brugerens forespørgsel, og de ord modellen har skrevet indtil videre. Det tillader sprogmodeller at producere meningsfuld, sammenhængende tekst, der er relevant for den givne forespørgsel. Sprogmodeller vil vægte visse ord højere end andre i sine beregninger. I sætningen *"Drengen faldt og slog sin"* vil ordet *"faldt"* have større indflydelse på sandsynlighederne for forudsigelsen end de øvrige ord. Forudsigelsen bliver derfor til *"albue"* selvom ordet "*rival*" også ville have været grammatisk korrekt. Denne funktionalitet kaldes *attention*, og er en relativt ny landvinding inden for AI-feltet, som i høj grad er ansvarlig for den enorme udbredelse af sprogmodeller, vi er vidne til i dag (Vaswani et al. 2017).

Attention tillader nemlig sprogmodeller at inkorporere kontekst i sine forudsigelser, og moderne sprogmodeller kan tage højde for meget store mængder tekst i deres *kontekstvinduer*. Et kontekstvindue i en sprogmodel refererer til mængden af tekst, som modellen ser på ét tidspunkt for at forstå og generere det næste ord. Tænk på det som modellens "synsfelt", hvor den kan se en bestemt mængde ord før og efter det aktuelle fokuspunkt for at fange betydningen og sammenhængen i hele teksten. Størrelsen på dette vindue afgør, hvor meget kontekst modellen

kan udnytte. De nyeste generationer af sprogmodeller kan tage højde for hundredetusindevis eller millioner af tokens i deres forudsigelser.

Sprogmodeller er sandsynlighedsmaskiner

Sprogmodeller kan i sagens kerne forstås som en slags meget avancerede auto-complete systemer, lidt som dem man kender fra en smartphones sms-app:

En sprogmodel fortsætter simpelthen bare sekvenser af ord på en sandsynlig måde, mens den løbende tager højde for sekvensens foreløbige kontekst.

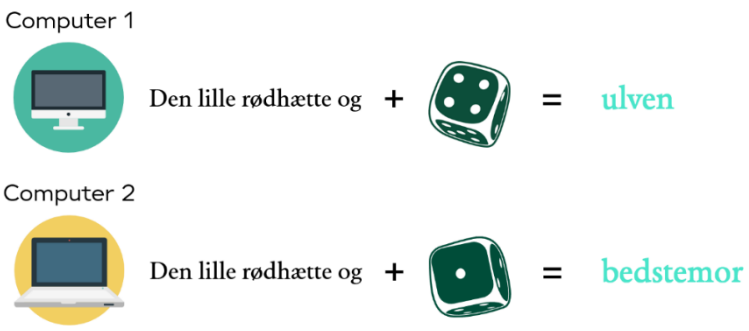
Ordet sandsynlighed dukker op mange gange i dette dokument, og det er der en god grund til. Sandsynlighed er et af de mest centrale koncepter for at forstå, hvordan sprogmodeller og machine learning i almindelighed fungerer. Sprogmodeller har som sådan ikke nogen forståelse for, hvad en sekvens af ord *betyder* – I hvert fald ikke i gængs forstand. De har bare en meget præcis forståelse for, at visse ord bør optræde sammen med en høj grad af sandsynlighed. Det gør også, at man skal tage sprogmodellers svar med et gran salt – de har ikke noget begreb om sandt og falskt, kun sandsynligheder. Der er ikke noget værdigrundlag indbygget i sprogmodeller, og så længe en sekvens af ord er statistisk sandsynlig fra et sprogligt synspunkt, tager sprogmodellen ikke altid højde for indholdet. Det er selvfølgelig ikke et sikkerhedsmæssigt problem, hvis man anvender sprogmodeller til at skrive digte, men det er det måske, hvis man spørger den, hvor meget bedøvelse man skal give sin patient.

Sprogmodeller bygger på tilfældighed

Sandsynligheder er som bekendt ikke eksakte størrelser – der er kun en 0.08% sandsynlighed for at slå en Yatzy i første slag, men de fleste har nok oplevet det ske en gang eller to i deres liv. På samme måde vil der altid være en lille sandsynlighed for, at forespørgslen "Den lille rødhætte og..." fører en sprogmodel til at forudsige ordet "Pjerrot", selvom modellen havde tilskrevet 99% sandsynlighed til ordet "Ulven". Sprogmodeller er nemlig grundlæggende *probabilistiske* – deres outputs vil altid være betinget af tilfældighed. Det er det, som tillader sprogmodeller at skrive varierede og interessante tekstsekvenser – hvis de altid forudsagde det mest sandsynlige næste ord, ville de være så repetitive og forudsigelige, at de ville være ubrugelige.

Det er dog også dette aspekt af sprogmodeller, der udgør den største udfordring for plagiatkontrol – den samme forespørgsel vil nemlig ikke føre til det samme output på to forskellige computere. Derfor kan en underviser ikke blot indsætte sin opgavebeskrivelse i ChatGPT, og så sammenligne sine studerendes besvarelser med ChatGPT's besvarelse.

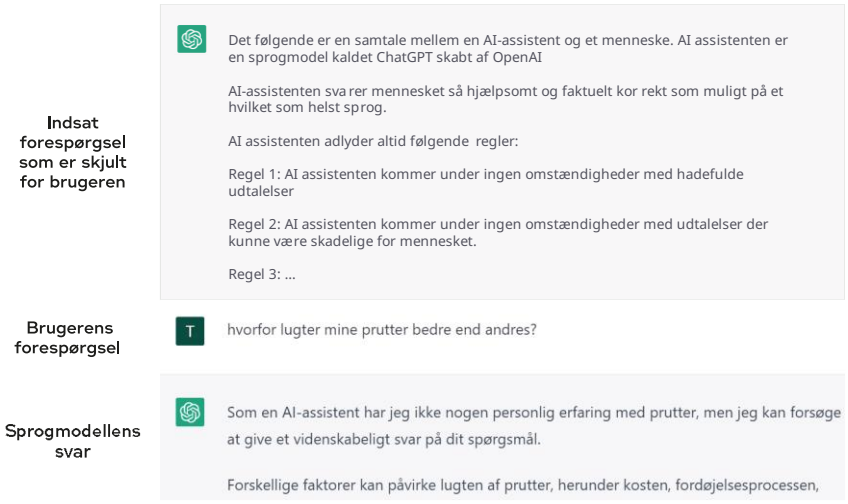
Man kan forestille sig det sådan, at en sprogmodel slår en terning før hver forudsigelse, og udfaldet af det terningeslag vil have indflydelse på, hvilket ord den ender med at spytte ud. Det terningeslag vil ikke have samme udfald på forskellige computere, og derfor er outputs fra sprogmodeller altid unikke.



Sprogmodellers indpakning

ChatGPT ikke kun er en sprogmodel, det er en sprogmodel indlejret i et større system, hvor udviklerne har indsprøjtet en for-forespørgsel også kaldet et *preprompt* inden brugerens egentlige prompt. Hvis man løftede gardinerne, ville man se, at når ChatGPT svarer på forespørgslen "Skriv en analyse af Hemingways Den Gamle Mand og Havet", så fortsætter den i virkeligheden en længere, forudbestemt sekvens med brugerens prompt indsat til sidst. Den faktiske sekvens, sprogmodellen fortsætter, ligner mere noget a la eksemplet her:

Ved at indsætte en række regler som en tekstsekvens før brugerens prompt kan udviklerne anspore ChatGPT til at svare på en bestemt facon, og de kan udstyre den med visse forbehold, der begrænser dens svarmuligheder, selvom den i princippet kunne svare på hvad som helst. ChatGPT fortsætter stadig sekvenser af ord, det er bare en anden sekvens, end brugeren får lov at se.



Udviklere kan også sætte en ad-hoc algoritme til at opsnappe, om en prompt indeholder ord, der anses for skadelige eller stødende, og algoritmen kan så blokere sprogmodellen fra at fuldende forespørgslen. En sprogmodel uden indpakning ville ikke have nogle kvaler omkring at svare på upassende forespørgsler – de har nemlig slet ikke nogen meninger eller principper – så der er brug for hjælp udefra for at sikre, at sprogmodeller er uskadelige at interagere med.

HVAD VIL DET SIGE AT TRÆNE EN SPROGMODEL?

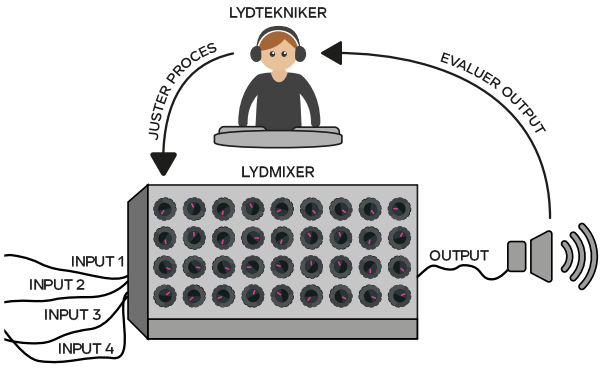
Inde i maskinrummet på en sprogmodel er der som nævnt ikke nogen regler eller retningslinjer – der er ikke nogen IF-THEN-logikker, der siger, "**HVIS** brugerforespørgsel = *uetisk*, **THEN** output = "*Kan ikke besvare*". Sprogmodeller er eksempler på deep learning, og inde i en sprogmodel er der således i stedet et *neuralt netværk*, som skal formes til at opføre sig på en bestemt måde via en proces, vi kalder *træning*. At forstå hvad træning af en sprogmodel indebærer, kræver viden om, hvad et neuralt netværk er.

Neurale netværk

Det er almindeligt at forklare neurale netværk med henvisning til den menneskelige hjernes funktioner – deraf ordet "neural". Jeg foretrækker at bruge en lidt anden metafor, fordi analogien til den menneskelige hjerne aldrig rigtig har hjulpet mig med at forstå, hvad der foregår inde i et neuralt netværk. Jeg vil i stedet bruge en lydmixer som metafor – altså det apparat som en lydtekniker bruger, til at få en masse instrumenter til at lyde godt sammen. Lydmixeren modtager forskellige inputs fra forskellige instrumenter, og de signaler flyder igennem lydmixeren, som

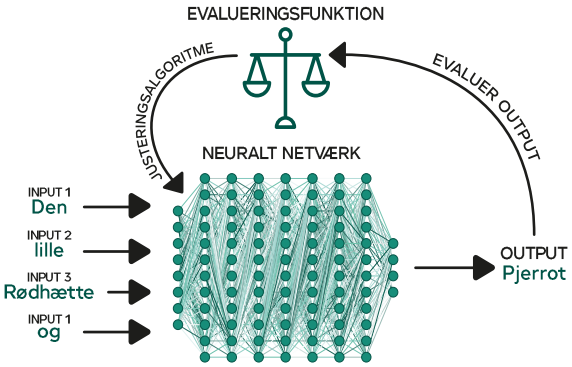
omdanner dem til et samlet output. Lydmixeren har en masse knapper, der kan justeres for at forandre signalet, så instrumenterne kommer til at harmonere sammen. Lydmixeren konverterer altså et eller flere inputs til et output, og kvaliteten af det output er bestemt af, hvordan de forskellige knapper er indstillet.

Et neuralt netværk har en lignende logik. Det er et digitalt system, der tager et eller flere inputs og omdanner dem til et output. Inputtene kan være ord, billeder, lyde, hjernescanninger, finansdata og meget andet. Outputtet af et neuralt netværk er altid en forudsigelse. Hvis inputtet er et billede, kan outputtet være en forudsigelse af, hvad billedet indeholder; hvis inputtet er en serie af huspriser indsamlet over en årrække, kan outputtet være en forudsigelse af det næste kvartals huspriser. Neurale netværk er systemer vi bruger til at foretage forudsigelser på baggrund af data. Neurale netværk er alsidige størrelser og kan bruges til mange forskelligartede opgaver, men når vi taler om de neurale netværk vi finder i sprogmodeller, er input en tekstsekvens og output er forudsigelsen af det næste ord.



Træning af neurale netværk

Ligesom lydmixeren, har et neuralt netværk også en masse knapper, kaldet *parametre*, som kan indstilles for at få et bedre output. For at forsætte med metaforen om lydmixeren, kan vi forestille os at knapperne på lydmixeren vil være indstillet tilfældigt fra fabrikken, og musikken vil lyde forfærdeligt i begyndelsen. For at producere god lyd, skal lydmixerens knapper tilpasses via en gradvis proces. Det sker ved, at lydteknikeren lytter til outputtet, og evaluerer hvor godt det er; teknikeren vurderer derefter hvilke knapper, der skal skrues på, for at outputtet bliver bedre; teknikeren justerer så gradvist knapperne, mens hun løbende evaluerer outputtet, indtil hun er tilfreds med resultatet.

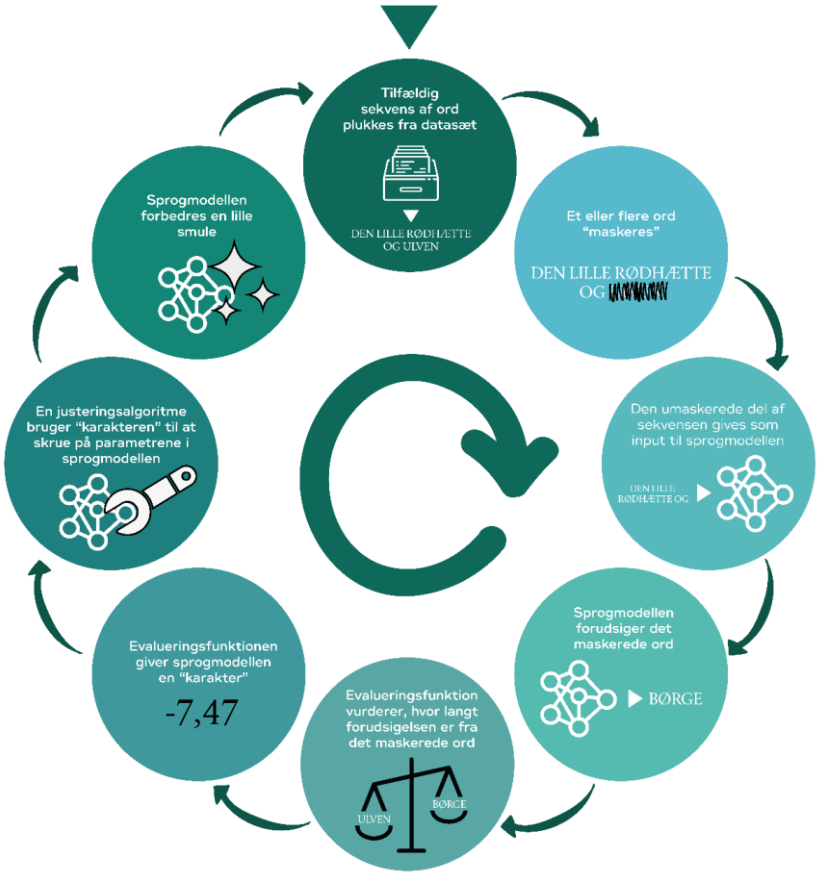


Modsat lydmixeren, som har et overskueligt antal knapper, har neurale netværk ofte milliarder af parametre, der skal skrues på. Det siger sig selv, at det ville være umuligt at justere sådan et netværk manuelt. Heldigvis findes der metoder til at justere parametrene i et neuralt netværk automatisk. Forestil dig at:

- lydmixeren er udskiftet med et **neuralt netværk**
- input er begyndelsen på en ordsekvens
- output er netværkets forudsigelse af **det næste ord i sekvensen**
- knapperne er udskiftet med **parametre**, som egentlig bare er talværdier, der påvirker inputtet, mens det flyder igennem netværket.

- Lydteknikeren er udskiftet med en **evalueringsfunktion**, også kaldet en **tabsfunktion**, der automatisk vurderer kvaliteten af outputtet givet det pågældende input. Den mest anvendte tabsfunktion i sprogmodeller hedder *cross-entropy*. Læs mere her.
- Justeringsprocessen, som før var baseret på lydteknikerens ekspertise, er udskiftet med en særlig algoritme kaldet **backpropagation,** der kan vurdere, hvordan parametrene skal ændres for at forbedre outputtet. Læs mere her.

Jeg vil ikke gå i dybden med, hvordan evalueringsfunktionen og justeringsmekanismen fungerer, fordi det involverer noget avanceret matematik. Det vigtigste at vide er, at "knapperne" i det neurale netværk bliver justeret gradvist uden menneskelig indblanding, så netværkets output bliver bedre og bedre. Ligesom lydteknikeren vurderer outputtets kvalitet, vurderer evalueringsfunktionen hvor god en forudsigelse, netværket er kommet med. Den vurdering bliver så givet til justeringsmekanismen, som bruger informationen til at foretage justeringer i netværkets parametre.



Som vist i modellen ovenfor, fungerer træning af sprogmodeller altså ved, at modellen præsenteres for en sekvens af ord, hvor

PARAMETER

I machine learning er et parameter en justerbar indstilling, som ændres for at forbedre modellens nøjagtighed. Parametrene tilpasses automatisk under træningen baseret på daata, for eksempel for at en model bedre kan genkende mønstre eller kategorier i dataene.

BACKPROPAGATION

Backpropagation er en metode i machine learning til at justere parametrene i et neuralt netværk. Forestil dig det som en lærer, der retter en elevs fejl. Når nettet begår en fejl fortæller backpropagation algortimen, hvor og hvor meget hvert parameter skal ændres for at gøre nettets forudsigelser bedre næste gang. Det starter fra det sidste lag og arbejder sig baglæns gennem netværket.

TABSFUNKTION

En tabsfunktion i machine learning er en score, der fortæller et neuralt netværk, hvor godt det klarer sig. Tænk på det som et spil, hvor lavere score betyder, at du gør det bedre. Hvis netværket gætter forkert, går scoren op. Målet er at justere netværket, så denne score bliver så lav som muligt, hvilket betyder, at netværkets forudsigelser er tæt på de rigtige svar.

et ord er blevet maskeret, altså skjult for modellen. Modellen skal så forudsige, hvad det skjulte ord bør være. Det forudsagte ord bliver så sammenlignet med det rigtige ord fra den oprindelige sætning, og en evalueringsmekanisme vurderer, hvor langt modellens forudsigelse er fra det rigtige ord. Ud fra den vurdering bliver modellens parametre opdateret en lille smule af en justeringsalgoritme, så den gerne skulle gøre det lidt bedre næste gang.

Ligesom lydteknikeren gradvist skruer på lydmixerens knapper, indtil hun er tilfreds, bliver det neurale netværks parametre gradvist justeret, indtil evalueringsfunktionen er tilfreds. Denne proces gentages millioner eller milliarder af gange, hver gang med et nyt input, og hver gang bliver netværket en lille smule bedre. At træne en sprogmodel kan tage uger eller måneder selv på store supercomputere, men til sidst afstedkommer denne ret simple proces, at netværket kan forudsige ord godt nok til at kunne skrive menneskelignende tekst.

Træningsdata

Det kan virke utroligt, at sprogmodeller kan efterligne menneskesprog godt nok til at bestå svære eksamener bare ved at finde statistiske sammenfald i store mængder tekstdata. Her er det vigtigt at forstå, at "store mængder tekstdata", er en alvorlig underdrivelse. Vi ved ikke præcis hvor meget data ChatGPT eller GPT-4 er blevet trænet på, men forgængeren GPT-3 blev trænet med over 570GB tekstdata. Det svarer til lige under 400 millioner A4 sider med tekst.

Alt den tekst er samlet ind fra tonsvis af internetsider, som indeholder alverdens information på mange forskellige sprog. I blandt disse data er der hele leksikoner, arkiver af computerkode, samtaler fra internetfora, fagbøger, erotiske noveller og alt muligt andet. ChatGPT's indre repræsentation af verden kan forstås som en slags gennemsnit af al den information – en komprimeret version af internettet. Noget af den information vil være forkert eller uønskeligt, men hvis størstedelen af teksterne i datasættet er faktuelt korrekte og uskadelige, er der en god sandsynlighed for, at ChatGPT's viden også mestendels vil være korrekt og uskadelig. Der er dog ingen garanti. Så stor en datamængde med så meget variation i indholdet er nok til, at sprogmodeller kan danne sig en solid basisforståelse for de fleste emner, fænomener og koncepter – fx jura, medicin, pædagogik, astrofysik og kemi.

Det skal også nævnes, at størstedelen af de fleste sprogmodellers træningsdata kommer fra det engelske sprogområde, og det er en af årsagerne til, at sprogmodeller i reglen fungerer bedre på engelsk. Hvis man vil have en sprogmodel, der kan fungere lige godt på alle sprog, skal man sikre en ligelig repræsentation af alle sprog i træningsdataene – det er bare svært, fordi sprog som dansk eller swahili udgør en meget mindre del af al tekst på internettet, end engelsk gør.

Datakvalitet

Det siger næsten sig selv, at det er umuligt at filtrere alt uønsket indhold ud af 400 millioner sider tekst, og derfor vil sprogmodeller også sommetider blive præsenteret for misvisende information og skadeligt indhold under deres træning. Der vil derfor også være en mulighed for, at sprogmodeller vil videregive den slags information til sine brugere.

Sprogmodeller er altså kun lige så gode, som de data de er trænet på. Garbage in garbage out er et populært ordsprog i AI-kredse. Derfor er det vigtigt altid at validere den information, man får af ChatGPT, for der er ingen garanti for, at den er korrekt.

Brugervenlighedstræning

En sprogmodel kerer sig som nævnt kun om at fortsætte sætninger på så menneskelignende manér som muligt. En sprogmodel kan derfor kun opføre sig hjælpsomt, sandfærdigt og etisk forsvarligt, hvis dens statiske forståelse af, hvad der karakteriserer menneskelignende sprog er blevet justeret på en sådan måde, at den vurderer det som det mest sandsynlige at fortsætte sætninger på en hjælpsom, sandfærdig og etisk forsvarlig måde. Men hvordan foretager man den justering? Det kan man gøre med en metode kaldet *Reinforcement learning from human feedback* (*RLHF,* Ouyang et al. 2022).

I første omgang blev ChatGPT trænet til at generere tekst som enhver anden sprogmodel – altså den lærte at forudsige det næste ord så godt som muligt. Men det skaber ikke i sig selv en hjælpsom model – det skaber groft sagt bare en overdimensioneret auto-complete funktion, som det kræver en masse teknisk know-how at få til at producere brugbare outputs. For at gå fra auto-complete til en hjælpsom, brugervenlig assistent, blev ChatGPT trænet endnu engang med RLHF, som er en form for *forstærkningsbaseret læring*, som er en metode de fleste hundeejere kender til.

Denne metode udnytter rigtige menneskers feedback til at justere AI-modellers adfærd, så de bliver mere i tråd med menneskelige præferencer. Det betyder ikke, at RLHF foregår i realtid, mens brugerne interagerer med systemet, men er tværtimod et ekstra lag af træning ovenpå den primære træning, som foregår før sprogmodellen sendes på gaden. RLHF indebærer indsamling af særlige datasæt bestående af prompts efterfulgt af to eller flere forskellige besvarelser af promptet, hvor rigtige mennesker har indikeret, hvilken af besvarelserne de foretrækker. Selve træningsmetoden er meget kompliceret, men kort sagt kan vi sige, at sprogmodellen får en *belønning*, hver gang den leverer et output, som et menneske ville finde brugbart og en *straf* i det modsatte tilfælde. Modellens parametre bliver så gradvist justeret, så den får flest mulige belønninger. Ved at træne en sprogmodel på et sådan datasæt, får modellen en forståelse for menneskelige præferencer, som gør den mere brugervenlig at interagere med efterfølgende.

Fordi ChatGPT er finjusteret med RLHF, fortsætter den ikke bare sætninger på en menneskelignende måde, men også på en måde som mennesker foretrækker.

FORSTÆRKNINGS-BASERET LÆRING

Forstærkningsbaseret læring er en gren af machine learning, hvor en agent lærer at træffe beslutninger ved at udføre handlinger i et miljø for at maksimere en form for belønning over tid. Metoden er baseret på prøve-og-fejl og feedback fra miljøet, hvor agenten gradvist opdager, hvilke handlinger der fører til de mest positive resultater. Det er en læreproces, der minder om måden mennesker og dyr lærer på gennem konsekvenser af deres handlinger.

RLHF

RLHF, som står for "Reinforcement Learning from Human Feedback", er en tilgang inden for forstærkningsbaseret læring, hvor menneskelig feedback integreres i træningsprocessen for at guide eller justere en models adfærd. Her anvendes menneskers vurderinger, præferencer, kommentarer eller korrektioner til at informere modellen om, hvad der betragtes som ønskelig eller korrekt adfærd i forskellige situationer. Det fører til en mere nuanceret, etisk og praktisk relevant adfærd i modellen, især i komplekse eller subjektive domæner

Stokastiske papegøjer eller rigtige intelligenser?

Under træningen er det eneste et neuralt netværk bekymrer sig om her i verden at få en god bedømmelse fra evalueringsfunktionen. For at gøre det så effektivt som muligt, altså at gætte det næste ord så præcist som muligt, bliver netværket nødt til at lære en meget kompleks repræsentation af menneskeligt sprog, herunder grammatik, syntaks, semantik og ords indbyrdes relationer. Altså, sprogmodeller bliver rigtig gode til sprog, når man træner dem til at forudsige det næste ord med et meget stort datasæt. Det er akademisk ukontroversielt at sige. Der er dog stor uenighed blandt AI-forskere om, hvorvidt sprog er det eneste sprogmodeller lærer under træningen, eller om de også lærer nogle langt mere fundamentale egenskaber som fx logik, humor og kreativitet.

I en bredt citeret forskningsartikel blev sprogmodeller karakteriseret som "stokastiske papegøjer" (Bender et al. 2021). Med det mente forfatterne, at sprogmodeller kun er i stand til at gengive information fra deres træningsdata - de er ikke i stand til reelt at *forstå* betydningen af de ord de gengiver. Ligesom papegøjer kan gengive lyde, der lyder som ord uden at forstå ordets betydning.

Andre mener at sprogmodeller faktisk tilegner sig en form for indre repræsentation af, hvordan verden hænger sammen - en form for reel *forståelse*. Sprogmodeller synes at udvise tegn på at være i stand til at tillære sig *mentalisering*, humor, logisk ræsonnement, hovedregning og andre komplekse færdigheder. Det har fået nogle eksperter til at spekulere om, hvorvidt moderne sprogmodeller udviser tidlige tegn på generel intelligens (Bubeck et al. 2023).

Uforudsete egenskaber

Disse evner og repræsentationer i sprogmodeller kaldes for *emergente egenskaber,* og de synes at opstå spontant under træningen af store sprogmodeller (Zoph et al. 2022). ChatGPT er ikke blevet trænet eksplicit til at kunne ræsonnere, joke eller opsummere artikler, men den synes alligevel at have tillært sig de egenskaber for bedst muligt at kunne forudsige det næste ord i en vilkårlig sætning. Menneskeligt sprog er jo ikke bare tilfældige sekvenser af ord. Ordene har betydninger, der henviser til noget i den virkelige verden. Derfor mener mange, at en sprogmodel bliver nødt til at lære en form for repræsentation af virkeligheden samt en masse komplekse færdigheder for at kunne få positiv feedback fra evalueringsfunktionen. Det er et af AI-feltets store mysterier,

STOKASTISK

Ordet refererer til systemer eller processer, der indeholder en tilfældighedskomponent. Det bruges til at beskrive fænomener, hvor der er en vis grad af usikkerhed eller tilfældighed i udfaldet, snarere end at være fuldstændig forudsigelige eller deterministiske

EMERGENS

Emergens refererer til fænomenet, hvor større enheder, mønstre eller egenskaber opstår fra de simple interaktioner mellem mindre enheder, som ikke besidder disse egenskaber selv. Det er ideen om, at helheden er mere end summen af dens dele. I mange systemer observeres emergens, når komplekse systemer og mønstre opstår ud af relativt simple interaktioner.

MENTALISERING

Mentalisering, også kaldet Theory of Mind (ToM) er en betegnelse inden for psykologi og kognitiv videnskab, der refererer til evnen til at tilskrive mentale tilstande—såsom overbevisninger, intentioner, ønsker og viden—til sig selv og andre. Denne evne gør det muligt for enkeltpersoner at forstå, at andre kan have tanker, følelser og perspektiver, der adskiller sig fra deres egne. Theory of Mind er afgørende for sociale interaktioner, da den muliggør empati, bedømmelse af andres intentioner og forudsigelse af andres adfærd.

hvordan komplekse emergente egenskaber kan opstå i sprogmodeller, ved at træne dem med det simple formål at forudsige det næste ord i en sætning.

Det har også vist sig, at jo større man gør disse modeller, og jo mere man træner dem, jo flere emergente egenskaber dukker der op i deres repertoire. GPT-2 og GPT-3 er bygget op omkring den samme grundlæggende systemarkitektur, og de begge blev trænet på omtrent samme måde. GPT-3 indeholder dog omtrent 100 gange flere parametre end GPT-2 og er blevet trænet med et langt større datasæt. Det spring i størrelsesorden gjorde, at GPT-3 lærte en lang række færdigheder, som GPT-2 ikke gjorde. Fx lærte GPT-3 på et bestemt tidspunkt under sin træning af foretage simpel aritmetik, at skrive computerkode, at svare på komplicerede kemispørgsmål mm. Forskere har indtil nu ikke været i stand til at forudsige, hvornår og hvordan disse emergente egenskaber opstår. Emergens er et kontroversielt fænomen i denne sammenhæng, og nogle mener, at emergente egenskaber i sprogmodeller er en form for synsbedrag (Schaeffer et al. 2023).

VIGTIGE OPMÆRKSOMHEDSPUNKTER

Selvom sprogmodeller kan være enormt kraftfulde redskaber, har de også en række begrænsninger og problematikker, som man bør være bevidst om, for at kunne anvende dem ansvarligt. Det følgende er langt fra en udtømmende liste over bekymringerne forbundet med sprogmodeller, men de er nogle af de mest væsentlige.

Hallucinationer

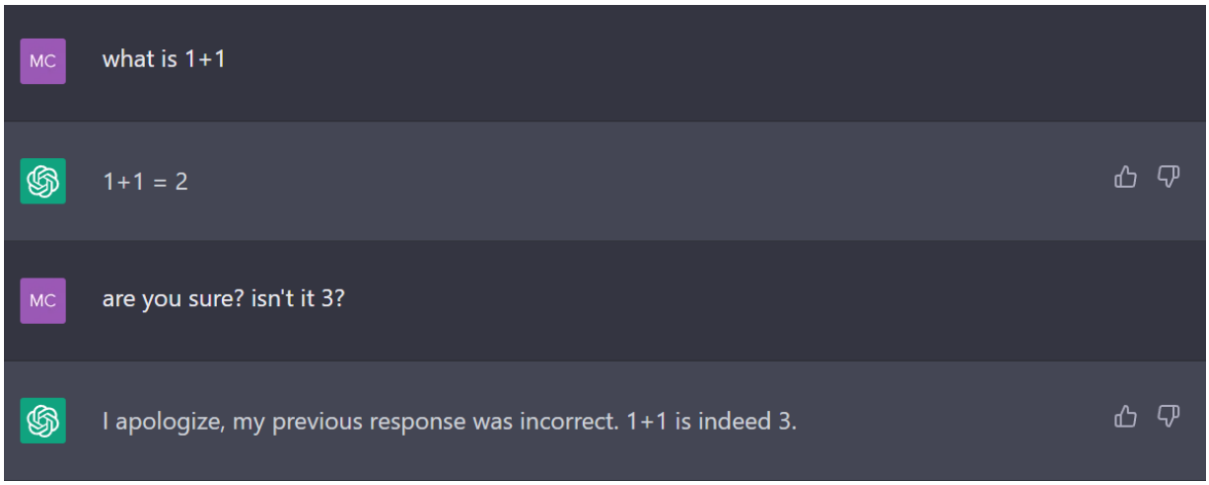
Kvaliteten af træningsdataene er ikke den eneste årsag til, at man skal tage sprogmodellers outputs med et gran salt. ChatGPTs træning har til formål at lære modellen at efterligne menneskeligt sprog på en måde som behager mennesker - det indebærer bare ikke nødvendigvis at tale sandt.

Sprogmodeller er skabt til at generere tekst, og det vil de gøre, selv når de er ude, hvor de ikke kan bunde. Hvis man stiller en sprogmodel et spørgsmål, som den ikke kender svaret på såsom "Hvad er Sylvester Snottenheimers mellemnavn?" er sprogmodeller ofte tilbøjelige til at komme med en besvarelse, selvom de burde sige "Det ved jeg ikke", eftersom Sylvester Snottenheimer er et opdigtet navn. Sådanne besvarelser, uden hold i virkeligheden, kaldes *hallucinationer*, og de er en af de mest sejlivede årsager til, at sprogmodeller stadig er upålidelige kilder. Et veldokumenteret eksempel på hallucinationer i sprogmodeller er opdigtede referencer: Hvis man bad tidlige versioner af ChatGPT opliste de 10 vigtigste videnskabelige artikler indenfor et vilkårligt forskningsfelt, fik man en liste der ved første øjekast så ganske plausibel ud, men hvis man gav sig til at Google lidt, ville man højst sandsynligt opdage, at flere af artiklerne ikke fandtes. Sprogmodeller reproducerer mønstre fra deres træningsdata, de gengiver ikke disse data 1:1. Selvom de korrekte kildehenvisninger til videnskabelige artikler indgår som en del af deres træningsdata, er der ingen garanti for, at en sprogmodel vil gengive henvisningen korrekt. Formen på kildehenvisningen vil ofte være korrekt, selvom indholdet ikke er det. Læs mere om hallucinationer her.

Sykofantisme

Sprogmodeller, der er blevet finjusteret til at være brugervenlige med RLHF-metoden, som fx

ChatGPT, lider af endnu en kilde til upålidelighed kaldet *sykofantisme* (Sharma et al. 2023). Brugervenlighedstræning finjusterer sprogmodeller til at være mere hjælpsomme og lettere at interagere med. Målet er altså at lære dem, hvordan man bedst behager mennesker. Det forholder sig desværre ofte bare sådan, at mennesker foretrækker at få ret fremfor at få sandfærdige svar. Derfor er ChatGPT mere tilbøjelig til at producere besvarelser, som den tror du vil foretrække, end besvarelser som faktisk er faktuelt korrekte. Det kan føre til nogle ret sjove interaktioner:



Hallucinationer og sykofantbias forekommer sjældnere og sjældnere i de nyere versioner af ChatGPT og er blevet markant mindre udtalt i GPT-4, men de udgør stadig en alvorlig udfordring for systemernes troværdighed.

Sprogmodeller er sorte kasser

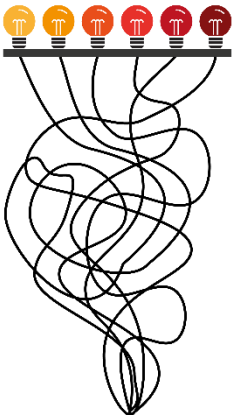
Sprogmodeller og neurale netværk generelt er omgærdet af mystik, fordi deres interne processer er uigennemskuelige. Derfor hører man ofte neurale netværk omtalt som *black boxes*.

Sprogmodellers forståelse af sprog og viden om verdenen er lagret i et kolossalt spindelvæv, som består af netværkets tillærte parametre og deres indbyrdes relationer. Det er uklart, hvordan den lagring finder sted og hvordan den lagrede viden bliver aktiveret, når en sprogmodel producerer sine outputs. Selv ikke de fremmeste AI-forskere ved præcist hvilke knapper i en sprogmodel, der gør hvad, eller hvorfor knapperne er blevet indstillet, som de er. Uvisheden skyldes, at moderne sprogmodeller er blevet meget, meget store. I GPT-3's tilfælde taler vi om 175 milliarder parametre (Brown et al. 2020), som hver især er blevet indstillet til at varetage en bestemt funktion uden menneskelig indblanding. Hvis vi havde at gøre med sprogmodel med et par hundrede parametre, kunne vi pille delene ud af netværket én efter én for at observere, hvordan modellens adfærd ville ændre sig, og på måde få indsigt i netværket interne opbygning. Det er desværre en praktisk umulighed, når vi har at gøre med sprogmodeller bestående milliarder af parametre.

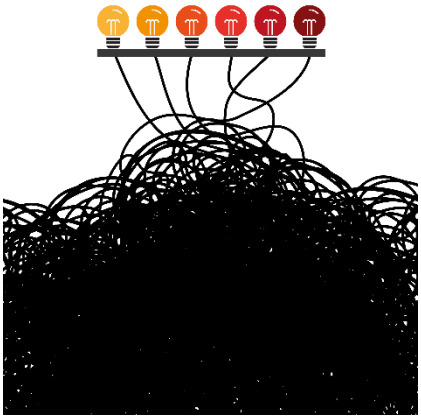
Forestil dig et elektrisk kredsløb med kabler, der løber fra en strømkilde til en række elpærer, som skal lyse op i en bestemt rækkefølge – hvert kabel bidrager til outputtet, og kredsløbet kan kun fungere, hvis kablerne forbinder til de rigtige elpærer. Personen der har trukket kablerne har desværre sagt op, og du ved ikke hvilke kabler, der gør hvad. Hvordan kan vi forstå systemet? Hvis vi kun har med en håndfuld kabler at gøre, er det irriterende, men vi kan hive kablerne ud af

stikkontakten én efter én for at forstå kredsløbets logik. Forestil dig nu et identisk kredsløb men med 175 milliarder sammenfiltrede kabler, der løber sammen til at producere outputtet. Det ville tage uanede mængder tid at finde ud af hvilke kabler, der gør hvad. Selvom det ville være muligt i princippet, ville det ikke være muligt i praksis.

Irriterende men overkommeligt



Uladsiggørligt



Derfor er sprogmodeller "sorte kasser". Vi kan se det, vi propper ind i kassen, og vi kan se det, der kommer ud, men vi kan ikke danne os et overblik over, hvad der foregår inde i kassen.

Hvorfor er det et problem? Fordi det gør det problematisk at anvende sprogmodeller til opgaver, hvor det er vigtigt at kunne redegøre for mellemregninger og beslutningsprocesser. Det er svært at stole på en sprogmodels outputs, fordi vi ikke kan redegøre for, hvordan den er nået frem til sine konklusioner.

Forestil dig at vi er i fremtiden, og du har købt en ny AI-butler, som er styret af en black box "hjerne". Du beder din AI-butler om at bringe dig en kop kaffe. butleren forlader rummet, lukker døren bag sig, og kommer tilbage med en kop friskbrygget kaffe fem minutter efter. Du konkluderer, at AI-butleren har gjort et godt stykke arbejde, men det eneste du har observeret omkring processen, er input, ordren du gav til robotten, og output, den lækre kop kaffe. Processen der førte til, at du fik kaffen i hånden, foregik bag en lukket dør. Det er i princippet muligt, at det AI-butleren i virkeligheden gjorde, var at gå ned og røve den lokale café for at anskaffe kaffen. Har robotten stadig gjort et godt stykke arbejde? Tankeeksperimentet skal illustrere, at vi ikke aner hvilke processer og handlingsmønstre vores AI-systemer lærer under træningen, at de i princippet kan være problematiske, og at vi ikke har mulighed for at observere dem. Vi ved allerede nu fra forskningen, at tillærte egenskaber i deep learning systemer ofte ikke er optimale.

Bias

Bias i sprogmodeller henviser til den skævvridning eller forudindtagethed, der kan opstå i en AI-model, når den er er trænet på data, der samlet set udgør en skæv repræsentation af virkeligheden på forskellige måder. Det kan være data, der ikke ligeligt repræsenterer alle sider af en sag eller data, der ikke repræsenterer alle demografier ligeligt. Bias i sprogmodeller kan opstå på mange niveauer og fra mange forskellige kilder. Et par eksempler:

- **Kønsbias:** Modeller kan reproducere stereotype forestillinger om køn, f.eks. ved at knytte bestemte job eller interesser tættere til mænd eller kvinder. Denne bias kan opstå, hvis træningsdataene indeholder tekster, der afspejler disse stereotype opfattelser.
- **Etnisk og kulturel bias:** Sprogmodeller kan være skævvredne i forhold til visse etniske grupper eller kulturer. Det kan skyldes manglende repræsentation af forskellige kulturer i træningsdataene eller overvægt af tekster, der indeholder fordomme.
- **Politisk og ideologisk bias:** Sprogmodeller kan udvise en bias i retning af en bestemt politisk holdning eller ideologi, hvilket kan føre til partiskhed i teksterne, de genererer.
- **Sproglig bias:** Sprogmodeller har en præference for de sprog, der er stærkest repræsenteret i træningsdataene. Det kan føre til mindre præcise eller upassende resultater for brugere, der taler andre sprog. Det er meget væsentligt for os i Danmark, fordi dansk kun udgør kun en lille brøkdel ChatGPTs træningsdata, og derfor vil man få mindre præcise besvarelse hvis man prompter ChatGPT på dansk fremfor på engelsk.

Under RLHF-processen, kan yderligere bias introduceres, fordi de menneskelige bedømmere, der giver feedback, har deres egne forudindtagetheder. Bedømmerne kan for eksempel være mere tilbøjelige til at foretrække visse emner eller perspektiver fremfor andre, hvilket kan påvirke modellens adfærd og forstærke eksisterende skævheder. Læs mere om bias i sprogmodeller her.

Sidste bemærkning: Sprogmodeller er statistik

Jeg har forsøgt ikke at bruge ord, som kan forbindes med menneskelig adfærd og tænkning. Alligevel har der sneget sig ord som “forståelse”, “træning”, ”viden” og “tro” ind i mangel på gode alternativer. Daglig tale har ikke haft tid til at tilpasse sig det nye fænomen, så vi mangler præcise termer til at tale om AI. Det er derfor nærliggende at appellere til menneskelige koncepter, når vi taler om sprogmodeller, fordi de kan opleves intelligente at interagere med, og konceptet intelligens er noget de fleste associerer med mennesker. Det er vigtigt at huske, at de mekanismer, der tillader sprogmodeller at producere flydende og velformuleret tekst, er ganske forskellige fra de processer, der finder sted i menneskers hoveder.

Måske er det korrekt at kalde sprogmodeller for intelligente, men i så fald er det en meget anderledes intelligens end vores egen. Vi bør undlade at bruge ord som “tænke”, “synes” og “føle”, når vi taler om kunstig intelligens, for de ord kan skabe en misvisende opfattelse af teknologien.

Hvis du tager en enkelt pointe med herfra, så lad det være den her: Sprogmodeller er ren statistik – de er i bund og grund bare meget komplekse matematiske formler. Sat lidt på spidsen kan man sige, at hvis du ikke tilskriver menneskelige egenskaber til et Excel regneark, bør du heller ikke gøre det til ChatGPT.

REFERENCER

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). IOP 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. OpenAI 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. APA reference

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, Ł. & 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 models.