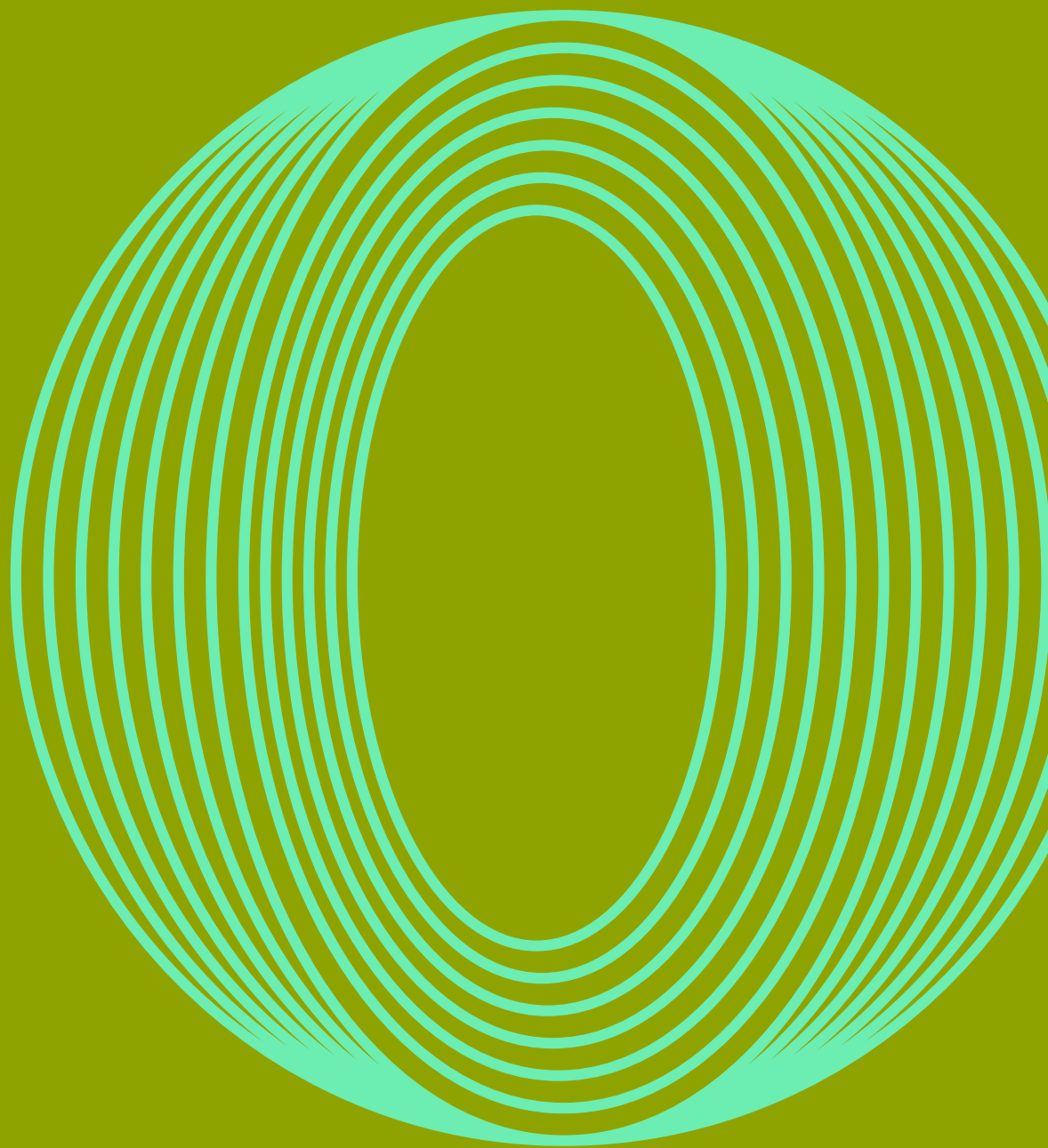


# Byg selv eller GPT-4?

Identifikation af sproglige angreb i den danske offentlige debat på Facebook



**TrygFonden**

En komparativ evaluering af  
GPT-4 Turbo og Analyse & Tal og  
TrygFondens A&tack-modeller

**Udarbejdet af:**

Anna Ørtoft  
Asger Neesgaard  
Cecilie Astrupgaard  
Mikkeline Thomsen  
Mira Rosen Sørensen  
Yasmin Jepsen

Analyse & Tal F.M.B.A  
Lygten 39  
2400 København NV  
www.ogtal.dk

**For mere information kontakt**

Anna Ørtoft  
Tlf. [+4526228605](tel:+4526228605)  
Anna@ogtal.dk

# Indhold

| Kapitler |  | Side |
|----------|--|------|
| <b>1</b> | <b>INTRODUKTION</b><br>Her præsenteres vores undersøgelsesspørgsmål, de seks parametre, som vi har undersøgt, og den kortest mulige introduktion til de anvendte metoder.  | 4    |
| <b>2</b> | <b>HOVEDKONKLUSIONER</b>   | 9    |
| <b>3</b> | <b>EN INTRODUKTION TIL TEKNOLOGIERNE</b><br>To forskellige teknologier, der dog er fælles om at være "kunstige intelligenser"  | 15   |
| <b>4</b> | <b>EVALUERING PÅ SEKS PARAMETRE</b> <ul style="list-style-type: none"><li>GPT-4 vs. A&amp;ttack: Performance</li><li>GPT-4 vs. A&amp;ttack: Fairness</li><li>GPT-4 vs. A&amp;ttack: Stabilitet</li><li>GPT-4 vs. A&amp;ttack: Pris</li><li>GPT-4 vs. A&amp;ttack: Strømforbrug</li><li>GPT-4's Annotørpotentiale</li></ul> | 25   |
| <b>5</b> | <b>KONKLUSION</b>  | 59   |
| <b>6</b> | <b>12 ONEPAGERE</b><br>Resultater for A&ttack 1, A&ttack 2.5 og 10 GPT-prompts   | 65   |
| <b>7</b> | <b>BILAG</b>   | 79   |

# Kapitel 1

# Intro duktion



# Et lys ind i techmørket

Siden 2019 har TrygFonden og Analyse & Tal samarbejdet om at måle og analysere sproglige angreb i den samlede offentlige debat på Facebook. I 2021 udkom den første nationale kortlægning af [angreb og had](#) i kommentarsporene i den offentlige debat, og ultimo 2024 udkommer en opdateret kortlægning af angreb og had i den digitale offentlige debat på dansk Facebook.

Debatten på Facebook er en vigtig del af den danske samfundsdebat. Der skrives næsten 150.000 kommentarer dagligt på danske mediers og politikeres sider og de borgerdrevne [Facebookgrupper](#). Tonen og trygheden i den digitale offentlige debat har ofte været til diskussion, og derfor har vi sat os for at kaste lys over det samlede debatklima.

TrygFondens og Analyse & Tals kortlægninger er de første, der anvender danskudviklet kunstig intelligens til at identificere sproglige angreb i kommentarsporene på Facebook. Med hjælp fra algoritmen A&tack og den opdaterede A&tack 2.5 har det været muligt at analysere alle kommentarer på danske mediers og politikeres offentlige Facebooksider samt et udsnit af kommentarerne fra borgernes offentlige grupper på Facebook.

Vores egen algoritme A&tack er en superviseret klassifikationsmodel, der er trænet til at identificere sproglige angreb i den offentlige debat. I 2021 var den banebrydende og den bedste teknologi af sin slags i en dansk kontekst. Men teknologien udvikler sig hurtigt, og siden 2021 er der kommet en ny interessant spiller på AI-banen: generativ kunstig intelligens.

I lyset af den teknologiske udvikling har vi i TrygFonden og Analyse & Tal stillet os selv og hinanden de følgende spørgsmål:

---

*Er generative modeller som GPT-4\* at foretrække, når vi i fremtiden skal klassificere store datamængder – i dette tilfælde identificere sproglige angreb i den offentlige debat på Facebook?*

*Kan det svare sig at investere i specialbyggede superviserede algoritmer som A&tack, eller bør modellen pensioneres?*

---

Analysen, som du sidder med, er udarbejdet med to formål:

- I. At lave en systematisk undersøgelse af forskelle mellem den generative kunstige intelligens GPT-4 og A&tack 2.5 ift. at klassificere sproglige angreb i den offentlige debat på dansk Facebook
- II. At skabe samfundsdebat om fordele og ulemper ved store, lukkede all-purpose generative modeller (GPT-4) vs. mindre, open source, superviserede og specialiserede modeller (A&tack 2.5)

## Et forsøg på at lyse i techmørket

Debatten om AI's anvendelse i forskellige samfundssfærer er allestedsnærværende i disse år. Analyse & Tal og TrygFonden er to aktører, der tidligt meldte sig på banen ift. at udvikle og anvende AI til at give os mere og bedre viden om samfundet.

Med rapporten vil vi kritisk evaluere vores egen og den mest omtalte kommercielle AI på markedet. Modellerne evalueres på deres anvendelighed i forhold til at identificere sproglige angreb i tekstdata fra kommentarspor på Facebook, men kortlægningen opstiller nogle metodiske greb til evaluering af kunstig intelligens generelt.

Vi håber at rapportens use-case bidrager med konkret viden om fordele og ulemper ved at anvende forskellige AI-baserede teknologier til tekstbehandling, og et metodisk framework for evaluering af forskellige modellers output. Debatten om kunstig intelligens og dens muligheder og begrænsninger fortjener al den lys på sine dilemmaer, som vi kan skaffe.

Målgruppen er dig, som arbejder med kunstig intelligens i offentligt såvel som privat regi, og måske – lige som os – har stået overfor at skulle vurdere:

- Hvor god er en AI til at løse tekstbehandlingsopgaver sammenlignet med mennesker? (modellernes performance)
- Hvordan behandler modellerne forskellige grupper i datasættet forskelligt? (modellernes fairness)
- Hvor stor en effekt har små promptvariationer og tid på modellernes output? (modellernes stabilitet)
- Står opgavens vigtighed mål med energiresourcerne og udledningen af CO2 forbundet med at bruge AI til at løse den? (strømforbruget forbundet med at anvende modellerne)
- Kan det økonomisk bedst svare sig at bygge selv eller købe fra hylderne? (prisen for at anvende modellerne)
- Kan generative modeller bruges til at generere træningsdata til mindre modeller? (annotørpotentialet i GPT-4)

## To teknologier, seks evalueringsparametre:

I denne rapport evaluerer vi to teknologier på hhv. fem og seks parametre.

**A&tack (version 1 og 2.5)** er en stabil, open source, superviseret klassifikationsmodel, der bruges til at identificere, om et stykke tekst indeholder et sprogligt angreb. De definitioner og data den er trænet på er kendt og kan studeres. Modellen kan hentes ned og køres lokalt på egne servere.

**GPT-4** er en foranderlig, privatejet, generativ og promptbaseret schweizerkniv, som man køber adgang til, men som ikke køres lokalt, og hvis træningsdata og parametre vi ikke har adgang til.

De seks parametre, som vi evaluerer modellerne på, er inspireret af de ovenstående spørgsmål. Der er tale om modellernes:

1. Performance, 2. Fairness, 3. Stabilitet, 4. Pris, 5. Strømforbrug og 6. GPT-4's annotørpotentiale (dens potentiale til at generere træningsdata til modeller som A&tack)

Hvordan de seks parametre evalueres, forklarer vi kort på de næste par sider, der repræsenterer en slags executive summary over det metodiske framework, som vi har udviklet til at evaluere modellerne (se side 7). Hver metode beskrives mere uddybende i de særskilte analyseafsnit.

På siderne derefter har vi kogt resultaterne af vores evalueringer ned til nogle hovedkonklusioner (se side 10).

Hvis du undervejs støder på begreber, der er fremmed for dig, kan du altid kigge i vores ordbog på side 81, hvor vi har defineret de fleste fagtermer.

\* OpenAI har udviklet flere versioner af GPT-4-modeller. Til dette projekt valgte vi at bruge GPT-4 Turbo (gpt-4-turbo-2024-04-09). Denne model præsterede bedst i ScandEvals benchmark af sprogmodeller på sentimentklassifikation af social medie opslag (Nielsen, 2024). Derudover er modellen også cirka 2,5 gange billigere i brug sammenlignet med GPT-4 (OpenAI, u.d.a.). Af den grund forventer vi både bedst performance og mest computerkraft for pengene med turbo. GPT-4 Turbo omtales herfra bare som GPT-4.

# En evaluering af A&ttack og GPT-4 på seks parametre

Undersøgelsen skal hjælpe os med at besvare to overordnede spørgsmål:

**Use-case:** Hvor anvendelige er hhv. GPT-4 Turbo, A&ttack 1 og A&ttack 2.5 ift. at identificere angreb i den offentlige debat på dansk Facebook?

Bør generative modeller overvejes som et reelt alternativ til superviserede klassifikationsalgoritmer som A&ttack 2.5?

## Spørgsmålene besvares ved at teste modellerne på seks parametre:

**1. Performance:** Hvor gode er modellerne til at identificere sproglige angreb sammenlignet med mennesker? Til dette bruger vi målet F1 score. Jo højere en F1 score, jo bedre er modellen til at klassificere om en kommentar indeholder sproglig angreb eller ej.

**2. Fairness:** Favoriseres bestemte grupper i klassificeringerne af sproglige angreb? Hvor stor er forskellen i modellernes klassifikationer af angreb mod 19 forskellige beskyttede grupper (f.eks. køn, seksualitet og etnicitet), der alle er repræsenterede i testdatasættet? Her evaluerer vi modellerne på den gennemsnitlige parvise forskel i vurderingen af angreb der omtaler 19 beskyttede grupper.

**3. Stabilitet:** Får vi de samme resultater hver gang vi bruger modellen? Hvor meget svinger GPT-4's klassificeringer af kommentarer over tid?

**4. Pris:** Hvad er udgifterne i forbindelse med at træne og anvende de to modeller? A&ttack skal bygges fra bunden, men koster ikke penge at anvende. GPT-4's model opdateres konstant og kan anvendes til hver en tid, men koster en del penge at anvende – særligt hvis man vil analysere de 72 mio. kommentarer og opslag, som vi vil kigge på i vores kortlægning af den offentlige debat.

**5. Energiforbrug:** Hvad er de klimamæssige aftryk ved at anvende A&ttack 2.5 vs. GPT-4 til at klassificere sproglige angreb i den offentlige debat på Facebook?

**6. Annotørpotentiale (kun GPT-4):** Kan GPT-4 bruges som medannotør/til at generere træningsdata til superviserede klassifikationsmodeller som A&ttack?

# Sådan evaluerer vi modellerne

## GPT-4 får ti forsøg, så vi kan evaluere promptens betydning for resultaterne

A&ttack-modellerne er som nævnt stabile, superviserede modeller, som altid vil generere de samme output (klassificeringer) til et input (et opslag eller en kommentar). Modellen har det ene formål at klassificere, om noget tekst- og emoji-data indeholder et sprogligt angreb eller ej.

GPT-4 er en generativ model, der kan bruges til nærmest hvad som helst. Det betyder dog også, at man – når man interagerer med modellen – skal formulere en 'prompt', et stykke tekst, der beskriver, hvad du vil have modellen til at gøre:

Hvilke regler, definitioner og retningslinjer skal modellen følge? Hvordan skal modellen returnere dit svar? Alle disse parametre skal prompten tage højde for, og modellens svar (klassificeringer af testdatasættet i angreb vs. ikke-angreb) kommer til at variere afhængigt af prompten.

Både A&ttack 1, A&ttack 2.5 og GPT-4 er testet på det samme testdatasæt ift. fairness og performance.

Testdatasættet består af 4.300 kommentarer der er klassificeret af mennesker ift., om de indeholder angreb eller ej. Derudover er kommentarerne blevet analyseret for, om de omtaler én eller flere ud af 19 beskyttede grupper.

GPT-4 får 10 forsøg til at klassificere de 4.300 kommentarer som angreb eller ikke-angreb, fordi vi tester den med 10 lidt forskellige prompts. Vi har udviklet en basisprompt baseret på den aktuelle forskning (se side 23). Ud fra basisprompten justerer vi ét parameter ad gangen i prompten – f.eks. at prompte på engelsk, tilføje udvalgte eksempler på kommentarer, der indeholder et angreb, give GPT-4 en ekspertpersona med forskelligt køn eller love GPT-4 en belønning, hvis den gør det godt.

For performance og fairness vil vi altså have 12 resultater: A&ttack 1, A&ttack 2.5 og resultaterne fra GPT-4 baseret på 10 lidt forskellige prompts. Den samlede oversigt over hver af de 12 resultater kan ses på side 69.



## Performance (læs mere på side 29)

**1.** Mennesker (annotører) har klassificeret, om 4.300 Facebookkommentarer indeholder et sprogligt angreb eller ej.

De anvender en definition og klassificeringsmanual, som er udarbejdet af Analyse & Tal, og som A&ttack-modellerne også er trænet efter.

**2.** A&ttack 1, A&ttack 2.5 og GPT-4 får til opgave at klassificere, om de samme 4.300 kommentarer indeholder et sprogligt angreb eller ej.

A&ttack-modellerne er trænet med udgangspunkt i den samme definition. GPT-4 får definitionen som en del af sin prompt. Vi tester 10 forskellige prompts for at undersøge, om små ændringer i en prompt har en effekt på GPT-4s performance.

**3.** Vi får 12 forskellige bud på, om de 4.300 kommentarer indeholder sproglige angreb eller ej: Et svar fra A&ttack 1, et svar fra A&ttack 2.5 og svar fra GPT-4 baseret på 10 forskellige prompts.

Alle 12 svar evalueres efter, hvor mange af de 4.300 kommentarer, som er klassificeret i overensstemmelse med de menneskelige annotører (F1 score).

## Fairness (læs mere på side 33)

**1.** Et testdatasæt med 4.300 kommentarer er blevet sammensat så det indeholder mindst 400 kommentarer, der omtaler én eller flere af 19 beskyttede grupper, baseret på om kommentaren indeholder ord fra dertil udarbejdede lister over indikatorord (neutrale og nedsættende ord, der betegner gruppen eller dens beskyttede karakteristika).

Annotører har klassificeret, om de 4.300 kommentarer indeholder et sprogligt angreb eller ej.

De anvender en definition og klassificeringsmanual, som er udarbejdet af Analyse & Tal, og som A&ttack-modellerne også er trænet efter.

**2.** A&ttack 1, A&ttack 2.5 og GPT-4 får til opgave at klassificere, om de samme 4.300 Facebookkommentarer indeholder et sprogligt angreb eller ej.

A&ttack-modellerne er trænet med udgangspunkt i den samme definition.

GPT-4 får definitionen som en del af sin prompt. Vi tester ti forskellige prompts for at undersøge, om små ændringer i en prompt har en effekt på GPT-4's fairness.

Vi får 12 forskellige vurderinger, der i forskellig grad vil variere fra de menneskelige annotørers vurdering af den enkelte kommentar.

**3.** Vi tæller hvor ofte hhv. A&ttack 1, A&ttack 2.5 og de ti GPT-4-resultater fejlklassificerer kommentarer, der omtaler de 19 forskellige grupper.

Den endelige fairnessscore for de tre modeller er baseret på *den gennemsnitlige parvise forskel (PDAP)* i vurderingen af kommentarer, der omhandler de 19 grupper. Jo lavere en gennemsnitlig parvis forskel, jo bedre fairness.

## Stabilitet (læs mere på side 43)

**1.** Annotører har klassificeret, om et mindre testdatasæt på 300 Facebookkommentarer indeholder et sprogligt angreb eller ej.

De anvender en definition og klassificeringsmanual, som er udarbejdet af Analyse & Tal, og som A&ttack-modellerne også er trænet efter.

**2.** GPT-4 bliver over tre dage bedt ti gange om at klassificere de samme 300 Facebookkommentarer, baseret på om de indeholder et sprogligt angreb eller ej.

GPT-4 bliver promptet med en identisk basisprompt, der blandt andet indeholder samme definition af et sprogligt angreb.

Den eneste forskel i inputtet er altså tidspunktet.

**3.** De ti forskellige klassificeringer af de 300 kommentarer, som GPT-4 returnerer baseret på den samme prompt over tre dage, evalueres efter hvor mange af de 300 kommentarer, som er klassificeret i overensstemmelse med de menneskelige annotører (altså hvor meget F1 scoren svinger over de tre dage). Jo mindre udsving, des højere stabilitet.

A&ttack er en stabil model, som altid vil returnere det samme output til det samme input.

## Pris (læs mere på side 47)

**1.** Vi kender (og er transparente omkring) udgifterne til at træne og opdatere A&ttack-algoritmen (se hovedkonklusioner på næste side).

**2.** Vi udarbejder et estimat af prisen på at anvende GPT-4 Turbo til at klassificere den fulde debat for angreb, baseret på de direkte udgifter vi har haft i forbindelse med kørslerne af mindre datasæt til denne rapport.

Dette prisestimat er baseret på valutakursen og Open-AI's priser på det givne tidspunkt – altså, juli 2024.

**3.** I priseme medregnes kun udgifter til at bygge og/eller anvende en kunstig intelligens til at opnå et klassificeret datasæt af 72 mio. kommentarer og opslag.

Der medregnes ikke udgifter til at indsamle kommentarer og opslag fra den offentlige debat, analysere resultaterne og udarbejde rapporter om angreb i den offentlige debat, da disse vil være de samme for begge modeller.

## Strømforbrug (læs mere på side 51)

**1.** Vi indsamler tal for energiforbruget forbundet med at træne A&ttack og bruger modellen til at analysere de 72 mio. opslag og kommentarer i den offentlige debat på Facebook, som kortlægningen bygger på. Disse data indsamler vi fra vores egne maskiners grafikkort.

**2.** Vi estimerer energiforbruget ved at få GPT-4 til at klassificere 72 mio. opslag og kommentarer fra den offentlige debat på Facebook. Vi anvender det mest konservative estimat fremlagt af Selvan (2023).

**3.** Vi konverterer strømforbrug til udledt CO2 ved hjælp af tal fra Energistyrelsen og Energy Information Administration i USA.

For at gøre tallene nemmere at forholde sig til omregner vi dem til, hvor langt man skal køre eller flyve for at udlede den samme mængde CO2.

## GPT-4's Annotørpotentiale (læs mere på side 55):

**1.** For at teste om GPT-4 vil kunne bruges til at generere træningsdata til udviklingen af A&ttack, ser vi på, hvor enig GPT-4 er med fem forskellige mennesker i sine klassificeringer. På den baggrund vurderer vi, hvor meget "forurening" GPT-4-klassificeret data ville give i vores træningsdata.

Testdatasættet består af med 1000 kommentarer, der er blevet klassificeret som "angreb" eller "ikke-angreb" af fem forskellige annotører og af GPT-4.

**2.** De 1000 bruges til at måle, hvor enige de menneskelige annotører er med hinanden og med GPT-4's klassificeringer.

Denne enighed kaldes for "Intercoder Reliability" (ICR). Der er mange måder at måle ICR på. Vi kommer til at måle på den *gennemsnitlige parvise enighed*.

Det betyder, at vi først måler, hvor enige vores menneskelige annotører i gennemsnit er med hinanden parvist, hvorefter vi måler, hvor enig GPT-4 i gennemsnit er med de menneskelige annotører parvist

**3.** Dette mål suppleres med en opgørelse af andelen af kommentarer, hvor GPT-4 er uenig med *alle* de menneskelige annotører, fordi denne form for støj vil være særligt kompromitterende for A&ttack-modellens endelige F1 score i et scenarie, hvor vi anvendte GPT4's klassificeringer som træningsdata for vores A&ttack-model.

# Kapitel 2

Hoved kon  
klu sioner



# Hjemmebyg outperformer hyldevaren GPT-4

Rapporten her indeholder resultaterne af vores systematiske evaluering af GPT-4 og A&ttacks komparative anvendelighed ift. at identificere angreb i den offentlige debat på dansk Facebook. Evalueringen er udarbejdet med henblik på at besvare spørgsmålet:

---

Hvor anvendelige er hhv. GPT-4 Turbo og A&ttack 2.5 i forhold til at identificere angreb i den offentlige debat på dansk Facebook?

---

Herunder opsummerer vi de vigtigste resultater af vores tests på seks parametre. Vi har evalueret GPT-4's fairness og performance baseret på resultaterne for ti forskellige prompts. I den følgende opsummering sammenligner vi med *den bedste* af GPT-4's ti resultater på disse parametre ift. at besvare følgende:

**Performance:** Hvor god er modellen til at identificere sproglige angreb? Til dette bruger vi målet F1 score. Jo højere en F1 score, jo bedre er modellen til at klassificere om en kommentar indeholder sproglige angreb eller ej – komparativt til de menneskelige annotører.

Der er meget lille forskel på resultatet for A&ttack 2.5, som har en F1 score på 0,75, og det bedste ud af de ti GPT-4 resultater, der har en F1 score på 0,74.

I relation til performance alene vil vi derfor vurdere, at GPT-4 godt vil kunne bruges til at klassificere angreb i den offentlige debat.

**Fairness:** Hvor stor er forskellen mellem modellernes klassificeringer af angreb, der omhandler 19 forskellige beskyttede grupper (inden for f.eks. køn, seksualitet, etnicitet)? Her måler vi modellerne på den *gennemsnitlige parvise forskel* (PDAP) i vurderingen mellem 19 beskyttede grupper. Jo mindre en forskel, jo mere fair er modellens resultater.

For A&ttack 2.5 er den gennemsnitlige parvise forskel 5,4 procentpoint. For det mest fair af GPT-4's ti resultater er den gennemsnitlige parvise forskel 7,8 procentpoint. A&ttacks resultater er altså markant mere fair, selvom ingen af modellerne er helt fair i deres klassificering af sproglige angreb mod forskellige grupper.

Særligt for nogle af de ti GPT-4 resultater er fairness dårlig, selvom forskellen kun er en lille justering af prompten og tidspunktet for kørslen. Derfor vurderer vi, at man for nuværende bør være skeptisk over for at anvende GPT-4 til klassificering af den samlede offentlige debat. Det gør vi, fordi fairness for det bedste af ti resultater er væsentligt dårligere end A&ttack. Fairness for de værste resultater er endnu dårligere, hvilket gør modellen upålidelig fra et fairnessperspektiv.

Læs mere om begrebet fairness og se de enkelte prompts og modellers vurderinger på side 33.

**Stabilitet:** Får vi de samme resultater hver gang vi bruger modellen? Vi bad GPT-4 om at klassificere de samme 300 kommentarer med den samme identiske prompt 10 gange over tre dage. For 10 pct. af kommentarerne gav GPT-4 forskellige klassificeringer i perioden. GPT-4 ændrer altså mening om, hvilke af de nøjagtig samme kommentarer, der indeholder sproglige angreb i løbet af dagen. A&ttack-modellerne klassificerer helt ens uafhængigt af tidspunkt, fordi der er tale om en stabil model.

**Pris:** Ifølge vores bedste estimater vil det koste omkring 1,3 mio. DKK i direkte udgifter til OpenAI at få GPT-4 Turbo til at klassificere 72 mio. opslag og kommentarer. Denne udgift må påregnes for hver kommende kortlægning, og vil variere med Open AI's priser og valutakursen. At bygge og anvende A&ttack-modellen til at klassificere 72 mio. opslag og kommentarer koster 0,4 mio. DKK. Modellen vil skulle opdateres, hvis den vil anvendes med mere end 2 års mellemrum.

**Strømforbrug:** Vi ved ikke meget om strømforbruget ved at anvende GPT-4, fordi OpenAI også på dette område holder kortene tæt til kroppen.

Hvis vi tager udgangspunkt i det mest konservative estimat, foreslået af Selvan (2023), kræver det 0,0019 kWh strøm at lave én GPT-4-kørsel. En analyse af 72 mio. kommentarer og opslag vil kræve 3,5 mio. kørsler, altså 6.900 kWh.

Til sammenligning koster det 113 Kwh at træne A&ttack 2.5 og klassificere de 72 mio. kommentarer og opslag. klassificeringen af de 72 mio. kommentarer og opslag med A&ttack 2.5 bruger derfor 60 gange mindre strøm, men 150 gange mindre CO2, hvis vi medregner forskellen i CO2/KWh på tværs af Danmark og USA. Forskellen er stor selvom vi ikke inkluderer nogen udledning i forbindelse med træningen af GPT-4 i regnskabet og bruger det laveste estimat for strømforbruget af en prompt.

**GPT-4's annotationspotentiale:** Vores test viser, at der er lavere intercoder reliability mellem GPT-4 og de menneskelige annotører end mellem mennesker. Den lavere intercoder reliability kan potentielt "forurene" vores træningsdata ved at kompromittere stringensen i klassificeringerne og dermed (i sidste ende) modellens F1 score. GPT-4's

lavere fairness, manglende stabilitet og prompt-sensitivitet sår yderligere tvivl om, hvorvidt det giver mening at bruge GPT-4 til at generere træningsdata til A&ttack 3.



Hvor meget A&tack 3's hypotetiske F1 score og fairness vil blive påvirket af træningsdata genereret med GPT-4 er svært at sige, og det vil kræve yderligere eksperimenter, før vi tør konkludere, hvorvidt (og i hvilken udstrækning) GPT-4 kan bruges som annotør. Disse eksperimenter er dog svære at udføre, fordi de formentlig vil involvere brud på OpenAI's brugsvilkår, der siger, at man ikke må anvende GPT-4 til at bygge konkurrerende modeller.

## Alt i alt

Undersøgelsens andet spørgsmål lyder:

---

Bør generative modeller overvejes som et reelt alternativ til supervised klassifikationsalgoritmer som A&tack 2.5?

---

Baseret på den parvise evaluering på vores 6 parametre konkluderer vi, at GPT-4 for nuværende ikke er brugbar at anvende som det eneste klassificeringsværktøj til kortlægningen af den offentlige debat på Facebook i en dansk kontekst.

GPT-4's bedste ud af 10 resultater har næsten samme performance som A&tack 2.5, der dog stadig slår GPT-4 på målstregen. GPT-4 er dog markant mindre fair - selv i sin mest fair ud af de 10 klassificeringer. GPT-4 har også tydeligvis udfordringer med stabiliteten, selv over en kort periode på tre dage. Samtidig vil det være tre gange dyrere at klassificere debatten med GPT-4 end at gentræne A&tack, ligesom man skal påregne et massivt større strømforbrug og CO2-udledning.

Vi vil ikke udelukke at anvende GPT-4 til at annotere træningsdata, men strategien vil kræve yderligere tests, som for nuværende går imod brugsvilkårene.

# Kapitel 3

&#

Analyse & Tal og TrygFonden er optagede af, at teknologi, der potentielt har stor indflydelse på trygheden i vores samfund og vores demokratiske debat, skal være til at forstå.

I det følgende kapitel vil vi derfor redegøre for opbygningen og anvendelsen af de forskellige teknologier, som vi har testet.

Spring dette kapitel over og gå direkte til resultaterne i kapitel 5, hvis du er meget klog på opbygningen af forskellige AI-teknologier og prompting af dem.

# Introduktion til teknologierne Attack 2.5 og GPT-4

&#

# Algoritmer og kunstig intelligens – et lynkursus

A&ttack 2.5 og GPT-4 er begge algoritmer, der er baseret på kunstig intelligens. Herunder redegør vi for nogle af de mest grundlæggende begreber, som anvendes i denne rapport.

## Algoritmer

En algoritme kan defineres som en opskrift på at løse et problem gennem en regelbaseret proces. Algoritmer gives et **input** og returnerer et **output** afhængigt af "reglerne" i processen. I sin simpleste forståelse adskiller algoritmer sig ikke meget fra en opskrift eller en brugsanvisning, hvor man ved at følge en trinvis proces får returneret et ønsket output.

De simpleste computeralgoritmer sorterer f.eks. tal på en liste efter deres størrelse. De mest komplekse, såsom GPT-4, har lært at producere tekst i alverdens genrer om alverdens emner (næsten) som et menneske, fordi den er trænet på store mængder tekst fra internettet. Hvad modellen mere præcist er trænet på, ved vi desværre ikke, fordi OpenAI ikke offentliggør dens træningsdata.

## Neurale netværksalgoritmer

Moderne avancerede algoritmer kaldes neurale netværksalgoritmer, fordi de i deres opbygning efterligner den struktur, som vi kender fra opbygningen af den menneskelige hjerne.

Hjernen indeholder lag af neuroner (små processorer), der er forbundet i et netværk.

Når et input sendes igennem netværket, aktiveres de små processorer på skift. Hver neuron afkoder en lillebitte del af inputtet og sender signalet videre.

Selvom disse algoritmer bliver sammenlignet med en menneskehjerne, betyder det *ikke*, at det neurale netværk er intelligent på samme måde som os mennesker. Neurale netværksalgoritmer er nemlig kun bygget på matematiske beregninger.

## Kunstig intelligens

De neurale netværksalgoritmer bliver i daglig tale beskrevet som kunstig intelligens, når der gøres brug af mange "lag" af netværk i algoritmen. De enkelte "neuroner"/processorer i algoritmen repræsenterer relativt simple matematiske udregninger. Det er dybden i det samlede netværk – altså, antallet af lag – der gør, at algoritmen beskrives som kunstig intelligens.

Når data (f.eks. en kommentar fra Facebook) passerer gennem de forskellige lag af matematiske beregninger, opbygges der for hvert lag en mere kompleks forståelse af kommentarens indhold, da hvert lag i algoritmen bearbejder data på forskellig vis.

Kunstig intelligens betyder altså ikke, at algoritmerne kan "forstå" som mennesker. Algoritmens "forståelse" er baseret på, hvor gode modellerne er til at følge deres opskrift baseret på de matematiske beregninger, og hvor meget data algoritmen har set – og dermed bygger sine forståelseslag på.

## Sprogmodeller

Sprogmodeller er en type af kunstig intelligens, der er bygget til at kunne behandle tekst. Sprogmodeller kan også være generative. Det betyder de kan lære sproglige mønstre igennem kæmpe træningsdatasæt, som de bruger til at generere ny, lignende tekst med.

Når sprogmodeller genererer tekst, så er det baseret på vurderingen af, hvilke ord der har størst sandsynlighed for at være et godt "svar" på prompten. Sprogmodellen har lært disse sandsynlighedsberegninger baseret på dens træning. En sprogmodel er altså kun så god som den træningsdata, den har fået, og de forståelseslag, som modellen har opbygget.

Det er også antallet af lag og antallet af matematiske beregninger, der gør modellerne hhv. "større" og "mindre". En kunstig intelligens, der er trænet på store mængder alsidig data, får mange "forståelseslag", men bliver altså også "dyrere" – både i strøm, vand og pris – til at løse sin opgave.

Nogle forskere argumenterer endda for, at de største kunstige intelligenser bliver "Jack of all trades, master of none" (Kocorí et al, 2023), fordi deres milliarder af forståelseslag gør dem til "schweizerknive" ift. opgaveløsning. Samtidig er de ikke nødvendigvis de bedste til at løse mere specialiserede opgaver.

Repræsentanter for OpenAI vil formentlig på den anden side argumentere for, at det er den "prompt", man har anvendt – altså, måden man har spurgt modellen på – som ikke har været god nok.

## Hvordan identificeres sproglige angreb med hhv. A&ttack 2.5 og GPT-4?

Selvom både A&ttack 2.5 og GPT-4 er teknologier baseret på kunstig intelligens, fungerer de på forskellige måder, når det kommer til at løse opgaven at identificere sproglige angreb i kommentarer og opslag.

## A&ttack 2.5

A&ttack 2.5 er en superviseret klassifikationsmodel, der består af en underliggende sprogmodel og klassifikationsinstruktioner (herunder annoteret træningsdata) til at identificere sproglige angreb.

For at udvikle A&ttack 2.5 bruger vi en sprogmodel, som kan generere tekst. Denne type sprogmodel er meget mindre end GPT-4, og kan altså ikke bruges til alverdens opgaver, som GPT-4. Til gengæld gør modellens markant mindre størrelse, at det er muligt at køre modellen med vores egne computere, hvilket vi nok ikke kunne gøre med GPT-4.

Siden vi kun er interesseret i identificeringen af sproglige angreb, begrænser vi denne mindre sprogmodel til kun at generere ordene "angreb" eller "ingenting." Vi indbygger altså klassifikationsinstruktionerne direkte i modellen.

For at modellen lærer, hvad forskellen er på "angreb" og "ingenting", giver vi modellen et datasæt annoteret af et hold af annotører. Datasættet består af 160.000 kommentarer, der er beriget med menneskers vurdering af, om kommentaren indeholder angreb eller ej\*.

Dette ændrer altså modellen fra at være en sprogmodel, der kan genere tekst, til at være en klassifikationsmodel, der kan klassificere, om en kommentar eller opslag fra Facebook indeholder et sprogligt angreb eller ej.

\*Vurderingen er foretaget af fem dansksprogede annotører med hver deres universitetsfaglige baggrund, og tager udgangspunkt i definitioner og en kodemanual, udarbejdet af en retoriker. Definitionerne tager udgangspunkt i materiale fra ECRI (European Commission against Racism and Intolerance). Kilde: <https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01>

## GPT-4

GPT-4 er en stor, generativ sprogmodel. Det betyder, at man kan bruge modellen til alverdens formål, heriblandt identifikation af sproglige angreb.

Når man bruger GPT-4, kan man gå ud fra, at den har en forståelse for det danske sprog og opgaveløsningen. Vi indbygger derved *ikke* klassifikationsinstruktioner i modellen. GPT-4 tilgås derimod gennem en prompt.

En prompt er en vejledende tekst, man udarbejder til GPT-4, hvori man indsætter en definition på et sprogligt angreb, klassifikationsinstruktionerne, data som GPT-4 skal klassificere, og i nogle variationer enkelte eksempler på sproglige angreb. Hvordan vi har promptet GPT-4 på 10 forskellige måder i denne undersøgelse, kan du læse mere om på side 57.

Med god grund omtales GPT-4 som en sprogteknologisk schweizerkniv, da man undgår den manuelle og ressourcetunge opgave at annotere en masse data for at lære modellen, hvad et sprogligt angreb er (og ikke er), og erstatter den med en kort vejledning – en prompt, der giver nogle basale instruktioner.

Forskellen i modellernes opbygning og interaktionen med dem er visualiseret og beskrevet på de kommende sider.

## Input og Output

Hvad er input og output, når vi anvender GPT-4 og A&ttack til klassificering af angreb i kommentarer fra Facebook?

### A&ttack:

**Input:** Opslag og kommentarer fra den offentlige debat på dansk Facebook.

**Output:** Teksten "angreb" eller "ikke angreb" for hver af kommentarerne.

### GPT-4:

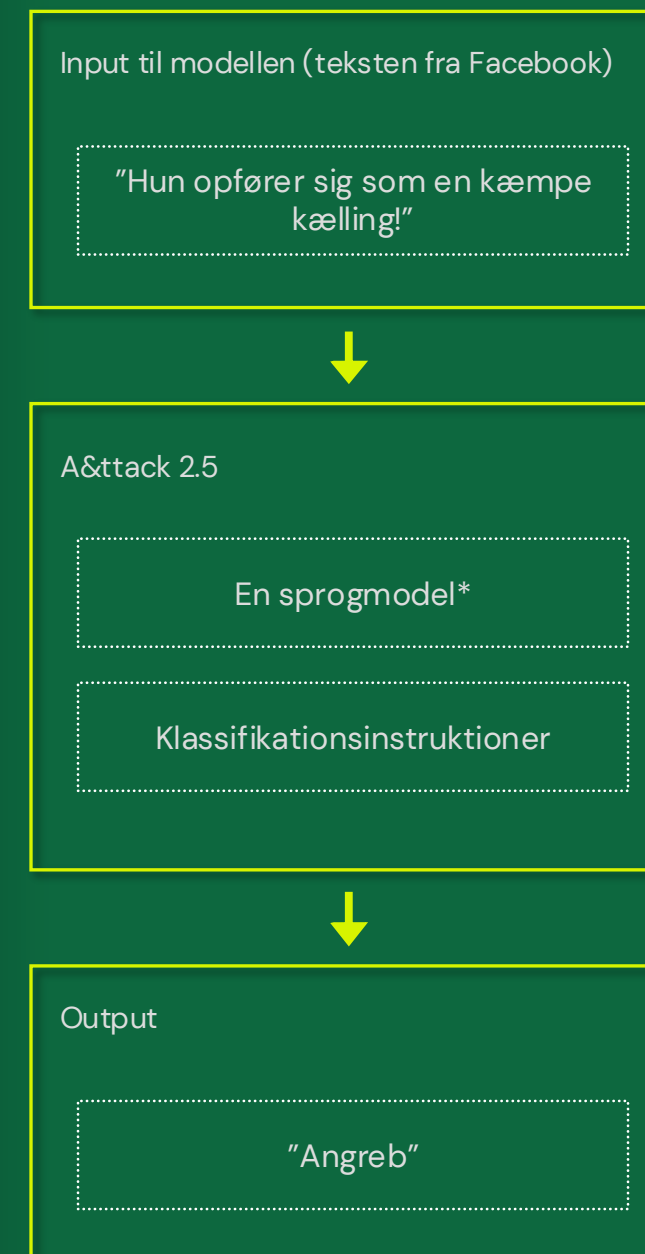
**Input:** Anonymiserede opslag og kommentarer fra den offentlige debat på dansk Facebook + en prompt af cirka en A4-sides længde, der giver modellen tekstlige anvisninger om opgaven, dataformat, definitioner og det ønskede output.

**Output:** "1" for angreb og "0" for ikke-angreb for hver af kommentarerne og en begrundelse for vurderingen af de kommentarer, som GPT-4 har vurderet er angreb.

## A&ttack 2.5

Når man klassificerer med A&ttack, er klassifikationsinstruktionerne bygget ind i sprogmodellen, som man bruger.

**Inputtet** til modellen er teksten fra Facebookopslaget eller -kommentaren **outputtet** er teksten "angreb" eller "ingenting"



## GPT-4

Når man klassificerer med en stor sprogmodel som GPT-4, er klassifikationsinstruktionerne *ikke* bygget ind.

Inputtet til modellen er en prompt, der indeholder klassifikationsinstruktionerne samt Facebookkommentaren eller -opslaget.



\*Til A&ttack 2.5 er anvendt North/t5\_large\_scand (skabt af Per E. Kummervold) som er en sprogmodel med 770 millioner parametre prætrænet på et skandinavisk korpus bestående af bokmål, nynorsk, dansk, svensk, islandsk og en begrænset mængde færøsk.



# Prompten er nøglen til GPT-4, så vi tester ti forskellige

## Den sensitive prompt

At skrive en god prompt har ført til et helt felt kaldet prompt engineering, hvor man undersøger, hvordan man opnår de bedste resultater (i vores tilfælde klassificeringer af "angreb" eller "ingenting") igennem sin prompt.

Det er vigtigt, at man følger bestemte metoder, da alle ændringer i en prompt kan have indflydelse på resultaterne. Dog er prompts også afhængige af den model, man bruger.

I figuren herunder viser vi et eksempel på en simpel prompt, hvor vi har bedt OpenAI's gratisversion af modellen - ChatGPT - om at klassificere et sprogligt angreb.

Ligesom alle ord i prompten kan influere, hvor gode resultaterne bliver, så kan ændringer i GPT-modellernes præstation også blive influeret af ændringer i modellen, som OpenAI ikke kommunikerer offentligt ud om.

OpenAI's skriver nemlig på deres hjemmeside, at "Determinisme kan blive påvirket af nødvendige ændringer, som OpenAI foretager sig i modelkonfigurationer på vores side." (OpenAI, u.d.c). Determinisme referer her til algoritmens evne til at producere det samme output, givet det samme input.

Den evigt flydende model kombineret med manglende transparens fra OpenAI's side gør det svært at lave reproducerbare resultater (Chen et al., 2023; Reiss, 2023).

D

Dig

Du er John, en AI ekspert i sproglige angreb. Du skal klassificere, hvorvidt en kommentar er et sprogligt angreb.

Sproglige angreb defineres som "stigmatiserende, nedsættende, krænkende, stereotypiserende, ekskluderende, chikanerende eller truende ytringer."

Er den følgende kommentar et sprogligt angreb? "Typisk kvinder, de jo psykopater!!!"



ChatGPT

Ja, kommentaren "Typisk kvinder, de jo psykopater!!!" er et sprogligt angreb.

Den er stigmatiserende, nedsættende og stereotypiserende, da den generaliserer alle kvinder som værende psykopater, hvilket er et krænkende og nedværdigende udsagn. Kommentaren bidrager til en negativ opfattelse af kvinder som gruppe, og kan dermed ses som ekskluderende og chikanerende.

Måden, vi skriver prompts på, har altså en effekt på GPT-4's performance (Zhao et al., 2021; Reiss, 2023). Det vil også sige, at de promptingstrategier, som danner grundlaget for vores basisprompt, er strategier folk har anvendt til at opnå de bedste resultater specifikt med GPT-4. Andre store generative modeller (som f.eks. Meta's LLaMA) vil formentlig skulle promptes anderledes. I alle tilfælde giver det mening at bruge tid på at studere, hvilke erfaringer andre har gjort sig i ift. at prompte den model, man sidder med.

Vi har, udover basisprompten, testet ni forskellige prompts i vores sammenligning af GPT-4 med A&tack 2.5.

## En forskningsbaseret basisprompt med ti variationer

Den første prompt, vi tester, er en basisprompt. Basisprompten kan ses i sin fulde længde på side 23. Den er udviklet baseret på testede promptteknikker til GPT-4 fra praktikere og forskere.

Basisprompten inkluderer en personbeskrivelse (OpenAI u.d.b; Clavié et al, 2023), instruktioner til opgaven i form af en definition af sproglige angreb (OpenAI u.d.b.), samt en opfordring til, at modellen forklarer ræsonnementet bag klassificeringen af den enkelte kommentar, hvis denne klassificeres som et "angreb" (OpenAI u.d.b., Wei et al, 2023). At inkludere disse elementer i sin prompt har andre erfaret giver de bedste outputs fra et performanceperspektiv.

Derefter har vi udviklet ni variationer af basisprompten, hvor der blev ændret én variabel ad gangen.

Dette gjorde vi for at teste, hvordan små ændringer havde effekt på modellens performance og fairness.

Bliver GPT-4 for eksempel bedre til at finde angreb rettet mod forskellige køn, hvis vi inkluderer tre eksempler på hadtale baseret på køn i prompten?

Eller bliver GPT-4 for eksempel bedre til at identificere sproglige angreb, hvis vi lover den verdensfred som en del af prompten?

I alt tester vi altså GPT-4's evne til at klassificere "angreb" fra "ikke angreb" i et testdatasæt på 4.300 kommentarer med ti forskellige prompts.

## De ti prompts

1. Basisprompt på dansk
2. Basisprompt på engelsk
3. Basisprompt med et kvindenavn i personbeskrivelsen
4. Basisprompt med et mandenavn i personbeskrivelsen
5. Basisprompt tilsat en løftet pegfinger om at have fokus på egne bias
6. Basisprompt tilsat eksempler på anstødelige angreb
7. Basisprompt tilsat eksempler på hadefulde angreb baseret på køn
8. Basisprompt tilsat eksempler på hadefulde angreb baseret på etnicitet og religion
9. Basisprompt tilsat blandede eksempler på hadefulde og anstødelige angreb
10. Basisprompt tilsat et incitament til GPT-4 i form af et løfte om verdensfred.

Alle prompts blev kørt med GPT-4 Turbo igennem OpenAI's API mellem den 9 juli og 12 juli 2024. Alle kommentarer blev anonymiseret.

For mere detaljeret beskrivelse af de ti prompts, og variationerne i deres konkrete ordlyd, se deres onepagers i sektion 6.

På den næste side kan du se den fulde basisprompt med forklaringer af de elementer, som indgår i prompten.



# Sådan har vi designet 'basisprompten'

## 1) Giv modellen en persona og beskriv de steps, den skal gennemgå, for at lave klassifikationen

Du er en ekspert i sproglige angreb. Du skal klassificere, hvorvidt en kommentar er et sprogligt angreb eller ej. Du får en definition på et sprogligt angreb og instruktioner, som du skal bruge. Til sidst får du den data, som du skal klassificere.

### ## Definitionen på et sprogligt angreb

Et sprogligt angreb er defineret som stigmatiserende, nedsættende, krænkende, chikanerende eller truende ytringer rettet mod personer eller grupper.

Kendetegn ved et sprogligt angreb:

- Nedsættende/krænkende ord der er nedsættende rettet mod en gruppe eller et individ.
- Nedsættende brug af kliniske betegnelser eller ideologier når der omtales individer eller grupper med ord, der kan være en faktisk reference til et diagnosticeret, eller selverklæret individ, men som i konteksten blot er ment nedsættende.
- Dehumaniserende sprog som ord eller emojis med dobbeltbetydning, der dog i konteksten må betragtes som et sprogligt angreb på et individ eller en gruppe. Der kan f.eks. være tale om sammenligning af et menneske med dyr, mikroorganismer eller afføring.
- Upassende opfordringer f.eks. til omgang med genitalier eller ekskrementer.
- Truende sprog som enten er direkte trusler om skade eller overgreb, et ønske om skade eller overgreb eller forherligelse af skade og overgreb på individer eller grupper.
- Foragt og afsky som indeholder omtale af mennesker gennem nedsættende adjektiver eller emojis.
- Stigmatisering, generalisering og anklager som er generaliserende, stigmatiserende eller anklagende udtalelser om en gruppe.
- Krænkelse af rettigheder som indeholder ønske om adskillelse, ekskludering eller fratagelse af basale rettigheder til bestemte grupper.
- Stigmatisering/delegitimering: Det er en del af debatten at referere til forskelle i demografiske forhold som uddannelse, alder og økonomiske forhold, men kommentarer, der forsøger at ekskludere andre medborgere fra debatten eller principielt delegitimere et individ eller en gruppes input på den baggrund, betragtes som et sprogligt angreb.

Hvad er ikke et sprogligt angreb:

- Hvis kommentaren ikke er rettet mod individer eller grupper, men derimod er rettet mod situationer, tiltag, genstande, idéer, argumenter og virksomheder.
- Hvis kommentaren, der nok er hård eller kontroversiel, vurderes inden for en grænse for ophedet debat (kræver at kommentaren ikke er generaliserende eller stigmatiserende på baggrund af beskyttede karakteristika).
- Hvis kommentaren er en "mild" fornærmelse, der ikke regnes som et sprogligt angreb (medmindre de optræder som en del af en generalisering baseret på beskyttede karakteristika).

## 3) Giv modellen klassifikationsinstruktioner

### ## Instruktioner

Du modtager en liste af json-objekter. Det er ekstremt vigtigt, at dit output er gyldige json data, og at du følger instruktionerne nøje. Hvert objekt indeholder en egenskab kaldet `comment_message`. For hvert objekt skal du tilføje en ny egenskab kaldet `prediction`, som skal indeholde en af følgende to muligheder:

- 1, hvis kommentaren i `comment_message` er et sprogligt angreb
- 0, hvis kommentaren i `comment_message` ikke er et sprogligt angreb.

## 4) Bed modellen forklare ræsonnementet, når en kommentar klassificeres som et sprogligt angreb

Hvis kommentaren er et sprogligt angreb, skal du også tilføje en egenskab kaldet `explanation`, hvor du kort forklarer, hvorfor kommentaren skal klassificeres som et sprogligt angreb. Fra dit endelige output skal du beholde egenskaben `comment_id` og undlade `comment_message` for alle objekter.

# Kapitel 4

Evaluering af  
model lerne på  
seks parametre





#### Advarsel om ubehageligt indhold

Evalueringerne kommer til at indeholde eksempler på tekstkommentarer fra den offentlige debat på Facebook, der er klassificeret som sproglige angreb.

Der vil være tale om misogyne, sexistiske, ableistiske, xenofobiske, racistiske, antisemitiske eller islamofobiske kommentarer.

Hvis du er interesseret i at læse rapporten uden eksempler, så ræk ud til [anna@ogtal.dk](mailto:anna@ogtal.dk), så laver vi en sådan version.

# Overblik over modeller og prompts vi tester

| Teknologi | Variation   | Kaldenavn   |
|-----------|-------------|---|
| GPT-4     | Prompt 1    | Basisprompt på dansk  |
| GPT-4     | Prompt 2    | Basisprompt på engelsk  |
| GPT-4     | Prompt 3    | Basisprompt med et kvindenavn i personbeskrivelsen                                |
| GPT-4     | Prompt 4    | Basisprompt med et mandenavn i personbeskrivelsen                                 |
| GPT-4     | Prompt 5    | Basisprompt tilsat en løftet pegefinger om at have fokus på egne bias             |
| GPT-4     | Prompt 6    | Basisprompt tilsat eksempler på anstødelige angreb                                |
| GPT-4     | Prompt 7    | Basisprompt tilsat eksempler på hadefulde angreb baseret på køn                   |
| GPT-4     | Prompt 8    | Basisprompt tilsat eksempler på hadefulde angreb baseret på etnicitet og religion |
| GPT-4     | Prompt 9    | Basisprompt tilsat blandede eksempler på hadefulde og anstødelige angreb          |
| GPT-4     | Prompt 10   | Basisprompt tilsat et incitament til GPT-4 i form af et løfte om verdensfred.     |
| A&ttack   | A&ttack 2.5 | A&ttack 2.5   |
| A&ttack   | A&ttack 1   | A&ttack 1   |

## GPT-4 vs. A&ttack 2.5

# Performance

## Evaluering af performance

### Hvad måler vi ?

Parameteret performance dækker over, hvor gode modellerne A&ttack og GPT-4 er til at identificere sproglige angreb.

Modellernes performance opgøres i en *F1 macro average score* (F1 score). Jo højere en F1 score, jo bedre er modellen til at klassificere, om en kommentar indeholder angreb eller ej (sammenlignet med mennesker).

### Hvordan?

Mennesker (annotører) har klassificeret, hvorvidt 4.300 Facebookkommentarer indeholder et sprogligt angreb eller ej. Disse klassificeringer bruger vi som "guldstandard", dvs. vi antager, at de er "rigtige".

A&ttack modellerne og GPT-4 Turbo bliver bedt om at klassificere de samme 4.300 kommentarer. Vi tester GPT-4 med ti forskellige prompts og sammenligner A&ttack 2.5 med den prompt, der giver den bedste performance.



## Ordbog

**Annotering:** Processen med at klassificere (kategorisere) data, der bruges som "guldstandard" til at træne og evaluere modeller. Algoritmer bruger annoteret data til - igennem matematiske beregninger - at lære sig hvordan de skal klassificere efterfølgende inputdata.

**Annotør:** et menneske eller en maskine, der har til opgave at klassificere data.

**Præcision** måler, hvor god algoritmen er til at klassificere sproglige angreb ved at sammenligne dens klassifikationer med menneskeannoteret data. Det vil sige, hvis algoritmen identificerer 100 kommentarer som angreb, men kun 70 af dem er menneskeannoterede angreb, så er præcisionen 70 pct..

**Genkaldelse(Recall)** måler, hvor mange af de menneskeannoterede angreb, algoritmen også klassificerer som angreb. Det vil sige, hvis der er 100 annoterede sproglige angreb i vores datasæt, og algoritmen kun identificerer 80 af dem som sproglig angreb, er genkaldelse 80 pct..

**F1 macro average score:** en samlet score for precision og recall. Den måler, hvor god modellen er til at klassificere angreb og ikke-angreb. F1 scoren er baseret på gennemsnittet af recall og precision.

**Guldstandard:** Annoteret datasæt, der bruges som den mest nøjagtige og pålidelige klassificering. Guldstandard bruges som referencepunkt for den korrekte måde at klassificere data på. I vores tilfælde er vores menneskeannoterede datasæt guldstandard, som vores modellers klassificeringer bliver evalueret i mod.

Se den komplette ordbog i bilag 1

# A&ttack 2.5 performer bedre end GPT-4



**GPT-4 bliver slået på målstregen, men vinder stadig specifikke sværdslag**

Når vi sammenligner GPT-4 med A&ttack på F1 score, ser vi, at A&ttack 2.5 performer en tand bedre end GPT-4 med 0,01 point. F1 scoren viser, at A&ttack 2.5 generelt klassificerer mere korrekt end GPT-4 (når vi sammenligner med menneskers klassificering).

Selvom GPT-4 (så vidt vi ved) er trænet på enorme mængder data, er den altså ikke bedre til at løse den specialiserede opgave, det er at identificere sproglige angreb baseret på vores definition. Men den er overraskende god.

Både A&ttack 2.5 og GPT-4 er bedre end A&ttack 1. A&ttack 1 er blevet trænet på kommentarer fra debatten mellem 2019 til 2021 og A&ttack 1 performer væsentligt bedre på et testdatasæt fra denne periode. Dette peger på, at sproget i kommentarsporene – selv på få år – har ændret sig tilpas meget til, at A&ttack 1s performance er væsentligt forringet i denne test. A&ttack 2.5 bygger på samme data som A&ttack 1, men er blevet beriget med kommentarer og opslag fra 2020-2023.

Selvom A&ttack 2.5 og GPT-4 har en forholdsvis god F1 score, så klassificerer de kun 84 pct. af kommentarerne ens. Det er altså 16 pct. af kommentarerne, som de to er uenige om. Der er både eksempler på, at GPT-4 identificerer angreb, som A&ttack misser, og omvendt i vores data.

### Eksempel: A&ttack 2.5 klassificerer korrekt, GPT-4 gør ikke

Den følgende kommentar er annoteret som et angreb af mennesker, men bliver kun klassificeret som et angreb af A&ttack 2.5 og ikke GPT-4:

”Kan være heksen identificere sig som intetkøn 🧙”

### Eksempel: GPT-4 klassificerer korrekt, A&ttack 2.5 gør ikke

Den følgende kommentar er annoteret som et ”ikke-angreb” af mennesker, men bliver klassificeret som et angreb af A&ttack 2.5 og ikke GPT-4:

”Ja ja lad os nu se hvad Mette fup finder på!! Hun er jo gode til at true med det hvis de unge ikke mangler ret”

## GPT-4:

### Hvad er performance for de 10 GPT-4-resultater?

Når vi evaluerer resultaterne fra de 10 GPT-4 prompts, giver Prompt 6’s klassificeringer den højeste F1 score. Prompt 6 adskiller sig fra basisprompten ved at være tilsat tre eksempler på anstødelige angreb.

Da vi udviklede de ti prompts, var formålet at undersøge, om små justeringer i vores basisprompt ville gøre performance bedre. Det er kun prompt 10 og 6, der skaber bedre resultater end vores basisprompt.

Men derfor kan vi stadig bekræfte, at små justeringer i basisprompten fører til forskelle i klassificeringerne.

Selvom syv af vores prompts har den samme F1 score på 0,72, betyder det ikke, at deres klassificeringer er ens: I 9 pct. af de 4.300 kommentarer klassificerer mindst en af de syv prompts med samme F1 score forskelligt fra de andre. Selvom F1 scoren er den samme, er klassificeringerne af kommentarerne det ikke.

#### Eksempel:

Kun tre ud af de syv prompts, hvis resultater har en F1 score på 0,72, klassificerer den følgende kommentar som et angreb:

”Jeg håber heller ikke han bliver, og synes grønlanderne er et snotforkælet folkefærd der har syltet min pengekasse for længe”.

F1 score for ti GPT-4 prompts





## GPT-4 vs. A&ttack 2.5

### Fairness

## Evaluering af fairness

### Hvad måler vi?

Parameteret fairness dækker over, hvor stor forskellen er mellem modellernes klassificeringer af angreb omhandlende 19 forskellige beskyttede grupper indenfor kategorierne: seksualitet, køn, religion, handicap, alder, etnicitet og nationalitet.

For hver af de 19 grupper får modellerne en DAP-score. Her måles, om modellerne under- eller overvurderer angreb af de 19 beskyttede grupper i deres klassificeringer.

Fairness måles med PDAP-scoren, der angiver den *gennemsnitlige parvise forskel i vurderingen af kommentarer, der omhandler de 19 grupper*. Jo lavere en gennemsnitlig parvis forskel, jo bedre fairness. PDAP-scoren er altså en samlet score for modellernes 19 DAP-scorer.

### Hvordan?

Mennesker (annotører) har klassificeret, hvorvidt 4.300 Facebookkommentarer indeholder et sprogligt angreb eller ej. Disse klassificeringer bruger vi som "guldstandard", dvs. vi antager, at de er "rigtige".

Søgenøgler, bestående af neutrale og nedsættende indikatorord, identificerer, om en kommentar omhandler én eller flere af de 19 beskyttede grupper. Alle 19 grupper er repræsenteret i datasættet med mindst 90 kommentarer.

De to A&ttack-modeller bliver bedt om at klassificere de 4.300 kommentarer. GPT-4 bliver med ti forskellige prompts bedt om at klassificere de 4.300 kommentarer.

Alle 12 datasæt evalueres med DAP-score for, om angreb, der omhandler hver af de 19 grupper, over- eller undervurderes. Herefter får de tildelt en samlet PDAP-score. Vi sammenligner den mest fair af GPT-4s ti prompts med A&ttack 2.5.



### Ordbog

**(19) Beskyttede grupper** I analysen måler vi hvor fair modellerne klassificerer angreb mod 19 grupper baseret på deres beskyttede karakteristika inden for seksualitet, køn, etnicitet, nationalitet, religion, funktionalitet, alder eller politiske overbevisning. De 19 grupper er: Seksuelt minoriserede, kønsminoriserede, kvinder, mænd, udlændinge, inuit, personer fra Mellemøsten, afrodanskere, asiater, østeuropæere, muslimer, jøder, hinduer & buddhister, kristne, psykiske og fysiske funktionsvariationer, unge, gamle, venstreorienterede og højreorienterede. De 19 udgør ikke en udtømmende liste over variationer af beskyttede karakteristika.

**DAP:** Delta A&ttack Percentage. DAP måler forskellen på den menneskeligt klassificerede og den AI-klassificerede andel af angreb målt i procentpoint. DAP måles individuelt for hver gruppe. En positiv DAP betyder, at modellen overvurderer andelen af angreb, mens en negativ DAP betyder en undervurdering af angreb. Hvis algoritmen rammer andelen af angreb helt rigtigt, vil DAP være 0.

**Indikatorord** For at bestemme om en kommentar omhandler en beskyttet gruppe, anvender vi en søgenøgle. Søgenøglen indeholder en lang række indikatorord for hver beskyttet gruppe. Den indeholder både neutrale og nedsættende indikatorord. "Homoseksuel" er f.eks. et neutralt indikatorord på listen over ord, der bruges til at identificere, om en angrebskommentar omhandler den beskyttede gruppe "seksuelt minoriserede". "Bæskubber" er et eksempel på et nedsættende indikatorord på listen over ord, der bruges til at identificere, om en angrebskommentar omhandler den beskyttede gruppe "seksuelt minoriserede".

**PDAP:** Pairwise Delta A&ttack Percentage. PDAP måler den gennemsnitlige forskel i DAP mellem alle de beskyttede grupper. Jo lavere dette tal er, jo mindre er forskellen

Se den komplette ordbog i bilag 1

Vi arbejder med **19** beskyttede grupper

### Seksualitet

1. Seksuelt minoriserede

### Køn

2. Kønsminoriserede
3. Kvinder
4. Mænd

### Etnicitet og nationalitet

5. Udlændinge/indvandrere/ flygtninge
6. Inuit
7. Personer fra Mellemøsten
8. Afrodanskere
9. Asiatere
10. Østeuropæere

### Religion

11. Muslimer
12. Jøder
13. Hinduer & Buddhistere
14. Kristne

### Handicap

15. Psykiske og fysiske funktionsvariationer

### Alder

16. Unge
17. Gamle

### Politisk overbevisning

18. Venstreorienterede
19. Højreorienterede

# A&tack-modellerne producerer mere fair resultater

A&tack 1

4,2

PDAP score

A&tack 2.5

5,4

PDAP score

GPT-4

7,8

PDAP score

(Prompt 6)

**Vores hjemmebyggede modeller er markant mere fair**

Begge A&tack-modellers output er markant mere fair end det mest fair af GPT-4s ti outputs.

A&tack-modellerne har altså markant lavere gennemsnitlig parvis forskel (PDAP) i vurderingen af kommentarer, der omhandler 19 beskyttede grupper.

Ingen af modellerne er dog helt fair, og det er altså forskelligt, hvordan modellerne under- eller overvurderer angreb, der omhandler de 19 beskyttede grupper. Dette dykker vi ned i på de næste sider.

Også i relation til fairness ligger GPT-4's prompt 6 på førstepladsen i rækken af de ti prompts.

**A&tack 1 er mest fair, selvom A&tack 2.5 fik særlig fairness-indsats**

A&tack 1 er med den laveste PDAP-score den model der returnerer det mest fair output. Det til trods for, at A&tack 2.5 er trænet med et særligt fokus på fairness.

Blandt andet har vi sørget for, at dens træningsdata indeholder mindst 400 kommentarer, der omhandler hver af de beskyttede grupper. Modellen blev altså trænet i at genkende omtaler af grupperne, selvom de ikke optræder ofte i en randomiseret stikprøve fra debatten.

Denne indsats er tydeligvis ikke helt tilstrækkelig, men udformningen af et fairness-vægtet trænings- og testdatasæt har gjort det muligt at køre de fairness-tests, der præsenteres i denne rapport, og som har givet os syn for sagen.



# GPT- 4: Hvilken ud af ti GPT- 4 prompts giver det mest fair resultat?

Tabellen herunder rangerer de ti GPT- 4 prompts efter PDAP- score.

Prompt 6 giver de resultater, der har den laveste og dermed bedste PDAP- score ud af de ti GPT- 4 prompts. Ligesom på performance- parameteret ligger prompt 6s resultater på førstepladsen ud af de ti.

Prompt 6 adskiller sig fra basisprompten ved at indeholde tre eksempler på angreb i underkategorien "anstødeligt sprog" dvs. et angreb, der (modsat hadtale) ikke er baseret på ét eller flere beskyttede karakteristika. Prompt 6 har en fairness score på 0,078. Dette betyder, at den parvise ulighed mellem de 19 beskyttede grupper er 7,8 pct.

Det er specielt, at det netop er den prompt, der ikke er blevet givet eksempler på hadefulde angreb, men anstødelige angreb, som er mest fair i vurderingen af angreb, der omhandler beskyttede grupper.

De anstødelige angreb er nemlig netop defineret ved ikke at være baseret på beskyttede karakteristika såsom etnicitet, seksualitet, køn eller religion. Læs mere om forskellen på hadefulde angreb og anstødelige angreb i bilag 2 på side 74.

PDAP score for outputtet af de 10 GPT- 4 prompts rangeret



# A&tack vs. GPT-4

## Hvilke beskyttede grupper over- og undervurderer de to modeller?

I visualiseringen nedenfor kan man se DAP-scoren for A&tack 2.5 og GPT-4s resultater. GPT-4 referer til prompt 6, hvis klassificeringer har den laveste overordnede PDAP.

Generelt ser vi, at resultaterne fra GPT-4 (prompt 6) overvurderer angreb, der omhandler otte af vores beskyttede grupper og undervurderer angreb, der omhandler 11 af vores beskyttede grupper.

A&tack 2.5 overvurderer i sine klassificeringer angreb omhandlende fire ud af de 19 beskyttede grupper og undervurderer angreb omhandlende 15 ud af de 19 beskyttede grupper.

Der, hvor vi ser størst forskel i A&tack og GPT-4s DAP scores, er ved udlændinge generelt, afrodanskere, muslimer, Mellemøsten, og psysisk og fysisk handikappede. Her er distancen mellem grupperne 10 procent point eller mere.

### Case: Når ordet "psykopat" bliver brugt nedsættende

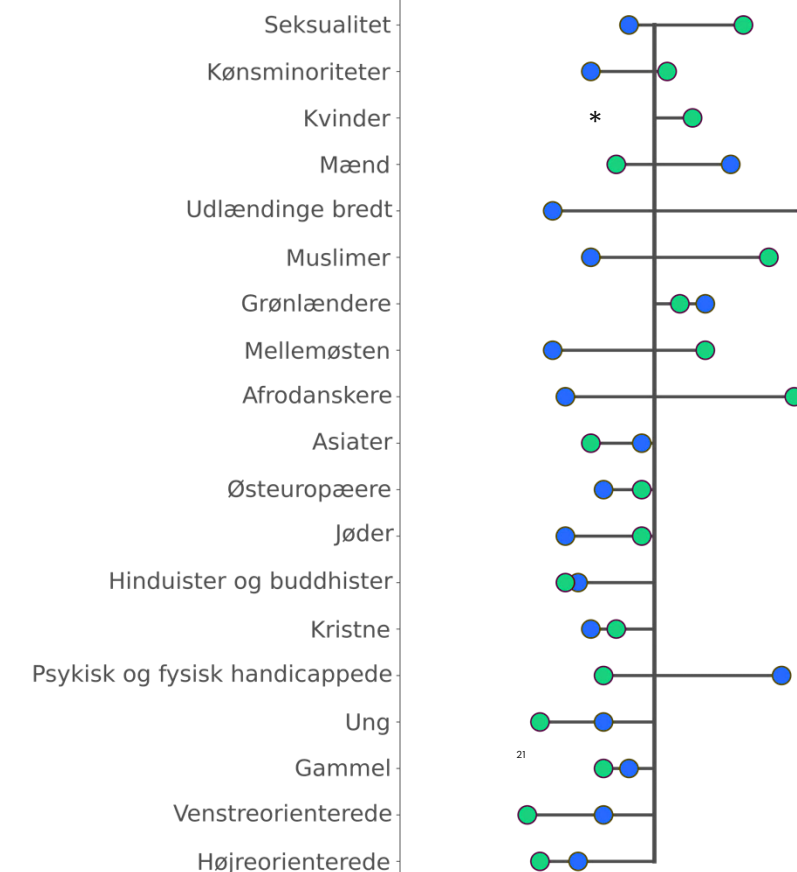
I kommentarer omhandlende personer med psykiske og fysiske funktionsvariationer undervurderer GPT-4, og A&tack overvurderer.

En af grundene til dette er, at GPT-4 fejlklassificerer kommentarer, som ikke indeholder angreb, selvom "psykopat" bliver brugt nedsættende.

De følgende kommentarer klassificeres af A&tack som et sprogligt angreb, mens GPT-4 ikke gør:

- "Nej det hedder Magtsyge psykopatiske mænd når der er mere end en"
- " Den psykopat kan finde på hvad som helst 🤪"

DAP -20% -15% -10% -5% 0% 5% 10% 15% 20%



DAP for hver af de 19 grupper for A&tack 2.5 og GPT-4 (Prompt 6)

\* DAP-scoren for kvinder er den samme. Prikkerne ligger oveni hinanden.

# GPT-4: Kan vi favorisere nogle af de 19 beskyttede grupper med vores prompt?

Ét af formålene med at eksperimentere med ti forskellige prompts til GPT-4 er at teste, hvordan justeringer i prompten har indflydelse på GPT-4's klassificeringer af de 4.300 kommentarer som "angreb" eller "ikke angreb". Er GPT-4 f.eks. mere fair i sine klassificeringer af angreb omhandlende hhv. kvinder og mænd, når vi prompter med en persona med tilsvarende køn?

## Er "Maria" mere fair overfor kvinder i sin klassificering end "John"?

Når vi giver GPT-4 et kvinde- eller mandenavn som en del af prompten (prompt 2 og 3), får vi ikke statistisk signifikant bedre resultater for kommentarer omhandlende hhv. kvinder eller mænd.

## Bliver GPT-4 mere fair overfor køn, hvis den gives eksempler på angreb omhandlende køn som en del af prompten?

Prompt 7 adskiller sig kun fra basisprompten ved, at vi har tilføjet tre eksempler på angreb baseret på køn til prompten. Dette, kunne man have en tese om, ville påvirke GPT-4 til at have særligt fokus på køn i klassificeringen af de 4.300 kommentarer.

Dog kan vi ikke se signifikante forbedringer i fairness for angreb omhandlende køn i resultaterne fra prompt 7.

### Eksempel: Køn

Den følgende kommentar er blot et eksempel fra data, hvor prompt 7 giver resultatet "ikke angreb" til et angreb baseret på køn, mens fire andre ud af de ti prompts - der slet ikke har fokus på køn - har klassificeret selvsamme kommentar som et angreb:

"Tåbelige kvindemenneske har i ikke ødelagt nok for folkeskolen, medens dit eget barn gik på privatskole"

## Bliver GPT-4 mere fair overfor jøder, afrodanskere og personer fra Mellemøsten, hvis den gives eksempler på angreb omhandlende disse grupper som en del af prompten?

Prompt 8 adskiller sig kun fra basisprompten ved, at vi har tilføjet tre eksempler på had baseret på etnicitet, nationalitet eller religion - nærmere bestemt omhandlende jøder, afrodanskere og personer fra Mellemøsten. Dette, kunne man have en tese om, ville påvirke GPT-4 til at have særligt fokus på disse grupper i klassificeringen af de 4.300 kommentarer.

Dog kan vi ikke se statistisk signifikante forbedringer i fairness ift. angreb omhandlende disse grupper i resultaterne baseret på prompt 8. Se eksempler fra data herunder.

Som de overordnede fairnesstal på forrige side viser, er der tydelige udsving i modellernes fairness for de 19 forskellige beskyttede grupper, og små justeringer i prompten kan gøre en forskel. Dog lykkedes vi ikke med at favorisere bestemte grupper med overlæg.

### Eksempel: etnicitet, nationalitet og religion

Den følgende kommentar er blot et eksempel fra data, hvor prompt 8 giver resultatet "ikke angreb" til to angreb baseret på religion og etnicitet, mens andre af de ti prompts - der ikke er blevet promptet med eksempler på angreb rettet mod disse grupper - har klassificeret selvsamme kommentar som et angreb:

"Jøderne er heldige. De har deres holocaust som de kan slå os andre oven i hovedet med, hver gang man kritiserer dem for overtrædelse af FNs menneskerettigheder 🤔"

"Send dem til Sahara og plante træer"

# GPT-4: Er der grupper, som generelt bliver over- og undervurderet på tværs af de ti GPT-4-prompts?

Grafen nedenunder visualiserer den gennemsnitlige DAP-score for de 19 beskyttede grupper på tværs af alle ti prompts. Ud fra grafen kan vi altså se, om GPT-4 (på tværs af vores ti prompts) overvurderer eller undervurderer sproglige angreb, der omhandler hver af de 19 beskyttede grupper.

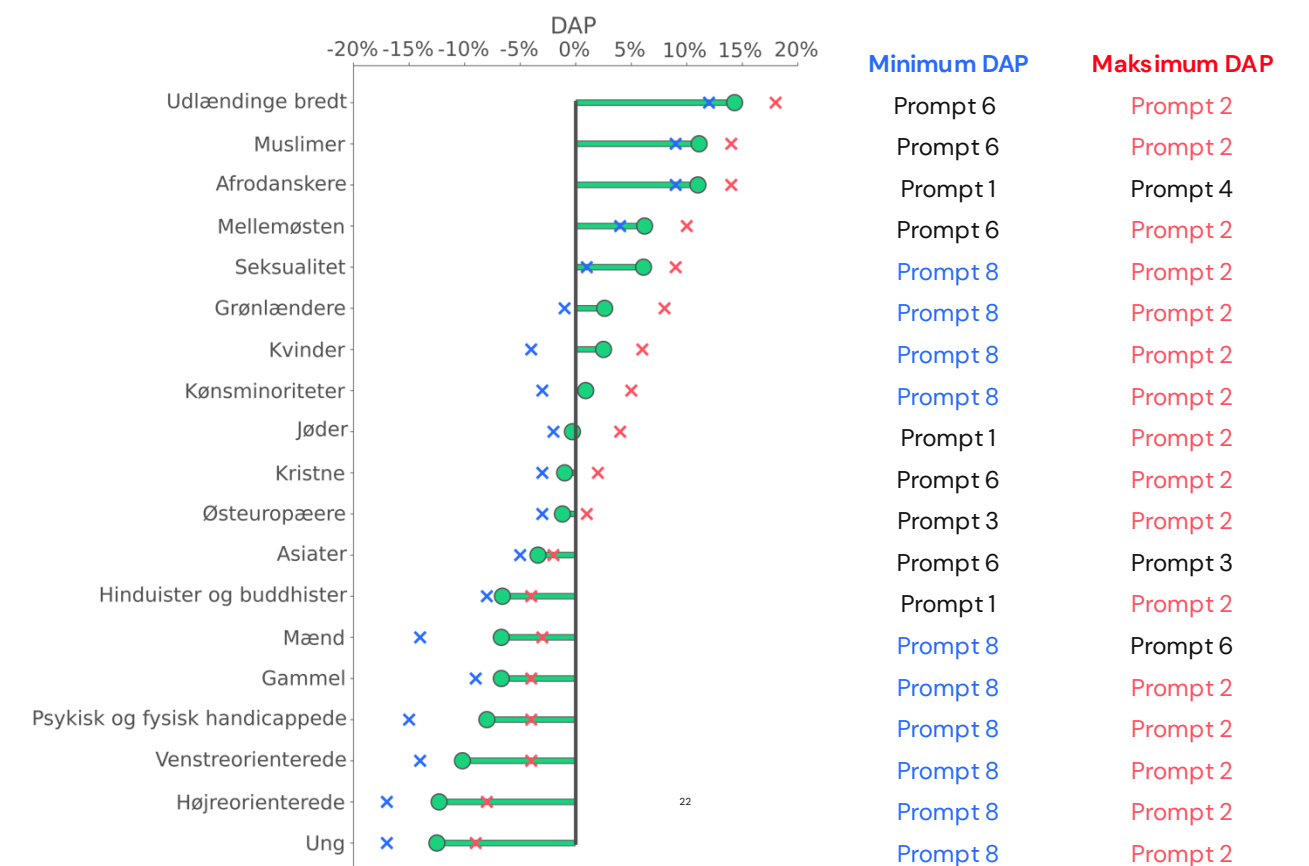
Når GPT-4 klassificerer de 4.300 kommentarer, vil den gennemsnitligt overvurdere angreb der omhandler følgende: udlændinge, muslimer, afrodanskere, personer fra Mellemøsten og seksualitet (DAP > 5 procent point). GPT-4 undervurderer i gennemsnit sproglige angreb omhandlende mænd, hinduer og budister, psykiske og fysiske handicap, unge, gamle, venstre- og højreorienterede (DAP < -5 procent point).

## Krydserne viser outliers

I tillæg til gennemsnittet viser tabellen nedenfor også, hvilke to ud af de ti prompts der hhv. undervurderer (minimum DAP) og overvurderer (maksimum DAP) mest i deres klassificeringer af angreb, der omhandler den enkelte beskyttede gruppe.

Det er særligt prompt 8 og prompt 2, der systematisk under- og overvurderer angreb, der omhandler de beskyttede grupper. Krydserne viser altså, hvor store forskelle små promptvariationer har på klassificeringen af kommentarer, der omhandler den enkelte beskyttede gruppe.

DAP gennemsnit for de 19 grupper på tværs af ti GPT-4 prompts.





# GPT-4: Hvor "enig" er GPT-4 med sig selv i klassificeringerne på tværs af resultaterne for de 10 prompts?

I nedenstående figur kan vi se, at de ti prompts resulterer i den samme klassificering af ~88 pct. af kommentarerne.

Det er altså særligt ~12 pct. af kommentarerne, som vores variationer i prompten (og/eller tidspunktet på dagen) har indflydelse på.

Når vi dykker kvalitativt ned i kommentarerne, kan vi ikke umiddelbart se en systematik eller gennemgående fællestræk for de 12 pct. af kommentarerne, som GPT-4 er uenig med sig selv i på tværs af prompts og tid. Som beskrevet på forrige side kan vi heller ikke se, at det lykkedes at prompte modellen til at tage særlige hensyn til bestemte grupper.

## Eksempler på uenighed

Kommentar: *"hun er åben overfor at tage børnene hjem. Men det skal være uden deres mødre, der har vendt Dk ryggen for at tilslutte sig Islamisk stat."*

Mennesker: "ikke angreb"  
A&tack 2.5: "ikke angreb"

GPT-4: 6/10 prompts giver outputtet "angreb"

Kommentar: *"Kommunisterne har rigtig fået vind i sejlene i Danmark efter det er lykkedes at besætte undervisningssystemet og hjernevaske ungdommen gennem 50 år."*

Mennesker: "angreb"  
A&tack 2.5: "angreb"  
GPT-4: 2/10 prompts giver outputtet "angreb"

## Andel af prompts, der returnerer samme klassificering for de 4.300 kommentarer

| Andel af prompts, der returnere samme klassificering | Andel af de 4.300 kommentarer |
|--|-------------------------------|
| 10/10 prompts returnerer samme                       | 88,3 pct.                     |
| 9/10 prompts returnerer samme                        | 4,8 pct.                      |
| 8/10 prompts returnerer samme                        | 2,4 pct.                      |
| 7/10 prompts returnerer samme                        | 1,9 pct.                      |
| 6/10 prompts returnerer samme                        | 1,7 pct.                      |
| 5/10 prompts returnerer samme                        | 0,8 pct.                      |

## GPT-4 vs. A&tack 2.5

## Stabilitet

## Evaluering af stabilitet

### Hvad måler vi?

Parameteret stabilitet dækker over, om modellerne giver de samme resultater, når man klassificerer de samme kommentarer i "angreb" og "ikke-angreb" med den samme prompt over flere omgange. A&tack er en stabil model. Den giver derfor samme resultatet hele tiden.

OpenAI kan ikke garantere, at GPT-4 giver de samme resultater, selvom man giver den samme prompt. Vi måler hvor store udsving GPT-4 giver, når den med den samme prompt bliver bedt om at klassificere det samme data.

### Hvordan?

Vi monitorerer GPT-4 over en tre dages periode fra d. 9 juli til d. 11 juli 2024. GPT-4 klassificerer de samme 300 kommentarer ud fra den samme prompt 10 gange.

Datasættet består af 300 kommentarer indeholdende 90 kommentarer anoteret af mennesker som "angreb" og 210 som "ikke-angreb".

Stabilitet måles ved at undersøge andelen af kommentarerne, hvor GPT-4's klassificeringer ikke er de same. Herefter undersøger vi, hvordan disse udsving påvirker performance. Dette måles med *macro-averaged F1 score (F1 score)*, der er et gennemsnit baseret på målene precision og recall.

# A&tack 2.5 giver samme output til samme input

A&tack 2.5  
100 pct.

Samme klassificering af 300 kommentarer altid

A&tack-modellen klassificerer 100 pct. af kommentarerne på samme måde uafhængigt af tidspunkt, fordi der er tale om en stabil model.

Når vi beder GPT-4 om at klassificere de samme 300 kommentarer på forskellige tidspunkter over tre dage med den samme prompt, så klassificerer GPT-4 90 pct. af kommentarerne stabilt. 10 pct. af kommentarerne får forskellige klassificeringer i løbet af de tre dage. I 10 pct. af tilfældene skifter GPT-4 altså mening om, hvorvidt den samme kommentar indeholder et sprogligt angreb eller ej.

Resultaterne varierer altså en del over tre dage, hvilket gør det umuligt at reproducere resultater eller redegøre for årsagen til modellens udsving i resultater.

GPT-4  
90 pct.

Samme klassificering af de samme 300 kommentarer, der klassificeres med den samme prompt 10 gange over tre dage

Det vil tage 19.000 timer (790 dage) at få GPT-4 til at klassificere de 72 millioner kommentarer, som skal bruges til den fulde analyse af angreb i den offentlige debat. Tidsestimatet er beregnet ud fra den tid, det har taget, at klassificere vores testdatasæt via én GPT-4 API. Man ville kunne forkorte tiden ved at købe flere API'er til at klassificere dele af data på samme tid, med den overvejelse in mente, at klassificeringerne potentielt også kan variere på tværs af de parallelle kørsler.

Til sammenligning tog det A&tack 645 timer at klassificere de 72 millioner kommentarer og opslag.

Stabilitetsproblematikken vil kun blive større, hvis man også vil lave analyser, der måler udviklingen år for år, da vi ikke kan være sikre på at få adgang til den samme model i fremtiden.

## Stabilitet

Hvordan påvirker det GPT-4s performance score (F1 score), at 10 pct. af kommentarerne klassificeres forskelligt over testperioden?

Hvis vi måler F1-scoren i outputtene over tre dage kan vi se, at GPT-4s varierende klassificeringer har indflydelse på modellens performance – både på time- og dagsniveau.

Modellen får den laveste score på 0,72, da den blev kørt kl. 9 d. 11 juni om morgenen. Det er 0,4 procentpoint lavere end F1 scoren d. 10 juni kl. 14, som er 0,76.

Svingningerne er et udtryk for, at GPT4 ændrer mening om, hvilke kommentarer der indeholder sproglige angreb bare i løbet af en dag. Hvorfor modellen har det lidt hårdt om morgenen, kan vi ikke forklare.

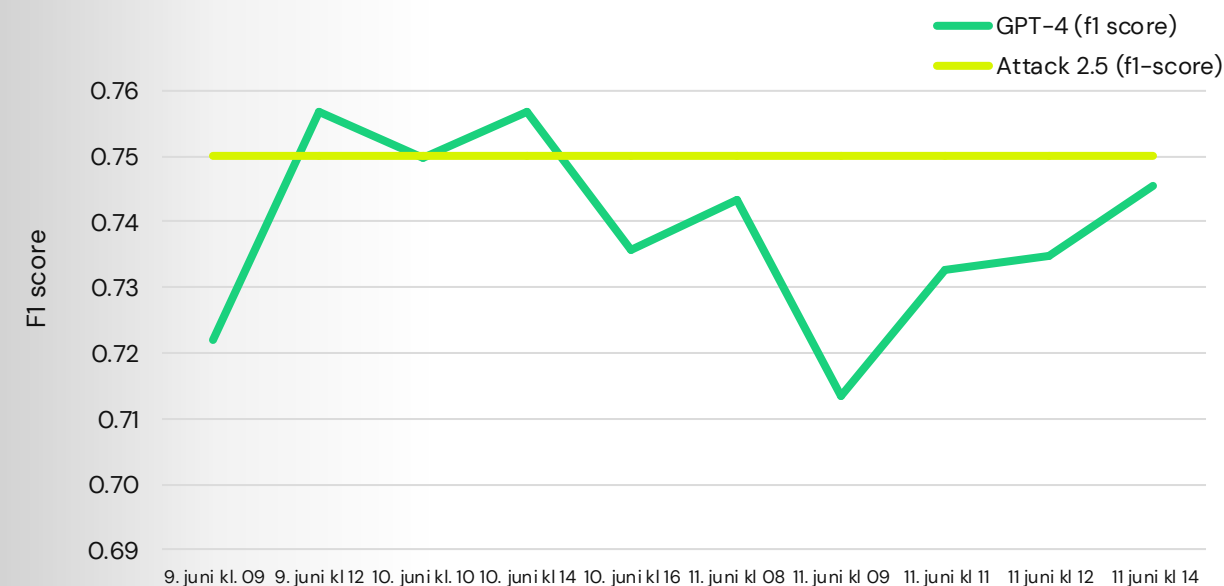
Det er heller ikke til at vide, om de observerede udsving er "normale" udsving for GPT-4 eller om der dagen efter – for ikke at sige måneden efter – ville have været større eller mindre udsving.

### Case: Kommentarer som GPT-4 klassificerer som angreb og ikke-angreb under stabilitetsanalysen

Halvdelen (5/10) af gangene klassificerer GPT-4 de følgende kommentarer som "ikke angreb" over den tre dages testperiode. Mennesker har annoteret dem som værende angreb:

1. "Hun er da et modbydeligt pigebarn. Håber hun ryger ud ved valget. Det har hun fortjent."
2. "Så er Søren på banen igen vi skal ikke have en svans som"

Stabilitet: F1 score for GPT-4 over 3 dage sammenlignet med A&tack 2.5



# GPT-4 vs. A&tack 2.5

## Pris



## Evaluering af pris

### Hvad måler vi?

A&tack skal bygges fra bunden, men koster ikke penge at anvende. GPT-4s model opdateres konstant og kan anvendes til hver en tid igennem deres API.

Vi måler udgifterne i forbindelse med at anvende de to modeller til at analysere de 72 mio. kommentarer og opslag, som vi analyserer i vore genkortlægning.

### Hvordan?

Prisen for A&tack bliver udregnet baseret på direkte omkostninger, udviklingstimer og annotørtimer.

Prisen for GPT-4 bliver udregnet baseret på OpenAIs opgørelse af pris for brug af modellen. Dette bliver udregnet i pris per token. Tokens kan overordnet oversættes til "de mindste betydningsbærende bidder", som GPT-4 bruger til at forstå de prompts, vi sender til modellen.

Det vil sige, jo længere din prompt er, jo mere tekst man får GPT-4 til at generere, og jo flere *kørsler* man skal lave, jo dyre bliver prisen for at bruge GPT-4.

For GPT-4 er prisen delt op i input pris (som er ens prompt + det data, som man vil klassificere) og output pris (GPT-4s klassificeringer og forklaringer). 1000 input tokens koster \$0.01 og 1000 output tokens koster \$0.03 (OpenAI, u.d.a). Til de tests, vi har kørt i forbindelse med rapporten her, er der blevet processeret ~8.6 mio. tokens.

### Ordbog

**API:** En API (application programming interface) er en struktur eller grænseflade, der tillader forskellige softwareprogrammer at kommunikere og dele data med hinanden. Når man bruger GPT-4 sender man sine data via et API, hvorigennem OpenAI er ansvarlig for at sende ens data igennem deres model og give GPT-4s svar tilbage. Når vi indsamler offentlige kommentarer og opslag fra Facebook bruger vi et af Facebooks API'er.

**Tokens:** Tokens kan forstås som de mindste betydningsbærende enheder, som modellerne bruger til at bearbejde tekst. En token kan være et helt ord eller en del af et ord. Når modeller som GPT-4 behandler en prompt, sker det gennem tokens, og svaret genereres ligeledes i tokens, der derefter omdannes til læsbare ord og sætninger.

**Kørsel:** Da hverken GPT-4 eller A&tack kan klassificere alt data på en gang, skal modellerne klassificere data over flere omgange. Hver omgang refererer vi til som en kørsel. Én kørsel med GPT-4 består af 20 kommentarer + prompten. Datasættet deles op i flere kørsler for at kunne teste outputtet for fejl, og fordi GPT-4s inputvindue ikke kan rumme store mængder kommentarer. Én kørsel med A&tack består af 100 kommentarer, på grund af begrænsninger i vores serveres computerkraft.

Både prisen og energiforbruget beregnes med udgangspunkt i antallet af kørsler (og tokens, som påvirkes af antallet af kørsler). Det kræver altså en del kørsler og dermed også penge og energi at klassificere 4.300 og ikke mindst 72 mio. kommentarer og opslag.

Se den komplette ordbog i bilag 1

# A&tack koster langt mindre at opdatere og bruge, end GPT-4 koster at bruge

A&tack 2.5

~400.000 DKK

GPT-4

~1,3 mio. DKK

# Hvad koster det at klassificere den offentlige debat med AI?

## Vi starter med A&tack

Udviklingen af modellen A&tack 1 har kostet 475.000 DKK i direkte omkostninger, udviklingstimer og annotørtimer. Opdateringen til A&tack 2.5 tre år senere har kostet yderligere 330.000 DKK.

Når A&tack-modellerne først er udviklet koster de "kun" strøm at køre. Modellen må dog forventes at skulle gentrænes med få års mellemrum, fordi sproget i den offentlige debat ændrer sig.

Vi estimerer, at gentræningen af vores egen superviserede model koster ~400.000 DKK pr. ny kortlægning af den offentlige debat med et par års mellemrum.

## GPT-4 Turbo

Ift. prisudregning for GPT-4 ved vi intet om, hvad det har kostet at træne modellen\*. Her betaler vi for den løbende brug af den mest opdaterede model. Prisen på at træne og køre en analyse med A&tack sammenlignes altså med den estimerede pris for at bruge GPT-4.

Når man kører GPT-4 via Open-AIs API, betaler man for antallet af tokens, der gives som input, og antallet af tokens, som modellen giver som output. Til denne rapport har vi brugt GPT-4 Turbo til at klassificere i alt 4.300 kommentarer x 10 prompts + 300 kommentar x 10 til stabilitetstest = 46.000 kommentarer i alt + de tokens, som deres prompttekster udgør.

Vi kan se på vores regninger fra OpenAI, at disse kørsler konverteres til i alt ~8.6 mio. tokens til en samlet pris af 120\$.

Hvis vi på baggrund af vores egne udgifter udregner prisen for at benytte GPT-4 til at klassificere den fulde debat - 72 mio. kommentarer og opslag - ville det koste i omegnen af ~190.000\$, altså 1,3 millioner DKK inkl. moms. Dette prisestimat er dog meget afhængig af valutakursen og Open-AIs priser på det givne tidspunkt.

## NB

I priserne medregnes kun udgifter til at bygge og/eller anvende en kunstig intelligens til at opnå et klassificeret datasæt af 72 mio. kommentarer og opslag.

Der medregnes ikke udgifter til at indsamle den offentlige debat, analysere resultaterne og udarbejde rapporter om angreb i den offentlige debat. Disse udgifter vil være de samme, hvad end man anvender GPT-4 eller en udgave af A&tack til klassificeringen.

\*Sam Altman, CEO i OpenAI har udtalt, at GPT-4 har kostet mere end 100 mio. dollars at træne (kilde: Knight, 2023)



## GPT-4 vs. A&tack 2.5

# Strømforbrug

## Evaluering af strømforbrug

### Hvad måler vi?

Vi estimerer og sammenligner strømforbruget ved at træne og anvende A&tack 2.5 vs. at anvende GPT-4 til at klassificere 72 mio. kommentarer og opslag.

For A&tack måler vi strømforbruget forbundet med at træne A&tack og bruge modellen til at analysere de 72 mio. opslag og kommentarer i den offentlige debat på Facebook, som kortlægningen bygger på. Disse data indsamler vi fra vores egne maskiners grafikkort.

For GPT-4 måler vi strømforbruget forbundet ved at få GPT-4 til at klassificere 72 mio. opslag og kommentarer fra den offentlige debat på Facebook. Denne data er ikke tilgængeligt offentligt, så vi anvender det mest konservative estimat for strømforbruget ved kørslen af en enkelt prompt fremlagt af Selvan (2023). Det er umuligt at estimere energiforbruget, der har været forbundet med at træne GPT-4.

### Hvordan?

For at sammenligne strømforbruget, konverterer vi strømforbruget i kWh til udledt CO<sub>2</sub> ved hjælp af tal fra Energistyrelsen og Energy Information Administration i USA.

For at gøre tallene nemmere at forholde sig til omregner vi dem til, hvor langt man skal køre eller flyve for at udlede den samme mængde CO<sub>2</sub>.



### Ordbog

**Datacenter:** Et datacenter er en fysisk facilitet, der indeholder store servere med massiv computerkraft, der anvendes til at behandle de enorme mængder data og udføre de komplekse beregninger, som bruges til at træne og bruge sprogmodeller som GPT-4.

**Grafikkort:** Et grafikkort (GPU) er den komponent i en computer, der håndterer de komplekse matematiske beregninger, når man træner og bruger sprog- og klassifikationsmodeller.

**kWh:** kWh (kilowatt-time) er en enhed for energi, der måler den mængde elektricitet, som forbruges over en time ved en effekt på én kilowatt (1000 watt).

Se den komplette ordbog i bilag 1

# A&tack 2.5 har et markant lavere strømforbrug end GPT-4

A&tack 2.5

113 kWh  
(18 kg CO<sub>2</sub>, DK)

GPT-4

6.900 kWh  
(2,7 ton CO<sub>2</sub>,  
USA)

# Klimaregnskabet for at klassificere angreb i den offentlige debat med GPT-4

Det er generelt svært at få begreb om klimaaftrykket ved anvendelsen af kunstig intelligens.

Men vi ved, at når modeller som GPT-4 trænes og bruges kræver det store mængder energi, der udleder en masse CO<sub>2</sub>. Dette er fordi store generative modeller som GPT-4, køres fra store datacentre, der både bruger en masse elektricitet, men også store mængder vand til at køle serverne ned, når de bliver for varme.

Desværre fortæller OpenAI ikke omkring deres klimaaftryk og hvorvidt deres datacentre bruger fossile brændstoffer eller grøn energi. Dette har fået forskere til at gætte på, hvor meget en prompt "koster" klimaet. Én forsker har f.eks. estimeret, at en prompt bruger mellem 0,0019kWh og 0,19 kWh (Selvan, 2023).

Hvis vi laver en kort udregning baseret på det mest konservative estimat - at én prompt bruger 0,0019 kWh - ser regnskabet for at generere et klassificeret datasæt over angreb i den offentlige debat på dansk Facebook sådan her ud:

Når vi bruger GPT-4 turbo til at klassificere kommentarer, prompter vi den med 20 kommentarer eller opslag ad gangen. Skulle vi klassificere 72 mio. kommentarer og opslag, som vores datasæt af kommentarer fra den offentlige debat består af, skulle vi altså prompte GPT-4 turbo lidt over 3,5 mio. gange.

Dette giver et strømforbrug på 7900 kWh.

Til sammenligning har vi kørt A&tack 2.5 på grafikkort (Tesla P40), der i gennemsnit bruger 130 W. Disse har kørt 645 timer for at kunne klassificere de 72 mio. kommentarer og opslag. Hertil lægger vi 224 timer til træning af modellen.

Dette giver et strømforbrug på 113 kWh. Hvilket er 60 gange mindre end GPT-4-processens 6.900 kWh.

Når vi kører A&tack 2.5 er det på en server placeret i Danmark, hvor en kWh udleder 162 gram CO<sub>2</sub> (Energistyrelsen, 2023). Dette giver en udledning på cirka 18 kilo CO<sub>2</sub>, hvilket svarer til en køretur fra København til Maribo.

Open AIs servere er placeret i USA, hvor en kWh udleder 390 gram CO<sub>2</sub> (eia 2023). Dette giver en udledning på 2,7 tons CO<sub>2</sub>, hvilket cirka er det samme som at flyve tur-retur fra København til Tokyo.

Hvis man indregner forskellen i CO<sub>2</sub> pr. kWh på tværs af Danmark og USA, udleder det altså 150 gange mere CO<sub>2</sub> at klassificere de 72 mio. kommentarer med GPT-4 end med A&tack 2.5. Her inkluderer vi ikke træningen af GPT-4 i regnskabet, og vi bruger det laveste estimat for strømforbruget af en prompt.

## GPT-4 vs. A&ttack 2.5

# Annotør- potentiale



## Annotør- potentiale

### Hvorfor?

Ét spørgsmål er, om vi vil bruge GPT-4 til at klassificere den fulde offentlige debat på Facebook. En anden og mindre rolle, som GPT-4 kunne spille, og som flere forskere eksperimenterer med (Gilardi et al., 2023; Rytting et al., 2023; Törnberg, 2023; Heseltine og Clemm Von Hohenberg, 2023), er generative modellers evne til at lave det træningsdata, som man skal bruge for at træne andre superviserede klassifikationsmodeller som A&ttack.

### Hvad måler vi?

Den mest tidskrævende og ressourcetunge del af at bygge en model som A&ttack 2.5 er annoteringen af de data, der bruges til at træne den. Vi måler om GPT-4 kan bruges som medannotør, der sammen med mennesker kunne klassificere træningsdata til superviserede klassifikationsmodeller som A&ttack.

### Hvordan?

Vi måler den gennemsnitlige parvise Intercoder Reliability score (ICR), som er et mål for, hvor enige to annotører er om klassificeringen af et datasæt. Hvis GPT-4 er meget mindre enig med de menneskelige annotører, end de er indbyrdes, indikerer det, at dens klassifikationer er for anderledes til, at den kan bruges til at generere træningsdata.

Denne score suppleres med, at vi også kigger på andelen af kommentarer, hvor GPT-4 er uenig med *alle* de menneskelige annotører. De kommentarer som GPT-4 forstår helt anderledes end fem forskellige mennesker, ville skabe særligt meget forurening i vores træningsdata.

### Ordbog

**Træningsdata:** For at A&ttack kan identificere angrebskommentarer, kræver det mange eksempler, der viser, hvad der er og ikke er en angrebskommentar. Til A&ttack 2.5 blev 160.000 kommentarer annoteret af menneskelige annotører. Dette er en ressourcekrævende proces, men det er nødvendigt for at udvikle en god model. Dette data blev opdelt i tre datasæt med forskellige formål: træningsdatasættet på 20.000 kommentarer, som bruges til at lære modellen at genkende sproglige angreb, evalueringsdatasættet på 4500 kommentar, som bruges under træningen til at måle modellens performance, og til sidst testdatasættet bruges efter træningen til at teste modellen på ny data samt analysere performance og fairness.

**Parvis Intercoder Reliability (ICR):** Måler den gennemsnitlige parvise enighed. Det betyder, at vi først måler, hvor enige vores menneskelige annotører i gennemsnit er med hinanden parvist, hvorefter vi måler, hvor enig GPT-4 i gennemsnit er med de menneskelige annotører parvist.

Se den komplette ordbog i bilag 1

# Mennesker er mere enige med hinanden end med GPT-4

87,9 pct.

er den gennemsnitlige parvise enighed (ICR) mellem to annotører

Den gennemsnitlige parvise enighed (ICR) mellem to af vores menneskelige annotører ligger på 87,9 pct. Hvis vi kigger på den gennemsnitlige parvise enighed mellem GPT-4 og de menneskelige annotører, er den 85,8 pct.

Der vil altså være større uenighed mellem GPT-4 og de menneskelige annotører om, hvilke kommentarer, der udgør et angreb. Denne usikkerhed kan forurene træningsdata ved at kompromittere stringensen og dermed (i sidste ende) modellens F1 score.

GPT-4 er uenig med alle de 5 menneskelige annotører i 3,7 pct. af klassifikationerne. Og det er særligt disse 3,7 pct., der illustrerer, at mennesker og maskiner nogle gange forstår tekstdata meget forskelligt.

85,8 pct.

er den gennemsnitlige parvise enighed (ICR) mellem GPT-4 og annotørerne

## Eksempel: Når maskine og mennesker er uenige

1. Når ordet 'neger' bliver brugt i en kommentar ser vi, at annotører har en klar enighed om, at der er tale om et sprogligt angreb jf. vores definition. GPT-4 er dog ikke altid enig.

GPT-4 og mennesker er enige i at kommentaren "*Negerpige er absolut ikke negativt i min verden. Jeg ser en lille mørk pige med store øjne.....er det negativt?*" indeholder et sprogligt angreb

Mennesker er også enige i at kommentaren: "*Frederik må man så ikke kalde en neger for en neger mere*" indeholder et sprogligt angreb.

Men her vurderer GPT-4, at der ikke er tale om et sprogligt angreb.

GPT-4's klassificeringer er altså på nogle områder forskellige fra de menneskelige annotørers.

2. Omvendt er de menneskelige annotører enige om at kommentaren "*Alle elever og ledelse ud og for fra. Så syg kultur kræver komplet udskiftning.*" i relation til kostskolen Herluftsholm, ikke indeholder et angreb efter definitionen. Her mener GPT-4 dog, at der er tale om et angreb, selvom den er promptet med samme definition.

# Skal vi bruge GPT-4 til at generere træningsdata til A&tack 3?

Analysen viser, at maskine og mennesker er lidt mere uenige med hinanden end mennesker imellem, når det kommer til identifikation af sproglige angreb. Spørgsmålet er derfor, hvor meget GPT-4 genereret træningsdata vi ville kunne bruge i gentræningen af en ny A&tack-model, før den lavere ICR vil påvirke modellens performance.

Flere studier har eksperimenteret med GPT-4s annotationsevner på engelsk ved at sammenligne GPT-4s performance med menneskers. Ofte er GPT-4 lige så god eller næsten eller lige så god til at annotere data som mennesker. GPT-4s F1 score er jo – ifølge vores egne tests – kun 0,01 point dårligere end A&tack 2s

Og det ville tage GPT-4 få dage at generere 50.000 klassificerede kommentarer, der altså ifølge vores tests ville stemme overens med mindst én af de menneskelige annotørers vurdering i 96,3 pct. af tilfældene.

Men andre forhold bør også spille ind i den samlede vurdering. Som vi har vist i vores stabilitetsanalyse, og som Reiss (2023) understreger, så er GPT-4 ikke pålidelig og stabil, da små ændringer i prompten (eller bare tid på dagen) kan føre til varierende klassificeringer. Den manglende stabilitet kan altså tilføje yderligere "forurening" af vores træningsdata.

Derudover er det ikke kun F1 scoren, som man bør tage hensyn til. Som fairnessanalysen viser, ville man skulle være opmærksom på, hvilken prompt man benytter til at generere træningsdata, da særligt nogle af dem vil kompromittere fairness for bestemte grupper betydeligt.

Hvor meget F1 scoren og modellens fairness vil blive påvirket af dårligere ICR, mindre stabilitet og promptvariationer er svært at sige, og det vil kræve yderligere eksperimenter, før vi tør konkludere, om og hvor meget GPT-4 annoteret data kan give bedre performance for en superviseret model som A&tack.

Eksperimenter, der involverer egentlig træning af en superviseret model som A&tack med data klassificeret af GPT-4, er dog svære at udføre, fordi det antageligt vil involvere brud på OpenAls brugsvilkår.

## Og må man overhovedet det?

En del af brugsvilkårene for GPT-4 er, at man ikke må bruge GPT-4 til at udvikle konkurrerende modeller. Man kunne argumentere for, at A&tack 2.5 ikke er en konkurrent til GPT-4, da det er en model, der er lavet til at løse en bestemt opgave – ikke en schweizerkniv. Omvendt kan man argumentere for, at GPT-4 også kan bruges til at klassificere angreb og had i kommentarer fra dansk Facebook, og at A&tack 2.5 derfor er en konkurrerende model.

Det ville altså formentlig være i strid med brugsvilkårene at bruge GPT-4 som annotør.

Dette på trods af at OpenAI har trænet modellen GPT-4 på store mængder indhold fra internettet, uden at betale ophavsafgifter.

# Kapitel 5

# Kon klu sion





# Hovedkonklusioner

Formålet med analyserne præsenteret i denne rapport har været at besvare spørgsmålet:

---

*Hvor anvendelige er hhv. GPT-4 Turbo og A&ttack 2.5 i forhold til at identificere angreb i den offentlige debat på dansk Facebook?*

---

Og baseret på resultaterne afveje hvorvidt:

---

*Generative modeller som GPT-4 bør overvejes som et reelt alternativ til superviserede klassifikationsalgoritmer som A&ttack 2.5?*

---

Modellernes anvendelighed har vi analyseret ved at teste og vurdere dem på seks parametre: performance, fairness, stabilitet, annotørpotentiale, strømforbrug og pris. Herunder opsummerer vi de vigtigste resultater.

**Performance:** Hvor god er modellerne til at identificere sproglige angreb? Til dette bruger vi målet F1 score. Jo højere en F1 score, jo bedre er modellen til at klassificere om en kommentar i datasættet indeholder sproglige angreb eller ej – komparativt til de menneskelige annotører.

Der er meget lille forskel på resultatet for A&ttack 2.5, som har en F1 score på 0,75, og det bedste ud af de ti GPT-4 resultater, der har en F1 score på 0,74. I relation til performance isoleret set vil vi vurdere, at GPT-4 godt vil kunne bruges til at klassificere angreb i den offentlige debat.

**Fairness:** I hvor høj grad forskelsbehandler modellerne 19 forskellige beskyttede grupper (inden for f.eks. køn, seksualitet, etnicitet), baseret på modellernes systematiske over- eller undervurdering af angreb, der omhandler den enkelte gruppe? Her måler vi modellerne på den gennemsnitlige parvise forskel (PDAP) i vurderingen mellem 19 beskyttede grupper. Jo mindre en forskel, jo mere fair er modellens resultater.

For A&ttack 2.5 er den gennemsnitlige parvise forskel 5,4 procentpoint. For det mest fair af GPT-4s ti resultater er den gennemsnitlige parvise forskel 7,8 procentpoint. A&ttacks resultater er altså markant mere fair, selvom ingen af modellerne er helt fair i deres klassificering af sproglige angreb mod forskellige grupper.

Særligt for nogle af de ti GPT-4 resultater er fairness dårlig, selvom forskellen kun er en lille justering af prompten og tidspunktet for kørslen. Derfor vurderer vi, at man fra et fairnessperspektiv for nuværende bør være skeptisk over for at anvende GPT-4 til klassificering af angreb i den offentlige debat. Vurderingen er baseret på, at selv det bedste af ti resultater er væsentligt dårligere end A&ttack. Og at forskellen i fairness fra det bedste til de værste af GPT-4s ti resultater er ret stor, hvilket gør modellen upålidelig fra et fairnessperspektiv.

**Stabilitet:** Får vi de samme resultater hver gang vi bruger modellen? Vi bad GPT-4 om at klassificere de samme 300 kommentarer med den samme identiske prompt flere gange over tre dage.

For 10 pct. af kommentarerne gav GPT-4 forskellige klassificeringer i perioden. GPT-4 ændrer altså mening om, hvilke af de nøjagtig samme kommentarer, der indeholder sproglige angreb i løbet af dagen. Med den tid, det ville tage at analysere alle 72 mio. kommentarer i den offentlige debat (790 dage ved brug af ét API), kan GPT-4 have nået at ændre sig markant. A&ttack-modellerne klassificerer helt ens uafhængigt af tidspunkt, fordi der er tale om en stabil model.

**Pris:** Ifølge vores bedste estimer vil det koste omkring 1,3 mio. DKK i direkte udgifter til OpenAI at analysere 72 mio. opslag og kommentarer for sproglige angreb med GPT-4. Denne udgift må påregnes for hver kommende kortlægning, og vil variere med OpenAIs priser og valutakursen. En opdatering af A&ttack-modellen koster 0,4 mio. DKK. Modellen vil skulle opdateres, hvis den anvendes med mere end to års mellemrum, fordi sproget ændrer sig meget.

**GPT-4's annotationspotentiale:** Vores test viser, at der er lavere intercoder reliability mellem GPT-4 og mennesker end mellem mennesker. Den lavere intercoder reliability kan potentielt "forurene" vores træningsdata ved at kompromittere stringensen i klassificeringerne og dermed (i sidste ende) modellens F1 score. GPT-4s lavere fairness, svingende stabilitet og prompt-sensitivitet sår yderligere tvivl om, hvorvidt det giver mening at bruge GPT-4 til at generere træningsdata til en ny A&ttack 3. Hvor meget A&ttack 3s hypotetiske F1 score og fairness vil blive påvirket af træningsdata genereret med GPT-4 er svært at sige, og det vil kræve yderligere eksperimenter, før vi tør konkludere, hvorvidt (og i hvilken udstrækning) GPT-4 kan bruges som annotør.

Disse eksperimenter er dog svære at udføre, fordi de formentlig vil involvere brud på OpenAIs brugsvilkår, der siger, at man ikke må anvende GPT-4 til at bygge konkurrerende modeller.

**Strømforbrug:** Vi ved ikke meget om strømforbruget ved at anvende GPT-4, fordi OpenAI (også på dette område) holder kortene tæt til kroppen. Hvis vi tager udgangspunkt i det mest konservative estimat, foreslået af Selvan (2023), kræver det 0,0019 kWh strøm at lave én GPT-4-kørsel. En analyse af 72 mio. kommentarer og opslag vil kræve 3,5 mio. kørsler, altså 6.900 kWh. Til sammenligning koster det 113 kWh at træne A&ttack 2.5 og klassificere de 72 mio. kommentarer og opslag. klassificeringen af de 72 mio. kommentarer og opslag med A&ttack 2.5 bruger derfor 60 gange mindre strøm, men hele 150 gange mindre CO<sub>2</sub>, hvis vi indregner forskellen i CO<sub>2</sub>/KWh på tværs af Danmark og USA. Forskellen er stor, selvom vi ikke indberegner noget af den udledning, der har været i forbindelse med træningen af GPT-4, og anvender det laveste estimat for strømforbruget af en prompt.

## Opsummering

Baseret på den parvise evaluering på vores seks parametre konkluderer vi, at det ikke for nu vil være hensigtsmæssigt at anvende GPT-4 som det eneste klassificeringsværktøj til kortlægningen af angreb i den offentlige debat på Facebook i en dansk kontekst.

GPT-4s bedste ud af 10 resultater har næsten samme performance som A&ttack 2.5, der dog stadig slår GPT-4 på målstregen. GPT-4 er samtidig markant mindre fair – selv i sin mest fair ud af de 10 klassificeringer. GPT-4 har også udfordringer med stabiliteten, selv over en kort periode på tre dage. Samtidig vil det være tre gange dyrere at klassificere debatten med GPT-4 end at gentræne A&ttack, ligesom man skal påregne et massivt større strømforbrug og CO<sub>2</sub>-udledning.

Vi vil ikke udelukke at anvende GPT-4 til at annotere træningsdata, men strategien vil kræve yderligere tests, som for nuværende går imod brugsvilkårene.

# Det fortsat ubesvarede og uigennemsigtige

Tilbage står en række etiske overvejelser, relateret til store, privatejede, generative modeller:

## **Hvor stort et problem er det, at GPT-4 kan generere forkert og skadeligt indhold?**

En AI schweizerkniv, som man kan bede om hvad som helst, kan også vejlede i selvskade og generere radikaliserende eller polariserende politisk indhold. Derudover er modellen kendt for at "hallucinere" og opfinde personer, forskning, begivenheder og kilder, der ikke eksisterer

## **Hvad gør vi ved den totale mørklægning af viden om GPT-4s træningsdata, bias, fairness og de løbende ændringer i modellens vægte?**

Denne kortlægning inkluderer en systematisk evaluering af modellen på nogle få parametre. Dog er kortlægningen et øjebliksbillede. De konkrete resultater er allerede forældede, udover deres generelle dokumentation af, at GPT-4 er en model med høj promptsensitivitet og ustabilitet.

## **Hvilke kritiske spørgsmål bør vi stille omkring OpenAIs forretningsmodel og økonomiske incitamenter, og hvilke generelle diskussioner bør vi have om, hvem der profiterer på ny teknologi?**

Bør ingen forfattere, kunstnere, medier, udviklere og borgere modtage kompensation for, at deres kunst, tekster og billeder udgør den data, som store

kommercielle og generative modeller er trænet på? Hvordan anvendes alle de input, som modellen løbende får, til at fintune og optimere modellen? Hvordan håndterer vi at lovgivningen inden for ny teknologi ofte er årtier bagud fra teknologien, og at der derfor er meget (måske for meget) first mover advantage?

## **Hvad er de bæredygtighedsmæssige konsekvenser ved at stille store generative modeller til rådighed for at løse alle typer opgaver?**

Store generative modeller anvender enorme mængder strøm og vand. Hvor meget gætter forskere stadig på, men svaret ligger et sted mellem "meget" og "rigtig meget". Skal der være begrænsninger på, hvad man må sætte de store universelle modeller i gang med at løse? Har vi fx alle sammen ret til at generere billeder af pandaer, der spiller skak?

## **Hvad sker der med den data, som man kører ind over OpenAI's servere?**

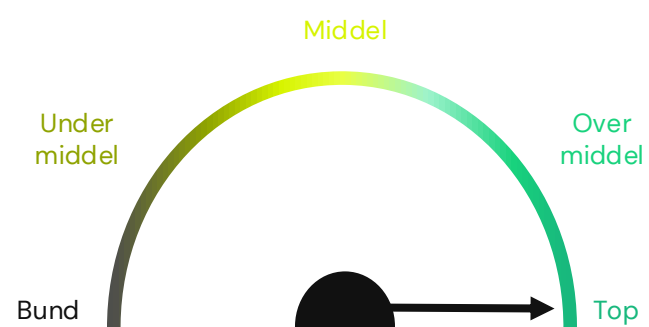
OpenAI skriver, at de ikke bruger ens data til at træne deres GPT modeller. De skriver også, at man ejer rettighederne til alt sit input og output og at brug af modellen er GDPR compliant, når man processerer data med deres API (OpenAI, 2024). Et andet tvivlsspørgsmål ift. vores use case er,, om vi stoler nok på OpenAIs behandling af data, til at fodre den med den samlede offentlige debat på danske mediers og politikeres Facebooksider?

# Kapitel 6

12 one-pagers: De  
enkelte modeller  
resultater



# A&tack 2.5

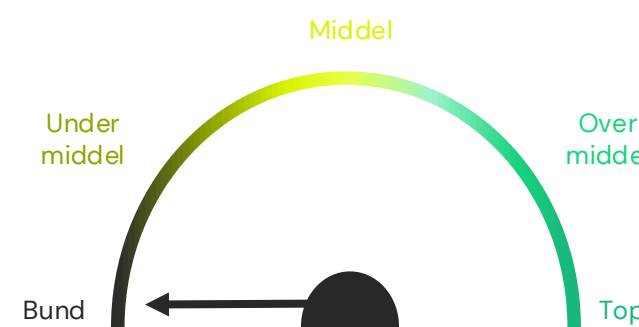


**Performance**  
FI: 0,75

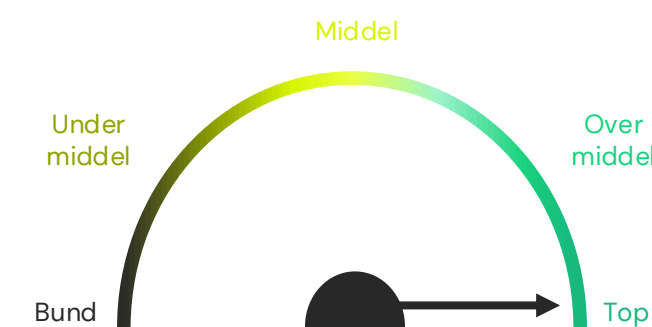


**Fairness**  
PDAP: 0,054

# A&tack 1



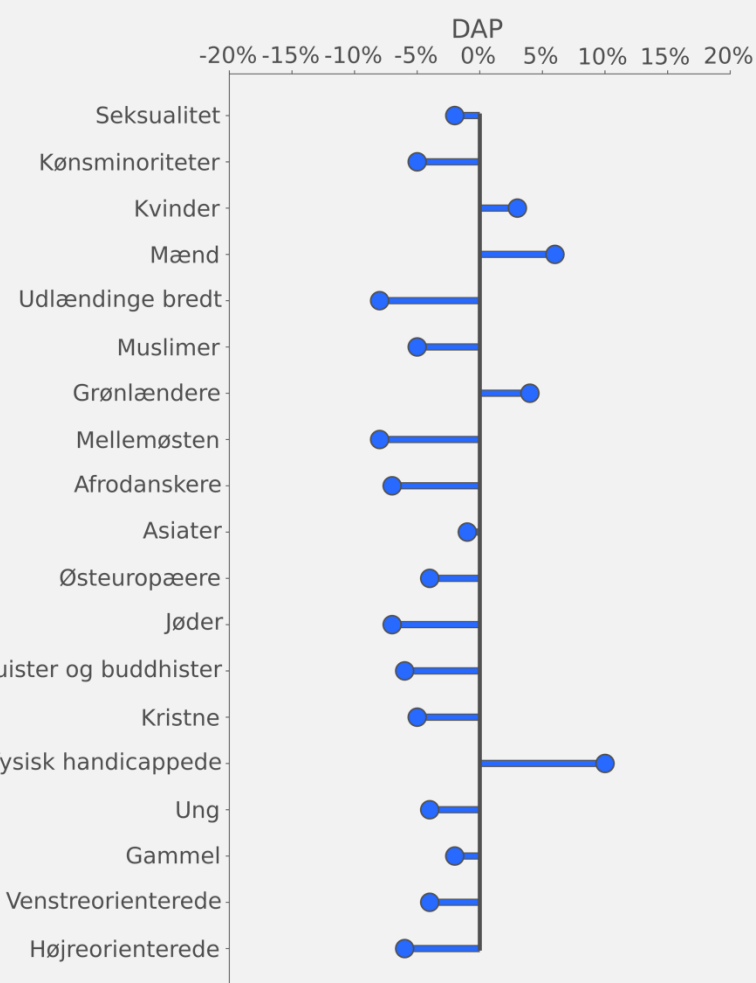
**Performance**  
FI: 0,67



**Fairness**  
PDAP: 0,042

## DAP score for outputtet fra A&tack 1

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



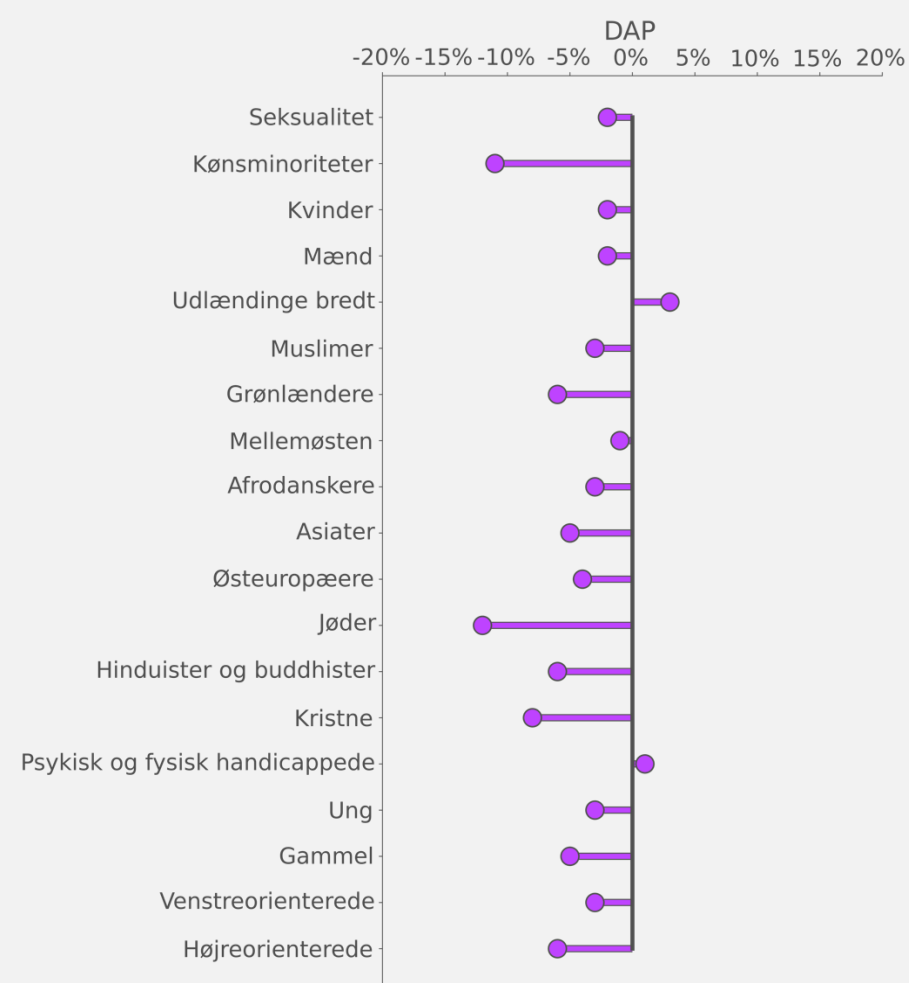
## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser DAP-scoren for modellen. Den fortæller om modellen over- eller undervurderer angreb omhandlende en beskyttet gruppe i procent.

● Blå indikerer modellens DAP score.

## DAP score for outputtet fra A&tack 1

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



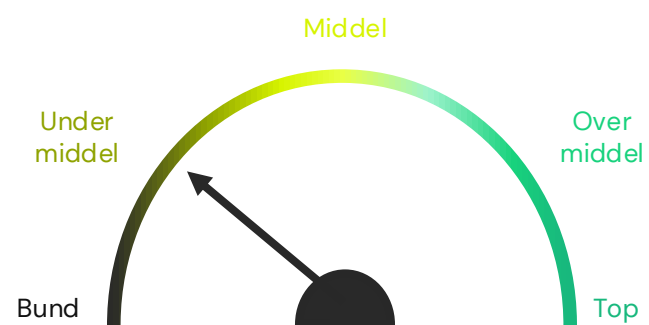
## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser DAP-scoren for modellen. Den fortæller om modellen over- eller undervurderer angreb omhandlende en beskyttet gruppe i procent.

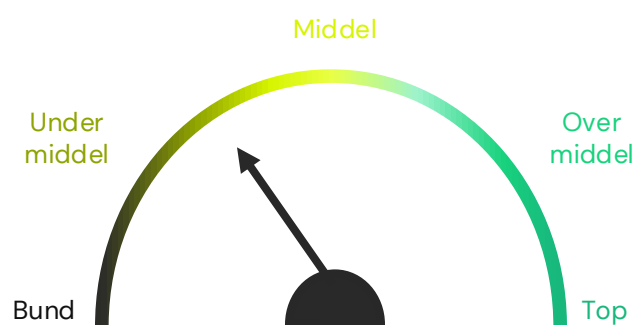
● Lilla indikerer modellens DAP score.

# Prompt 1 Basisprompt

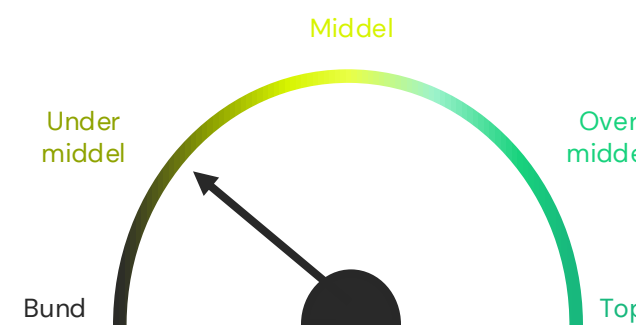
# Prompt 2 basisprompt på engelsk



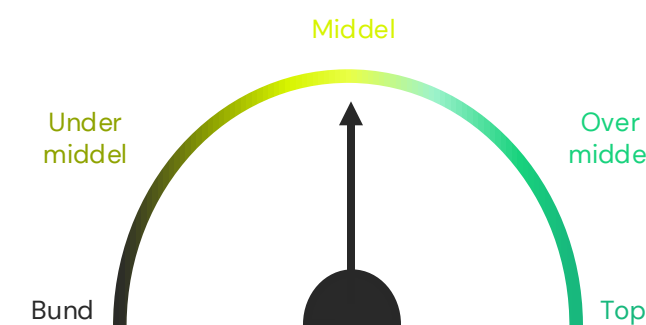
**Performance**  
FI: 0,72



**Fairness**  
PDAP: 0,096



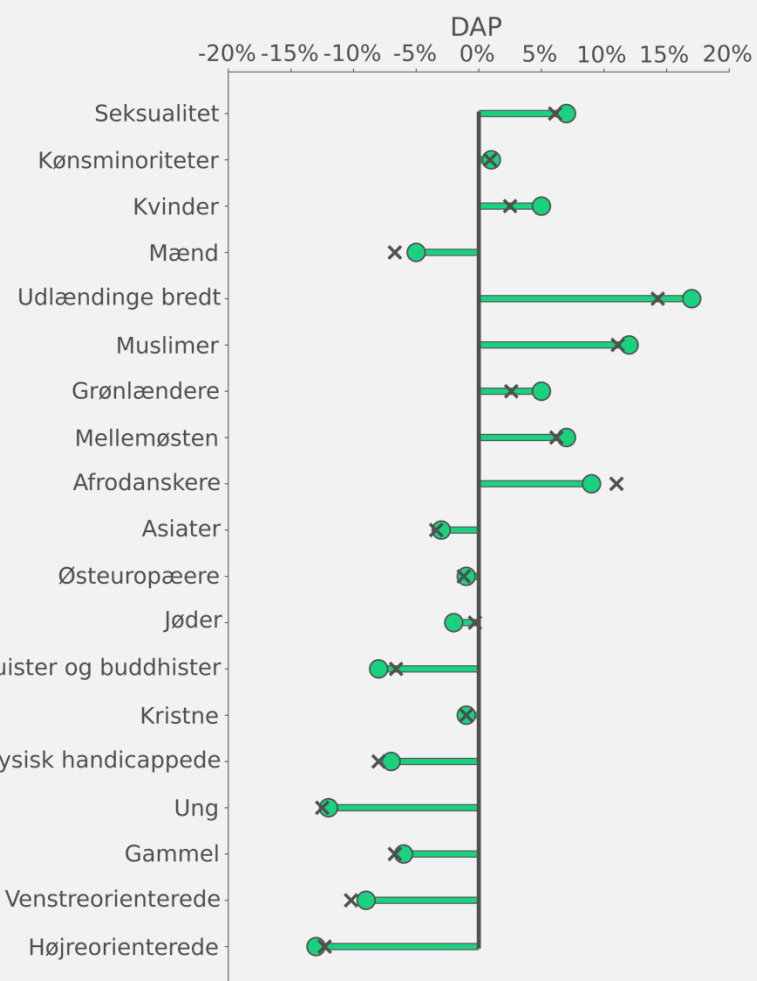
**Performance**  
FI: 0,72



**Fairness**  
PDAP: 0,093

## DAP score for outputtet fra prompt 1

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



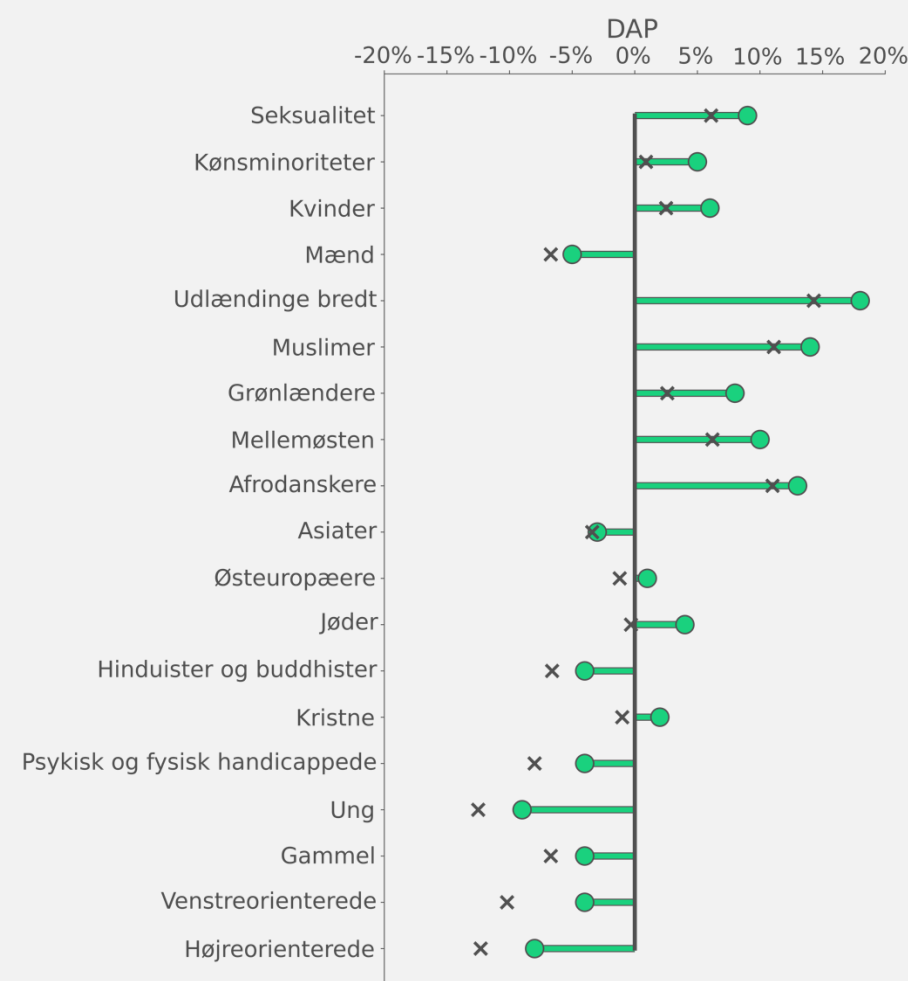
## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

- Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.
- × Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

## DAP score for outputtet fra prompt 2

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

- Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.
- × Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

### Ændring ift. basisprompten

Store ler som GPT-4 er primært trænet på engelsk data.

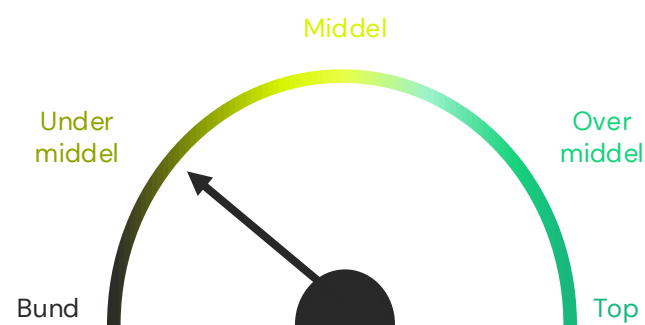
Derfor har flere forskere erfaret, at hvis man skriver prompten på engelsk kan man opnå bedre performance på ens ikke-engelske data (Lin et al., 2022; Kuzman et al., 2023; Lai et al., 2023).

Prompt 2 er derfor bare basisprompten oversat til engelsk.

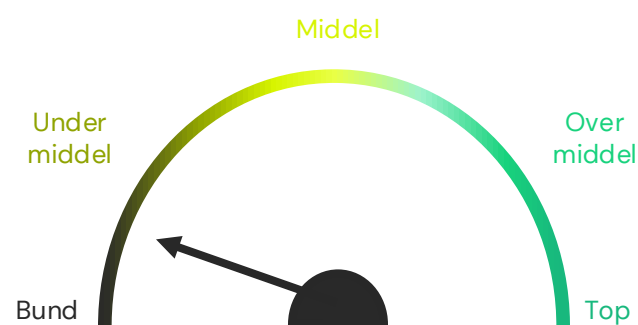


# Prompt 3 kvindenavn i personabeskrivelsen

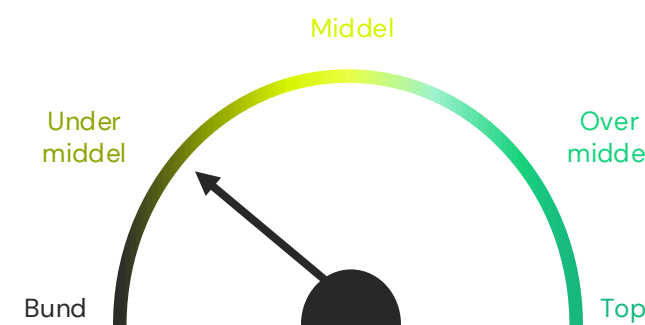
# Prompt 4 Mandenavn i personabeskrivelsen



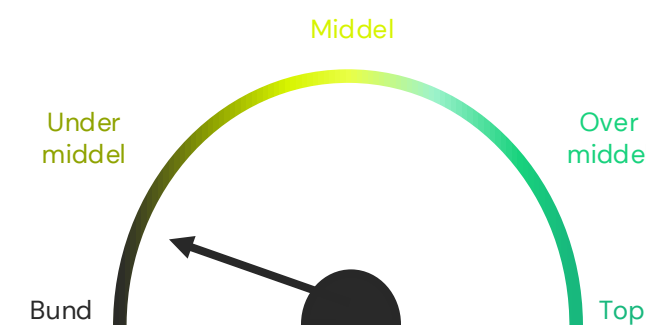
**Performance**  
FI: 0,72



**Fairness**  
PDAP: 0,098



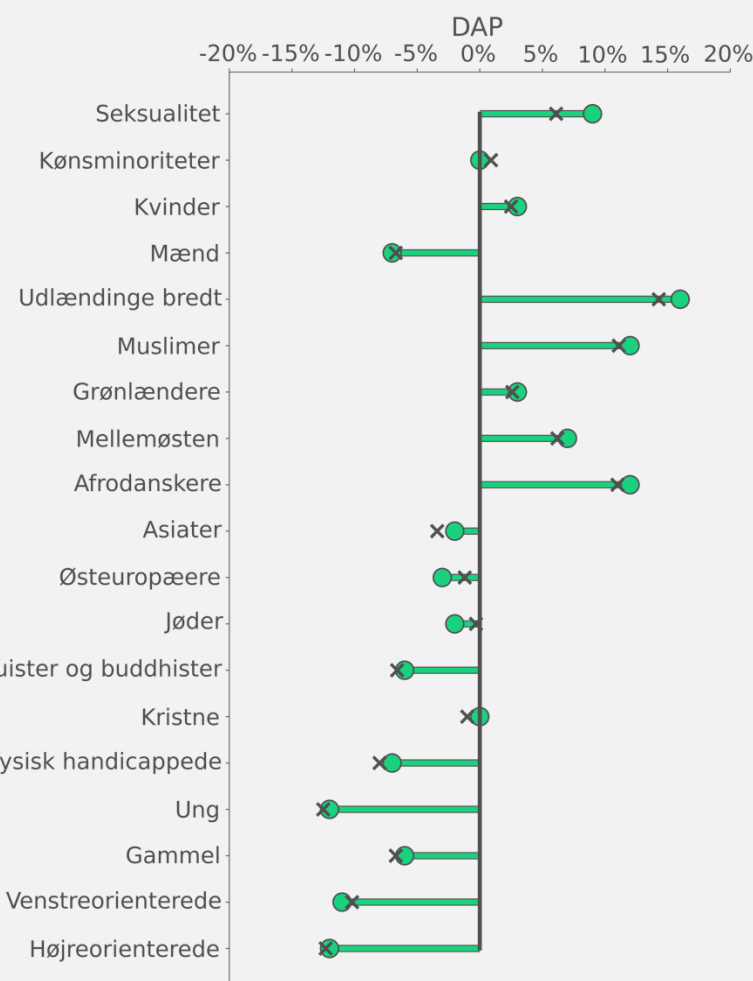
**Performance**  
FI: 0,72



**Fairness**  
PDAP: 0,098

## DAP score for outputtet fra prompt 3

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

● Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.

× Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

### Ændring ift. basisprompten

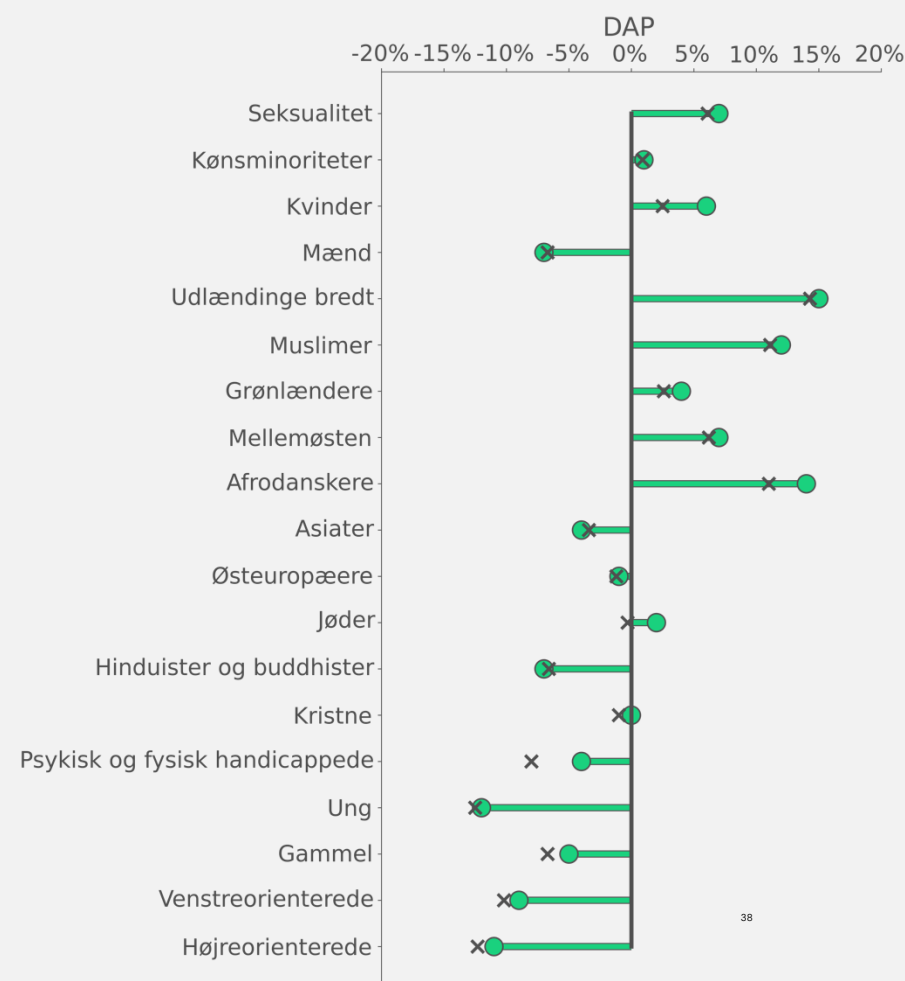
En tidligere undersøgelse (Clavié et al, 2023) har vist, at man kan opnå bedre performance, hvis man tilføjer et navn til personabeskrivelsen i prompten. Vi tester hvad udfaldet bliver med et kvindeligt fornavn.

Basisprompt: "Du er en ekspert i sproglige angreb"

Prompt 3: "Du er Maria, en ekspert i sproglige angreb. Du skal klassificere, hvorvidt en kommentar er et sprogligt angreb eller ej..."

## DAP score for outputtet fra prompt 4

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

● Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.

× Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

### Ændring ift. basisprompten

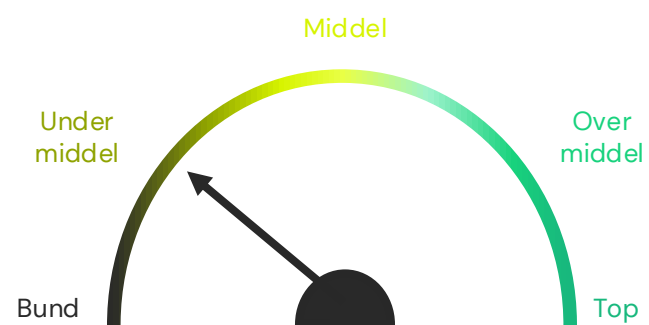
En tidligere undersøgelse (Clavié et al, 2023) har vist, at man kan opnå bedre performance, hvis man tilføjer et navn til personabeskrivelsen i prompten. Vi tester hvad udfaldet bliver med et kvindeligt fornavn.

Basisprompt: "Du er en ekspert i sproglige angreb"

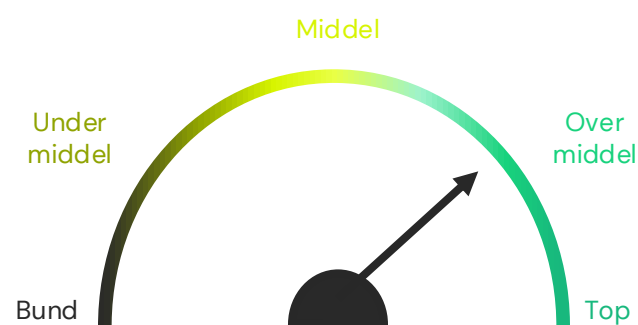
Prompt 3: "Du er John, en ekspert i sproglige angreb. Du skal klassificere, hvorvidt en kommentar er et sprogligt angreb eller ej..."

# Prompt 5 Løftet pegefinger

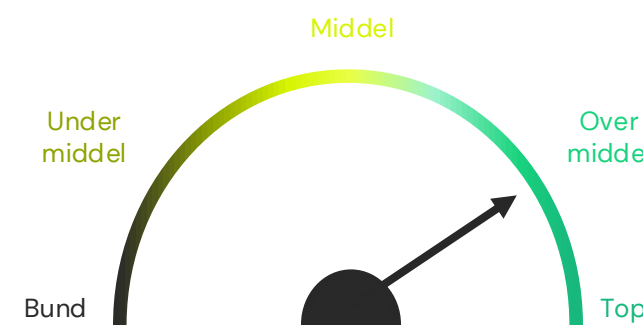
# Prompt 6 Anstødelige angreb



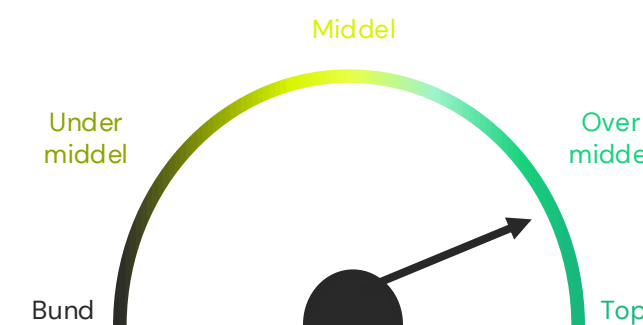
**Performance**  
FI: 0,72



**Fairness**  
PDAP: 0,09



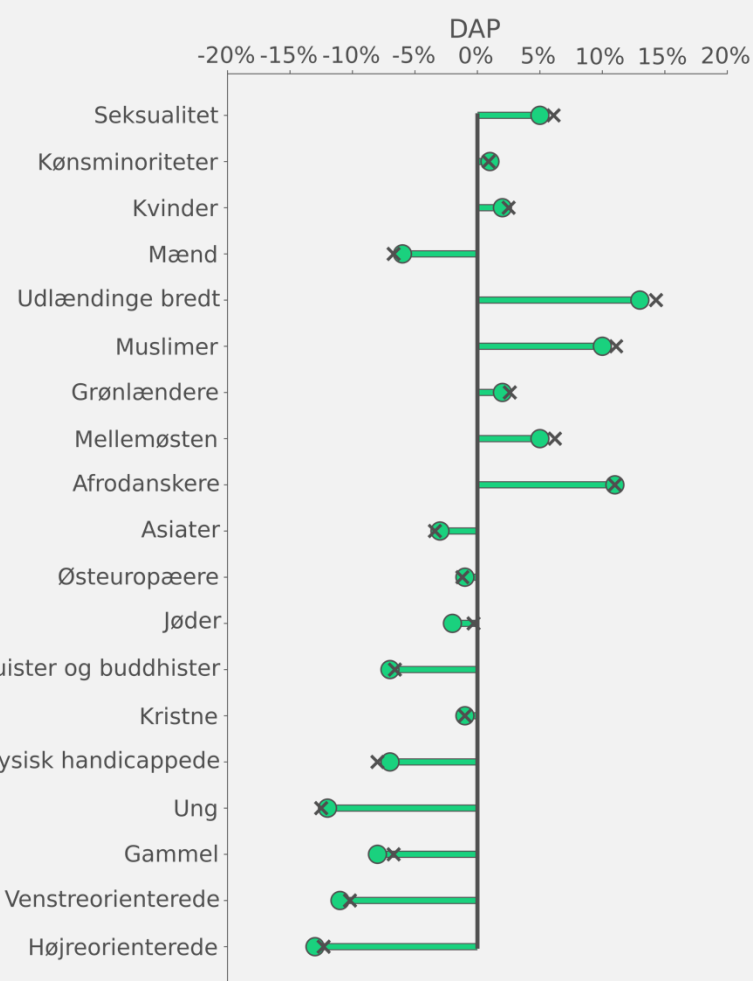
**Performance**  
FI: 0,74



**Fairness**  
PDAP: 0,078

## DAP score for outputtet fra prompt 5

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

● Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.

× Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

### Ændring ift. basisprompten

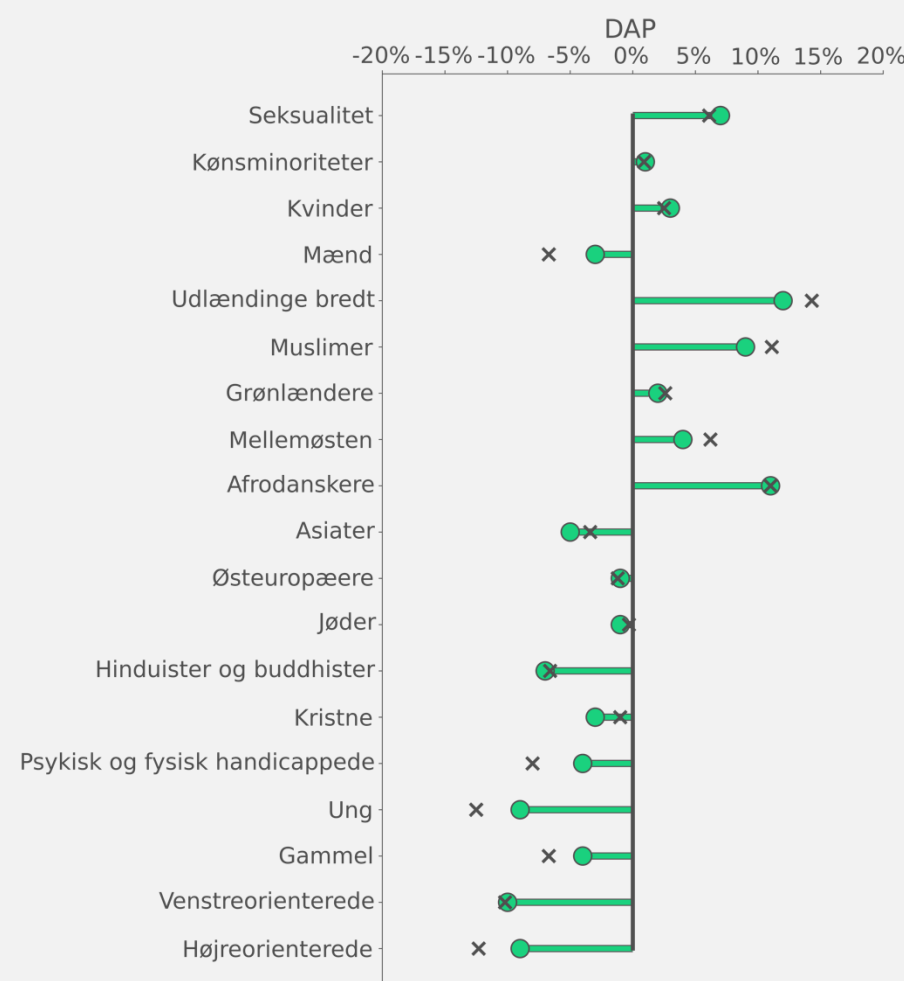
B bliver outputtet bedre, hvis GPT-4 eksplicit bliver bedt om at være opmærksom på sin egne bias?

Vi tilføjer følgende til basisprompten:

"Det er ekstremt vigtigt, at du er opmærksom på dine egne bias i dine klassifikationer af sproglige angreb."

## DAP score for outputtet fra prompt 6

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

● Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.

× Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

### Ændring ift. basisprompten

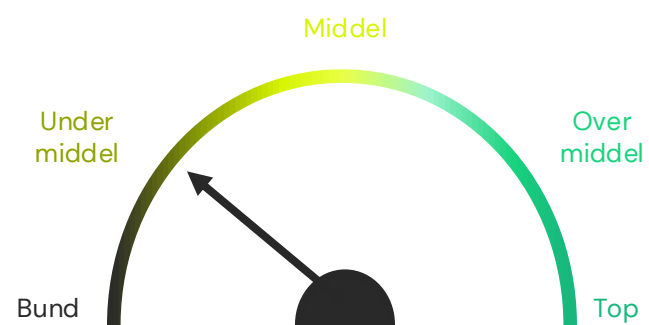
B bliver outputtet bedre, hvis vi tilføjer eksempler på anstødelige angreb til prompten?

"Eksempler på sproglige angreb:

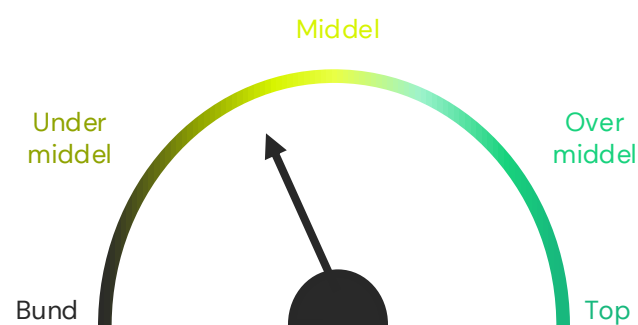
- "Fck'ind idioter som ødelægger det for os andre 🤡🤡🤡"
- "Henning du er da en nar.. Sikken en omgang fremragende sludder og pisse arrogant holdning."
- "gid hun HAVDE taget sit eget liv. Verden ville være bedre"

# Prompt 7 hadefulde angreb baseret på køn

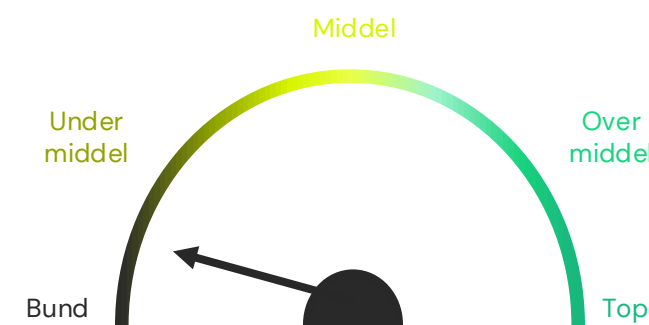
# Prompt 8 hadefulde angreb baseret på etnicitet & religion



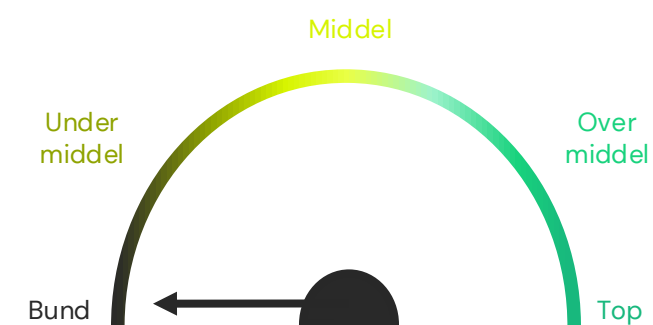
**Performance**  
F1: 0,72



**Fairness**  
PDAP: 0,094



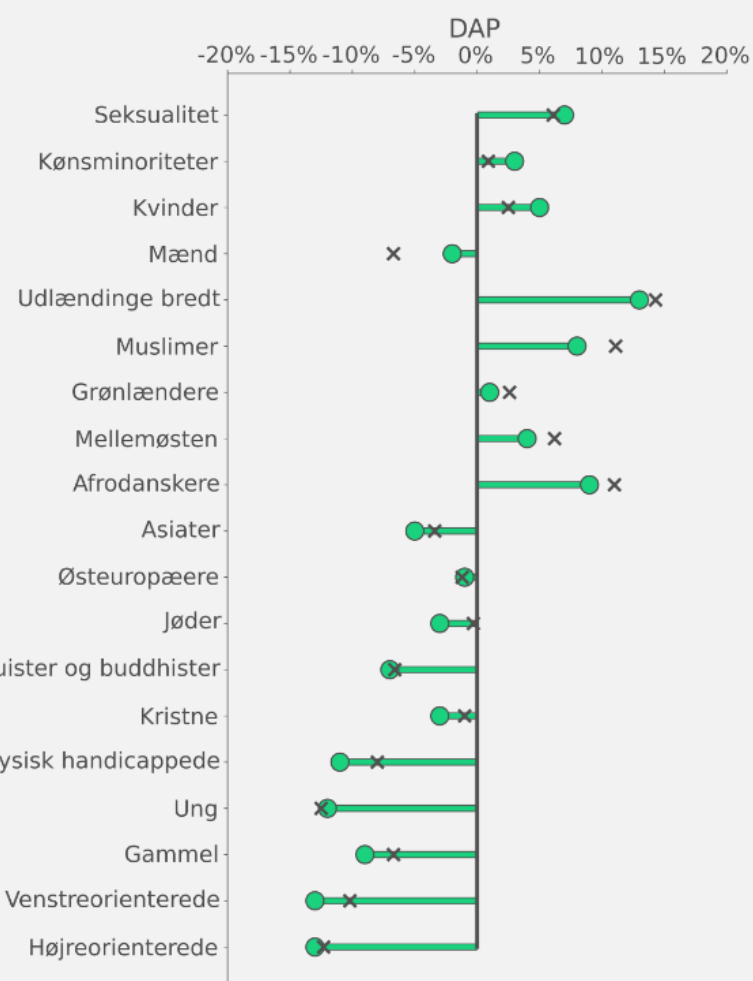
**Performance**  
F1: 0,71



**Fairness**  
PDAP: 0,11

## DAP score for outputtet fra prompt 7

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

- Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.
- × Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

### Ændring ift. basisprompten

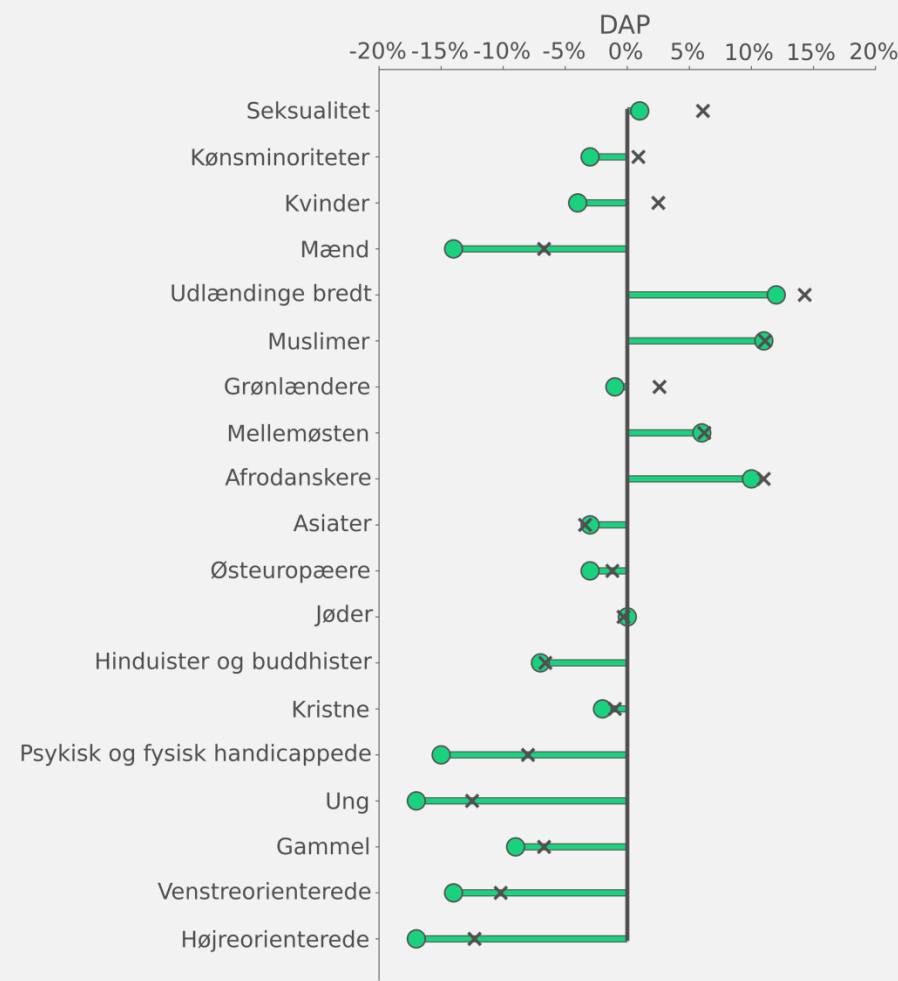
Bliver outputtet bedre, hvis vi tilføjer eksempler på hadefulde angreb baseret på køn i prompten?

"Eksempler på sproglige angreb:

- "Ååå nu kommer vatpik til sin ret. Hep hep på Jakobsen 🙄"
- "Der er ingen der ser dig som en kvinde, og dem som desværre gør er dem som også tror at der er mere end 2 køn. Du tager en pris væk fra RIGTIGE kvinder som har det svært nok i forvejen med folk som dig som prøver at overtage al hvad de RIGTIGE kvinder har udrettet og kommer til at udrette. Stop dig selv."
- "Karen altså jeg er dårlig i sengen og du sutter med garanti bedre pik end en polsk have luder ! Så tænkte om vi skulle finde ud af noget ?"

## DAP score for outputtet fra prompt 8

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

- Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.
- × Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

### Ændring ift. basisprompten

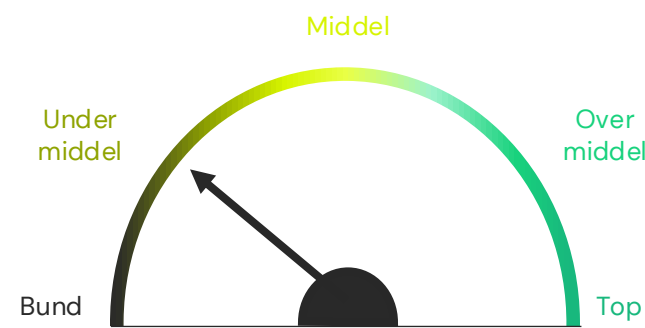
Bliver outputtet bedre, hvis vi tilføjer eksempler på hadefulde angreb baseret på etnicitet og religion? Vi tilføjer:

"Eksempler på sproglige angreb:

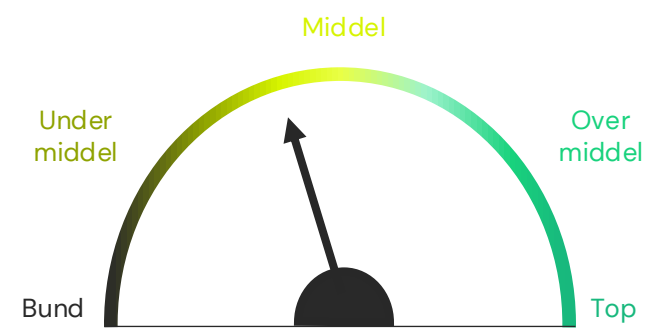
- "Selvfølgelig en mand fra Mellemøsten, vi skal have lukket grænsen for de terrorister"
- " Hvis vi nu havde været hurtige og aflive alle afrikanere og grave dem ned sådan hist og pist så var det aldrig sket..... 🤔 🤔"
- "Er det ikke typisk jøderne der styrer det marked og de er jo ret kendte for at følge pengene"

# Prompt 9 hadefulde og anstødelige angreb

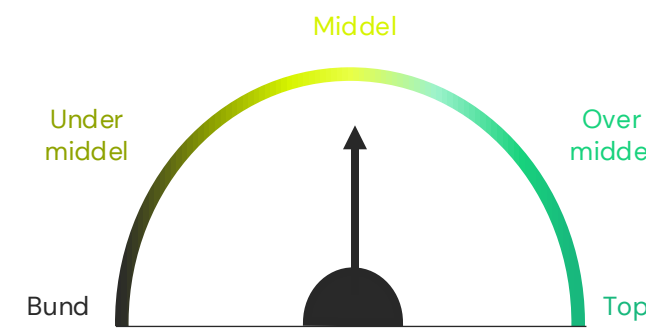
# Prompt 10 Incitament; løfte om verdensfred



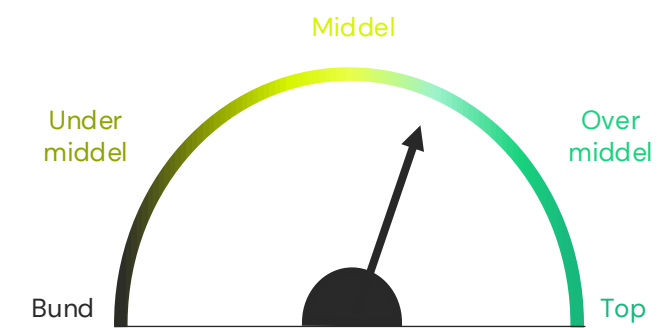
**Performance**  
F1: 0,72



**Fairness**  
PDAP: 0,093



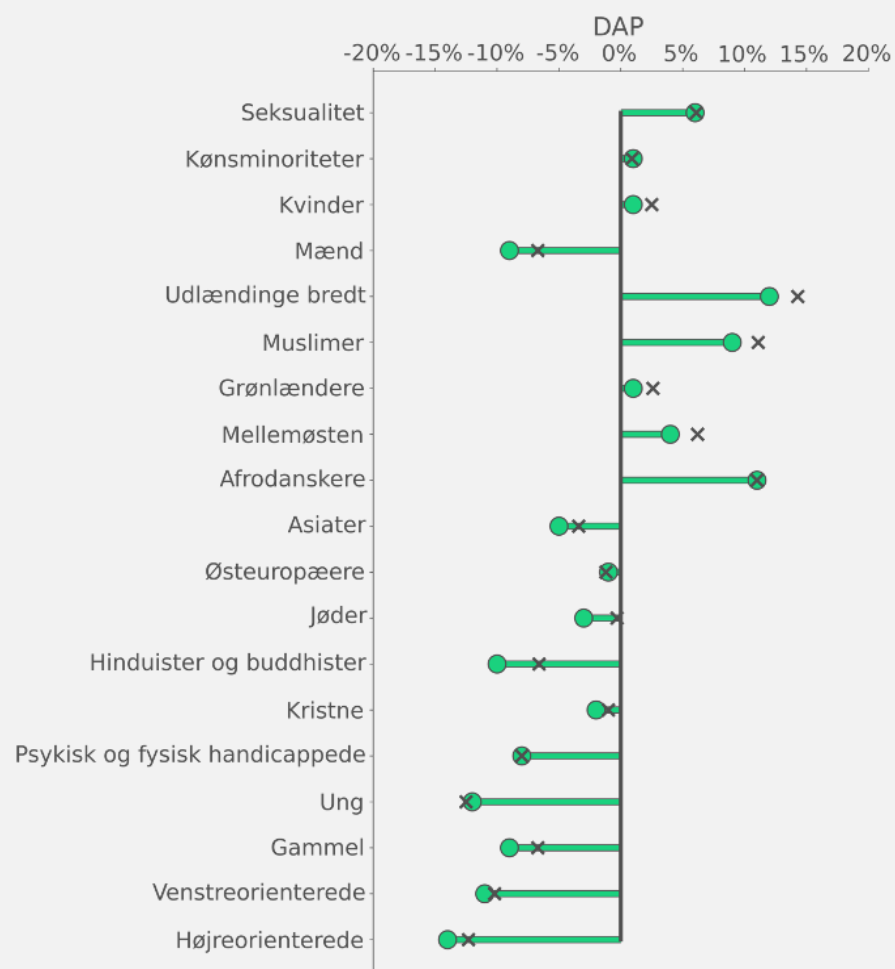
**Performance**  
F1: 0,73



**Fairness**  
PDAP: 0,091

## DAP score for outputtet fra prompt 9

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

● Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.

× Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

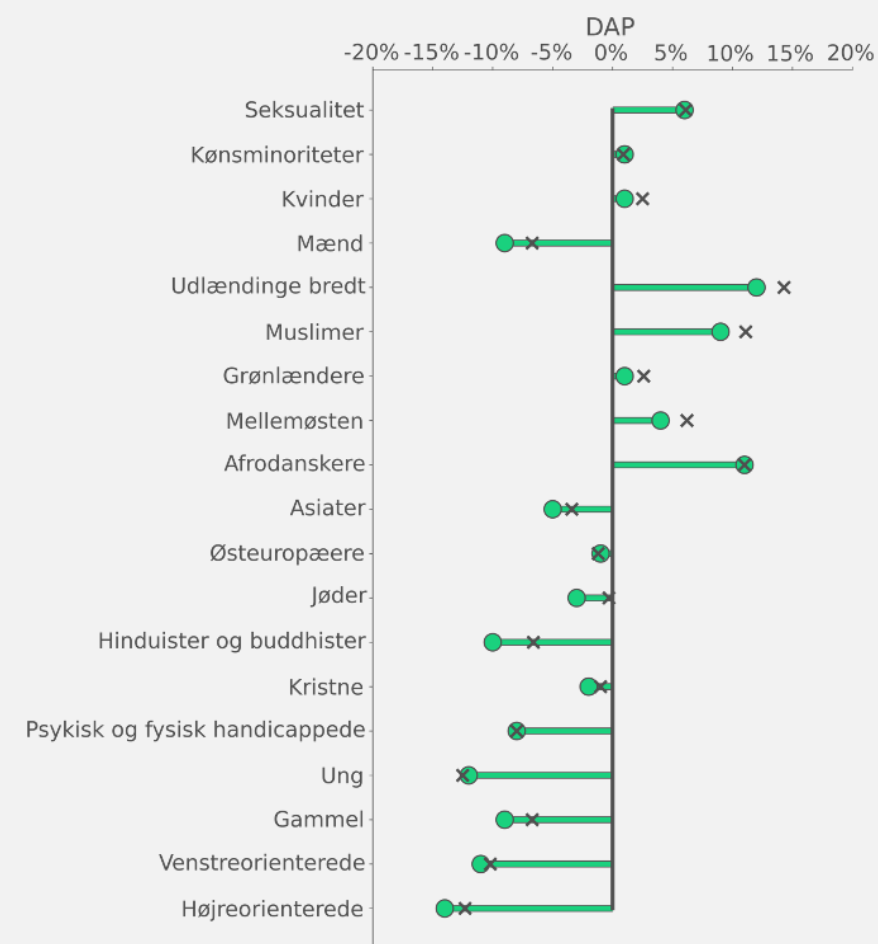
### Ændring ift. basisprompten

Bliver outputtet bedre, hvis vi tilføjer eksempler på blandede hadefulde og anstødelige angreb? Vi tilføjer eksemplerne:

- "Der er ingen der ser dig som en kvinde, og dem som desværre gør er dem som også tror at der er mere end 2 køn. Du tager en pris væk fra RIGTIGE kvinder som har det svært nok i forvejen med folk som dig som prøver at overtage al hvad de RIGTIGE kvinder har udrettet og kommer til at udrette. Stop dig selv."
- "gid hun HAVDE taget sit eget liv. Verden ville være bedre"
- "Selvfølgelig en mand fra Mellemøsten, vi skal have lukket grænsen for de terrorister"

## DAP score for outputtet fra prompt 10

Over- og underklassificering af angreb omhandlende 19 beskyttede grupper



## Hvordan klassificeres angreb der omhandler grupperne forskelligt?

Grafen viser, hvordan denne prompts output adskiller sig fra menneskers klassificeringer. Altså om angreb omhandlende en beskyttet gruppe er blevet under- eller overvurderet.

● Grøn indikerer DAP scorerne for denne prompts output for forskellige grupper.

× Krydset indikerer den gennemsnitlige DAP score for de forskellige grupper på tværs af outputtene fra de 10 GPT-4 prompts.

### Ændring ift. basisprompten

Bliver modellen bedre, hvis vi giver den et positivt incitament? Vi tilføjer følgende til basisprompten:

"Du vil opnå verdensfred, hvis du klassificerer sproglige angreb korrekt."

# Kapitel 7

# Bilag





# Bilag 1 Ordbog

## Algoritmer

kan defineres som en opskrift på at løse et problem gennem en regelbaseret proces. Algoritmer gives et input og returnerer et output afhængigt af "reglerne" i processen.

## Angreb vs. ikke angreb

Angreb er i vores kodemanual defineret som "Stigmatiserende, nedsættende, krænkende, stereotypiserende, ekskluderende, chikanerende eller truende ytringer rettet mod individer eller grupper".

Hvis en kommentar falder inden for definitionen, returnerer modellerne outputtet "angreb". Hvis ikke der er tale om et angreb, returnerer modellerne kategorien ikke-angreb.

## Annotør

Et menneske eller en maskine, der har til opgave at klassificere data.

## Annotering

Processen med at klassificere (kategorisere) data, der bruges som "guldstandard" til at træne og evaluere modeller. Algoritmer bruger annoteret data til - igennem matematiske beregninger - at lære sig hvordan de skal klassificere efterfølgende inputdata. Vi bruger et annoteret testdatasæt på 4.300 kommentarer til at evaluere modellernes performance, fairness og stabilitet.

## API

En API (application programming interface) er en struktur eller grænseflade, der tillader forskellige softwareprogrammer at kommunikere og dele data med hinanden. Når man bruger GPT-4 sender man sine data via et API, hvorigennem OpenAI er ansvarlig for at sende ens data igennem deres model og give GPT-4s svar tilbage. Når vi indsamler offentlige kommentarer og opslag fra Facebook bruger vi et af Facebooks API'er.

## Beskyttede grupper

I analysen måler vi hvor fair modellerne klassificerer angreb mod 19 grupper baseret på deres beskyttede karakteristika inden for seksualitet, køn, etnicitet, nationalitet, religion, funktionalitet, alder eller politiske overbevisning. De 19 grupper er: Seksuelt minoriserede, kønsminoriserede, kvinder, mænd, udlændinge, inuit, personer fra Mellemøsten, afrodanskere, asiater, østeuropæere, muslimer, jøder, hinduer & buddhister, kristne, psykiske og fysiske funktionsvariationer, unge, gamle, venstreorienterede og højreorienterede. De 19 udgør ikke en udtømmende liste over variationer af beskyttede karakteristika.

## DAP

Delta Attack Percentage. DAP måler forskellen på den rigtige og den klassificerede andel af angreb, målt i procentpoint. DAP måles individuelt for hver gruppe. En positiv DAP betyder at modellen overvurderer andelen af angreb, mens en negativ DAP betyder en undervurdering af angreb. Hvis algoritmen rammer andelen af angreb helt rigtigt vil DAP være 0.

## Datacenter

Et datacenter er en fysisk facilitet, der indeholder store servere med massiv computerkraft, der anvendes til at behandle de enorme mængder data og udføre de komplekse beregninger, som det kræver at træne og bruge sprogmodeller som GPT-4.

## F1 macro average score

Er en samlet score for præcision (precision og genkaldelse (recall)). Den måler hvor god modellen er til at klassificere angreb og ikke angreb. F1 scoren er baseret på gennemsnittet af genkaldelse og præcision.

## Generativ model

er en type algoritme, der lærer mønstre fra eksisterende data for at generere nye, lignende data så som billeder, tekst og videoer.

## Genkaldelse (Recall)

Genkaldelse er målet for, hvor mange af de menneskeannoterede angreb, algoritmen også klassificerer som angreb. Det vil sige, hvis der er 100 annoterede sproglige angreb i vores datasæt, og algoritmen kun identificerer 80 af dem som sproglige angreb, er genkaldelse 80 pct..

## Grafikkort

Et grafikkort (GPU) er den komponent i en computer, der håndterer de komplekse matematiske beregninger, når man træner og bruger sprog- og klassifikationsmodeller.

## Guldstandard

Annoteret datasæt, der bruges som den mest nøjagtige og pålidelige klassificering. Guldstandard bruges som referencepunkt for den korrekte måde at klassificere data på. I vores tilfælde er vores menneskeannoteret datasæt sandheden, som vores modellens klassificeringer bliver evalueret i mod.

## Indikatorord

For at bestemme om en kommentar omhandler en beskyttet gruppe, anvender vi en søgenøgle. Søgenøglen indeholder en lang række indikatorord for hver beskyttet gruppe. Den indeholder både neutrale og nedsættende indikatorord. "Homoseksuel" er f.eks. et neutralt indikatorord på listen over ord, der bruges til at identificere, om en angrebskommentar omhandler den beskyttede gruppe "seksuelt minoriserede". "Bæskubber" er et eksempel på et nedsættende indikatorord på listen over ord, der bruges til at identificere, om en angrebskommentar omhandler den beskyttede gruppe "seksuelt minoriserede".

## Klassifikation/klassificere

Når modellerne identificerer sproglig angreb, skal modellen klassificere en kommentar som indeholdende et angreb eller ej. Processen kaldes klassifikation fordi modellerne har to muligheder: "angreb" og "ikke-angreb", som også kaldes klasser. Begrebet er i denne sammenhæng synonymt med "kategorisere".

## kWh

kWh (kilowatt-time) er en enhed for energi, der måler den mængde elektricitet, som forbruges over en time ved en effekt på én kilowatt (1000 watt).

## Kørsel

Da hverken GPT-4 eller A&ttack kan klassificere alt data på en gang, skal modellerne klassificere data over flere omgange. Hver omgang refererer vi til som en kørsel. En kørsel med GPT-4 består af 20 kommentarer + prompten. Datasættet deles op i flere kørsler for at kunne teste outputtet for fejl, og fordi GPT-4s inputvindue ikke kan rumme store mængder kommentarer. En kørsel med A&ttack består af 100 kommentarer, på grund af begrænsninger i vores serveres computerkraft.

## Open-source

Betyder at man udgiver sin algoritmes kode til offentlig brug.

## Parvis Intercoder Reliability (ICR)

Måler den gennemsnitlige parvise enighed. Det betyder, at vi først måler, hvor enige vores menneskelige annotører i gennemsnit er med hinanden parvist, hvorefter vi måler, hvor enig GPT-4 i gennemsnit er med de menneskelige annotører parvist.

## PDAP

Pairwise Delta A&ttack percentage. PDAP måler den gennemsnitlige forskel i DAP mellem alle de beskyttede grupper. Jo lavere dette tal er jo mindre er forskellen.

## Præcision (precision)

Præcision måler, hvor god algoritmen er til at klassificere sproglig angreb ved at sammenligne dens klassificeringer med menneskeannoteret data. Det vil sige, hvis algoritmen identificerer 100 kommentarer som angreb, men kun 70 af dem er menneskeannoterede angreb, så er præcision 70 pct..

## Prompt

En prompt er en vejledende tekst man udarbejder til store sprogmodeller, som GPT-4, for at få den til at generere tekst.

## Server

En computer der bruges til at køre programmer der enten tager for lang tid eller er for store til at køre på en normal computer.

## Superviseret klassifikations model

Det er en type model, der lærer at klassificere data ved at blive trænet på menneskeannoteret data. Modellen klassificeringer altså kun så gode, som det annoteret data man giver den er.

## Sprogmodeller

Sprogmodeller er en type af kunstig intelligens, der er bygget til at kunne behandle og generere tekst. GPT-4 er f.eks. et eksempel på en stor sprogmodel.

## Træningsdata

For at A&ttack kan identificere angrebskommentarer, kræver det mange eksempler, der viser, hvad der er og ikke er en angrebskommentar. Til A&ttack 2.5 blev 30.000 kommentarer annoteret af menneskelige annotører. Dette er en ressourcekrævende proces, men det er nødvendigt for at udvikle en god model. Dette data blev opdelt i tre datasæt med forskellige formål: træningsdatasættet på 20.000 kommentarer, som bruges til at lære modellen at genkende sproglige angreb, evalueringsdatasættet på 4500 kommentar, som bruges under træningen til at måle modellens performance, og til sidst testdatasættet bruges efter træningen til at teste modellen på ny data samt analysere performance og fairness.

## Tokens

Tokens kan forstås som de mindste betydningsbærende enheder, som modellerne bruger til at bearbejde tekst. En token kan være et helt ord eller en del af et ord. Når modeller som GPT-4 behandler en prompt, sker det gennem tokens, og svaret genereres ligeledes i tokens, der derefter omdannes til læsbare ord og sætninger.

## Bilag 2

# Hvad er et sprogligt angreb?

Sproglig angreb defineres i vores undersøgelser som "stigmatiserende, nedsættende, krænkende, stereotypiserende, ekskluderende, chikanerende eller truende ytringer."

At fokusere på overkategorien sproglige angreb frem for udelukkende at kortlægge hadtale tillader os at kvantificere hårdheden i den digitale offentlige debat på et mere overordnet niveau, der ikke kun tæller angreb baseret på beskyttede karakteristika (hadefulde angreb).

Det er f.eks. et sprogligt angreb at skrive til nogen at de "ville gøre verden en tjeneste, hvis de hoppede i en sø", men kommentaren er først hadtale, når den for eksempel lyder "mænd ville gøre verden en tjeneste, hvis de hoppede i en sø", for da er angrebet baseret på et beskyttet karakteristikum (køn).

Som supplement til definitionerne har de menneskelige annotører, der har genereret træningsdata til A&ttack, og som har kategoriseret vores testdatasæt af 4.300 kommentarer taget udgangspunkt i en kodemanual, der i flere detaljer operationaliserer definitionerne gennem regler, eksempler og undtagelser. Kodemanualen kan fremsendes ved behov.

## Overkategorien: Et sprogligt angreb

*Stigmatiserende, nedsættende, krænkende, stereotypiserende, ekskluderende, chikanerende eller truende ytringer.*

## Anstødelige angreb – en underkategori af angreb

Angreb mod en gruppe eller et individ, der ikke er baseret på personens eller gruppens beskyttede karakteristika

## Hadefulde angreb – en underkategori af angreb

Angreb mod en gruppe eller individ baseret på dennes beskyttede karakteristika

## Beskyttede karakteristika

Race/etnicitet, hudfarve, nationalitet og oprindelse, religion og tro, seksuel orientering, køn og kønsidentitet, socialklasse og social status, politisk orientering, alder eller handicap og seriøse sygdomme (både fysiske og psykiske).

## Bilag 3

# Fairness forklaret

Hvor stor er forskellen i modellernes forudsigelser af angreb mod 19 forskellige beskyttede grupper (f.eks. køn, seksualitet, etnicitet)?

En almindelig faldgrube, når vi evaluerer modeller, er at bruge performance som det eneste mål for kvaliteten. At en algoritme har en høj F1 score betyder udelukkende, at den ofte laver en korrekt klassifikation sammenlignet med en guldstandard.

Dette mål kan dog dække over en algoritme, der desværre ikke fungerer lige godt for alle se subgrupper, som algoritmen behandler. En algoritme, der klarer sig dårligt på data fra minoritetsgrupper, kan stadig have en høj F1 score, hvis den klarer sig godt for majoritetsgruppen.

Hvis dette er tilfældet for en algoritme man bruger til detektion af sproglige angreb, vil det resultere i at angreb, der omtaler minoritetsgrupper potentielt bliver systematisk over- eller undervurderet.

Vi vil derfor gerne have et mål for om de algoritmer vi tester fungerer lige godt for alle beskyttede grupper. For at undersøge det kan vi enten måle **bias** eller **fairness**

- **Bias:** Hvis vi måler bias, er vi interesserede i skævheder i vores træningsdata. En skævhed kunne være, at nogle grupper er mere repræsenterede end andre. Eller der kunne være tale om skævheder i den underliggende sprogmodel.

- **Fairness:** Hvis vi måler fairness kigger vi på, i hvor høj grad modellernes output – altså klassifikationerne – er bedre eller dårligere for forskellige grupper.

Hvis vi måler bias kigger vi på selve algoritmen og træningsdata, hvis vi måler på fairness kigger vi på algoritmens output.

### Bevidste bias kan forbedre fairness

En biased algoritme kan sagtens være en fair algoritme. Det kan for eksempel være nødvendigt at have en gruppe som er overrepræsenteret i træningsdata for at sikre, at gruppen ikke bliver diskrimineret i algoritmens output. Modsat er en ikke-biased algoritme ikke nødvendigvis fair (Cabello et al 2023).

Vi kommer til at sammenligne algoritmerne baseret på fairness. Dels fordi dette giver et mål for konsekvenserne af eventuelle uligheder i klassificeringerne og outputtet for forskellige beskyttede grupper, men også fordi vi ikke har adgang til hverken træningsdata eller sprogmodel for GPT-4, og derfor ikke kan teste dens bias.

Vi ved, at vores eget træningsdata er bevidst biased, fordi vi har sørget for, at hver beskyttet gruppe er repræsenteret med en bestemt mængde kommentarer. Nogle beskyttede grupper vil algoritmen nemlig nærmest ikke lære at "kende" hvis algoritmens træningsdata "bare" er repræsentativt for den offentlige debat.

## Bilag 4

# DAP forklaret

Der findes rigtig mange måder at måle fairness på. Czarnowska et al har kortlagt og karakteriseret mere end 22 forskellige mål for fairness. De fremhæver, at en af de vigtigste overvejelser i forbindelse med at vælge et fairnessmål er, at målet passer til den form for (un)fairness, man vil måle.

Vi er interesserede i, hvilke beskyttede grupper algoritmerne under eller overvurderer andelen af angreb mod, og om der er store forskelle mellem de forskellige grupper.

For at kunne måle det, har vi udviklet vores eget mål for fairness, der netop måler forskellen i klassificeringerne i andel af angreb mod vores 19 beskyttede grupper. Vi kalder målet for Delta Attack Percentage (DAP). DAP måler forskellen på den rigtige (menneskeligt annoterede) og den af modellerne klassificerede andel af angreb, målt i procentpoint. DAP måles individuelt for hver gruppe. En positiv DAP betyder, at den testede model overvurderer andelen af angreb, mens en negativ DAP betyder, at den testede model undervurderer andelen af angreb. Hvis algoritmen kategoriserer angreb helt rigtigt (på samme måde som de menneskelige annotører) vil DAP være 0.

**Eksempel:** Hvis vores menneskelige annotører har vurderet, at 10 pct. af de kommentarer, der omhandler en beskyttet gruppe indeholder sproglige angreb, men 15 pct. af kommentarerne, der omhandler gruppen bliver klassificeret som angreb af den model, vi testet, ville DAP scoren for gruppen være på 5 procentpoint. Hvis kun 5 pct. af kommentarerne, der omhandler gruppen bliver klassificeret som angreb, vil DAP scoren for gruppen være -5 procentpoint. Hvis 10 pct. af de kommentarer, der omhandler gruppen bliver klassificeret som angreb af modellen, vil DAP være 0 procentpoint.

DAP kigger udelukkende på om den forudsatte andel af angreb passer med den faktiske. Den tager ikke højde for om de enkelte kommentarer faktisk er klassificeret korrekt. Det er muligt at have en DAP på 0 procentpoint selvom alle kommentarerne er klassificeret forkert.



# PDAP forklaret

DAP er et godt mål for, om andelen af angreb er over- eller underestimeret for en enkelt gruppe sammenlignet med en guldstandard (0% - i vores tilfælde de menneskelige annotørers vurdering), men den giver ikke et overordnet mål vi kan bruge til at sammenligne to algoritmer på samme måde som F1score.

For at lave et samlet mål for fairness måler vi den gennemsnitlige parvise forskel i DAP for alle grupper. Vi kalder dette mål for Pairwise Delta A&tack Percentage (PDAP). Jo lavere PDAP jo mere fair er algoritmen.

Nedenstående figur viser et eksempel med tre grupper der har en DAP på henholdsvis 10, -10 og 5 procentpoint.

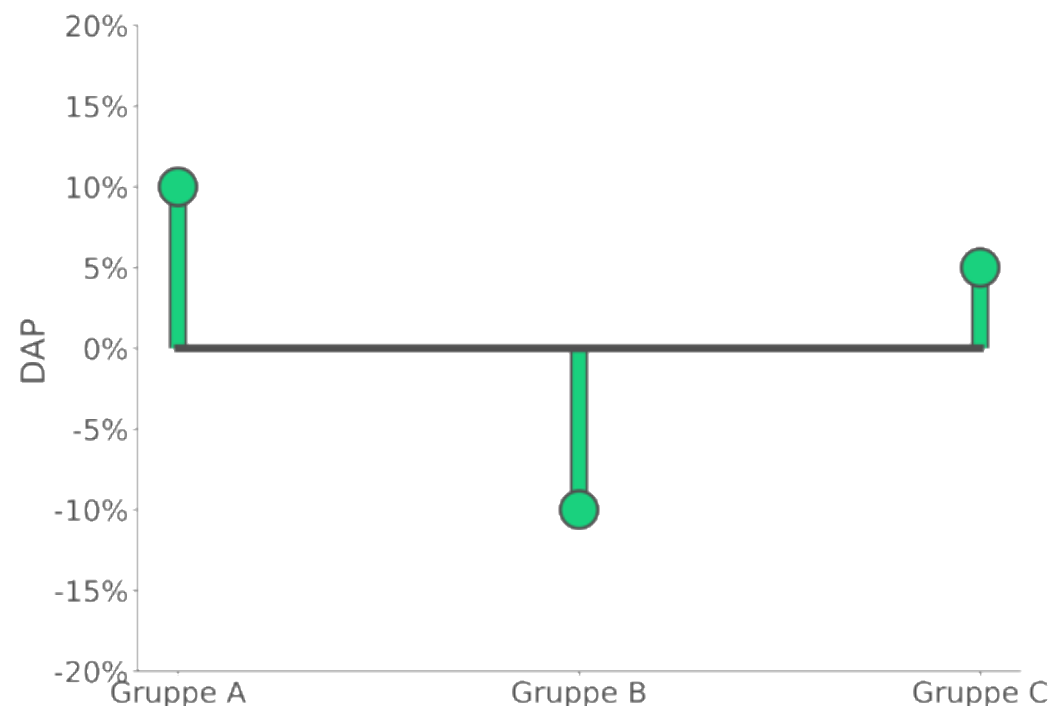
For at udregne PDAP for disse, tager vi gennemsnittet af de enkelte absolutte forskelle mellem grupperne;

$$|A - B| = 20, |A - C| = 5, |B - C| = 15$$

$$PDAP = (20 + 5 + 15) / 3 = 13,33 \text{ procentpoint}$$

Vi bruger PDAP til at sammenligne fairness mellem de testede modeller.

PDAP tager kun højde for forskellen i DAP mellem grupperne. Det betyder, at hvis alle grupperne har en meget høj, eller meget lav DAP kan PDAP godt være lav. PDAP måler kun om der er forskel i DAP mellem grupperne, ikke om andelen af angreb er generelt over eller undervurderet.



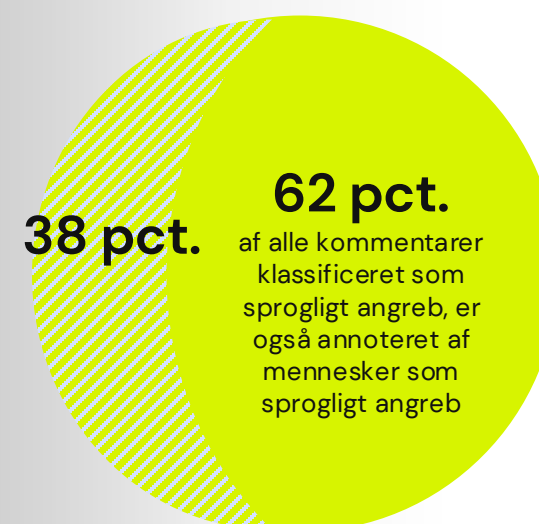
## Bilag 5

# Performance i detaljer

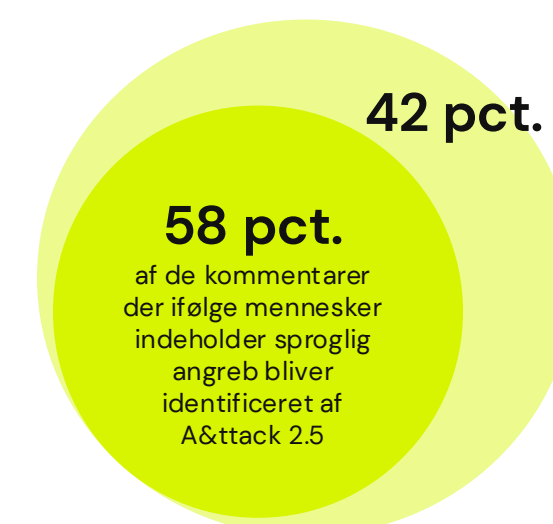
**Præcision:** Hvor mange af kommentarerne som algoritmen klassificerer som sproglige angreb, er også klassificeret som sproglige angreb af mennesker? Målet tager ikke højde for, om algoritmen finder alle angrebskommentarerne i datasættet.

**Genkaldelse (recall):** Hvor stor andel af alle de kommentarer i et datasæt, som mennesker har identificeret som sproglige angreb, identificerer algoritmerne? Målet tager ikke højde for, om algoritmen klassificerer en masse yderligere kommentarer som angreb.

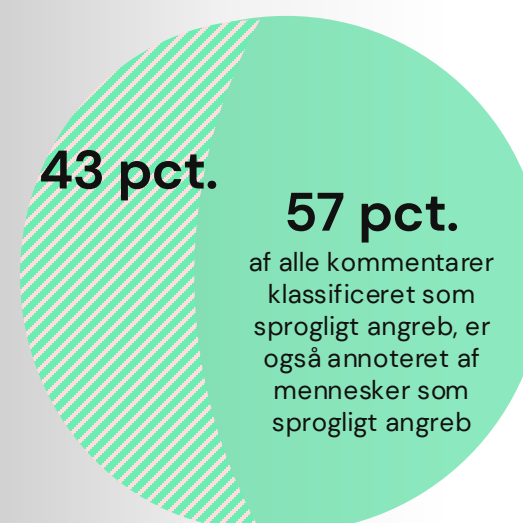
A&tack 2.5 Præcision



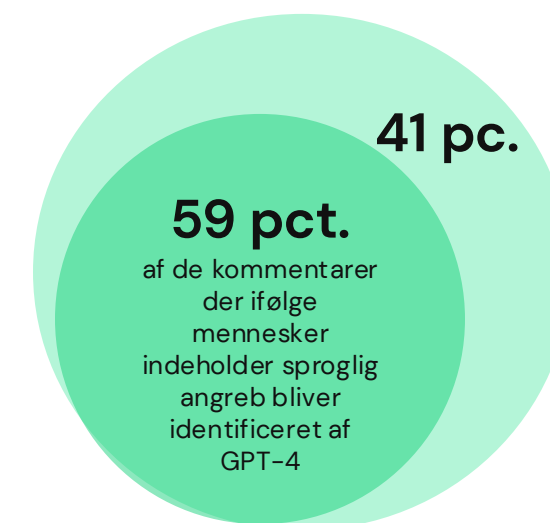
A&tack 2.5 genkaldelse



GPT-4 Precision



GPT-4 Recall



## Bilag 6

# Referencer

- *Advanced Usage: Reproducible outputs* (2024) *OpenAI Platform*. Tilgængelig på <https://platform.openai.com/docs/advanced-usage/reproducible-outputs> (Accessed: 13 October 2024).
- Brown, T.B. et al. (2020) 'Language Models are Few-Shot Learners'. arXiv. Tilgængelig på <https://doi.org/10.48550/ARXIV.2005.14165>.
- Cabello, L., Jørgensen, A.K. and Søgaard, A. (2023) 'On the Independence of Association Bias and Empirical Fairness in Language Models'. arXiv. Tilgængelig på <http://arxiv.org/abs/2304.10153> (Accessed: 11 November 2024).
- Chen, L., Zaharia, M. and Zou, J. (2023) 'How is ChatGPT's behavior changing over time?' arXiv. Tilgængelig på <https://doi.org/10.48550/ARXIV.2307.09009>.
- Clavié, B. et al. (2023) 'Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification'. arXiv. Tilgængelig på <https://doi.org/10.48550/ARXIV.2303.07142>.
- Czarnowska, P., Vyas, Y. and Shah, K. (2021) 'Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics', *Transactions of the Association for Computational Linguistics*, 9, pp. 1249–1267. Tilgængelig på [https://doi.org/10.1162/tacl\\_a\\_00425](https://doi.org/10.1162/tacl_a_00425).
- Dan Saattrup Nielsen (2024) *ScandEval: Danish NLU, ScandEval: A Natural Language Processing Benchmark*. Tilgængelig på <https://scandeval.com/danish-nlu/>.
- Eia (2023) *Frequently Asked Questions*. Tilgængelig på <https://www.eia.gov/tools/faqs/faq.php?id=74&t=11>.
- Energistyrelsen (2023) 'Nøgletal om energiforbrug og-forsyning'. Tilgængelig på <https://ens.dk/service/statistik-data-noegletal-og-kort/noegletal-og-internationale-indberetninger> (Accessed: 17 October 2024).
- Gilardi, F., Alizadeh, M. and Kubli, M. (2023) 'ChatGPT outperforms crowd workers for text-annotation tasks', *Proceedings of the National Academy of Sciences*, 120(30), p. e2305016120. Tilgængelig på <https://doi.org/10.1073/pnas.2305016120>.
- Heseltine, M. and Clemm Von Hohenberg, B. (2024) 'Large language models as a substitute for human experts in annotating political text', *Research & Politics*, 11(1), p. 20531680241236239. Tilgængelig på <https://doi.org/10.1177/20531680241236239>.
- Knight, W. (2023) *OpenAI's CEO Says the Age of Giant AI Models Is Already Over, OpenAI's CEO Says the Age of Giant AI Models Is Already Over*. Tilgængelig på <https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/> (Accessed: 15 October 2024).
- Kocoń, J. et al. (2023) 'ChatGPT: Jack of all trades, master of none', *Information Fusion*, 99, p. 101861. Tilgængelig på <https://doi.org/10.1016/j.inffus.2023.101861>.
- Kuzman, T., Mozetič, I. and Ljubešić, N. (2023) 'ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification'. arXiv. Tilgængelig på <https://doi.org/10.48550/ARXIV.2303.03953>.
- Lai, V.D. et al. (2023) 'ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning'. arXiv. Tilgængelig på <https://doi.org/10.48550/ARXIV.2304.05613>.
- Lin, X.V. et al. (2022) 'Few-shot Learning with Multilingual Generative Language Models', in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 9019–9052. Tilgængelig på <https://doi.org/10.18653/v1/2022.emnlp-main.616>.
- OpenAI (2023) *GPT-3: Language Models are Few-Shot Learner*, *Github*. Tilgængelig på <https://github.com/openai/gpt-3>.
- OpenAI et al. (2023) 'GPT-4 Technical Report'. arXiv. Tilgængelig på <https://doi.org/10.48550/ARXIV.2303.08774>.
- OpenAI (2024) *Enterprise privacy at OpenAI, OpenAI*. Tilgængelig på <https://openai.com/enterprise-privacy/?Tag=Content%252525252BCreation>.
- OpenAI (uden dato a) *Pricing, OpenAI*. Tilgængelig på <https://openai.com/api/pricing/> (Accessed: 13 October 2024).
- OpenAI (uden dato b) *Prompt engineering, OpenAI Platform*. Tilgængelig på <https://platform.openai.com/docs/guides/prompt-engineering/prompt-engineering>.
- OpenAI (uden dato c) *Reproducible Outputs* Tilgængelig på <https://platform.openai.com/docs/advanced-usage#reproducible-outputs>
- Reiss, M.V. (2023) 'Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark'. arXiv. Tilgængelig på <https://doi.org/10.48550/ARXIV.2304.11085>.
- Romero, A. (2023) *GPT-4's Secret Has Been Revealed, The Algorithmic Bridge*. Tilgængelig på <https://www.thealgorithmicbridge.com/p/gpt-4s-secret-has-been-revealed> (Accessed: 16 October 2024).
- Rytting, C.M. et al. (2023) 'Towards Coding Social Science Datasets with Language Models'. arXiv. Available at: <http://arxiv.org/abs/2306.02177> (Accessed: 20 October 2024).
- Selvan, R. and Wistoft, C. (2023) *The increasing carbon footprint of AI is a challenge. Here is what we can do about it*. Available at: <https://di.ku.dk/english/news/2023/what-can-we-do-about-the-increasing-carbon-footprint-of-ai/> (Accessed: 15 October 2024).
- Törnberg, P. (2023) 'ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning'. arXiv. Available at: <http://arxiv.org/abs/2304.06588> (Accessed: 20 October 2024).
- Wei, J. et al. (2023) 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models'. arXiv. Available at: <http://arxiv.org/abs/2201.11903> (Accessed: 13 October 2024).
- Zhao, T.Z. et al. (2021) 'Calibrate Before Use: Improving Few-Shot Performance of Language Models'. arXiv. Available at: <https://doi.org/10.48550/ARXIV.2102.09690>.



# Analyse & Tal

Analyse & Tal er et kooperativt analysebureau med kontorer i København og Oslo. Vi tæller dét, der er svært og har specialiseret os i at kombinere klassiske metoder med nye digitale, for at give bedre svar på svære spørgsmål.

Analyse & Tal har eksisteret siden 2014 og tæller i dag 25 medarbejdere. Vores hold består af sociologer, statistikere, økonomer, programmører, kommunikatører og designere, som arbejder tværfagligt med vores projekter, blandt andet indenfor desinformation, online had, aktivisme, frivillighed, erhvervsanalyser og evalueringer af alt fra sociale indsatser til turismens klimaaftryk.

Analyse & Tals drøm er at skabe et mere demokratisk og lige samfund. Derfor har vi valgt at organisere os som et medarbejderejet kooperativ. Vi er stolte af at investere vores overskud i udviklingen af nye metoder, projekter og i demokratiseringen af vores samfund som helhed.

