Ud af Mange, En.

En undersøgelse af effekterne af model, temperatur og prompting på varians af LLM-generede syntetisk spørgeskema data i en dansk kontekst.

Lavet af:

Simon Taudal Hede
Studienr: 20205726
shede20@student.aau.dk

Cand.scient.pol. – Social Data Science

10. semester. / 4. semester kandidaten.

Specialeopgave

Omfang:

142840 tegn / 59.5 sider

Vejleder:

Anvendelse af Generativ AI:

I forbindelse med dette projekt er generativ AI blevet brugt som forskningsinstrument og subjekt af analysen. Derudover er det anvendt til sparring undervejs i skriveprocessen, som online søgemaskine og i forbindelse med programmering til at løse fejlkoder.

AALBORG
UNIVERSITET

Indholdsfortegnelse:

0. Abstract

Surveys are a key source of data within social science research, however declining response rates have spurred interest in Large Language Models as a source of synthetic data. A hurdle shown by initial research is the lack of variance in model output when compared to human data. This study examines how newer models, and changes in temperature and prompt affect this truncated variance. I utilize *Silicon Sampling*, prompting models to adopt personas based on Danish demographic and response data from *Tryghedsmålingen* 2024 to generate and compare 1,000 personas with a real human baseline. The models tested are GPT-3.5-turbo, GPT-4.1 and o3, at temperature 1 and 2, and with a 'flawed' and 'improved' prompt. I find that models perform worse than initial research had led to expect, with truncated variance and mean bias in both aggregate and item-level distributions. Newer models appear less capable of creating a human-like item-level histograms, possibly caused by stronger alignment. Temperature is either inapplicable, insufficient or cause high rates of response failure. Effects of prompt prove difficult to predict, with results varying across model and temperature. This implies recent LLM developments have not resolved the issue of truncated variance, and that the common methods are currently inadequate. Finally I discuss the possible implications and issues with LLM-generated synthetic data within science and society at large.

1. Indledning

Politologien og socialvidenskaben, samt private og offentlige organisationer og foreninger, benytter sig i høj grad af spørgeskemaundersøgelser til at danne sig et indtryk af holdninger i samfundet. Over tid har antallet af respondenter været faldende, hvilket gør det sværere, dyrere og langsommere at indsamle tilstrækkelig data (California Health Interview Survey, 2024). Dette har fået forskere inden for politologien og socialvidenskaben til at søge efter alternative tilgange til dataindsamling. En af disse alternativer er Large Language Models, herefter refereret til som LLM, som ChatGPT, for hvis LLM kan svare som menneskelige respondenter, ville det potentielt kunne reducere mange af de forskellige omkostninger og udfordringer som eksistere i forbindelse med dataindsamling via spørgeskemaundersøgelser, for eksempel i form af tidsbesparelser eller dannelsen af et mere repræsentativt dataset (Argyle et al., 2023, 2025; Bail, 2024; Bisbee et al., 2024; Dominguez-Olmedo et al., 2024; Eggleston, 2024; McIntosh et al., 2024; Rystrøm et al., 2025; Röttger et al., 2024).

Det er her værd at pointere at det ikke er dette projekts intention at LLM modeller vil kunne totalt erstatte menneskelige respondenter i almene spørgeskemaundersøgelser eller meningsmålinger. Jeg opfatter det derimod som et potentielt metodisk værktøj som ville kunne hjælpe fremtidige analyser ved at udvide eller supplere eksisterende empirigrundlag. Et eksempel på dette kunne være data dannet ud fra et naturligt eksperiment, men som er for småt til at kunne anvendes i forbindelse med konventionelle kvantitative metoder. I tilfælde af at LLM modeller kan bruges på praktisk lige fod med mennesker, ville de her potentielt kunne skalere data op uden at ødelægge de analytisk relevante mønstre som ønskes undersøgt. Generelle undersøgelser ville potentielt også kunne gøre brug af lignende metoder til at reducere antallet af menneskelige respondenter som kræves for at lave analyser, og på den måde gøre processen billigere, et argument som er fremsat af flere forskere inden for emnet, også af forskere, som er mere tilbageholdende over brugen af LLM modeller inden for social- og samfundsvidenskaben (Bisbee et al., 2024).

Trods forbehold, som uddybes i diskussionsafsnittet, er der også mulige anvendelser af LLM modeller inden for social- og samfundsvidenskaben som bruger få eller ingen menneskelige respondenter. Dette er pilotstudier, som potentielt ville kunne afprøve spørgsmål eller metoder med minimal omkostning og ventetid, hvilket kunne gøre manglen på reel validering det værd.

Anvendelsen af LLM generede data kan også blive en nødvendighed. Eggleston (2024) skrev en artikel i *Journal of Survey Statistics and methodology*, hvor han undersøger og finder belæg for hypotesen, at den faldende responsrate på spørgeskema er forårsaget af spørgeskema-udmattelse, hvor folk der har forholdt sig til et spørgeskema, er mindre tilbøjelig til at deltage i det næste (Eggleston, 2024).

Dette kan være et svært problem at løse idet spørgeskemaer bliver anvendt af både det offentlige og det private til alt fra folkesundhedsundersøgelser og vælgermålinger til kundeprofilering og tilfredshedsundersøgelser.

Eggleston pointerer i artiklen, at juridiske indgreb fra regeringsniveau kan være nødvendige for at mindske den samlede responsbyrde (Eggleston, 2024, s.1154). Intentionen med sådanne indgreb ville derfor være at øge sandsynligheden for at befolkningen besvarer vigtige undersøgelser om for eksempel folkesundhed, frem for relativt trivielle undersøgelser som holdninger til diverse produkter. I sådan et scenarie kan syntetisk genererede data være det eneste alternativ, og hvis sådanne ændringer ikke udføres kan det betyde at så få undersøgelser besvares at det er nødvendigt med metoder som extrapolerer data ud fra et frø af menneskelige respondenter.

På trods af at anvendelsen af LLM stadig er relativt ny inden for statistisk social- og samfundsvidenskab, er nogle metoder begyndt at falde på plads. Relevant for dette projekt er der '*Silicon Sampling*', en metode dannet af Argyle et al. (2023), hvori modellen gives kontekst i form af en persona, typisk dannet ud fra relevant demografisk eller politisk data fra reelle mennesker, som derefter præsenteres for et sæt af spørgsmål eller opgaver som tilsvarende mennesker også har løst. Resultaterne af disse undersøgelser af modellernes anvendelighed som syntetisk data-generator har været blandede til negativ (Argyle et al., 2023, 2025; Bisbee et al., 2024; McIntosh et al., 2024; Röttger et al., 2024). Et af de hyppige problemer

som opstår, er at modellerne ikke opfanger den diversitet og varians der eksisterer inden for en given population. Trods medianen af modellens responser typisk er tæt på det menneskelige, så er den udfordret når det kommer til ekstremerne (Bisbee et al., 2024). Dette kan medføre en række forskellige udfordringer, som bias i form af konformitet til stereotyper og eksklusion af niche synspunkter, men det peger også på en grundlæggende begrænsning i forhold til anvendelsen af LLM som metode for syntetisk data-generation, idet modellerne ikke svarer som mennesker i aggregat.

Årsagen til (mit valg af) fokus på hvad der her er beskrevet som 'variansproblemet', frem for andre relevante LLM-problematikker i forbindelse med syntetisk data-generation som hallucinationer og generaliserbarhed er at variansproblemet er efter min læsning underbelyst. Som Argyle et al. (2025) nævner i en nyere artikel, er der flere studier som har stødt på, at LLM har svært ved at indfange varians i konteksten af syntetisk data-generation, men dette har været i forbindelse med undersøgelsen af andre effekter eller områder inden for brugen af LLM.

Men LLM udviklingen bevæger sig hurtigt, og flere af disse studier er et par år gamle, hvilket betyder at nye og mere sofistikerede modeller nu er tilgængelige. Bare under dannelsen af dette projekt udgav OpenAI deres GPT-4.5 og GPT-5 modeller, og opdaterede deres GPT-4o model (OpenAI, 2025a, 2025d). Der er derudover også værktøjer tilgængelige via OpenAIs API til at justere på modellens opførsel og råderum. Det er derfor værd at kigge på flere af de nye modeller og metoder for at undersøge, hvordan de påvirker variansproblemet.
Dette projekt har derfor følgende konkrete problemformulering:

Hvordan påvirker nyere Large Language Modeller, temperaturindstillinger og forbedringer i prompten modellernes evne til at svare med menneskelignende varians i aggregat?

Til det formål vil dette projekt fokusere på tre tilgange, for at belyse variansproblemet, med forventning om at finde teknikker til at minimere det. Den grundlæggende metode er Silicon Sampling, hvor udvalgte spørgsmål fra den seneste Tryghedsmåling fra 2024 vil blive anvendt som det menneskelige data

komponent og demografiske kontekst til dannelsen af personaer (Andersen et al., 2024).

Den første tilgang er at afprøve en række modeller, herunder GPT-3.5-Turbo, GPT-4.1 og o3. Dette er for at kunne sammenligne udviklingen mellem modellerne over tid, og særligt om o3 modellens evne til at danne en kæde af tanker kan medvirke til at løse problemet.

Den anden tilgang er at sammenligne forskellige temperaturindstillinger, hvilket gøres på de samme modeller, så de to analyser kan sammenlignes. Her vil standard og maksimum temperaturindstillinger undersøges, dette er ud fra den logik, at den begrænsede varians kan opstå på grund af begrænset råderum for modellen.

Den tredje tilgang er at sammenligne to prompt, denne tilgang opstod naturligt som del af analysen og sammenligner alle tre modeller ved begge temperaturer, så resultaterne her også kan sammenlignes med de forrige tilgange. Forskellen mellem de to prompt er en række rettelser som uddybes i metodeafsnittet.

2. Teori

Dette afsnit dækker projektets teoretiske grundlag i form af begrebsafklaring, hvor centrale begreber defineres, og der foretages en kort gennemgang af den eksisterende forskning indenfor brugen af LLM til besvarelse af spørgeskemaundersøgelser med henblik på syntetisk datageneration.

2.1. Begrebsafklaring

2.1.1. Large Language Model

LLM eller Large Language Models kan forstås på forskellige måder i forskellige kontekster. Den følgende forklaring holder sig grundlæggende til dette projekts problemfelt. Dette betyder at denne forklaring er ikke ment som en fyldestgørende formidling af disse modellers grundlæggende funktion eller arkitektur, men en lægmandsforklaring af de dele som er mest relevant for dette projekt.

En LLM er en type sprogmodel, en algoritme, hvis input eller output kan være generelt forståeligt tekst. Sprogmodeller er vigtige, fordi de er et værktøj der tillader computere at interagere med naturlig tekst. Moderne sprogmodeller er typisk det der hedder *generative pre-trained transformers*, eller GPTs. Disse modeller fungerer ved at gætte på det næste ord i en tekst ud fra de forrige ord i teksten. Deres gæt er informeret ud fra et neuralt netværk som, ved at være blevet trænet på store mængder af tekst, opfanger ikke kun de grammatiske mønstre i naturligt skriftligt sprog, men også mønstrene af betydning.

Når modellen er blevet promptet og skal generere et output, er det ved at danne en liste af mulige ord, hvoraf nogle er mere sandsynlige end andre. Modellen vælger et ord, og danner en ny opdateret liste af mulige ord, ud fra prompten og det første ord, den valgte. Dette gentager modellen indtil et endepunkt er nået. Størrelsen på listen og sandsynligheden for hvert ord på den er bestemt af en række parametre, herunder temperatur, som samlet set hedder decoder-strategien (IBM, 2023; Zhao et al., 2025, pp. 1-3).

Decoder-strategien vil typisk være indstillet til at undgå altid at vælge det mest sandsynlige ord. Hvis decoder-strategien altid fremtvang det mest sandsynlige ord, ville modellens endelige output være deterministisk. Variansproblemet kan opstå idet den liste decoder-strategi skaber potentielt også kan inkludere svar som reelt ikke

passer ind i konteksten af prompten og det output som er blevet genereret indtil da. Hvis rækkefølgen på ordene modellen skaber inkludere sådan irrelevante ord før ord som er gyldige, men blot usandsynlige, for eksempel minoritets svar, så kan forsøget på at ekskludere de irrelevante ord fjerne de relevante men usandsynlige ord også. Givet irrelevante svar vil undgås, anvendes typisk en decoder-strategi som danner en mere begrænset liste.

2.1.2. Temperatur

Temperatur er en type random sampling strategi som er brugt i LLM decoder-strategier. Temperatur er et tal positivt tal over 0 og typisk på eller under 2. Når en LLM skal vælge et ord danner den som nævnt ovenfor en liste af mulige ord, hver med et tal som viser hvor vigtigt modellen mener det ord er ud fra konteksten. Dette tal er input i en formular, hvis output er sandsynligheden for at det ord bliver valgt. Temperaturindstillingen indgår i formularen i form af en division af det tal modellen har bestemt.

I praksis betyder dette, at en temperatur tæt på 0 skalerer disse tal op, hvilket gør at ord som modellen har dømt vigtige bliver mere sandsynlige, og ord som er mindre vigtige er mindre sandsynlige. En temperatur på 2 skalerer derimod modellens tal ned, hvilket udjævner sandsynlighedsfordelingen i listen. En temperatur på 1 efterlader tallet uændret (Zhao et al., 2025, p. 27).

2.1.3. Variansproblemet

Variansproblemet er et fænomen, hvor LLM overvurderer sandsynligheden for mere-sandsynlige forekomster, og undervurder sandsynligheden for de mindre-sandsynlige forekomster. På denne måde formår modellen ikke at rumme det reelle spænd af variation, der kan være i menneskelige besvarelser, hvor der ofte vil være enkelte atypiske udfald. Variansproblemet er på sin vis ikke noget der eksisterer i en enkelt LLM besvarelse, men, i hvordan en LLM svarer på sammenlignelige eller identiske spørgsmål, særligt sammenlignet med hvordan en gruppe af mennesker ville svare.

LLM output er ofte relativt gode til at opfange median eller den hyppigste besvarelser af en population, men har svært ved at opfange den korrekte fordeling af besvarelser i en population. Den overvurderer hvor mange der svarer det mest typiske eller gennemsnitlige, og enten undervurdere eller fraholder sig fra at give

svar som agerer outliere, eller blot er atypisk relativt til normalen (Argyle et al., 2023, Bisbee et al., 2024).

Variansproblemet kan på denne måde minde om underdispersion, hvor der er mindre variation i data end forventet, og mode-collapse, hvor en machine-learning model kun giver enkelte typer af svar frem for det fulde spænd af svar der eksisterer i træningsdata. Jeg har valgt alligevel at beskrive fænomenet som variansproblemet fordi årsagerne til at det opstår og dermed hvordan det løses ikke nødvendigvis er de samme som ved underdispersion eller mode-collapse. Derudover lægger begrebet 'variansproblemet' også større vægt på fænomenets udfordringer i en samfundsvidenskabelig kontekst, hvor stereotypificering typisk ikke er ønskeligt.

2.2. Eksisterende Forskning

Forskningen på brugen af LLM inden for socialvidenskaben har været omfattende, særligt inden for brugen af LLM til spørgeskemabesvarelse, men har også lidt af akademisk vokseværk i den forstand at det har været fokuseret på empiri-dannelse frem for teori-dannelse. Dette er ikke nødvendigvis underligt, givet den hurtige udvikling på området. Men det betyder, at der er en mangel på anerkendte teoretiske værktøjer og tilgange inden for området (Argyle et al., 2025).

Det er ikke intentionen med dette afsnit at gennemgå LLM forskningens historie fra start til slut, heller ikke i konteksten af samfundsvidenskabelig forskning, men derimod at give et indtryk af de muligheder og udfordringer som berører variansproblemet.

Den 21 februar, 2023 udkom artiklen "*Out of One, Many: Using Language Models to Simulate Human Samples*" i skriftet *Political Analysis*, skrevet af Lisa P, Argyle, et al. (2023). Denne artikel foreslog at bruge Large Language Models som GPT-3 til at agere spørgeskema respondenter. Dette kom ud fra den observation, at modellen kunne påtage sig en bestemt rolle, og på den måde svarer, som om den tilhørte en udvalgt befolkningsgruppe. Denne egenskab blev identificeret og understøttet af det artiklen beskriver som *Algorithmic Fidelity* (Argyle et al., 2023, s.339), hvilket de definerer som graden hvor ved en model kan agere ud fra de komplekse mønstre og forhold mellem ideer, holdninger og socioøkonomisk kontekst som afspejler reelle menneskelige befolkningsgrupper. Det primære kriterium for at en model kan siges at have høj Algorithmic Fidelity er, at dens besvarelser ikke lader sig skelne fra menneskelige besvarelser, altså at den består en form for socialvidenskabelig Tûring test (Argyle et al., 2023).

Dette, at en models besvarelse ikke kan skelnes fra en menneskelig besvarelse, er stadig et faktum for mange af de tekstbaserede LLM-besvarelser. OpenAIs forsøgte at løse dette potentielle problem i form af et værktøj som havde til formål at kunne skelne LLM-genereret tekst fra menneskelig tekst, men dette værktøj blev efterfølgende tilbagetrukket, grundet lav nøjagtighed (OpenAI, 2023a).

En høj Algorithmic Fidelity løser på den måde det grundlæggende problem, der har eksisteret ved at anvende chatbots til syntetisk data-generation, at deres træningsdata har bestået af vilkårlige udklip af internettet. Det betyder, at deres holdninger sjældent har afspejlet nogen repræsentativ demografi. Men hvis LLM kan

bringes til at agere som udvalgte målgrupper når den gives den relevante kontekst, ville man kunne bruge modellens svar til at studere de tilsvarende menneskelige befolkningsgrupper. Dette danner den metode, de kalder *Silicon Sampling* (Argyle et al., 2023, s.340), hvor modellen gives den relevante kontekst til at påtage sig rollen som en bestemt befolkningsgruppe, hvorefter modellen løser en række opgaver. I artiklen holdes disse resultater derefter op imod tilsvarende menneskelige besvarelser. Resultaterne viser, at modellens svar følger mange af de samme menneskelige mønstre, hvilket også fungerer som en validering af modellens resultater (Argyle et al., 2023).

På baggrund af artiklens analyse viser de, hvordan man alene ved brug af modellens resultater ville kunne drage de samme konklusioner som ved de menneske-generede resultater. På baggrund af dette argumenterer de for, at det er muligt at bruge syntetisk generet data i tilfælde, hvor menneskelige data ikke kan opnås (Argyle et al., 2023).

Den 17 maj, 2024 udgav *Political Analysis* en artikel online af Bisbee et al. (2024), som undersøgte en række af de udfordringer som ligger i anvendelsen af Silicon Sampling. Ved brug af *The American National Election Survey,* giver Bisbee et al. (2024) modellen og metoden et best case scenarie, hvor den har både demografisk og politisk kontekst, i form af 10 forskellige parametre, inklusivt respondentens politisk identitet, det vil sige, om de var registreret demokrat eller republikaner. Modellen skal derefter oplyse respondentens holdning til en målgruppe som et tal fra 0 til 100, på samme måde som de reelle respondenter gjorde (Bisbee et al., 2024).

Som Argyle et al. (2023) viste, er modellen god til at fange medianen, særligt i den aggregerede data, dog fremhæver Bisbee et al. (2024) at allerede her har modellen en tendens til systematisk at reducere variansen. Dette problem bliver kun større når de demografiske og politiske grupper underinddeles. Her bliver modellens svar mere extreme og karikerede, særligt grundet den manglende evne til korrekt at vise den fulde varians i svar. Dette mønster af overkonfidens i modellens besvarelser medfører også, at den undervurderer hvor mange respondenter, som skal bruges for at opfange en ændring i affektiv polarisering siden 2012. I konteksten af studiet viser de, at modellen her estimerer forkert med næsten faktor 10, hvilket er en udfordring for anvendelsen af syntetisk generet data til pilotstudier og planlægning af fremtidig forskning (Bisbee et al., 2024).

Overfor artiklen fra Argyle et al. (2023), udfordre Bisbee et al. (2024) som sådan ikke den underliggende logik ved Algorithmic Fidelity eller Silicon Sampling, men understreger derimod, at der stadig er lang vej igen før syntetisk generet data vil kunne supplere eller erstatte data fra reelle mennesker.

Den 8 april, 2025 udkom et preprint på *ArXiv.org* af Argyle et al. (2025). Dette preprint, om end ikke et metastudie, forholder sig til status på brugen af LLM inden for politisk og samfundsvidenskabelig forskning over de sidste par år, hvilke erfaringer det har medført og giver en række opfordringer til at forbedre forskningen indenfor dette område (Argyle et al., 2025).

De beskriver hvordan LLM er forsøgt anvendt til mange forskellige opgaver inden for den videnskabelige proces, inklusiv opgaver som før ikke var mulige. I flere af disse situationer har modellens evner været imponerende, men de viser også at modellen ofte har fejlet, selvom opgaverne har været stærkt sammenlignelige med nogen, den før har løst (Argyle et al., 2025).

Denne variation i anvendelsen og begrænsning i generaliserbarhed i forhold til andre socialvidenskabelige metodiske værktøjer, samt manglen på metodiske standarder og benchmarks har sandsynligvis medvirket til den fragmenterede holdning til anvendelsen af LLM inden for området, som vist ved de to ovenstående artikler.

For at minimere denne forvirring og gøre fremtidige resultater mere anvendelige til den kumulative forskning foreslår Argyle et al. (2025) at der bør lægges mere vægt på inferens i en klassisk socialvidenskabelig forstand, hvor det handler om at kunne generalisere ud fra sine fund ved at vise, at det observerede mønster, effekt eller mekanisme er gældende i mere end en enkelt situation (Argyle et al., 2025).

Det er i deres øjne derfor vigtigt at forholde sig til både modellens succeser og dens fejl. Fejl bør ikke over- eller underfortolkes, forstået på den måde, at hvis modellen fejler, er det vigtigt at forsøge at afklare hvordan, om det var fordi LLM ikke kan løse den type opgaver, om det er manglende træning i den givne model, om alignment forstyrrer processen, eller det er en fejl i hvordan modellen er prompted. Der lægges også vægt på validering af modellernes resultater, særligt i tilfælde hvor resultaterne virker lovende. Dette er både for at sikre at resultaterne er reelt gældende, men også for at afgrænse deres omfang, for, som vist ved de foregående

artikler, betyder et positivt resultat, f.eks. at modellen kan opfange et samlet median af en udvalgt population, ikke nødvendigvis, at den kan klare en lignende opgave, f.eks. opfange varians eller median i underindelte populationer (Argyle et al., 2025).

Disse artikler fremhæver derfor en række overvejelser, som er relevante i forbindelse med dette projekt og denne analyse. Silicon sampling er den metode som vil blive anvendt, og ud fra disse artikler er der en forventning om at modellen vil kunne estimere populationernes samlede median, men med lavere end menneskelig varians. Det forventes også, at hvis populationen underinddeles, vil modellens evne til at gengive både median og varians forværres. Også tilfælles med disse artikler undersøger dette projekt selv-rapporterede holdninger og danner personaerne til silicon sampling metoden ud fra relevante demografiske variabler inkluderet i et menneskeskabt dataset.

Til forskel fra artiklerne fokusere dette projekt på Danmark, og anvender derfor dansk data. Også til forskel fra de forrige studier fokuserer dette projekt på spørgsmål om tillid, givet at det er bredt samfundsmæssigt relevant både i Danmark og i andre lande. Ved at undersøge Danmark og tillid frem for amerikansk politik rykker analysen væk fra den tidligere best case tilgang til en mere typical case tilgang. Dette har det henblik at få et indtryk af hvor meget, hvis noget, modellens evne til at gengive menneskelige responser i aggregat forværres. Det betyder også at det er vigtigt at undersøge eller overveje, i det omfang modellen fejler, hvorfor det sker og hvilken effekt nyere modeller, højere temperatur og en bedre prompt har på modellens evne til at generere syntetisk data med menneskelignende varians og median.

3. Metode

Det følgende afsnit omhandler projektets metode. Først præsenteres Tryghedsmålingen, hvilke afhængige og uafhængige variabler der anvendes og hvorfor. Dernæst beskrives hvilke modeller og temperaturindstillinger der er blevet anvendt og hvorfor. Endelig gennemgås selve analysedesignet.

3.1. Valg af Data: Tryghedsmålingen 2024

Tryghedsmålingen er en landsdækkende spørgeskemaundersøgelse, som undersøgerer danskeres holdninger og opfattelse af tryghed i en række forskellige forbindelser som familie, vold og kriminalitet, økonomi, samt deres tillid til politiet, medier og politikere. Spørgeskemaundersøgelsen blev udført af YouGovs Danmarks panel i perioden den 12 februar til den 4 april 2024, med henblik på at indsamle et repræsentativt snit af den danske befolkning over 18 år. 5725 respondenter færdiggjorde undersøgelsen, hvoraf omkring 2000-3500 har deltaget i et eller flere tidligere Tryghedsmålinger, som tillader et begrænset element af longitudinelt design, trods målingen hovedsagligt er et tværsnitsdesign. Data omkring tillid er suppleret fra Rockwool-projektets undersøgelse Borgerne og Lovene.

Tryghedsmålingens repræsentativitet er dog begrænset af sit digitale format, som i mindre grad kan nå ud til svage ældre og hjemløse. Blandt respondenterne er kvinder overrepræsenteret med 56,7% versus 43,3% for mænd, og yngre aldersgrupper er også underrepræsenteret. Dette medføre at Tryghedsmålingen gør brug af vægtning på baggrund af Danmarks Statistik for køn, alder, bopæls-region og uddannelse.

3.1.1. Demografisk Kontekst

For at kunne anvende Silicon Sampling er det nødvendigt at give modellen kontekst for den population som den skal agere som. Med henblik på dette har jeg valgt følgende demografiske variabler, vist i Tabel 1, hentet fra Tryghedsmålingens dataset. Disse variabler er valgt ud fra både tilgængelighed og ud fra, at tidligere studier inden for anvendelsen af LLM til syntetisk data-generation har valgt tilsvarende variabler.

Variable	Betydning	Svarmuligheder
Alder	Alder på respondenten	Tal på eller over 18, gennemsnit på 51,4. højeste respons er 98
Køn	Respondentens køn	Binært mand eller kvinde
Uddannelse	Respondentens højest gennemførte uddannelse	8 svarmuligheder
Civilstatus	Respondentens civiltilstand, f.eks. Om de er gift.	8 svarmuligheder
Familie-Status	Om respondenten f.eks. bor alene, med forældre eller med ægtefælle.	7 svarmuligheder
Børn	Hvor mange børn der er i respondentens hushold.	6 svarmuligheder, fra 0 til 5 eller mere.
Personlig Indkomst	Hvilken indkomst har respondenten	13 svarmuligheder, inkl. Ved ikke og ønsker ikke at svare.
Parti til næste valg	Hvilket parti/hvordan respondenten vil stemme til næste folketingsvalg.	20 svarmuligheder, inkl. Muligheder som ‘ har ikke stemmeret’ og ‘vil stemme blankt’. ‘Ved ikke’ udgør det hyppigste svar.
Parti til forrige valg	Hvilket parti/hvordan respondenten stemte ved det sidste folketingsvalg (2022).	19 svarmuligheder, enkelte muligheder er slået sammen og ‘husker ikke’ er introduceret.
Bystørrelse	Størrelsen på den by respondenten bor i.	9 svarmuligheder.
Kommune	Hvilken Kommune respondenten bor i.	98 svarmuligheder.

Tabel 1: *Demografisk Kontekst*. Viser de 11 demografiske variabler anvendt til dannelsen af personaer, variablernes betydning samt antallet af svarmuligheder med eventuelle uddybninger. Manglende værdier forekom primært i bystørrelse (n=344).

3.1.2. Valg af Spørgsmålsbatteri

For at begrænse omkostningerne i forbindelse med udførelsen af analysen har jeg valgt at holde mig til et batteri af udsagns-spørgsmål fra Tryghedsmålingen. For dette projekt har jeg derfor valgt at arbejde med et batteri bestående af i alt 9 udsagn som alle på forskellig vis omhandler oplevelsen af tillid.

Anvendelsen af Tryghedsmålingen, og dette batteri omkring tillid, frem for generelt offentligt tilgængeligt data eller generelle spørgsmål om tillid, kommer af at LLM modellerne i høj grad er trænet på offentligt tilgængeligt data. Dette betyder at ved anvendelsen af offentlig data kunne modellen forsøge at svare på spørgsmål ud fra dens hukommelse, frem for at ræsonnere sig til svaret. Batteriet er også valgt på baggrund af responsformatet på udsagnene. Givet LLM er trænet på sprog-data, har jeg en forventning om at modellen typisk har en stærkere forståelse for sproglige beskrivelser frem for matematiske beskrivelser, altså at modellen bedre kan forholde sig til og svare nøjagtigt på et spørgsmål hvor den kan svare f.eks. 'Enig' frem for spørgsmål hvor den skal give et tal. (Mahendra, et al., 2025)

Tabel 2 viser de 9 udsagn fra det udvalgte spørgsmålsbatteri. Her har de haft 6 svarmuligheder: "Helt enig", "Delvist enig", "Hverken Enig eller Uenig", "Delvist Uenig", "Helt Uenig" og "Ved ikke/Ikke relevant/Ønsker ikke at svare".

#	Spørgsmålsformulering	Nøgle
1	Det er de færreste, der kan holde på en vigtig hemmelighed.	Hemmelighed
2	Folk tager typisk ansvaret, hvis deres børn har gjort noget galt.	Ansvar Børn
3	De fleste er normalt ærlige, når de fortæller om noget, de har oplevet.	Ærlighed
4	Det er efterhånden de færreste, der sætter en ære i at holde, hvad de lover.	Løfte
5	Det er meget få, der ikke betaler, når de tager noget fra en bod ved vejen.	Betaling
6	De fleste vil gerne hjælpe andre, også selvom de ikke kender dem.	Hjælpsomhed
7	Korruption er meget sjældent i Danmark.	Korruption
8	Næsten ingen bruger vold, bortset fra i nødstilfælde.	Nød Vold
9	Man kan stole på at de allerfleste voksne vil opføre sig ansvarligt/ordentligt, hvis de har ansvaret for andres børn.	Ansvar Voksne

Tabel 2: *Spørgsmålsformulering*. Viser de 9 udsagn fra spørgsmålsbatteriet omkring tillid, samt en nøgle som er et ord der opsummerer hvert udsagnet.

3.1.3. Tillidsindeks

Givet dette projekt omhandler hvor godt modellen kan indfange menneskelig median og varians er der en række udfordringer ved dette batteri. Svarene er kategoriske, en af mulighederne kan ikke gøres intervalskaleret og at svare ‘enig’ betyder ikke ‘mere tillid’ på tværs af alle spørgsmålene.

For at løse den første udfordring har jeg valgt at omdanne svarende til en intervalskala, hvor “helt enig” er 1, “delvist enig” er 2, “hverken enig eller uenig” er 3, og så videre.

Derefter har jeg valgt at ekskludere svarmuligheden “Ved ikke/Ikke relevant/Ønsker ikke at svare” fra modellens svarmuligheder grundet denne intervalskalering, og en forventning om at LLM vil have en forhøjet tilbøjelighed til at vælge det udfald hvis muligheden bliver givet (Rupprecht et al., 2025).

For at løse den sidste udfordring med de modsatrettede udsagn, 1 om hemmeligheder og 4 om at love, kan jeg bare konvertere et “helt enig” (1) svar til “helt uenig” (5) og dernæst konvertere “delvist enig” (2) til “delvist uenig” (4), og så lade “hverken enig eller uenig” (3) forblive uændret.

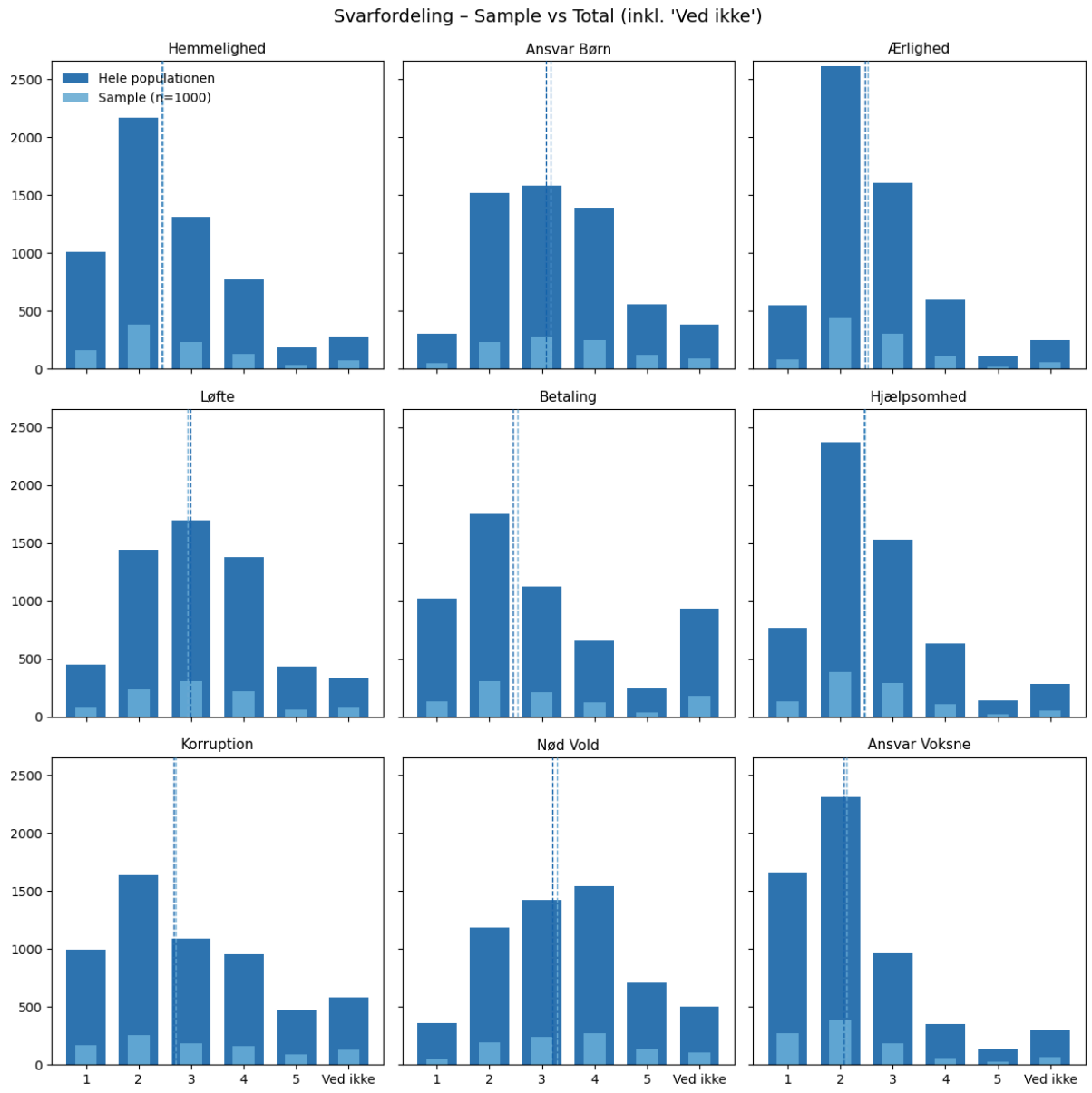
Efter alle 9 udsagn om tillid er blevet intervalskaleret, kan jeg lave en aggregatværdi, som herefter refereres til som tillidsindekset. Vær opmærksom på at en værdi af 1 på dette tillidsindeks betyder at personen i høj grad oplever tillid, mens en værdi af 5 betyder at personen i lav grad oplever tillid.

Dette tillidsindeks kan derfor bruges til at se, samlet set, hvordan modellen klarer sig i forhold til menneskene. De underliggende spørgsmål eksisterer stadig og vil også indgå i analysen for at se, hvorfor modellen svarer som den gør i aggregat, og hvilke mønstre der er i dens besvarelser.

Konverteringen af kategoriske værdier til interval værdier, der danner en aggregeret værdi er ikke uden visse omkostninger for resultaterne af analysen, idet værdierne ikke har tilsvarende reelle svarmuligheder. Antallet af kategoriske svarmuligheder, at spørgsmålene alle omhandler samme tema, tillid, og at analysen i mindre grad går på det substantielle indhold af spørgsmålet frem for fordelingen af svar, medvirker alle til at konverteringen ikke invalidere analysen eller den resulterende konklusion.

3.1.4. Summary Statistics

For at give et hurtigt indblik i svarfordelingen mellem samplet bestående af de 1000 respondenter, som vil blive anvendt i analysen og den samlede population af hele Tryghedsmålingen, se figur 1 nedenfor. Vær opmærksom på at denne figur indeholder ‘ved ikke’ svar muligheden, som er ekskluderet fra analysen, medianen er dannet kun ud fra svarmulighederne 1-5.



Figur 1: *Svarfordeling - Sample Vs Total (inkl. 'Ved ikke')*. Viser svarfordelingen ved hvert af de 9 udsagn for samplet (n=1000) der vil bruges i analysen som er vist i lyseblå, i forhold til det totale antal af respondenter i populationen (n=5725) som er vist i mørkeblå. X-aksen ved hvert bar plot viser svarmulighederne, hvor 1 = helt enig, 2 = delvist enig, 3 = hverken enig eller uenig, 4 = delvist uenig, 5 = helt uenig, Ved ikke = Ved ikke. Y-aksen er antallet af respondenter. Lodrette stiplede linjer viser medianen af sample og total population, ekskluderet 'ved ikke' svarmuligheden.

3.2. Afhængige Variabler

For den kommende analyse vil der være tre afhængige variabler. Variansandel, Medianafstand og Fejlrate. Dannelsen af disse variabler er konkret beskrevet under Analysedesign i afsnittet længere ned i metodeafsnittet.

3.2.1. Variansandel

Variansandel er et positivt tal over nul som viser hvor tilsvarende variansen i modellens samlede besvarelse er i forhold til mennesket, hvor en værdi på 1 betyder at modellens samlede varians er den samme som den samlede menneskelige varians. Det er muligt, men forventet usandsynligt, at observere en variansandel på over 1, da dette ville være udtryk for at modellen har opfanget en større samlet variation end den tilsvarende menneskelige sample.

Vær opmærksom på, at variansandel ikke viser om modellen har givet de samme svar som menneskene, men blot om den har kunne indfange den tilsvarende varians. Svarfordelingen på de underliggende spørgsmål vil undersøges undervejs i analysen, men på et rent kvalitativt plan.

3.2.2. Medianafstand

Medianafstand er et tal mellem 0 og 5 som viser afstanden mellem medianen af modellens samlede besvarelse og den samlede menneskelige besvarelse. Et 0 ville betyde, at modellen har nøjagtigt den samme median som den tilsvarende gruppe mennesker. Et 1 ville betyder at medianen for modellens besvarelse er et interval trin væk fra den menneskelige median.

Ligesom ved variansandel er dette ikke et mål for, om modellen svarer ‘rigtigt’ i forhold til menneskene, men blot hvor langt modellens median, samlet set, er fra den menneskelige.

3.2.3. Fejlrate

Fejlraten er lidt anderledes end de to forrige variabler. Det er et procenttal for, hvor mange procent af modellens forsøg på at generere et sæt af 9 besvarelser som fejler. Den er vigtig grundet interaktionen mellem de uafhængige variabler. En fejlrate på 0% ville betyde at alle modellens forsøg er klaret, og en fejlrate på 100% ville betyde at ingen af de svar modellen har sendt tilbage har været anvendelige responser.

De nøjagtige årsager til at modellen fejler kan variere, og er ikke inden for rammerne af dette studie. Det er potentielt muligt at reducere fejlraten ved at køre den samme test flere gange indtil modellen har givet et funktionelt svar for alle respondenter, men denne løsning forventes urealistisk for større studier, så derfor opfattes en høj fejlrate som et fund i sig selv.

3.3. Uafhængige Variabler

For den kommende analyse vil der være tre uafhængige variabler: Model, temperatur og prompt. Jeg har valgt tre modeller, to temperaturindstillinger og to forskellige prompt. Idet o3 modellen ikke kan anvende temperaturindstillingen betyder dette, at jeg kører i alt 10 test. En for hver kombination af de tre variabler, minus to for test af temperatur på modellen som ikke er kompatible med den indstilling.

3.3.1. Valg af Model

Jeg har valgt at arbejde med 3 modeller, med henblik på at undersøge modelfunktionalitet og kapacitet på variansproblemet. Jeg fraholder mig dog fra at direkte sammenligne modelparametre grundet disse ikke altid er offentligt tilgængelige samt, at det ikke viser det fulde billede af hvad der differentiere modeller, hvilket ofte også involverer træning og alignment, som forventes i mindre grad kan kvantificeres og direkte sammenlignes. Jeg har valgt at holde mig til modellerne lavet af OpenAI grundet tidligere forskning inden for varians i syntetisk generet data har anvendt GPT-modeller (Argyle, et al., 2023; Bisbee, et al., 2024).

Derudover tillader OpenAIs API batching, hvilket reducerer omkostninger og simplificerer gentagne spørgsmål, hvilket er særligt relevant for hvordan jeg udfører min analyse.

GPT-3.5 Turbo:

Dette er en ældre model, udgivet den 1. Marts, 2023 (OpenAI, 2023b). Denne model er beskrevet som en legacy model på OpenAIs model documentation. Dette gør den egnet som et aproximat for status af offentligt tilgængelige LLM fra den tid, og lader den derfor agere som en kontroltest i forhold til de nyere modeller når det kommer til at kigge på hvordan model kapacitet påvirker variansproblemet (OpenAI, n.d.-b).

GPT-4.1:

Dette er en nyere model, udgivet den 14. April, 2025 (OpenAI, 2025b). Denne model er beskrevet af OpenAI som deres nuværende (26/05/2025) flagship model, og kan tilgås både via API og via chatgpt.com. Dette forventes at gøre den egnet som en indikation af nuværende, bredt-anvendelige modeller. Jeg kunne også have valgt GPT-4o modellen for dette formål da den er standarden hvis man åbner chatgpt.com, men givet dette projekt er rettet mod forskning som forventes at anvende API adgang, er GPT-4.1 set som mere passende for den nuværende standard model (OpenAI, n.d.-c).

o3:

Dette er en stærkere model, som udkom den 16. april, 2025 (OpenAI, 2025c). o3 er det OpenAI kalder en *reasoning model* (OpenAI, n.d.-e), hvilket betyder at den danner en intern kæde af tanker som den så svarer ud fra. Det gør modellen bedre til at håndtere komplekse problemer, eller problemer hvis løsning kræver flere trin. Det har dog den omkostning, at det både er langsommere og dyrere at bruge o3 modellen end de tidligere nævnte. Denne model er valgt både for at se om reasoning arkitektur giver forbedringer i forhold til variansproblemet, men også fordi den forventes at være en forbedring over GPT-4.1 modellen når det kommer til ræsonnement evne (OpenAI, n.d.-e; OpenAI, n.d.-d).

3.3.2. Valg af Temperatur

For dette projekt har jeg valgt at anvende to temperaturmål, et på ‘1’ og et på ‘2’. Temperatur på ‘1’ er valgt ud fra en forventning om at det udgør standardindstillingen via OpenAIs API. Temperaturindstillingen på ‘2’ er valgt fordi den er den maksimale, og forventes derfor at vise den største stigning i varians.

Jeg afgrænser mig fra at analysere temperaturmål under 1, fordi der ikke er nogen umiddelbar forventning om at det skulle forbedre variansproblemet. Derudover søger jeg ikke efter en optimalt temperatur for dannelsen af menneskelig varians, ud fra en forventning om at sådan et mål ville afhænge af faktorer som model, dataset, prompt og opgave, og derfor ikke ville kunne generaliseres behjælpsomt til anden forskning.

3.3.3. Ændring af prompt

Det var oprindeligt ikke min intention at afprøve effekten af prompt på varians grundet en ambition om at holde fokus klart ved at holde det snævert. Jeg opdagede dog et par fejl efter at have kørt min analyse første gang. Disse fejl var en misforståelse af 'børn' variablen, som viste sig at være antallet af børn i husholdet og ikke antallet af børn, respondenten er far, mor eller værge til. Jeg havde derudover også overset 'kommune' variablen, som jeg opfatter som relevant demografisk information for modellen. Endelig bestod min indledende kode af et simplere svarformat for modellen, hvor den skulle svare med et tal frem for tekst.

Med andre ord kunne min oprindelige prompt opfattes som 'dårlig', idet den gav modellen forkert information, undlod relevant information og fik den til at svare i et format som den forventes at være svag i.

Jeg valgte derfor at gøre min analyse om, med disse mangler rettet til. Det tillod mig at sammenholde de to analyser for at se hvordan disse ændringer i prompt har påvirket modellens svar. Det er ikke muligt at sige hvordan og i hvilket omfang hver enkelt individuel ændring af prompten har påvirket modellens svar. Jeg opfatter det dog stadig som et gyldigt analytisk design, idet alle ændringerne kan forventes at forbedre modellens resultater. Dette punkt vender jeg tilbage til i diskussionsafsnittet.

Prompten vist i analyse design afsnittet omkring prompt-generation er den nye 'rettede' prompt, den gamle prompt og det resulterende eksempel kan findes i appendiks 4.

3.4. Analysedesign

Dette afsnit dækker projektets analyse design, ved først at gennemgå hvordan respondenter er blevet samplet ud fra det fulde dataset, derefter beskrives hvordan den prompt som sendes til modellerne bliver genereret, herefter hvordan modellens responser omdannes til en variabel med varians frem for kategorier og hvordan den sammenholdes med de menneskelige svar. Herefter forklares hvordan ændring af model, temperatur og prompt udføres for at isolere disse variabler for analysen.

3.4.1. Sampling strategi

For dette projekts analyse har jeg valgt at anvende en sample size på 1000 respondenter. Et højere antal ville have øget omkostninger i forbindelse med hver test, og et lavere antal ville gøre effekten af ændring i model, temperatur eller prompt mindre tydelig på tværs af test. For hver test der er kørt har det været de samme 1000 respondenter som er blevet testet, dette er gjort for at sikre den interne validitet ved at undgå at ændringer i resultater fra den ene test til den anden skyldes støj fra samplingen.

Disse 1000 respondenter er dog ikke tilfældigt udvalgt fra den oprindelige tryghedsmåling, de er derimod udvalgt ved først at splitte de 5725 respondenter i to grupper, en på 70% (4007) og en anden på 30% (1718). Dette split er gjort ud fra deres svar på 'parti til næste valg' variablen vist ovenfor i Tabel 2. Denne variabel er valgt grundet tidligere forskning på området har undersøgt partitilhørsforhold, og en forventning om at der er en relativt spredt fordeling i folks svar ud fra deres politiske holdning.

De 1000 respondenter er derefter udvalgt fra 70% (4007), hvilket sikre at der er ikke er noget overlap med de 30% (1718), som bruges til valideringen, som beskrevet i underafsnit **3.4.4. Validering**.

3.4.2. Prompt-generering

Det næste trin har så været at skabe det som sendes til modellen, hvilket består af to komponenter, et er instruktionen, hvilket giver modellen generel kontekst og fortæller den hvad den skal gøre. Dette ændrer sig ikke fra respondent til respondent. Det andet er input, hvilket er selve prompten, som er den konkrete opgave modellen skal løse.

Dette afsnit inkluderer udklip fra den Python Notebook som er anvendt til analysen. Jeg har forsøgt at formatere det så det er læseligt uden programmeringserfaring, men der vil stadig være enkelte dele kode undervejs. Der vil være et eksempel på det endelige input som sendes til modellen i slutningen af prompt-genereringsafsnittet.

Instruktionen som er sendt til modellen lyder som følger:

```
"Du er deltager i en opinionsundersøgelse (Tryghedsmålingen 2024).
Dette er en dansk undersøgelse, fortaget i 2024, som undersøger
danske borgers tryghed."
"Du skal simulere respondenten nedenfor i forbindelse med forskning."
```

Den anden del, prompten, er i sig selv underinddelt og består af i alt 4 komponenter: personaen, svarmuligheder, spørgsmålene og deres nummerering, og endelig svar formatet.

Personaen er der, hvor de uafhængige variabler kommer i spil, det er en skabelon som beskriver et individ, med 'tomme felter' for diverse variabler som automatisk udfyldes med information som svarer til en af de 1000 udvalgte respondenter. Persona delen ser derfor ud som følger:

```
f"Du er en {row.age}-årig {row.gender} med {row.education}, "
f"{row.marital}, {row.family}, har {row.children} børn i hjemmet, "
f"en personlig indkomst på {row.income} kr., "
f"du stemte på {row.party_id_2022} ved sidste valg, og planlægger at
fstemme på {row.party_id}. "
f"Du bor i en {row.city_size} størrelse by i {row.municipality}."
```

Dernæst er der svarmulighederne. Som nævnt tidligere er det her, muligheden for at svare 'ved ikke' ekskluderes:

```
"Hvor enig eller uenig er du i følgende udsagn?"
"Svar for hvert udsagn med en af følgende fem svarmuligheder:"
"Helt enig\n"
"Delvist enig\n"
"Hverken enig eller uenig\n"
"Delvist uenig\n"
"Helt uenig\n"
```

Så oplyses selve spørgsmålene:

```
"Det er de færreste, der kan holde på en vigtig hemmelighed",
"Folk tager typisk ansvaret, hvis deres børn har gjort noget galt",
"De fleste er normalt ærlige, når de fortæller om noget, de har
oplevet",
"Det er efterhånden de færreste, der sætter en ære i at holde, hvad
de lover",
"Det er meget få, der ikke betaler, når de tager noget fra en bod ved
vejen",
"De fleste vil gerne hjælpe andre, også selvom de ikke kender dem",
"Korruption er meget sjældent i Danmark",
"Næsten ingen bruger vold, bortset fra i nødstilfælde",
"Man kan stole på at de allerfleste voksne vil opføre sig
ansvarligt/ordentligt, hvis de har ansvaret for andres børn"
```

Som så nummereres og sættes op i punktform med den følgende linje kode:

```
udsagnstekst = "\n".join(f"{i+1}. {t}" for i, t in enumerate(udsagn))
```

Endelig er der så svarformatet, som beskriver hvordan modellen skal kommunikere de svar, den har valgt til hvert spørgsmål. Her er det valgt at den skal være i et JSON-format, hvilket gør det nemmere at indsamle dens svar fra batchfilen, der sendes tilbage fra API'en, når alle 1000 respondenter er blevet simuleret. Denne del er som følger:

```
"Giv dit svar i JSON-format nøjagtigt som nedenfor, "
"med nøglerne i denne rækkefølge:\n"
'{"hemmelighed": "<kategori>", '
'"ansvar_børn": "<kategori>", '
'"ærlighed": "<kategori>", '
'"løfte": "<kategori>", '
'"betaling": "<kategori>", '
'"hjælpsomhed": "<kategori>", '
'"korruption": "<kategori>", '
'"nød_vold": "<kategori>", '
'"ansvar_voksne": "<kategori>"}\n\n'
"Hvor <kategori> er et af: "
"'Helt enig', 'Delvist enig', 'Hverken enig eller uenig', "
"'Delvist uenig', 'Helt uenig'."
```

Disse dele, persona, svarmuligheder, spørgsmål og nummerering, og svarformat, danner så et endeligt input, et eksempel er vist her:

```
Du er en 74.0-årig Kvinde med Erhvervsfaglig uddannelse, Gift, Jeg er
samboende/gift/registreret partnerskab og har ingen hjemmeboende
børn, har 0 børn i hjemmet, en personlig indkomst på 100.000 til
199.999 kr. kr., du stemte på A. Socialdemokratiet ved sidste valg,
og planlægger at stemme på F. SF - Socialistisk Folkeparti. Du bor i
en En by med over 40.000 indbyggere størrelse by i Køge.

Hvor enig eller uenig er du i følgende udsagn?Svar for hvert udsagn
med en af følgende fem svarmuligheder:Helt enig
Delvist enig
Hverken enig eller uenig
Delvist uenig
Helt uenig


1. Det er de færreste, der kan holde på en vigtig hemmelighed
2. Folk tager typisk ansvaret, hvis deres børn har gjort noget galt
3. De fleste er normalt ærlige, når de fortæller om noget, de har
oplevet
4. Det er efterhånden de færreste, der sætter en ære i at holde, hvad
de lover
5. Det er meget få, der ikke betaler, når de tager noget fra en bod
ved vejen
6. De fleste vil gerne hjælpe andre, også selvom de ikke kender dem
7. Korruption er meget sjældent i Danmark
8. Næsten ingen bruger vold, bortset fra i nødstilfælde
9. Man kan stole på at de allerfleste voksne vil opføre sig
ansvarligt/ordentligt, hvis de har ansvaret for andres børn

Giv dit svar i JSON-format nøjagtigt som nedenfor, med nøglerne i
denne rækkefølge:
{"hemmelighed": "<kategori>", "ansvar_børn": "<kategori>",
"ærlighed": "<kategori>", "løfte": "<kategori>", "betaling":
"<kategori>", "hjælpsomhed": "<kategori>", "korruption":
"<kategori>", "nød_vold": "<kategori>", "ansvar_voksne":
"<kategori>"}

Hvor <kategori> er et af: 'Helt enig', 'Delvist enig', 'Hverken enig
eller uenig', 'Delvist uenig', 'Helt uenig'.
```

Eksemplet viser at denne tilgang har en mulig svaghed i at de data, som kommer fra Tryghedsmålingen, ikke nødvendigvis passer ind i selve teksten rent grammatisk. Dette forventes dog ikke at ville fundamentalt ændre modellens svar. løsningerne på dette problem ville enten være omfattende, f.eks. kunne prompten gøres totalt dynamisk prompt, så den omkringliggende sætning ændres alt efter hvilket svar der er givet i en demografisk variabel. Alternativt kunne formuleringen af de demografiske variabler ændres så de passede ind i prompt formatet, men dette ville have introduceret validitetsproblemer.

For den gamle prompt med det numeriske svarformat, er ikke foretaget nogen ændring i den overordnede instruktions prompt, i spørgsmålene eller deres nummerering. De steder der er foretaget ændringer, er i persona, svarmuligheder og svarformat, som nævnt tidligere kan den gamle prompt finde i appendiks 4.

3.4.3. Variabeldannelse

Efter modellen har kørt og batch filen er hentet, er det næste trin at indsamle og rense dens svar og derefter danne en aggregeret variabel, som kan sammenholdes med de menneskelige svar i forhold til median og varians.

Indsamlingen af modellens svar er relativt simpelt med JSON-formatet. Med udgangspunkt i at de svar, modellen har skullet vælge mellem, har været tekst, har det været praktisk nødvendigt at udføre standard datarensningsmetoder som omdannelse til små bogstaver og stemming for at undgå at 'helt enig' og 'Helt Enig!" opfattes som to forskellige svar.

Hvis modellens svar ikke har været formateret korrekt, så det ikke har været muligt at indsamle et gyldigt svar fra 'respondenten', så er både modellen og det tilsvarende menneskelige svar blevet ekskluderet for den test. Dette er gjort med henblik på at sikre, at der er det samme antal syntetiske og reelle svar ved hver test, for at undgå at en ændring i variation skyldes et mismatch i størrelse mellem de to grupper. Fejlraten for modellen og dens indstillinger er dog i sig selv et vigtigt resultat, og vil derfor oplyses undervejs i analysen. Inden den aggregerede variabel, som viser en slags samlet 'tillid' for en given respondent, er det også nødvendigt at konvertere spørgsmål 1, om folk kan holde hemmeligheder, og spørgsmål 4, om folk holder hvad de lover, fordi ved disse spørgsmål svarer 'helt enig' til lavere tryghed og ikke højere.

Der dannes herved et datasæt, som består af de respondenter som har haft en succesfuld syntetisk besvarelse, hvad den menneskelige besvarelse har været til hvert spørgsmål, samt hvad de syntetiske svar har været. Hvis et menneske har svaret 'ved ikke' er det her listet som 'NaN', hvilket betyder at respondentens svar på netop det spørgsmål ikke er inkluderet i beregningen af gennemsnit. Dette betyder at det reelt set er umuligt for modellen at opnå 100% match på den menneskelige variation, men det er heller ikke målet ved dette projekt, og forventer ikke at have medført nogen nævneværdig ændring på de statistiske mønstre.

Ud fra dette udregnes følgende værdier:

1) Størrelsen på datasættet, det vil sige hvor mange gange ud af de 1000 respondenter, har modellen kunne give et anvendeligt svar. Dette danner fejlratevariablen.

2) Median og varians værdien for en respondents reelle besvarelse og for den syntetiske.

3) Hvor stor en andel den syntetiske varians udgør af den reelle. Dette danner variansandelvariablen.

4) Størrelsen på afstanden fra den reelle median til den syntetiske. Dette danner medianafstandsvariablen.

For at teste effekten af de uafhængige variabler på de afhængige blev en batch med 1000 respondenter sent til hver af de tre udvalgte modeller, først ved temperatur 1 med den oprindelige prompt, dernæst temperatur 2 for GPT-3.5-Turbo og GPT-4.1 også med den oprindelige prompt. Dernæst blev denne proces gentaget, blot med den nye prompt.

Det endelige resultat er derfor 10 test, hvor en test kan sammenlignes på enten model, temperatur eller prompt med andre test, med undtagelse af o3 modellen som ikke acceptere temperaturændringer. Dette tillader derfor også et indblik i hvordan disse afhængige variabler påvirker hinanden.

3.4.4. Validering

Det vil være værd at undersøge, om de fund, der er dannet undervejs i analysen, stadig er gældende, hvis der anvendes et nyt set af respondenter. For at undersøge dette vil der udføres en eller flere valideringstests, hvor en kombination af uafhængige variabler gentages, men med 1000 nye respondenter taget ud af de 30% (1718) af den samlede population, som blev sat til side til formålet som beskrevet i **3.4.1. Sampling Strategi**. Disse gentagne tests vil så blive sammenlignet med de forrige tests, og hvis flere test gentages vil de også sammenlignes indbyrdes. Dette foretages ud fra den logik, at hvis de mønstre som er observerede i den oprindelige analyse kan gentages med et nyt set af respondenter, er det et tegn på at de mønstre sandsynligvis er robuste. Nøjagtigt hvilke tests der vil gentages og hvorfor, vil blive beskrevet i selve analyseafsnittet.

4. Analysen

Dette afsnit beskriver projektets analyse. Det starter med et overblik over de udførte tests samt en kort præsentation af de vigtigste fund. Derefter er der en systematisk gennemgang af resultaterne af samtlige test. Derefter laves en sammenfatning af disse resultater med fokus på hver afhængig variabel. Endeligt foretages en validering af udvalgte test. Undervejs i testgennemgangen vil der præsenteres udvalgte visualiseringer.

Alle visualiseringer og yderligere data som prisestimat, nøjagtige processeringspunkter, Python Notebooks og output batch files for samtlige test kan findes i appendiks1-3 og appendiks 5-20.

Det er vigtigt for læsningen af dette afsnit at være opmærksom på at der ikke som sådan er et mål for hvornår modellen har gjort noget 'godt nok' i forhold til variansandel, medianafstand og fejlrate. Givet dette projekt ikke forsøger konkret at løse variansproblemmet, eller andre problemer i forbindelse med brugen af LLM til syntetisk datageneration, men derimod at undersøge påvirkningen af model, temperatur og prompt på variansproblemet, er det ikke min intention at opstille hårde kriterier for hvornår modellen har præsteret 'godt nok'. Fokus er på at beskrive de relative forbedringer og forværringer i forhold til det ideale resultat, og resultaterne af de andre test. Det bedst mulige resultat givet de variabler der undersøges, ville være en variansandel på 100%, en medianafstand på 0 og fejlrate på 0%.

4.1. Overblik

Tabel 3 beskriver resultaterne af hver af de 10 udførte tests, samt navn på hver test. Navnet er dannet ud fra de uafhængige variabler, med model først, efterfulgt af temperaturen, hvis det kan justeres, og endeligt hvilken prompt der har været brugt. På denne måde er den første test, udført med GPT-3.5-Turbo modellen, ved temperatur 1 og ved brug af den numerisk-baserede prompt kaldet Old-T1-N. Afsnit **4.1.2. Konkrete Resultater** indeholder Tabel 4 som viser de samme test, men vist kvantitativt for sammenligning af konkrete tal.

Numerisk Prompt	**GPT-3.5-Turbo**	**GPT-4.1**	**o3**
Temp 1	Old-T1-N	New-T1-N	Strong-N
	Parametre er tættest på tidligere forskning. Resultat er dårlig variansandel og medianafstand, men har fejlrate på nul.	Ny model øger variansandel lidt og sænker medianafstand meget, fejlrate er lidt over nul.	Tænke modellen har meget stræk varians relativt til de andre test, men også meget skæv median afstand. Fejlrate på nul.
Temp 2	Old-T2-N	New-T2-N	-
	Øges temp. ved den gamle model stiger fejlraten katastrofalt. Variansandel og Medianafstand kan derfor ikke sammenlignes.	Øges temp. Ved den nye model øges variansandel, medianafstand og fejlrate.	
Tekst Prompt	**GPT-3.5-Turbo**	**GPT-4.1**	**o3**
Temp 1	Old-T1-T	New-T1-T	Strong-T
	Tekst prompt øger variansandel og medianafstanden en del, fejlrate øges svagt. Har den mest menneskelige svarfordeling	Tekst prompt sænker her variansandelen meget og øger medianafstand. Fejlrate falder lidt.	Tekst prompt får variansandel til at falde og medianafstand til at stige, fejlrate forbliver uændret.
Temp 2	Old-T2-T	New-T2-T	-
	Tekst prompt øger den i forvejen høje fejlrate. Variansandel og Medianafstand stadig ikke anvendelig.	Tekst prompt sænker her variansandelen, men her øger den både medianafstanden og fejlraten.	

Tabel 3: *Resultat Overblik, deskriptivt*. Viser testnavn og projektets resultater ved hvert test, kort opsummeret i tekst. Øverste set af 5 test viser test gjort med den gamle prompt med det numeriske svarformat, nederste 5 er for den nye med tekst-baseret svarformat. Vandret ændres model, og lodret, internt ved hver prompt, ændres temperatur. Ændring i model ændre farvekodningen, mens ændring i temperatur ændre farveintensitet, ingen visualiserings ændring ved ændring af model.

4.1.1. Primære fund

Det generelle fund af samtlige test er at modellerne præstere dårligt i forhold til idealet og forventningerne dannet af tidligere forskning. De nøjagtige årsager til dette kan være svære at afgøre, men anvendelsen af danske data og spørgsmål rettet mod det danske samfund frem for amerikansk data og spørgsmål rettet mod det amerikanske samfund har sandsynligvis været mere udslagsgivende end forventet. Dette forhindrer dog ikke analysen fra at lave en række andre vigtige observationer.

På tværs af tests finder jeg, at de tre uafhængige variabler påvirker hinanden på flere forskellige måder, der gør det svært totalt at adskille effekten af en variabel fra de andre. For eksempel finder jeg at effekten af høj temperatur afhænger kraftigt af om det er GPT-3.5-Turbo eller GPT-4.1.

Trods dette er der dog nogle enkelte mønstre, som viser sig i hver af de tre uafhængige variabler. Generelt så har nyere modeller højere aggregeret variansandel end ældre modeller, men medianafstanden svinger meget fra model til model, og fejlraten er relativt upåvirket af denne variabel. Selve svarfordelingen inden for hvert udsagn er dog typisk mindre menneskelig ved nyere modeller end den gamle GPT-3.5-Turbo. Temperatur øger variansandel, men på bekostning af højere medianafstand og fejlrate, denne effekt er kraftig på den gamle model, og mere behersket på den nye, og kan ikke appliceres på o3. Derudover forbedre det ikke på alle de underliggende facetter af variansproblemet. Endelig viser prompttestene, at den numeriske prompt generelt fungerer bedre i aggregat end den ‘forbedrede’ tekst-baserede prompt. Dette er dog med undtagelse af GPT-3.5-Turbo, som kan danne de mest menneskelige svarfordelinger ud af alle test i dette projekt ved brug af den nye prompt.

Samlet set betyder dette, at de værktøjer og metoder, som er tilgængelige på nuværende tidspunkt, ikke giver nogen klar indikation på hvordan variansproblemet bør løses, eller om det overhovedet kan løses.

4.1.2. Konkrete Resultater

Test Navn	Model	Temp	Prompt	Varians andel	Varians sand	Varians LLM	Median afstand	Median sand	Median LLM	Fejl rate
Old-T1-N	3.5	1	Num	**17.1%**	0.367	0.063	**0.321**	2.792	2.470	**0%**
Old-T2-N	3.5	2	Num	~~85.3%~~	~~0.371~~	~~0.316~~	~~0.218~~	~~2.787~~	~~2.569~~	***61.6%***
New-T1-N	4.1	1	Num	**20.5%**	0.366	0.075	**0.139**	2.791	2.653	**0.7%**
New-T2-N	4.1	2	Num	**24.4%**	0.373	0.091	**0.146**	2.790	2.644	**6.1%**
Strong-N	o3	-	Num	**42.9%**	0.367	0.157	**0.540**	2.792	2.252	**0%**
Old-T1-T	3.5	1	Text	**28.7%**	0.367	0.105	**0.409**	2.792	2.383	**0.3%**
Old-T2-T	3.5	2	Text	~~42.3%~~	~~0.318~~	~~0.134~~	~~0.107~~	~~2.699~~	~~2.592~~	***94.2%***
New-T1-T	4.1	1	Text	**11.4%**	0.366	0.042	**0.202**	2.791	2.590	**0.5%**
New-T2-T	4.1	2	Text	**20.4%**	0.364	0.074	**0.191**	2.793	2.601	**12.2%**
Strong-T	o3	-	Text	**40.4%**	0.367	0.148	**0.656**	2.792	2.136	**0%**

Tabel 4: *Resultat Overblik, Kvantitativt*. Viser resultaterne af de 10 test udført i analysen. Afhængige variabler markeret med fed skriftype, uanvendelige tal med overstregning. Venstre del viser testnavn og uafhængige variabler, højre del viser varians, median og fejlrate. Ved varians og fejlrate er der vist både menneskelig og model resultater. Vær opmærksom på at modelnavnene er forkortet ned til versionsnumre i de tilfælde hvor navnet var for langt til tabellen.

Tabel 4 viser de test, der er foretaget samt hvad deres resultater er. Første søjle viser navnet på testen, som er dannet og grupperet ud fra de afhængige variabler, som udgør de næste tre søjler. Derefter er der tre søjler, der beskriver varians, hvor den første viser variansandel, efterfulgt af den sande menneskelige varians og derefter variansen i LLM besvarelserne. De næste tre søjler viser det tilsvarende for medianen. Den sidste søjle viser så fejlraten for denne test. De vigtigste tal er fremhævet med fed skrifttype, og tal som opstår i forbindelse med katastrofalt høj fejlrate er overstreget og skriften er givet en lysere farve.

4.2. De 10 Test

Dette afsnit består af en systematisk gennemgang af de 10 test, some er udført for at belyse projektets problemformulering. Testende kommer i samme rækkefølge som ved tabellen vist i **4.1.2. konkrete resultater**. Hver test præsenterer de afhængige og uafhængige variabler, samt figurer af den aggregerede model respons versus den aggregerede menneskelige respons, og af modellens svarfordeling versus den menneskelige svarfordeling, hver figur har en medfølgende tolkning. To test er undtaget figurer og videre tolkning, idet deres høje fejlrate gør deres andre resultater uanvendelige. Som nævnt er der yderligere figure over underinddeling af den aggregerede response ved hver test, fordelt på køn, alder og uddannelse, disse kan findes i appendiks 6-19.

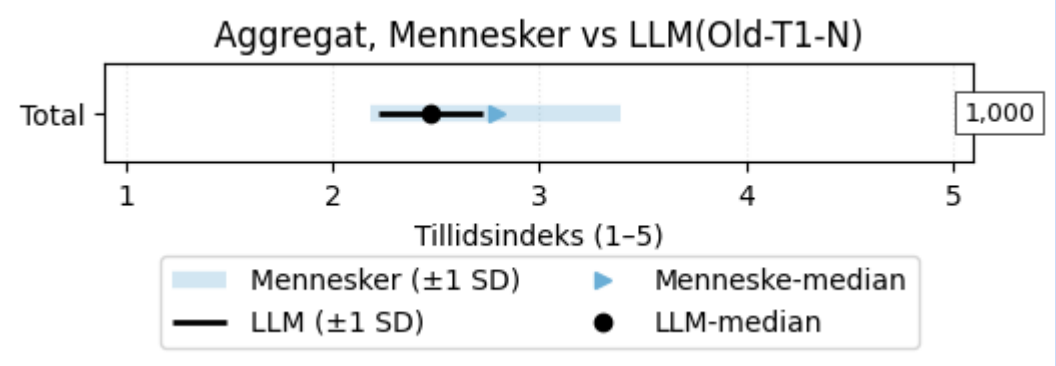
For at holde billedtekst hvor hver figur i det kommende afsnit, er der her en kort forklaring af figurerne.

Aggregatfiguren, se Figur 2 for eksempel, viser den samlede respons for modellen relativt til det menneskelige sample, vist ved +/- 1 standardafvigelse og median. Modellen er sort, med rund medianmarkør, og menneskene er lyseblå med trekantet medianmarkør. X-aksen viser tillidsindeks, hvor 1 er høj tillid og 5 er lav tillid. Y-aksen består kun af det totale sample, med n vist til højre. De underindelte figure som kan findes i appendiks 6-19 følger dette format, blot med de underindelte kategorier op af Y-aksen, frem for totale sample, her viser tallet til højre antallet af respondent/model par i hver undergruppe.

Svarfordelingsfiguren, se Figur 3 for eksempel, viser 9 plots over svarfordelingen for modellen relativt til det menneskelige sample, startende med udsagn 1 i øverst venstre hjørne. Modellen er sort og menneskene i lyseblå. X-aksen viser svarmulighederne, hvor 1 = ‘helt enig’, 2 = ‘delvist enig’, 3 = ‘hverken enig eller uenig’, 4 = ‘delvist uenig’, 5 = ‘helt uenig’. Y-aksen er antallet af respondenter. De stiplede lodrette streger viser medianen for henholdsvis modellen og menneskene ved hvert udsagn.

4.2.1. Old-T1-N

Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
Old-T1-N	GPT-3.5-Turbo	1	Numerisk	**17.1%**	**0.321**	**0%**

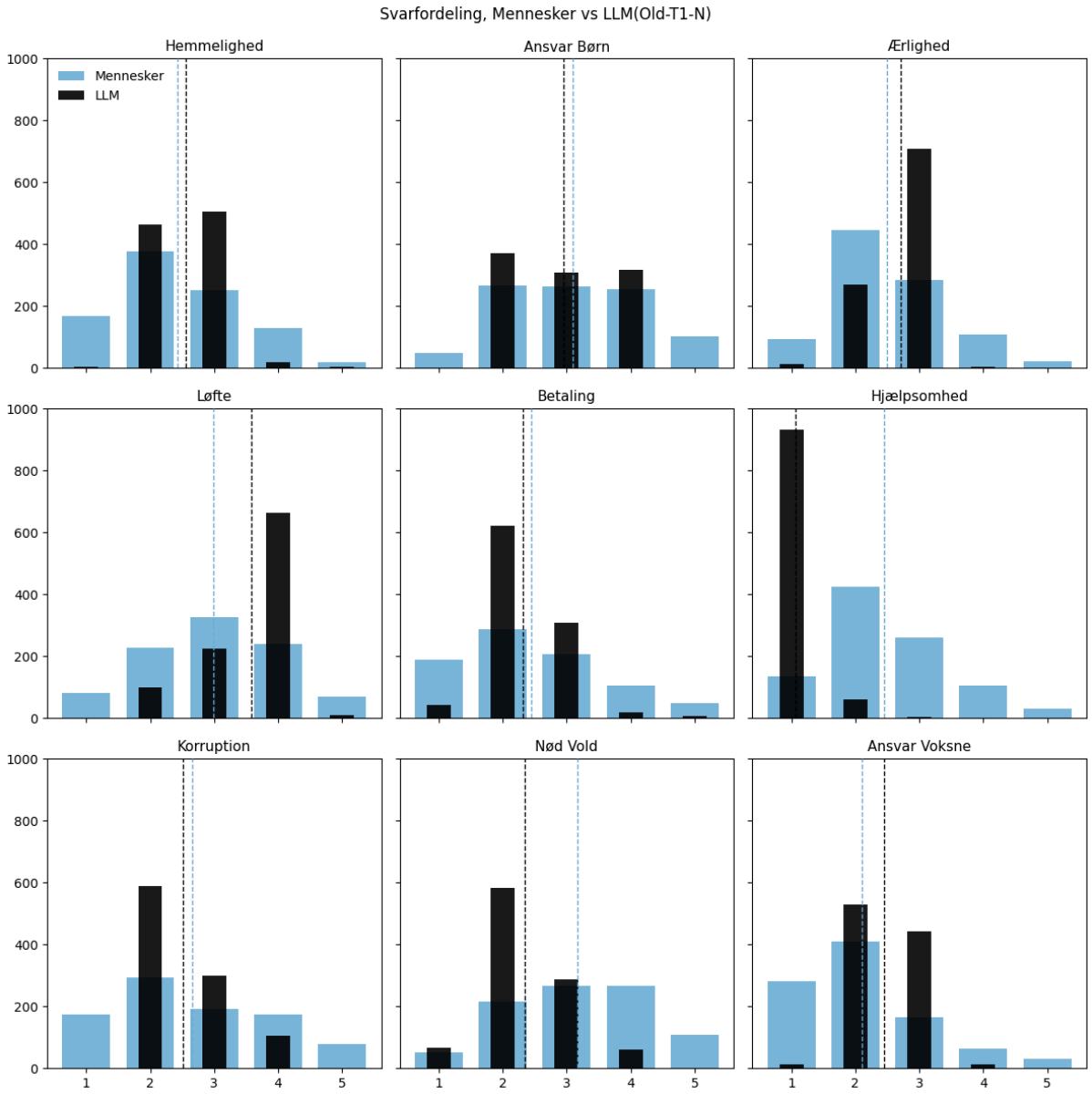


Figur 2: *Aggregat, Mennesker vs LLM(Old-T1-N)*. Viser aggregatfiguren for Old-T1-N testen.

Den første test bestod af at køre GPT-3.5-Turbo ved standard temperatur og ved brug af den gamle prompt, som anvender numerisk svarformat. Modellen fejler ikke ved nogle af de 1000 respondenter, og har derfor en 0% fejlrate.

Efter forventningerne viser denne test en lav variansandel på 17% af den menneskelige. Modsat forventningen har modellen også en relativt høj afstand mellem medianerne. Begge disse mønstre kan ses i visualiseringen ovenfor. Dette betyder at modellen ikke kun undervurderer variansen i tillid inden for samplet, den overvurderer også, hvor tillidsfulde de er.

Et andet interesandt mønster kan ses i visualiseringerne for køn, alder og uddannelse som kan findes i appendiks 6. Det slående her er, at modellens svar ligger omkring det samme område på tillidsindekset, uanset køn, alder og uddannelse. Samtidig kan man se hvordan menneskene i disse underkategorier ligger mere spredt i forhold til hinanden. Dette tyder på at disse demografiske variabler i ringe grad påvirker hvilke svar modellen ender på. Årsagen til at modellens varians er så begrænset kan undersøges nærmere med følgende visualisering over svarfordelingen for modellen og menneskerne inden for hvert spørgsmål:



Figur 3: *Svarfordeling, Mennesker vs LLM(Old-T1-N)*. Viser svarfordelingsfiguren for Old-T1-N testen.

Figur 3 viser at LLM besvarelserne har en tendens til at samle sig omkring de midtersøgende værdier, og med undtagelse af dens svar til udsagn 6 om hjælpsomhed, så er det sjældent den vælger at svare 1 “helt enig” eller 5 “helt uenig”. Dette kan også ses i, at dens svarfordeling har en tendens til at være meget begrænset, den vælger typisk et, enkelte steder to svar for næsten hele samplet, dens svarmønster ved udsagn 2 er den eneste undtagelse, hvor den her er fordelt på tre svarmuligheder, hvilket er tættere på den menneskelige fordeling, relativt til hvordan den har svaret på de andre udsagn.

Figuren viser også, at en stor del af medianafstanden skyldes modellens svar på udsagnet om hjælpsomhed, dog er afstanden mellem medianerne inden for de

individuelle udsagn heller ikke jævnt god. Ved udsagn 4 undervurderer den samlet set samplets tillid, og ved udsagn 8 overvurderer den tillid. Trods de menneskelige respondenter også sjældent svarer 1 og 5 ved udsagnene, er det svar modellen mest hyppigt giver inden for et udsagn kun det samme som det menneskelige ved udsagn 2, 5, 7 og 9, altså fanger den i ringe grad det mest typiske svar inden for hvert spørgsmål.

Samlet set viser denne test at den gamle GPT-3.5-Turbo model, ved standard temperatur og den numerisk baserede prompt, har en lav variansandel som skyldes en blanding af manglende ekstremværdier og en tendens til kun at vælge en eller to besvarelser ved hvert udsagn. Den skæve medianafstand skyldes en blanding af et ekstremt skævt svar til udsagn 6, og en manglende evne til at gengive det mest typiske svar inden for det menneskelige sample.

4.2.2. Old-T2-N

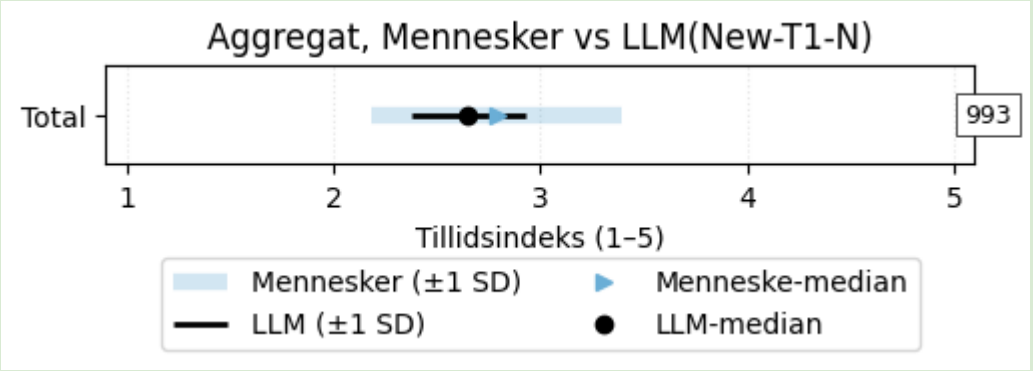
Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
Old-T2-N	GPT-3.5-Turbo	2	Numerisk	~~***85.3%***~~	~~***0.218***~~	***61.6%***

Den næste test består af at skrue op for temperaturindstillingen, men bibeholder model og prompt. Umiddelbart viser den en øget variansandel og også en forbedret medianafstand, dog kan dette hænge sammen med den ekstremt høje fejlrate i modellens besvarelser. Denne høje fejlrate reducere sammenlignligheden af medianafstand og variansandel, givet det er uklart om et mønster ligger gemt i hvilke respondenter modellen fejlede ved. Disse respondenter kunne testes igen indtil modellen har givet anvendelige outputs til alle respondenter, men i min optik ville dette dække over den højre fejlrate som i sig selv er et resultater, i det den viser omkostningen og udfordringerne af den øgede temperatur.

Visualiseringerne for denne test kan ses i appendiks 7, men udelades fra analyseafsnittet af hensyn til længde, layout og visuel støj.

4.2.3. New-T1-N

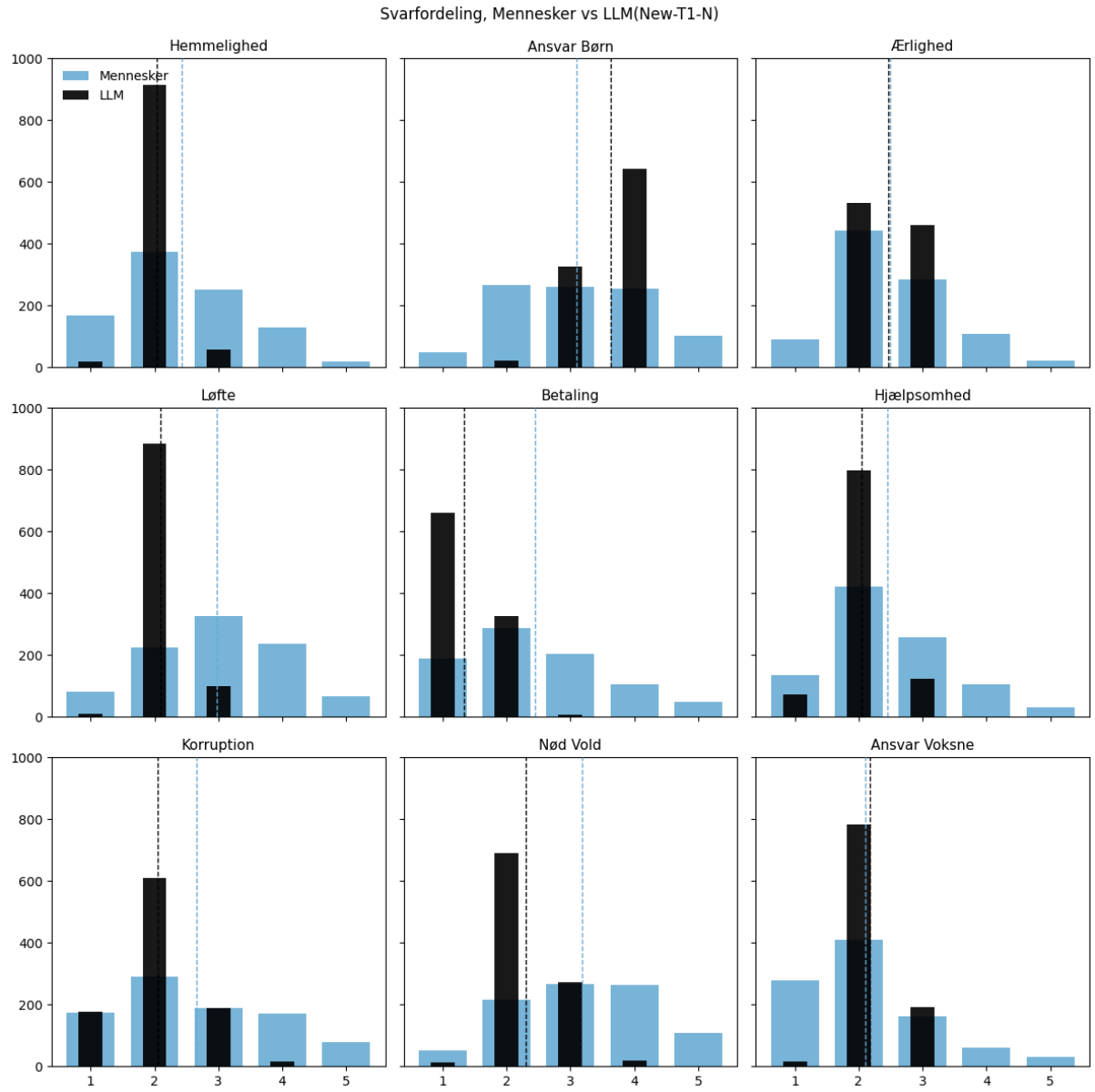
Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
New-T1-N	GPT-4.1	1	Numerisk	**20.5%**	**0.139**	**0.7%**



Figur 4: *Aggregat, Mennesker vs LLM(New-T1-N)*. Viser aggregatfiguren for New-T1-N testen.

Den næste test består af den nyere GPT-4.1 model, ved standard temperaturindstilling og prompten med det numeriske svarformat. Ligesom den første test som anvendte standard temperatur har denne model en lav fejlrate på 0.7%, hvilket ikke forventes at have nogen effekt på anvendeligheden af modellens resultater eller brug.

Denne nyere model har en forbedret medianafstand på 0.139, hvilket er det næst laveste ud af alle 10 test, kun overgået af test Old-T2-T, som til gengæld har den højeste fejlrate. New-T1-N testen har også en højere variansandel end den forrige lav temperatur test, dog er den stadig lav ved kun 20.5% af det menneskelige. Det er derfor interessant at se om denne forbedring i variansandel stammer fra at den nyere model har løst nogle af de problemer som blev beskrevet i forbindelse med test Old-T1-N.



Figur 5: *Svarfordeling, Mennesker vs LLM(New-T1-N)*. Viser svarfordelingsfiguren for New-T1-N testen.

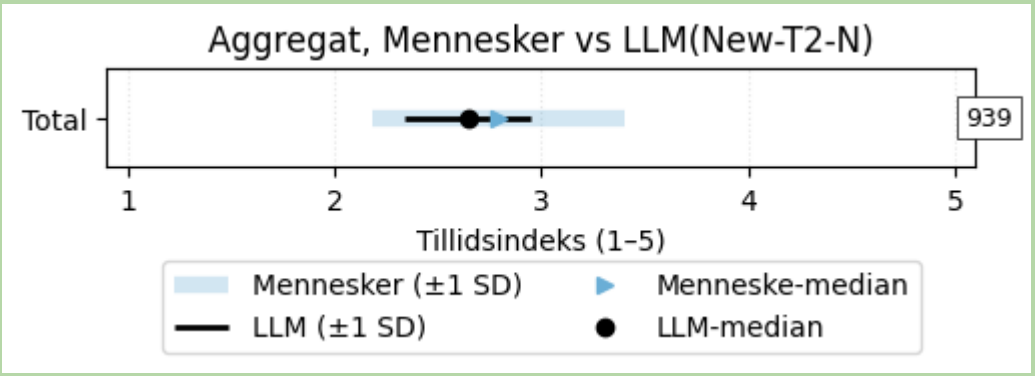
Som vist i denne visualisering er flere af de gamle mønstre stadig gældende. Det er stadig sjældent, den vælger ekstremværdierne og dens svarfordeling er faktisk blevet mere begrænset relativt til den første test, trods den totale variansandel er svagt forbedret. Der er enkelte steder hvor modellens svar inden for et spørgsmål tæt matcher antallet af menneskelige besvarelser, for eksempel er antallet af menneskelige og LLM respondenter der har svaret 1 og 3 ved udsagn 7 om korruption tæt på identisk. Mere end tidligere virker det også til, at modellen undervurderer antallet af negative respondenter, med undtagelse af udsagn 2 hvor den overvurderer det i stedet.

De fejl ved test Old-T1-N som medførte en relativt høj medianafstan er også stadig gældende, dog i en svagt anden form. Modellens svar til udsagn 6 er mindre ekstremt, men udsagn 5 om betaling er derimod skæv, hvor den før var en af modellens bedre besvarelser, dog er dens skævhed ikke lige så ekstrem som ved den forrige tests svarfordeling ved udsagn 6. Den nyere models hyppigste svar matcher det menneskelige i 5 ud af 9 udsagn, en forbedring på en.

Dette tyder på at de forbedringer der er vundet ved at anvende GPT-4.1 frem for GPT-3.5-Turbo, i konteksten af dette projekt, ligger i aggregatet, og at modellens reelle besvarelser til de enkelte udsagn potentielt er endnu skævere ved den nyere model frem for den gamle.

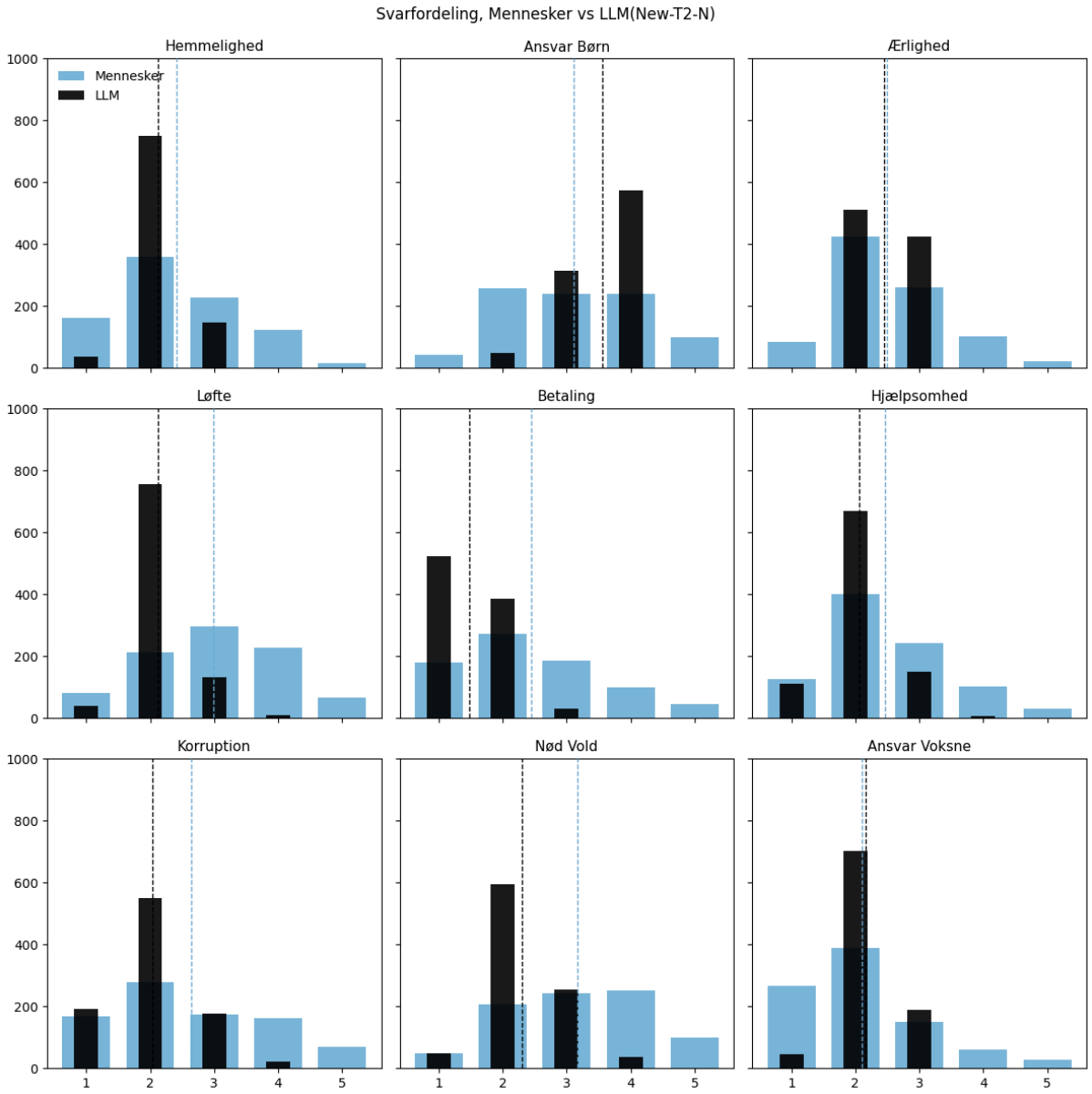
4.2.4. New-T2-N

Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
New-T2-N	GPT-4.1	2	Numerisk	**24.4%**	**0.146**	**6.1%**



Figur 6: *Aggregat, Mennesker vs LLM(New-T2-N)*. Viser aggregatfiguren for New-T2-N testen.

Den næste test består af at øge temperaturindstillingen for GPT-4.1 modellen, stadig ved den samme prompt. Variansandelen stiger til 24.4%, en forbedring relativt til New-T1-N testen på omkring 19%. Medianafstanden er også øget, dog kun med omkring 5% til 0.146. Tilsvarende den forrige test med øget temperatur ser vi igen en øget fejlrate som nu er på 6.1%, dette forventes dog ikke at invalidere sammenligneligheden af denne tests resultater.



Figur 7: *Svarfordeling, Mennesker vs LLM(New-T2-N)*. Viser svarfordelingsfiguren for New-T2-N testen.

Svarfordelingen minder her meget om den forrige test med GPT-4.1 modellen. Den har stadig samme tendens som de forrige til kun at vælge et eller to svarmuligheder til hvert udsagn på tværs af samplet, den undervurderer de ekstreme og de negative besvarelser, med undtagelse af udsagn 5 og 7 hvor den overvurderer antallet af ekstreme besvarelser, og ved udsagn 2 overvurdere den stadig antallet af negative besvarelser.

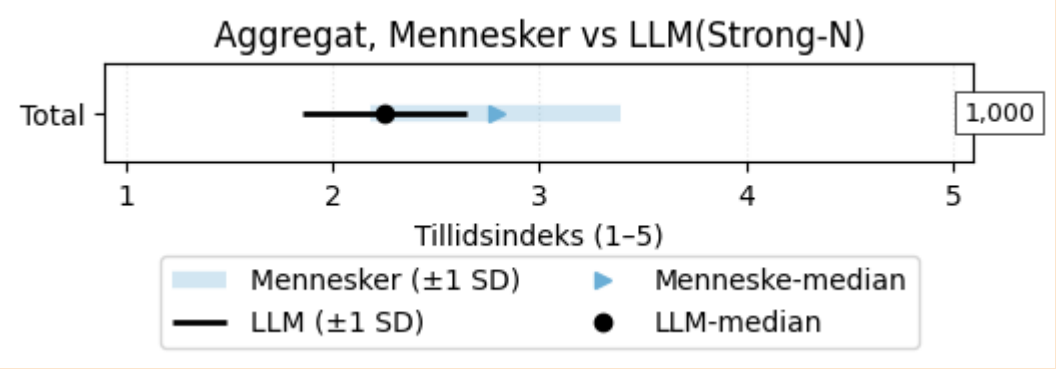
Den reelle forbedring ved denne test relativt til den forrige, New-T1-N, er at den høje temperatur får modellen til ikke at vælge det samme svar inden for hver test lige så ofte. Ved den lavere temperatur var modellens hyppigste svar inden for

hvert udsagn over 600 ud af 1000 i 8 ud af 9 tilfælde. Ved høj temperatur er det i 5 ud af de 9 tilfælde, og den kommer aldrig over 800. Disse respondenter er derimod fordelt på de andre besvarelser, dog ikke i et omfang der reelt opbløder modellens svarfordelingsmønster.

Samlet set har temperatur hjulpet lidt på variansandelen, men dens svarfordeling indeholder stadig de samme fejl som den oprindelige test.

4.2.5. Strong-N

Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
Strong-N	o3	-	Numerisk	**42.9%**	**0.540**	**0%**

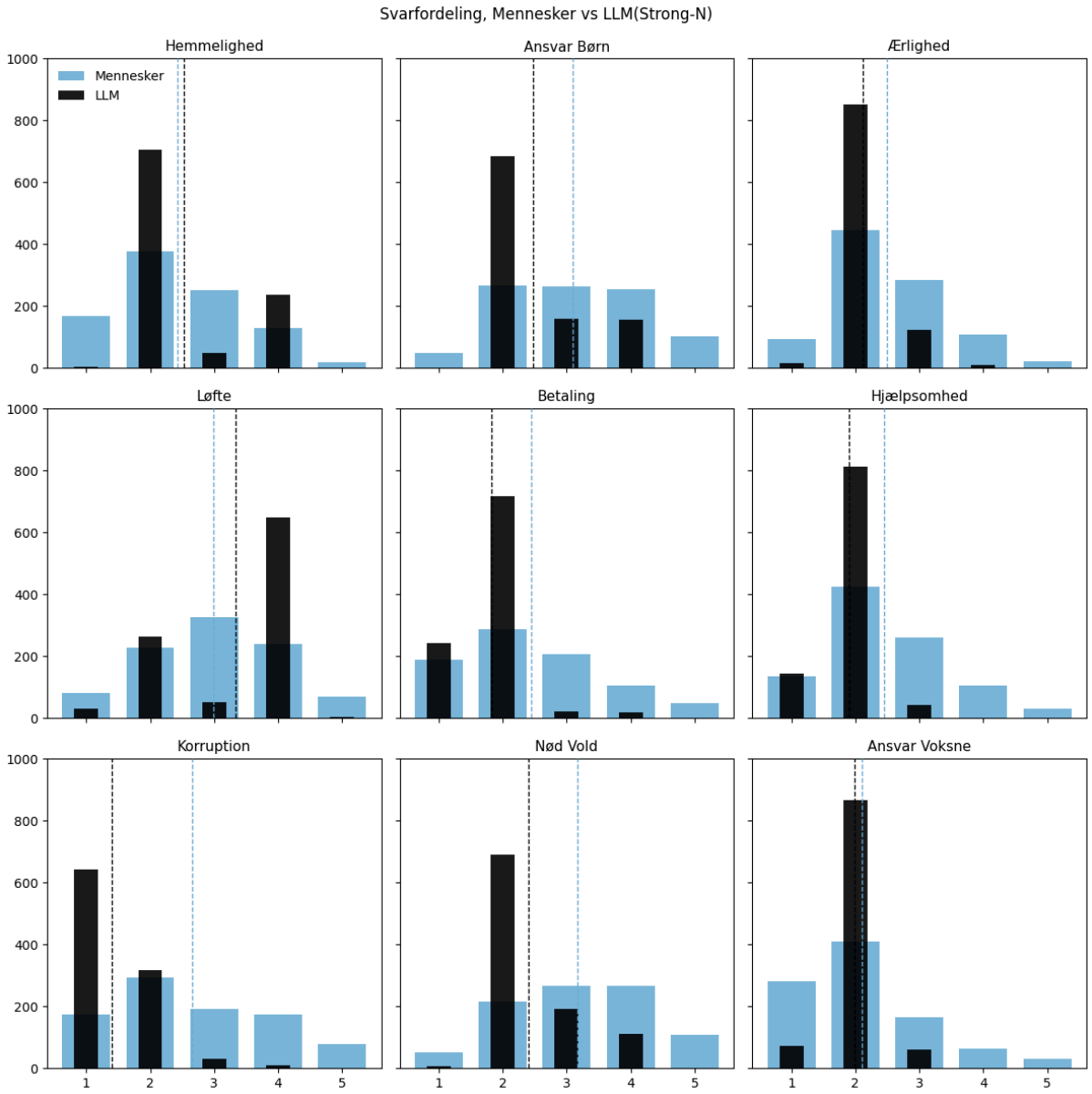


Figur 8: *Aggregat, Mennesker vs LLM(Strong-N)*. Viser aggregatfiguren for Strong-N testen.

Den næste test anvender o3 modellen, og er den sidste test som anvender prompten med det numeriske svarformat. I modsætning til samtlige forrige tests kan temperaturindstillingen af denne model ikke ændres. o3 modellen formår at give valide output for alle 1000 respondenter, og den har derudover den højeste varians andel på 42.9%, kun overgået af Old-T2-N testen, som i modsætning havde en høj fejlrate.

Unikt ved denne test er dog dens høje medianafstand på 0.54, som overgår alle forrige test, hvor af den næsthøjeste er den oprindelige Old-T1-N test på 0.321. Denne høje variansandel og medianafstand kan ses i den ovenstående visualisering af aggregat værdierne. Jeg vil igen her også gøre opmærksom på visualiseringerne opdelt på køn, alder og uddannelse, som kan findes i appendiks 10. Her kan man se at modellen i højere grad en tidligere vælger forskellige svar inden for hver undergruppe, hvilket er et positivt tegn i den forstand at det tyder på at modellen i højere grad anvender de demografiske data.

Umiddelbart virker det til at o3 modellen håndterer denne opgave meget anderledes end de forrige modeller, så det er interessant at se, om svarfordelingen kan give nogen forklaring på nøjagtigt hvordan og måske hvorfor:



Figur 9: *Svarfordeling, Mennesker vs LLM(Strong-N)*. Viser svarfordelingsfiguren for Strong-N testen.

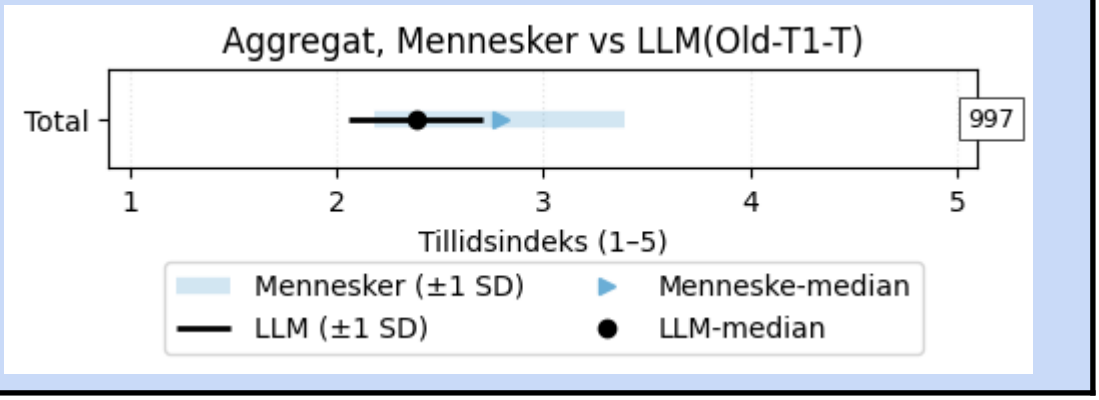
Det virker umiddelbart til at o3 modellen har meget de samme fejl som de forrige. Der er et eller to svarmuligheder den vælger for næsten hele samplet. Med enkelte undtagelser undervurdere den antallet af ekstreme og negative besvarelser. En forskel mellem denne model og de forrige er, at den generelt afholder sig fra at svare 3, altså 'hverken enig eller uenig'. Til gengæld er den svagt mere villig til at svare 4 og 1. Dette giver modellens svarfordelinger 'dale' omkring svarmulighed 3, hvilket er meget fjernt fra den menneskelige svarfordeling, og svarfordelingen i de andre test hvor modellens næsthyppigste svar trods alt var ved siden af det hyppigste selvom forskellen ofte var stor. Dette virker dog stadig til samlet set at forbedre variansen i

modellens svar, dog ikke i retning af noget menneskelignende der løser variansproblemet.

Denne model har dog en forbedring, nemlig at dens mest hyppige besvarelse passer til den mest hyppige menneskelige besvarelse ved 6 ud af 9 udsagn.

4.2.6. Old-T1-T

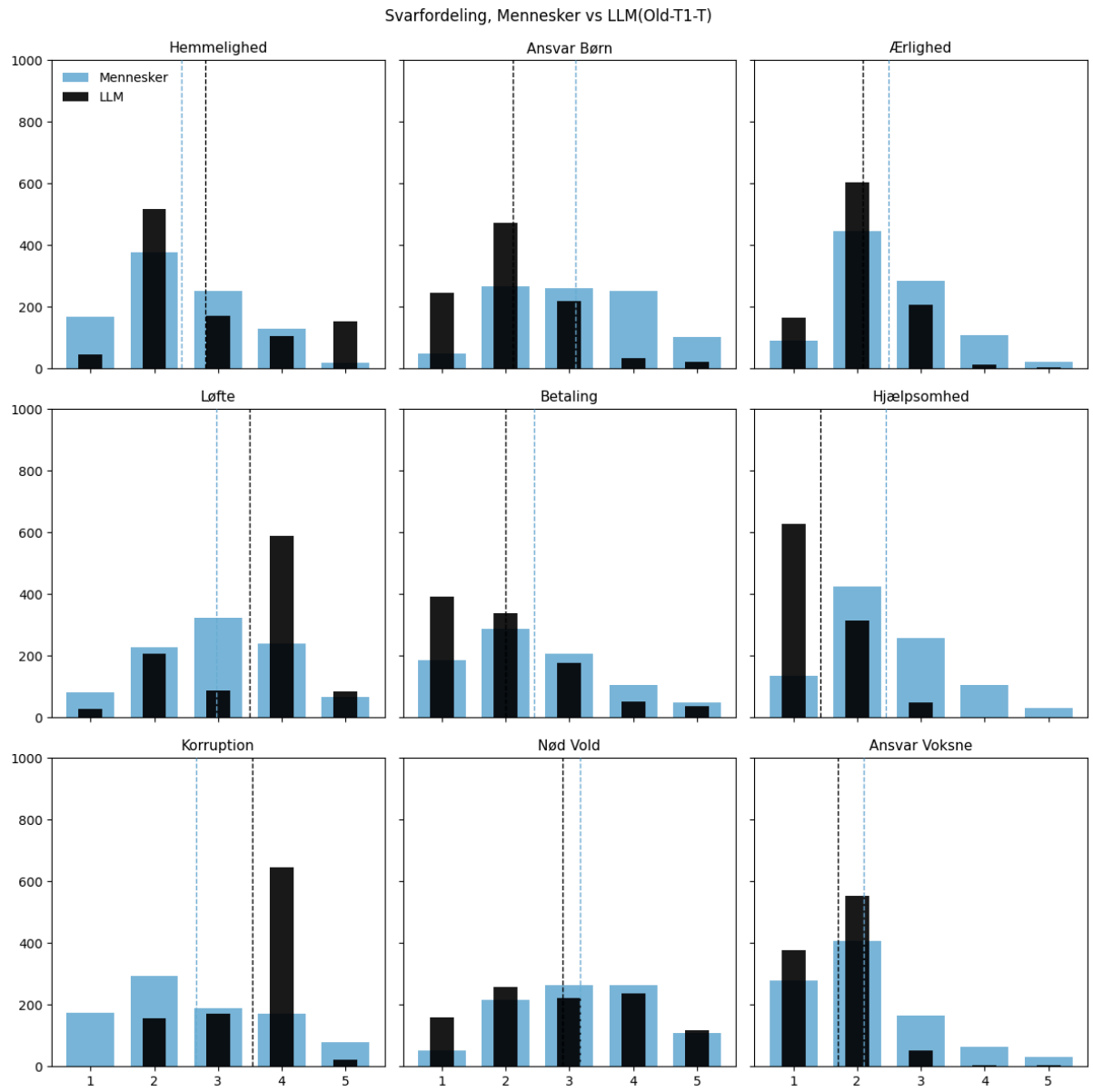
Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
Old-T1-T	GPT-3.5-Turbo	1	Tekst	**28.7%**	**0.409**	**0.3%**



Figur 10: *Aggregat, Mennesker vs LLM(Old-T1-T)*. Viser aggregatfiguren for Old-T1-T testen.

Dette er den første test som anvender prompten med det tekst-baserede svarformat, inklusionen af kommunevariablen og rettelsen af børn variablen. Som ved den første test i analyseafsnittet er den udført med den gamle GPT-3.5-turbo model, ved temperaturindstilling 1. I modsætning til Old-T1-N har denne test en fejlrate på 0.3%.

Variansandelen er her steget til 28.7%, hvilket betyder den nye prompt har her medført en relativ forbedring på 67.8%, set i aggregat. Medianafstanden ved denne test er på 0.409, hvilket derimod er en forværring på 27.4% i forhold til den tilsvarende test med numerisk prompt. Den nye prompts effekt på variansandelen kan alligevel virke lovende, men hvis figur 10 ovenfor, som viser de samlede svar for denne test og mennesker, sammenholdes med figur 2 ved Old-T1-N testen, kan man se, at den øgede varians opstår fordi modellen i højere grad har være svaret at den er tillidsfuld. Dette var i forvejen den retning modellen overvurderede, så det medfører derfor også den øgede medianafstand. Hvis figurene for køn, alder og uddannelse også sammenholdes, kan det ses at den nye prompt virker til at have gjort modellen mere følsom over for de demografiske variabler, idet der er større variation mellem grupperne i Old-T1-T end i Old-T1-N testen. Som før er det interessant at se hvordan den nye prompt har påvirket modellens svarfordeling:



Figur 11: *Svarfordeling, Mennesker vs LLM(Old-T1-T)*. Viser svarfordelingsfiguren for Old-T1-T testen.

Her er der en række interessante fund. Man kan se, at modellen er mindre tilbøjelig til at vælge et eller to svarmuligheder inden for et udsagn for hele samplet, og enkelte steder har den en relativt jævn fordeling, ikke kun på de moderate svar, men også på de ekstreme, næsten unikt i forhold til samtlige test. Indtil videre er denne model villig til at svare, at folk ikke har meget lidt tillid, særligt i forbindelse med udsagn 1, 4 og 8 hvor den faktisk overvurderer den menneskelige svarfordeling. Udsagn 1, 5 og 8 er her også interessant fordi modellen faktisk har formået at opfange en relativt menneskelignende fordeling. Jeg minder i denne forbindelse om, at svarmulighed 6, 'ved ikke' ikke har været et gyldigt svar for modellen, hvilket er

med til at forklare at modellens samlede svar ikke kan passe perfekt på det menneskelige.

Figur 11 her viser også at noget af grunden til at denne test har en højere variansandel men også højere medianafstand er fordi denne test konsekvent undervurderer antallet af respondenter som svarer 3, hvorimod Old-T1-N testen typisk overvurderede det.

4.2.7. Old-T2-T

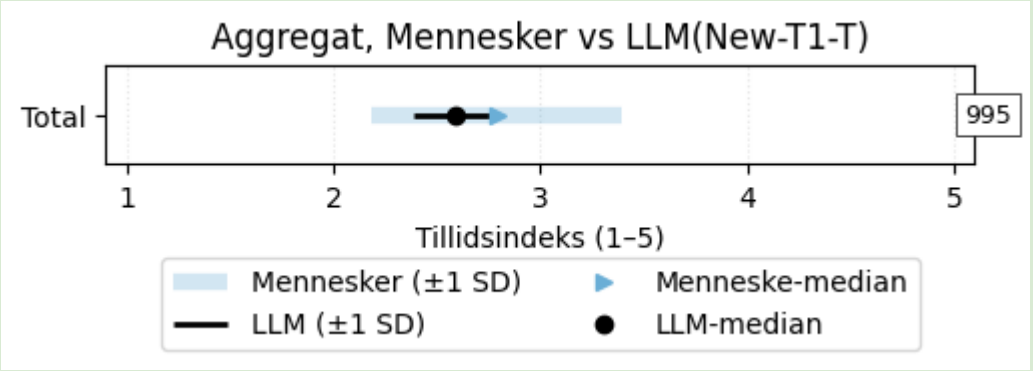
Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
Old-T2-T	3.5	2	Tekst	~~***42.3%***~~	~~*0.107*~~	***94.2%***

Denne test er den sidste som anvender GPT-3.5-Turbo, den er udført ved temperaturindstilling 2 og med den tekst-baserede prompt. Som ved den anden høj-temperatur test af den gamle model har denne også lidt et stort tab af anvendelige resultater, men her er fejlraten oppe på 94.2%, hvilket gør de resterende data fra denne test uanvendelige.

Det umiddelbare resultat fra denne test er derfor at noget ved den nye prompt, muligvis det tekst-baserede svarformat, forstyrrer modellens evne til at give et anvendeligt svar, og at denne svaghed forværres ved høj temperatur.

4.2.8. New-T1-T

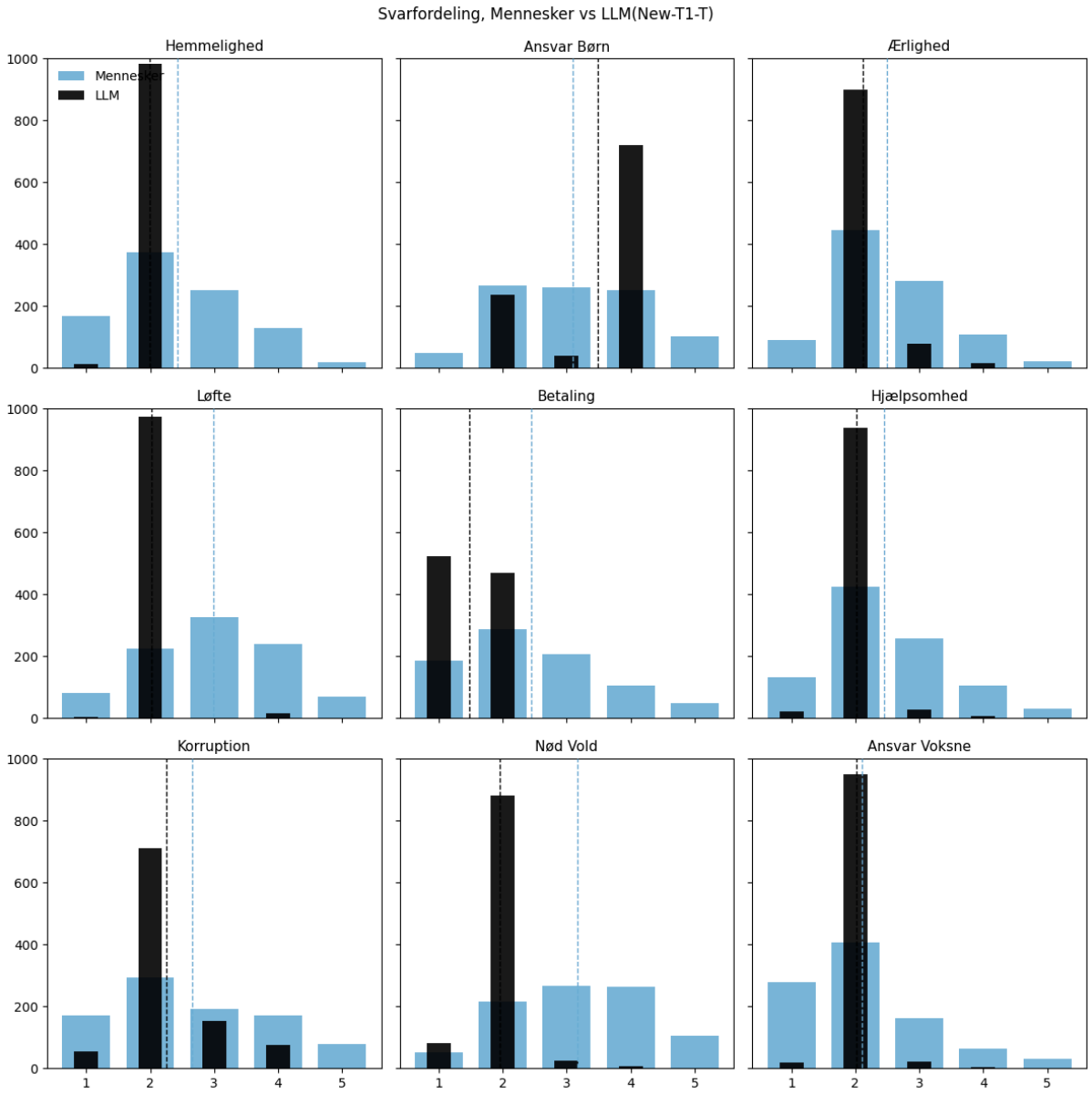
Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
New-T1-T	GPT-4.1	1	Tekst	**11.4%**	**0.202**	**0.5%**



Figur 12: *Aggregat, Mennesker vs LLM(New-T1-T)*. Viser aggregatfiguren for New-T1-T testen.

Denne test anvender GPT-4.1 modellen, ved temperaturindstillingen på 1, men med den nye tekst-baserede prompt. Som ved den tilsvarende test med numerisk prompt er fejlraten under en procent, dog lavere end ved den anden test.

Mest slående ved denne test er den lave variansandel på kun 11.4%, den laveste på tværs af alle test i dette projekt. Samtidigt er medianafstanden også steget relativt til den tilsvarende test med den numeriske prompt. Årsagen til dette kan findes i svarfordelingen:

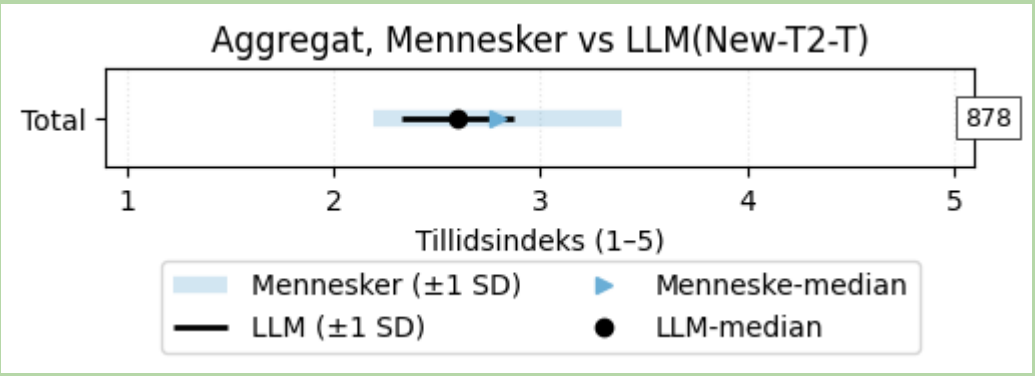


Figur 13: *Svarfordeling, Mennesker vs LLM(New-T1-T)*. Viser svarfordelingsfiguren for New-T1-T testen.

Ved 7 ud af 9 udsagn er 'delvist enig' det svar, modellen har givet for næsten alle respondenterne, og selv ved de udsagn hvor modellen har valgt flere af et andet svar er 'delvist enig' det næst hyppigste. Så udover at modellen her har en ekstremt ujævn svarfordeling, er den ujævn på samme måde på tværs af de fleste udsagn. Den laver altså den samme type fejl som de forrige modeller, men i et endnu større omfang. I modsætning til Old-T1-T testen, som viste at den nye prompt fik modellen til at ændre sine svarmønstre, virker det her til at den nye prompt har forstærket modellens svarmønstre. Dette kan ses i, at de hyppigste svar inden for hvert udsagn er de samme her som ved New-T1-N, og med undtagelse af udsagn 5, er de steget.

4.2.9. New-T2-T

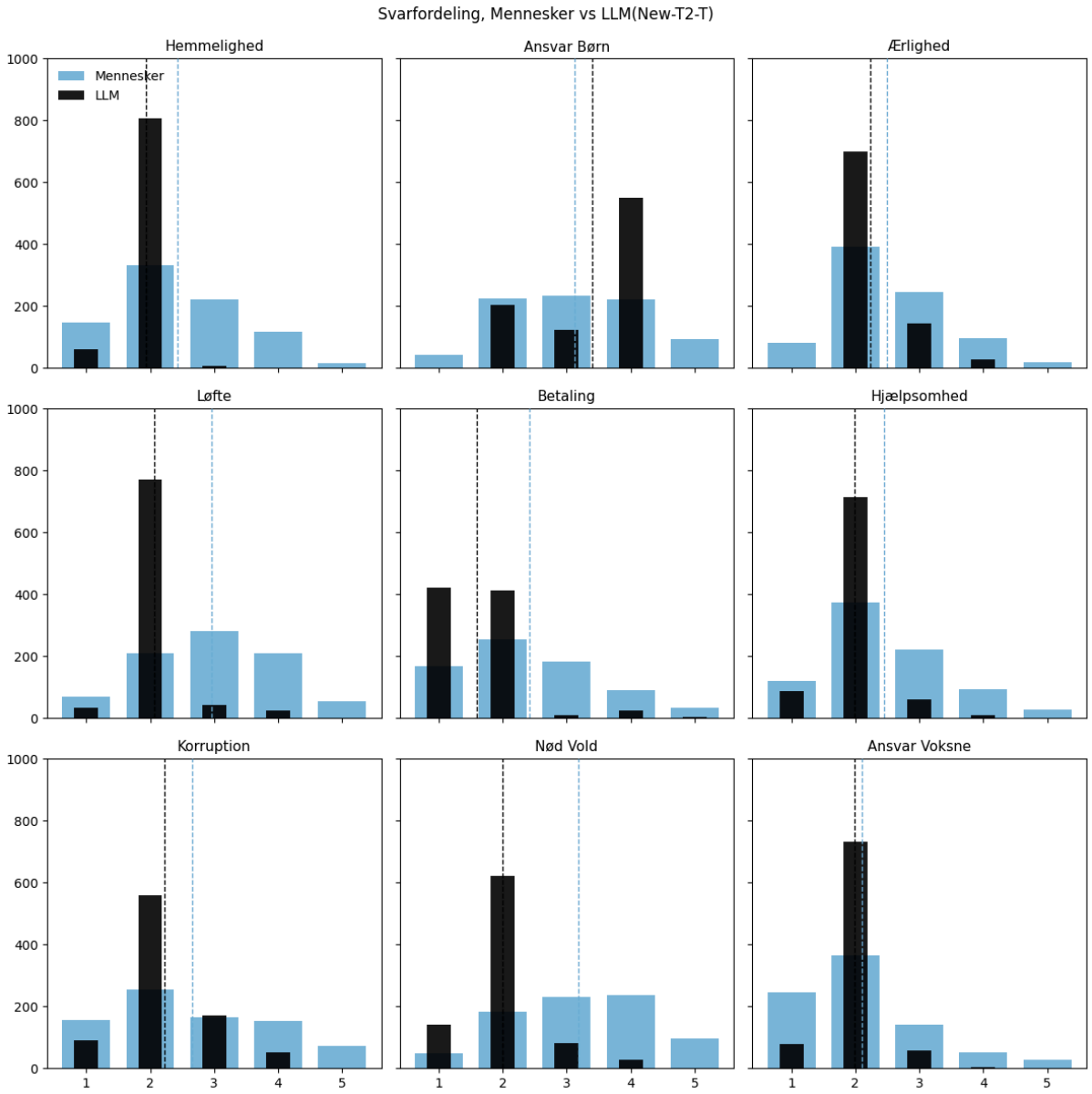
Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
New-T2-T	GPT-4.1	2	Tekst	**20.4%**	**0.191**	**12.2%**



Figur 14: *Aggregat, Mennesker vs LLM(New-T2-T)*. Viser aggregatfiguren for New-T2-T testen.

Denne test anvender også GPT-4.1 modellen og den tekst-baserede prompt, men med en temperaturindstilling på 2. Ved denne test er der en fejlrate på 12.2%, en fordobling i forhold til den tilsvarende test med den numeriske prompt. Også i forhold til New-T2-N testen er der her en forringelse af både variationsandel og medianafstand. Henholdsvis er variationsandel på 20.4%, en forringelse på 16.3% og medianafstand her på 0.191, en 30.8% forringelse. Denne test er dog en forbedring på New-T1-T testen, særligt i forhold til variansandel som er steget med 78.9%. Forbedringen på medianafstand er dog minimal.

New-T2-N testen tyder på, at de fejl som opstår i GPT-4.1 modellen grundet den nye prompt, kan opblødes lidt med temperatur, men ikke i et omfang der reelt set løser det. Det samme kan ses i svarfordelingsfiguren for denne test:

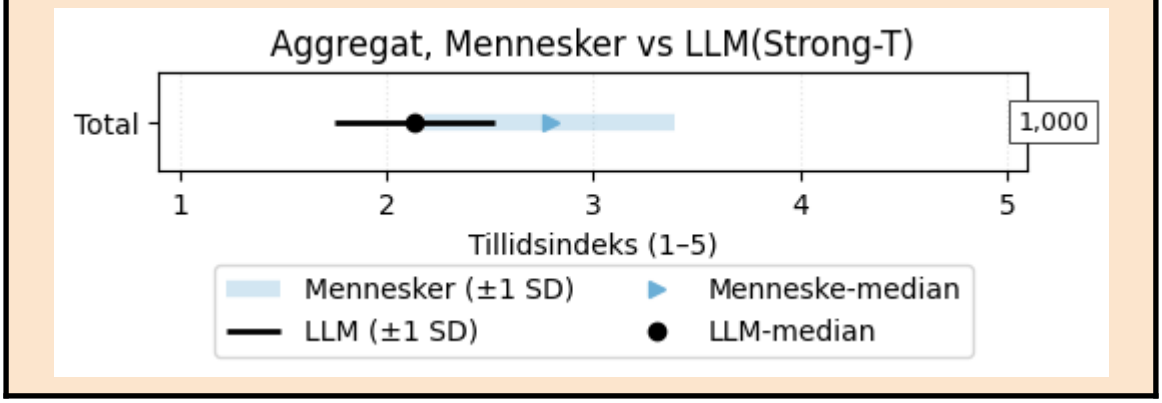


Figur 15: *Svarfordeling, Mennesker vs LLM(New-T2-T)*. Viser svarfordelingsfiguren for New-T2-T testen.

Forbedringen af den øgede temperatur er altså relativt begrænset, og er i et vist omfang også sløret af den øgede fejlrate.

4.2.10. Strong-T

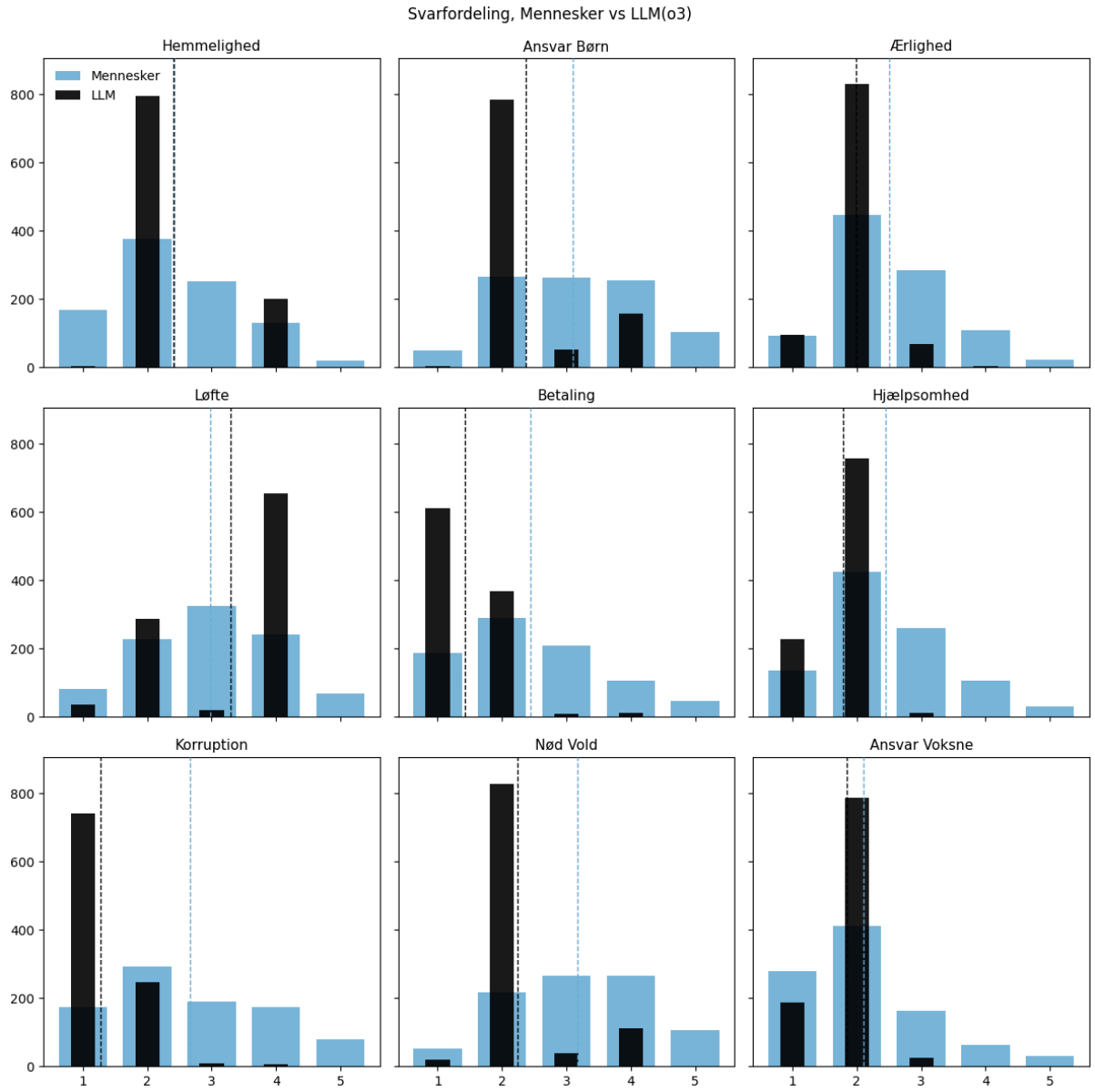
Test Navn	Model	Temp	Prompt	Variansandel	Medianafstand	Fejlrate
Strong-T	o3	-	Tekst	**40.4%**	**0.656**	**0%**



Figur 16: *Aggregat, Mennesker vs LLM(Strong-T)*. Viser aggregatfiguren for Strong-T testen.

Denne test anvender igen o3 modellen, som ikke opererer med temperaturindstillinger, her med den tekst-baserede prompt. Den har igen en fejlrate på 0%, så resultaterne her skulle være stærkt sammenlignelige med Strong-N testen, og i forhold til den test er der her en forværring af henholdsvis variansandelen med et fald på 5.8% og medianafstand med en stigning på 21.4%.

Denne test holder dog det samme mønster som den forrige o3 test, med en relativt høj variansandel, men også en høj negativ medianafstand. Hvis der ses på de opdelte samplefigurer, som kan findes i appendiks 15, så kan man igen se, at o3 modellen virker til i højere grad at anvende de demografiske variabler, men ikke på en måde der hjælper den med at ramme den menneskelige median. Her har effekten af prompten altså været relativt begrænset.



Figur 17: *Svarfordeling, Mennesker vs LLM(Strong-T)*. Viser svarfordelingsfiguren for Strong-T testen.

Svarfordelingen her er igen meget sammenlignelig med den tilsvarende test med den forrige prompt. Modellen laver fortsat den samme type fejl ved at vælge en eller to svar for næsten hele samplet, den undertrykker svarmulighed 3, overvurderer positive og moderate responser og undervurderer de negative og ekstreme.

4.3. Opsummering

Dette afsnit sammenfatter kort resultaterne af de 10 test, struktureret efter de tre uafhængige variabler, startende med effekten af modeller.

4.3.1. Modelvinklen

Effekten af nyere modeller på variansen er ikke entydigt, og afhænger i høj grad af andre parametre, som temperatur og prompt. Hvis der fokuseres på aggregeret variansandel, så forbedres den generelt ved nyere modeller i forhold til ældre. GPT-3.5-Turbo har relativt høj medianafstand, men variansen i dens svarfordeling kan i høj grad afhænge af prompten. Derudover er denne model kraftigt påvirket af temperatur, hvor høj temperatur medføre katastrofalt høje fejlrater. GPT-4.1 giver en forbedring på medianafstand, men om det var en forbedring på variansproblemet afhænger af både prompt og temperatur. Endeligt divergere o3 modellen ved at både medianafstanden og variansen begge stiger.

Hvis der fokuseres på fordelingen af svar internt til hvert udsagn, så har de nyere modeller en tendens til at undgå både ekstreme og neutrale svar, hvilket giver deres svarfordelinger en mere ujævn karakter, hvorimod den ældre model, som også generelt ekskluderede ekstreme værdier, var mere tilbøjelig til at danne relativt bløde fordelinger med mindre spring fra en svarmulighed til den anden ved både at inkludere neutrale svar, men også flere forskellig svar med den tekst-baserede prompt.

Hvis der kigges på de underindelte figurer, så virker det til at nyere modeller i højere grad anvender de demografiske variabler, der er oplyst i prompten, men ikke på en måde som hjælper dem med at danne en menneskelige svarfordeling.

Helt grundlæggende, så nej, nyere modeller løser ikke variansproblemet, og selv stærke modeller kan ikke danne en menneskelig svarfordeling i hverken aggregat eller inden for de enkelte udsagn. De nyere modeller er generelt mere robuste og virker til at være bedre til at anvende de data der er i prompten, men de er også mindre fleksible på den måde, at temperatur og prompt i mindre grad påvirker dem, relativt til den ældre model.

4.3.2. Temperaturvinklen

Effekten af temperatur er mere klar, men støder ind i flere problemer. Ved GPT-3.5-Turbo medfører øget temperatur at modellen i ringe grad kan udføre den opgave den er stillet, særligt ved den skriftlige prompt, og ved GPT-4.1 er effekten af øget temperatur på variansandel relativt svag, med undtagelse af den skriftlige prompt. Endelig divergerer o3 modellen igen ved at den ikke acceptere temperatur som input parameter, muligvis fordi den er en thinking model.

Øget temperatur virker også kun til at løse dele af det, som tilsammen udgør variansproblemet. Typisk reducerer den øgede temperatur mængden af gange modellen vælger den samme svarmulighed inden for et udsagn, men typisk skubber det også aggregat medianen længere væk fra den menneskelige.

Helt grundlæggende, så ja: Høj temperatur øger modellens varians, dog ikke i et omfang der reelt løser variansproblemet, og kan potentielt medføre problemer med medianafstand og særligt fejlraten. Derudover er effekten af temperatur mindre i nye modeller frem for gamle, og de mest avancerede modeller kan ikke anvende det.

4.3.3. Promptvinklen

Effekten af at skifte til den ‘forbedrede’ prompt med det tekst-baserede svarformat er mere uklar. I aggregat virker den nye prompt til generelt at være en forværring, idet den medfører lavere variansandel samt højere medianafstand og fejlrater, særligt blandt de nyere modeller og ved højere temperatur. Den ældste model, GPT-3.5-Turbo, ændrede sin svarfordeling grundet den nye prompt, i modsætning til de nyere modeller, hvor de underliggende mønstre generelt var de samme. Dette kunne tyde på at måden hvorpå modellen skal svare ifølge prompten, eller de informationer den har til rådighed, påvirker dens evne til at variere svarene.

Nøjagtigt hvad det er ved den tekstbaserede prompt versus den numeriske der medførte at den gamle model kunne danne en, relativt til de andre test, mere menneskelig varians og svarfordeling er svært at sige, uden en mere omfattende og grundig undersøgelse. Samtidigt viser resultaterne også, at den gamle model med den tekstbaserede prompt stadig i ringe grad kunne opfange den menneskelige varians ved alle udsagn, og var stadig tilbøjelig til at overvurdere enkelte svarmuligheder, og undervurdere negative responser. Derudover er der ikke nogen klar indikation i underindelte data af, at modellen forstår, at der er forskel mellem

disse grupper. Derudover medførte den 'forbedrede' tekstbaserede prompt, som nævnt tidligere, en forværring blandt de nyere modeller.
Dette er generelt mod forventningen af de ændringer, som blev udført fra den numeriske til den tekst-baserede prompt, og med undtagelse af stigningen i fejlrate i forbindelse med øget temperatur som kan skyldes det øgede sproglige udfaldsrum i tekst frem for tal, så opfatter jeg stadig logikken bag ved ændringerne som saglig.

Helt grundlæggende, så ja, prompt har en effekt på variansen, men uden videre forskning inden for promptingteknikker eller en omfattende trial-and-error process, ser jeg ikke prompting som en potentiel løsning af variansproblemet, givet den tilgang dette projekt har anvendt.

4.4. Validering

Det er værd at dobbelttjekke om disse mønstre i effekterne af model, temperatur og prompt gør sig gældende ved andre respondenter end de 1000 som er anvendt til de ovenstående test. Grundet praktiske begrænsninger er det ikke min intention at gennemføre samtlige test på ny med 1000 nye respondenter.
Udvælgelsesprocessen for valideringstest har været som følger:
o3 modellerne er med flere længder de dyreste at køre, så de test bliver ikke gentaget. Grundet de høje fejlrater er det ikke min intention at gentage test af GPT-3.5-Turbo modellen ved høj temperatur. De test, som anvendte den tekst-baserede prompt, havde tendens til højere fejlrater, og oplevede kun forbedringer i forhold til varians andel på den ældste model ved lav temperatur, så med undtagelse af den bliver de heller ikke gentaget.

Dette efterlader fire test som jeg opfatter kan være værd gentage, disse er Old-T1-N, New-T1-N, New-T2-N, og Old-T1-T. Med disse test kan jeg delvist validere de observerede mønstre i forbindelse med model og temperatur.

Det er grundlæggende min forventning at finde tilnærmelsesvis de samme tal i valideringstesten som ved de oprindelige test. Hvis de divergerer kraftigt kan det være et tegn på enten at typen af respondenter påvirker modellen, eller at kun et syntetisk svar ikke nødvendigvis kan afspejle en menneskelig respondent.

Test Navn	Model	Temp	Prompt	Varians andel	Varians sand	Varians LLM	Median afstand	Median sand	Median LLM	Fejl rate
Old-T1-N	3.5	1	Num	**17.1%**	0.367	0.063	**0.321**	2.792	2.470	**0%**
New-T1-N	4.1	1	Num	**20.5%**	0.366	0.075	**0.139**	2.791	2.653	**0.7%**
New-T2-N	4.1	2	Num	**24.4%**	0.373	0.091	**0.146**	2.790	2.644	**6.1%**
V-Old-T1-N	3.5	1	Num	**17.1%**	0.358	0.061	**0.298**	2.771	2.473	**0%**
V-New-T1-N	4.1	1	Num	**23.7%**	0.359	0.085	**0.133**	2.771	2.637	**0.2%**
V-New-T2-N	4.1	2	Num	**27.8%**	0.351	0.098	**0.148**	2.773	2.626	**5.3%**
Old-T1-T	3.5	1	Tekst	**28.7%**	0.367	0.105	**0.409**	2.792	2.383	**0.3%**
V-Old-T1-T	3.5	1	Tekst	**34.0%**	0.358	0.122	**0.372**	2.770	2.398	**0.2%**

Tabel 5: *Valideringsoverblik, Kvantitativt*. Viser resultaterne af de 4 valideringstest, med de tilhørende 4 test udført i analysen. De øverste set af 6 test er med numerisk prompt, de to nederste er med tekst-baseret prompt. Afhængige variabler markeret med fed skriftype. Venstre del viser testnavn og uafhængige variabler, højre del viser varians, median og fejlrate. Ved varians og fejlrate er der vist både menneskelig og model resultater. Vær opmærksom på at modelnavnene er forkortet ned til versionsnumre i de tilfælde hvor navnet var for langt til tabellen.

Tabel 5 viser resultaterne for de relevante oprindelige test, som beholder det samme testnavn, samt valideringstestene, som er markeret med "V-" prefix. Vær opmærksom på, at de 1000 respondenter som er brugt til valideringstestene ikke er de samme som ved de oprindelige test. Refererer til **4.1.2. Konkrete Resultater** for forklaringen af tabellens generelle layout.

I forhold til det forrige sæt af 1000 respondenter har dette en lavere reel medianværdi og samtidig er variansen også lavere. Hvis de forrige generelle mønstre holder, så burde valideringerne klare sig bedere end de forrige, hvilket er det, som tabel 5 viser, dog med to undtagelser. Den første er medianafstand i V-New-T2-N er svagt højere end ved den oprindelige New-T2-N test, dog kan fejlraten ved begge test gøre det svært at afgøre omfanget af denne ændring, særligt fordi at fejlraten ved den oprindelige test medførte, at den menneskelige median blev lavere, hvorimens fejlraten ved valideringstesten gjorde den højere. Den anden er variansandel ved V-Old-T1-N, som dog er nær identisk med Old-T1-N testen. Se appendiks 5 for de nøjagtige værdier for disse test.

Hvor der opstår en ændring i valideringstestene i forhold til de oprindelige test er ved variansandel ved GPT-4.1 modellerne og tekst prompt testen. Noget af dette skyldes sandsynligvis den mere snævre og tillidsfulde sammensætning af de nye 1000

respondenter. Noget af det kan også være et tegn på, at der er variation blandt individuelle svar per respondent for modellen, og at den anvendte metode med et syntetisk svar for en menneskelig respondent ikke nødvendigvis er fuldt tilstrækkeligt for et nøjagtigt indblik i effekterne af model, temperatur og prompt. Det at jeg anvender en aggregatværdi dannet ud af ni udsagn burde hjælpe på dette, men måske ikke nok til at sikre perfekt reliabilitet ved analysen.

Trods denne divergens er det samme mønster gældende i valideringstest som ved de oprindelige test: Inden for den numeriske prompt har GPT-4.1 højere variansandel og lavere medianafstand end GPT-3.5-Turbo, så den forbedring i model holder. Samtidig bliver fejlraten fra V-New-T1-N til V-New-T2-N højere, samtidig med at det øger median afstand og varians andel, hvilket støtter det tidligere fund omkring øget temperatur. Derudover reproducere V-Old-T1-T den samme forbedring af svarfordeling som den tekst-baserede prompt dannede i forbindelse med den oprindelige Old-T1-T test, og set i sammenhold med Old-T1-N testen har validerings testen også stærkt øget variansandel men også øget medianafstand og fejlrate, sammenligneligt med ændringen fra Old-T1-N til Old-T1-T.

5. Konklusion

Dette afsnit sammenfatter projektets analyse og besvarer forskningspørgsmålet, som blev præsenteret i indledningen. Afsnittet starter med at gentage problemformuleringen, efterfulgt af en forklaring af hvordan det er blevet undersøgt. Besvarelsen af problemformuleringen kommer i form af en præsentation af analysens fund, først med fokus på mønstrene fundet i aggregat og dernæst i svarfordelingen og endeligt præsenteres et uventet fund. Afsnittet afrundes med en opsummering af de samlede fund, og deres implikationer for problemfeltet.

5.1. Problemformulering

Som indledning til konklusionsafsnittet vil projektets problemformulering her gentages:

Hvordan påvirker nyere Large Language Modeller, temperaturindstillinger og forbedringer i prompten modellernes evne til at svare med menneskelignende varians i aggregat?

5.2. Delt resultat

Resultatet af dette projekt er todelt. Som vist i analyseafsnittet er der flere dimensioner og svar på effekten af model, temperatur og prompt på modellernes evne til at skabe menneskelignende varians. Det skyldes, at effekterne af disse tre variabler er påvirket af hinanden, og desuden afhænger af, om man ser på det brede aggregat, eller svarfordelingen. Altså, hvordan modellen svarer for alle 1000 respondenter i testen, eller på modellens svarfordeling relativt til den menneskelige svarfordeling ved hvert udsagn. I første omgang vil jeg fokusere på det brede aggregat, derefter svarfordelingen.

5.2.1. Aggregat

Set i aggregat så vil de nye modeller typisk øge variansen. Omfanget af forbedring er dog stærkt varierende, med en lille forbedring fra GPT-3.5-Turbo til GPT-4.1, men en stor forbedring når der testes på o3 modellen. Dette er dog ikke et stabilt mønster, idet GPT-3.5-Turbo med den tekst-baserede prompt fik en højere variansandel end samtlige GPT-4.1 test, dog stadig værre end o3. I modsætning fik

GPT-4.1 med den tekst-baserede prompt et værre resultat end nogen anden test, inklusiv GPT-3.5-Turbo med den numeriske prompt.

Hvis fejlrate og modellens evne til at matche den menneskelige median også tages i betragtning, kompliceres det yderligere. GPT-3.5-Turbo ved lav temperatur og o3 modellen var begge meget robuste med lave fejlrater, men disse test havde også relativt høje medianafstande, særligt o3 modellen. De nyere GPT-4.1 modeller var mindre robuste i forhold til fejlrate, dog ikke i et omfang der invaliderede anvendeligheden af resultaterne, selv ved høj temperatur og de havde de laveste medianafstande af alle anvendelige test. Til forskel havde den ældre model ved høj temperatur ekstremt høje fejlrater i sine test, og har derfor ikke givet yderligere anvendelige resultater.

Effekten af temperaturen på varians i aggregat er mere entydig. Øges temperaturen øges variansen, men det har visse omkostninger. Dette kommer i form af en forøget fejlrate, som nævnt var dette værst for GPT-3.5-Turbo modellen. Øget temperatur kan også medføre en øget medianafstand. Dette er dog uklart i aggregat perspektivet og vil uddybes i svarfordelings perspektivet. Hvor meget det hjælper modellen at øge temperaturen er svært at give et estimat for grundet GPT-4.1 testene var de eneste som gav anvendelige resultater inden for temperaturmålet. Tendensen virker dog til at være, at fejlraten stiger hurtigere end variansandelen øges, så selv hvis temperatur kunne øges over 2, ville det ikke nødvendigvis resultere i anvendelig data.

Effekten af den ‘forbedrede’ tekst-baserede prompt i forhold til den numeriske, set i aggregat, er også relativt entydig, dog med en undtagelse. Generelt har ændringerne fra den gamle numeriske prompt til den nye tekst-baserede prompt haft negative konsekvenser for både variansandel, medianafstand, og fejlraten. Undtagelsen opstår i forbindelse med den ældre GPT-3.5-Turbo model, ved lav temperatur og den tekst-baserede prompt. Her stiger variansandelen meget, relativt til samme test med den gamle numeriske prompt, dog stiger medianafstand og fejlrate også.

5.2.2. Svarfordeling

Hvis der derimod ses på selve svarfordelingen inden for hvert udsagn, henholdsvis de menneskelige og syntetiske respondenter har besvaret, så er der markant anderledes mønstre på spil. Helt grundlæggende ser vi også, at noget af årsagen til at variansproblemet opstår er, at modellen vælger en eller to af de mulige svarmuligheder på tværs af hele samplet. Her ser vi også at modellerne, på tværs af næsten alle test, afholder sig fra de mest ekstreme svarmuligheder. Dette passer meget skidt på de menneskelige svarmønstre som typisk følger bløde buer, med små men betydelige grupper af respondenter, som svarer 1 eller 5, altså de ekstreme svarmuligheder, mens de fleste ligger omkring de moderate eller neutrale svar. Selv de største grupper inden for et givent udsagn er sjældent over 40% af de 1000 udvalgte respondenter.

I kontekst af svarfordelingerne er effekten af model variabel næsten modsat konklusionen fra aggregat vinklen. Her forværrer nyere modeller evnen til at danne menneskelignende varians. Dette sker idet GPT-4.1 og o3 modellerne i højere grad end GPT-3.5-Turbo har tendens til at vælge enkelte svar for hele samplet. Her ser vi også at de nyere modeller, særligt o3 modellen, undgår at vælge den neutrale svarmulighed, samtidigt med at den typisk undgår de ekstreme, hvilket tvinger modellen ud i de moderate svarmuligheder. Dette medføre også at o3 modellen, trods den høje aggregat variansandel, har en meget ujævn svarfordeling med ‘dale’ i midten hvor modellen undgår det neutrale svar, hvilket ikke passer på nogen af de menneskelige svarfordelinger.

I kontrast er den ældre model mere tilbøjelig end de andre til at vælge den neutrale svarmulighed, hvilket hjælper den med at danne en menneskeligende svarfordeling, på trods af den generelt ikke vælger de ekstreme svarmuligheder reduceres dens variansandel. Vi ser her også, at dens manglende evne til at opfange den menneskelige median i aggregatet, skyldes en tendens til generelt at overvurdere respondenternes tillid, dog var dette en udfordring for alle modellerne.

Effekten af temperaturindstillingen er dog relativt uændret i svarfordelingsperspektivet relativt til aggregatet. Her ser vi at variansen øges ved at den hyppigste svarmulighed inden for hvert udsagn er mindre ved høj temperatur end ved lav. Denne effekt er dog meget begrænset ved GPT-4.1, som er de eneste

gyldige resultater ved høj temperatur. Derudover løser høj temperatur ikke modellernes anden udfordring, nemlig tendensen til at overvurdere tilliden. Som nævnt før medfører den høje temperatur også markant øget fejlrate, dog er dette svært at se i svarfordelings perspektivet.

Effekten af prompten er også tydeligere i svarfordelings perspektivet. Her ser vi at den grundlæggende ændrer mønsteret i GPT-3.5-Turbos svarfordeling, mens den virker til at forstærke de tendenser som skaber variansproblemet ved de nyere modeller. Dette er særligt tydeligt ved GPT-4.1 ved lav temperatur og den tekst-baserede prompt, som vælger den samme svarmulighed for hele samplet i 7 ud af 9 tilfælde, og dermed har den laveste varians på tværs af alle test.

Ændringen i GPT-3.5-Turbos svarfordeling har i høj grad hjulpet den til at vælge et svar på tværs af flere svarmuligheder, og i nogle tilfælde alle 5, hvilket giver den de mest menneskelignende svarfordelinger på tværs af alle test.

5.3. Opsummering

Set i helhed er effekten af nyere modeller, temperatur og prompt på modellernes evne til at danne menneskelig varians flertydigt og internt afhængigt. Hvor nyere modeller kan give indtrykket af forbedret varians, særligt o3 modellen set i aggregat, så er realiteten at de i tilsvarende eller højere grad laver den samme fejl som den ældre model, idet de vælger et eller to svarmuligheder for hele samplet ved hvert udsagn.

Temperatur hjælper med at løse netop dette problem ved at fordele nogle af respondenterne fra den hyppigste svarmulighed til de andre, dog medfører dette ofte en svag forværring af modellernes tendens til at overvurdere tilliden blandt respondenterne. Derudover er den ældste model uanvendelig ved høj temperatur grundet katastrofalt høj fejlrate, og o3 modellen kan ikke anvende temperatur.

Ændringen til den tekst-baserede prompt forværrer generelt modellernes evne til at variere deres svar, samtidigt med at den også forværrer fejlraten og modellernes evne til at opfange medianen. Undtagelsen er dog den ældre GPT-3.5-Turbo model ved lav temperatur, som ændrer svarmønster og præsterer markant bedre ved den nye prompt end den numeriske.

Så på trods af at GPT-3.5-Turbo har noget af den værste performance inden for varians i aggregat, trods at varians øges ved høj temperatur og trods at den tekst-baserede prompt forværrer variansen ved alle andre test, så er kombinationen af GPT-3.5-Turbo, lav temperatur og den tekst-baserede prompt det som skaber den mest menneskelignende varians, ud af de test som er udført i forbindelse med dette projekt.

5.4. Uventet fund

I forbindelse med analysen er der også gjort et uventet fund som er relevant for problemfeltet men ikke er dækket af problemformuleringen, hvilket er at næsten alle tests udført i forbindelse med dette projekt har præsteret værre end forventet. Ud fra litteraturen på området havde jeg en forventning om at modellerne ville kunne gengive den menneskelige median, særligt i aggregat, men være udfordret på varians og når aggregatet underinddeles i f.eks. Køn, alder og uddannelse. I flere test var modellens svar mere end en standardafvigelse væk fra den menneskelige median (Argyle, et al., 2023; Bisbee, et al. 2024)

Forventningerne omkring modellens udfordringer i forhold til varians holder, idet at variansen i modellernes svar højst kom op på 42.9% af den menneskelige varians, og at modellerne kun i begrænset omfang varierede svarene ud fra de demografiske variabler og derfor ofte gav meget ensformige svar på tværs af hele samplet. Trods denne overraskende svage performance har analysen stadig gjort en række fund, som er relevante for problemfeltet og som besvarer selve problemformuleringen.

5.5. Implikationer

Set i helhed, betyder disse resultater at sprogmodeller sandsynligvis stadig er lang vej fra at kunne erstatte menneskelige respondenter i spørgeskemaundersøgelser. Spørgsmålet bliver derfor om denne type opgave er noget LLM grundlæggende ikke kan klare, eller om noget andet ligger i vejen.

Hvis ikke det var for resultaterne af testen med den ældre model og den tekst-baserede prompt ville jeg umiddelbart have ment at denne type af syntetisk data generation ikke var noget LLM ville kunne bruges til, grundet de grundlæggende mønstre der danner variansproblemet i høj grad vedblev på tværs af de forskellige ændringer. Idet at den test viste, at det er muligt for modellen at danne

en menneskelignende svarfordeling giver det et spinkelt men reelt håb for at LLM kan anvendes til dette formål. Så hvis denne type opgave ikke er uløselig for LLM, hvorfor 'fejlede' modellerne så, i så godt som alle 10 test?

En mulig årsag, er at det er et model problem. Det at GPT-3.5-Turbo blev forbedret ved den nye prompt og de to andre modeller ikke gjorde, kunne indikere dette. Men selv det bedste resultat fra den model gav kun en varians andel på 34% af det menneskelige. Jeg har her kun testet på OpenAI modeller. Det kunne være at nogle af modellerne fra Google, Anthropic eller nogle af de andre LLM firmaer kan præstere bedre på netop dette område end GPT-3.5-Turbo.

En anden årsag kan være, at det er et alignment problem. Alignment er hvordan modellerne er trænet til at opføre sig, for eksempel hvordan den bør svare på visse typer af opgaver, eller hvad den ikke må svare på. Et muligt tegn på dette kan være det, at svarfordelingen inden for hvert udsagn blev skævere ved de nye modeller, og særligt det at o3 modellen fraholdt sig fra at give neutrale svar. Dette kan være fordi OpenAI har erfaret, at brugere ikke er interesserede i neutrale 'enten eller' svar, og de manglende ekstrem værdier kan også være et forsøg på at undgå at modellen udtrykker radikale eller skadelige holdninger, men derimod bør holde sig til de ukontrovertielle moderate synspunkter. Modellernes tendens til at give et eller to svar for hele samplet ved et udsagn kan også komme af alignment, hvis OpenAI har forsøgt at gøre modellen mere sandhedssøgende og konsistente på tværs af formuleringer, ved at træne dem til at reducere variationen i svar.

Endelig kan det også være et prompt problem, altså at måden begge de to prompter, jeg testede var skrevet på, gjorde at modellerne blev skubbet i retning af de samme svar, trods variationerne i de demografiske variabler. Det kan også være at antallet af demografiske variable var for mange eller ikke gav modellen den rette information. Det at jeg dog fik enkelte menneskelignende svarfordelinger med en prompt som samtidigt reducerede performance ved andre modeller syntes dog at modsige prompt-forklaringen, i hvert fald som den eneste forklaring.

Ud fra dette projekts resultater virker alignment forklaringen til at være den med størst forklaringsfaktor, men det vil kræve yderligere forskning med afklaringen af dette som mål for at kunne sige det med sikkerhed.

Fra en socialvidenskabelig synsvinkel er dette dog et relativt godt tegn for anvendelsen af LLM til syntetisk data generation. Så længe opgaven ikke er noget

LLM fundamentalt ikke kan løse, så er udfordringen for dette forskningsfelt at finde de rette metoder og omstændigheder der vil kunne tillade en model at løse opgaven. Dette betyder ikke, at løsningen kommer snart, bliver nem at anvende eller overhovedet er garanteret, men blot at resultaterne af dette projekt ikke indikerer, at variansproblemet er uløseligt.

Resultaterne af dette projekt tyder dog også på at variansproblemet ikke løser sig selv, forstået på den måde, at løsningen ikke opstår som konsekvens af at løse andre problemer med LLM eller brugen af LLM til syntetisk data generation. Denne forståelse kommer af hvordan udviklingen fra den ældre til de nyere modeller ikke medfører forbedringer i modellernes evne til at danne en menneskelignende svarfordeling i de test som blev udført i analysen. Dette betyder at hvis variansproblemet skal løses, vil det højst sandsynligt kræve målrettet forskning på området.

En anden indikation af projektets resultater er, at selv hvis der findes en løsning, eller blot noget som hjælper med at begrænse variansproblemet, så kan den hastige udvikling på LLM området som helhed forårsage ændringer som gør den løsning ineffektiv. Vi ser i dette projekt at øget temperatur hjalp GPT-4.1 modellen, selvom det også medførte andre problemer og kun hjalp på dele af variansproblemet. Men som også er vist så kan o3 modellen slet ikke anvende den indstilling. Netop dette element af variansproblemet kan ende med at løse sig selv hvis udviklingen på LLM området rammer et plateau, men det er uklart hvornår det kunne ske. En mere engageret løsning kunne være at træne en bestemt LLM model til formålet, dog er det endnu uklart hvor omkostningsfuldt eller hvor stor en forbedring det ville være.

Projektets resultater understøtter også en anden observation som er gjort inden for brugen af LLM til forskning, inden og uden for socialvidenskaben, nemlig at det er meget svært at forudsige hvilke opgaver en LLM kan løse, og at deres evne til at løse opgaver inden for et område, eller af en bestemt karakter, ikke nødvendigvis er nogen indikation af, at den kan løse andre opgaver inden for samme område, eller af lignende karakter. Dette kan blandet ses i hvordan ændringen i prompt, hvor den tekst-baserede prompt indeholdt en række ændringer som burde have hjulpet modellen, i stedet forværrede dens resultat i samtlige test undtagen den ene test hvor prompten virker til at have gjort en vital positiv forskel. For den fremtidige

forskning på området, både inden for brugen af LLM til generation af syntetisk survey data, men sandsynligvis også forskning som gør brug af LLM gennem prompting mere bredt, betyder dette, at det er vigtigt at afprøve flere modeller, indstillinger og særligt prompt.

Variansproblemet er dog ikke det eneste problem som dette projekt har fundet tegn på. Som beskrevet ovenfor var der et uventet fund i form af modellernes lave præstation, særligt når det kom til at genskabe en menneskelig median. Forskningen, som dette projekt bygger på, viste resultater hvor LLM kunne opfange den menneskelige median, men kom til korte når det kom til varians. Resultaterne af dette projekt er derimod at modellerne, i varierende grad, fejler på begge fronter. Nøjagtigt hvorfor dette sker ved jeg ikke, men et muligt svar er at der er anvendt dansk data frem for amerikansk, og at modellens opgave er rettet mod Danmark frem for USA. Den tidligere forskning har fokuseret på amerikansk politik, som i øjeblikket er ekstremt polariseret, på en måde, der skiller populationen i omkring to til tre grupper, altså Demokrater, Republikanere og Independents. Dette gør de demografiske variabler meget stærke. Særligt som Bisbee et al. (2024) studiet viser, er inklusionen af politiske variabler noget der i høj grad hjælper modellen. Hvis man ved at nogen stemte republikansk, så er det ikke svært for modellen at gætte personens holdninger til demokrater.

I modsætning er dansk politik meget mindre polariseret, der er langt flere partier og meget mere vælgervandring. Dette gør et stykke information som hvad en person stemte på til sidste valg en meget svagere variable i en dansk kontekst. Derudover kan spørgsmål om tillid, som de udsagn anvendes i dette projekt, være noget der i mindre grad er træningsmateriale for, særligt på dansk men også helt generelt. Folk på nettet og generelt i tekst er sandsynligvis mere eksplicitte omkring deres egne og andres politiske holdninger. Derimod kan det være mere uklart eller subtilt, om folk har tillid eller ej ud fra tekst på nettet. Hvis dette er en del af forklaringen på, hvorfor modellerne ikke kunne opfange medianen i forbindelse med dette projekt, så kunne det også tyde på en stor begrænsning ved anvendelsen af LLM i forbindelse med syntetisk data generation til spørgeskemaer. Hvis modellerne kun kan svare eller forholde sig ordentligt til emner som mennesker har givet klare udtryk for på tekst som så også skal indgå i modellernes træningsdata, så begrænser det anvendeligheden af LLM til et meget mere snævert set af områder.

En anden problematik, som dette projekt også fremviser, og som også er omtalt i forskningen som projektet bygger på, er at man skal være meget varsom med hvilke konklusioner man drager ud fra hvilke dele af modellens data. Om nyere modeller forbedrer eller forværrer variansen inden for hver test afhænger af, om man ser på svarene i aggregat eller inden for hvert udsagn. Et element dette projekt ikke har for øje er, i hvor høj grad modellerne svarer 'rigtigt'. Det er ikke svært at se at en model som o3 er meget ved siden af, når den giver det samme svar for næsten hele samplet ved flere udsagn. Men når GPT-3.5-Turbo danner en menneskelignende fordeling, i hvor høj grad passer svarene fra en af modellens respondenter på svarene fra det tilsvarende menneske? Hvis der er god sammenhæng er der ikke et problem, men hvis det ikke passer på det menneskelige, altså hvis modellen blot har tildelt respondenterne tilfældige svar men fordelt på en måde der passer til den menneskelig fordeling, så ville dataen se god ud på overfladen, men hvis man stiller spørgsmål til de underliggende mønstre, ville de svar man får ikke passe på det der ville være i det menneskelige data.

Som dette projekt og den tidligere forskning på området har vist, er variansproblemet kompliceret og har flere lag og facetter. For at kunne anvende syntetisk survey data til mere end overfladisk analyse er det vigtigt, at modellerne ikke kun giver et retvisende svar i aggregat, men også i svarfordelinger, og trods det ligger uden for rammerne af dette projekt, vil det også kræve, at de syntetiske data vedligeholder mønstrene, som findes i det menneskelige data. Dette betyder også at løsningen på variansproblemet højst sandsynligt ikke vil findes et enkelt sted, men at de forskellige dele vil kræve hver deres løsning, noget til at sikre svarfordelingen, noget til at øge modellernes evne til at genskabe de menneskelige svar mønstre, noget til at minimere den bias eller skævhed, der kan være i modellens svar.

Den endelige konklusion af dette projekt er derfor at LLM, som de er nu og ud fra de metoder, som er anvendt af forskningen på området, endnu ikke er egnet til at generere syntetisk data til surveyforskning. Projektet viser også at udviklingen inden for LLM ikke virker til at have løst de problemer som gør dem uegnede til data generations opgaven, og at de værktøjer som burde være egnet til at løse det, altså temperatur, kun i begrænset grad hjælper og nemt kan danne flere problemer, end

de løser. Projektet viser også, at måden hvorpå modellerne promptes har både vitale og uforudsigelige effekter. Derudover understøtter projektet også den del af den tidligere forskning, som viser, at modellernes præstation eller egnethed til et område eller opgave kan nemt afhænge af faktorer som er lette at overse. Projektets resultater peger på, at en del af variansproblemet muligvis skyldes måden modellerne er alignet på. Men også, at variansproblemet sandsynligvis vil kræve mere end en løsning, og at disse løsninger ikke vil opstå uden målrettet arbejde på området.

6. Diskussion

Projektets diskussionsafsnit indleder med yderligere overvejelser af projektets fund inden for forvaltning, med fokus på aktuelle forslag og tiltag af digitaliseringsministeriet. Derefter er der en kort undersøgelse og bud på årsagen til skævhed der eksisterer i de underindelte grupper. Dernæst er der en præsentation af de overvejelser og afvejninger, der er gjort i forbindelse med ændringerne fra den numeriske til den tekst-baserede prompt. Herefter kommer to underafsnit som forholder sig mere bredt til anvendelse af LLM, henholdvis inden for forskning, med fokus på reproducerbarhed, og for samfundet, med fokus på eksternaliteterne af videre udvikling og anvendelse af LLM. Det sidste underafsnit er en kort afrunding af diskussionsafsnittet samt overvejelser om fremtiden.

6.1. Yderligere Implikationer

Konklusionsafsnittet præsenterede en række implikationer af dette projekts resultater for socialvidenskaben og særligt survey forskningen. Det er dog værd at overveje, hvad disse resultater betyder for andre nærtliggende eller vigtige områder, som brugen af LLM inden for forvaltning i danmark. Jeg vil igen her gerne understrege, at det kan være svært, men ikke umuligt, at generalisere fund i forbindelse med studier af LLM, selv til områder eller opgaver der virker nært beslægtede (Argyle et al., 2025). Disse implikationer skal derfor forstås som det projektets resultater viser er sandsynligt, og ikke det, som kommer til at ske.

Grundet de potentielle effektiviseringsmuligheder og en ambition om at være forløber inden for kunstig intelligens, og herunder LLM, ligger det danske digitaliseringsministerium vægt på at udvide brugen heraf, ikke bare inden for det private, men også det offentlige (Digitaliseringsministeriet, 2024).

I denne forbindelse præsenterer Digitaliseringsministeriet selv at kunstig intelligens kan bruges til sagsbehandling inden for byggeansøgninger (Digitaliseringsministeriet, 2024, p.11), og trods at det beskrives som “høj risiko” i ministeriets strategiske indsats, er der beskrevet muligheden for at bruge det til bestemme adgang til uddannelsesinstitutioner og i forbindelse med rekruttering og jobsamtale (Digitaliseringsministeriet, 2024, p.22). Vi ser også at danske kommuner allerede forsøger sig med brugen af kunstig intelligens til at skrive høringsreferat af lange høringssvar (Digitaliseringsstyrelsen, n.d.). Inden for uddannelsessektoren er

der også kommet et forslag fra Niels Yde, formand for Danske Erhvervsskoler og -Gymnasier, om at kunstig intelligens vil kunne øge kvaliteten af bedømmelsesarbejdet og den skriftlige censur. Her jævnføres brugen af kunstig intelligens inden for sundhedssektoren til journalisering og diagnosticering (Rasmussen, 2025).

Sat i konteksten af dette projekt, og den tidligere forskning som det bygger på, indeholder disse typer af opgaver, altså klassificerings opgaver og referatskrivning, en række mulige udfordringer. Som vist i figurene i appendiks 6-19 for modellernes svar versus de menneskelige svar, fordelt i undergrupper som køn, alder og uddannelse, så har modellerne en tendens til både at undervurdere forskellen mellem to grupper, særligt blandt de ældre modeller, og i det omfang den fanger at der eksistere en forskel, passer denne forskel ikke på det reelle mønster. Dette kunne betyde, at en model som er sat til at skulle bedømme kompletheden af byggeansøgninger, kun i begrænset grad lægger vægt på de rette elementer i ansøgningen, og i det omfang den anvender indholdet til at danne et svar, selv bare til at støtte en menneskelig afgørelse, vil den ikke anvende det korrekt indhold til opgaven.

En anden, men relateret, problematik kan opstå i forbindelse med opgaver som referatskrivningen, nemlig at enkelte svar fra modellen, eller svar set i meget bredt aggregat, kan virke meget overbevisende. Modellerne er gode til at skrive grammatisk korrekt og sammenhængende tekst, hvilket typisk læses som sagligt og kompetent, men i hvilket omfang referaterne også fanger substansen af den, tekst de er lavet ud fra, er mindre klart og kan netop være svært at bedømme ud fra en ren menneskelig vurdering (Argyle et al., 2025). Disse referater kunne muligvis også blive påvirket af de mønstre, dette projekt fandt i de nyere modeller, hvis det er dem som bliver anvendt. Det kunne være, at modellerne ville være mindre tilbøjelige til at inkludere eller retvisende referere til dele af et høringssvar som udtrykker radikale, ekstreme eller på anden vis atypiske holdninger eller formuleringer. Samtidig kunne det modsatte ske, at modellen ikke skriver neutralt nok, men er skubbet ud i at tage stilling gennem eksempelvis ordvalg eller formuleringer.

Hertil skal det tilføjes, at det ikke er umuligt at få en model til at give ekstreme eller neutrale svar, hvis det er det, den bliver bedt om. Dette bringer den næste udfordring i spil, nemlig at der kan være stor forskel mellem hvilken prompt der

anvendes, derudover er det svært at sige hvad der gør en prompt bedre end en anden. Hvis modellens råd og vejledning til sagsbehandleren afhænger af subtile formuleringer i prompten, hvordan laver man så en fair prompt? Kan man som borger sige, at man har fået en fair og lovmæssig sagsbehandling hvis ens afgørelse ville have været anderledes med en minimal ændring i prompt? Denne problematik berør også den del af sagsbehandling og offentlig forvaltning som handler om skøn. En lovpligtig del af jobbet for mange sagsbehandlere er det individuelle skøn, altså at skøn ikke underlægges regel. Selv hvis modellerne udelukkende udøver en vejledende rolle for sagsbehandlerne, er der risiko for at det underminerer det individuelle skøn, særligt hvis det at give et andet svar end modellen, kræver yderligere arbejdstimer fra sagsbehandleren. Dette kan blive et problem, hvis de nødvendige extra arbejdstimer ikke er tilgængelige fordi effektiviseringen vundet ved modellerne forventer at hver enkelt sagsbehandler bruger kortere tid per sag.

Dette lægger sig op ad en anden problematik, som er den bias, der er i modellens svar, I dette projekt var median af modellen samtlige test mere tillidsfulde end gennemsnit af menneskene, danskere var stereotypificerede til at være tillidsfulde. I forvaltningskontekst kan en sådan stereotypificering være til forhindring for en fair og korrekt sagsbehandling. Det er vigtigt at pointere, at mennesker selvfølgelig ikke er fri for bias og stereotypificering, men de kan være opmærksomme på det selv og fordele opgaverne internt for at begrænse påvirkningen af denne bias. Selv hvis sagsbehandlerne ikke er klar over deres egne bias, så er der stadig en vis form for lighed i det, at sagsbehandlere har mere tilfældig og forskelligartet bias, hvor en sagsbehandler kan være mere positiv over for en bestemt gruppe, kan en anden sagsbehandler kan være mere negativ. Ved modellerne vil denne bias potentielt være meget mere ensartet og konsistent, hvilket på samfundsmæssigt plan kan være mere problematisk end det mere menneskelige bias.

Implementeringen af LLM er dog ikke uden mulige gevinster, selv når de ovenstående bekymringer og risici tages i betragtning. Nogle af disse argumenter bygger på mine konkrete resultater, mens andre bygger på mere grundlæggende træk i LLM og de mulige anvendelser inden for forvaltning som er beskrevet i starten af dette afsnit. Det at mennesker har mange forskellige bias betyder ikke nødvendigvis, at de samlet set er neutrale og retfærdige, og det er svært at sige i

hvilket omfang en given afgørelse er påvirket af positiv eller negativ bias. Det at modellerne har en mere ensartet bias kan medføre problemer, som beskrevet ovenfor, men det er samtidigt også mere målbart, robust, og hvis danmark anvender sin egen model, vil det også være noget man i højere grad vil kunne bevidst justere. En anden måde, resultater indikere at modellerne kan hjælpe på bias og øge lighed er at de i mindre grad er påvirket af de demografiske variabler. Det at modellen ikke lader sig påvirke af dem kunne betyde, at borgere fra forskellige demografiske grupper får mere ens behandling. Dette er måske ikke altid eftertragtet, under visse omstendigheder kan det være vigtigt at sagsbehandlerens skøn tager højde for køn, alder, uddannelse, eller lignende, men i mange andre omstændigheder er det bedre hvis de kan abstrahere fra det. I denne forbindelse er den større udfordring måske at modellerne, i det omfang de er påvirket af de demografiske variabler, ikke er tættere på menneskernes svar. Det er her vigtigt at pointere at der også kan være begrænsninger ved at generalisere modellens evne ud i at bruge demografiske variabler til at udtrykke tillid, og så bruge dem til opgaver i forbindelse med sagsbehandling.

En anden brug af modellerne er som vidensbank. Det ville kræve særlige træningsdata og sandsynligvis også målrettet alignment arbejde, men hvis de indlæres med diverse relevante kommunale særregler, bekendtgørelser, domspraksis og lignende, ville sagsbehandlere, særligt nyansatte, have nogen de kan spørge til råds og derved sænke byrden forbundet med oplæring. Dette ville også gøre disse arbejdspladser mere institutionelt robuste, idet viden og ekspertise i mindre grad kan gå tabt på grund af personale, der skifter job eller går på pension, men det kan også gøre det nemmer at fyre personale, givet deres personlige viden og erfaring bliver mere bredt tilgængeligt.

En anden fordel ved implementeringen af LLM i forvaltnings- og sagsbehandlingsprocesser er, at alt der sendes til og fra dem, bliver logget og ligger digitalt. På denne måde hjælper modellerne med at øge journalføring, hvilket kan gøre det nemmere og dermed hurtigere at følge op på klagesager eller dokumentere og rette processuelle problemer.

Som beskrevet i konklusionsafsnittet er der tegn på, at variansproblemet ikke løser sig selv, så hvis Danmark vil have indflydelse på, hvordan løsningen på dette og andre problemer med LLM, kommer til at se ud, er det måske nødvendigt for staten at engagere sig aktivt i brugen og udviklingen af LLM. Dette er særligt

gældende hvis det viser sig, at variansproblemet skyldes alignment, altså hvordan modellen er trænet og indstillet, fordi dette til en hvis grad er kulturelt præget. Hvis vi derfor ikke er med til at styre alignment, særligt for en model som bruges til forvaltning i danmark, så kan vi blive underlagt kulturelle normer og holdninger som vi ikke klart kender til og måske ikke er enige med. Modsvaret til dette er, at vi så burde afholde os fra brugen af LLM, måske særligt inden for forvaltning, men som beskrevet her er der også måder hvorpå LLM åbner muligheder for forbedring.

6.2. Er der et amerikansk mønster?

Et af dette projekts fund har været, at modellerne i mindre grad end forventet har kunnet fange den menneskelige median i aggregatgrupperne. Som beskrevet tidligere kan dette muligvis skyldes, at dette projekt har anvendt dansk data og har rettet opgaven mod Danmark, hvor den tidligere forskning på området i højere grad har anvendt amerikansk data. Noget der ligger til grund for denne opfattelse er måderne hvorpå o3 modellen fordeler undergrupperne på særligt alder og uddannelse, fordi her divergerer modellen fra de menneskelige respondenter på en næsten spejlvendt måde. Blandt menneskene er der en positiv sammenhæng mellem alder og tillid, men modellen viser en negativ. På uddannelsen kan man se, hvordan modellerne gætter der er høj tillid blandt de kortuddannede, særligt erhvervsuddannelser og almen gymnasial uddannelse, mens blandt menneskene var de grupper nogle af de mindst tillidsfulde.

Se appendiks 10 og 15 for test Strong-N og Strong-T for figurer på undergrupper.

Idet den tidligere forskning fokuserede på USA, og dette projekt fokuserer på Danmark, er det så muligt at denne skævhed opstår fordi modellerne i højere grad følger et amerikansk mønster? Hvis dette var tilfældet, burde amerikanske opinionsundersøgelser vise modellens skævhed. Ifølge tal fra Pew Research i 2025, er der en positiv sammenhæng mellem alder og tillid i USA, og det samme for uddannelse, altså det samme mønster som Danmark, og ikke hvad modellen viser (Andersen et al., 2024, Silver et al., 2025).

Men hvad kan så forklare denne skævhed?

En mulig forklaring er, at modellerne reelt har en underliggende bias, som den applicerer på opgaver, men at denne bias ikke nødvendigvis er amerikansk, men global. Et studie af Kovač et al. (2023) peger på, at modellerne ikke har et bestemt bias, men deres bias afhænger af flere forskellige faktorer, heraf tilsyneladende irrelevante ændringer i promptens formulering eller formatering (Kovač et al., 2023). Et andet studie, udarbejdet af Wright et al. (2024), beskriver hvordan LLM har tendens til at bruge de samme fraser og argumenter på tværs af forskellige prompter (Wright et al., 2024). Endelig er der også et studie af Cao et al. (2025), som viser at stereotyperne i et sprog kan sprede sig til andre (Cao et al., 2025).

Tilsammen betyder dette, at skævheden, som opstår i modellernes svar i dette projekt, ikke nødvendigvis er et tegn på at modellerne er mere amerikanske, men at noget ved prompten skubber modellen i retning af bestemte stereotyper, altså at unge er tillidsfulde og ældre er skeptiske, og at kortuddannede og PhD uddannede begge har relativt høj tillid i forhold til resten. Studiet af Cao et al. (2025) foreslår også, at de sprog som er svagt repræsenteret i træningsdata, kan være mere tilbøjelige til at opsamle stereotyper fra andre lande. I det omfang dansk er svagt repræsenteret kunne dette være en yderligere forklaring på hvorfor modellerne virker til at fejlræsonnere, når det kommer til de relative holdninger i undergrupperne (Cao et al., 2025, Kovač et al., 2023, Wright et al., 2024).

6.3. Promptændringer

Som beskrevet i metodeafsnittet opstod ændringen i prompt fra den numeriske til den tekst-baserede organisk undervejs i udarbejdelsen af projektet. Dette begrunder jeg ved at tolke de rettelser som udgør ændringerne fra den numeriske til den tekst-baserede som forbedringer, men som analysen viser, havde det, med en enkelt undtagelse, den modsatte virkning. Det er derfor værd at overveje, hvordan ændringerne kan have forværret resultaterne. Dette ændrer ikke på de konklusioner, som projektet drager på baggrund af prompt-delen af analysen.

Ændringen af 'børn' variablen, fra den oprindelige fejltolkning af 'antal børn respondenten er far, mor eller værge for', til 'antal børn i respondentens hushold', blev gjort ud fra en forventning om at den korrekte tolkning i højere grad ville hjælpe modellen med at gengive de menneskelige personaer. Det har dog den svaghed, at præciseringen kan have forvirret modellen, idet den rettede variabel sandsynligvis har en lavere forklaringskraft end fejltolkningen. Altså fortæller det mere om en person at sige de har to børn, end at sige de bor i samme hushold som to børn. Dette er understøttet af at gennemsnitsalderen for Tryghedsmålingen, som vist i metodeafsnittet, er 51,4 år. Det kan derfor også være sandsynligt, at hvis en person på 51 år har svaret at de bor sammen med to børn, så er de sandsynligvis børn, som respondenten er forældre til, altså at fejlfortolkningen har været tæt nok på sandheden. Til modsætning, hvis gennemsnitsalderen havde været omkring 20 år, kunne børn i husholdet i højere grad overlappe med yngre søskende, og så ville de to sandsynligvis divergere.

Inklusionen af kommunevariablen blev gjort ud fra en forventning om at modellen kunne bruge viden om de danske kommuner til at hjælpe den med at gengive de menneskelige personaer. Problemet her er, at prompten i forvejen inkluder et mål for befolkningen i byen de bor i. Præciseringen af at nogen for eksempel bor i en mindre by, men i en kommune som er navngivet eller kendt for sin storby, kan i virkeligheden have fået modellen til at fokusere på kommunenavnet frem for befolkningstallet, som måske har været mere informativt. En anden mulig udfordring er, at nogle danske kommuner sandsynligvis ikke udgør en stor nok del af modellernes datagrundlag til at den har kunnet bruge kommune navnet, og derfor har inklusionen reelt bare været støj for modellen.

Endelig var der ændringen af svarformatet, gjort ud fra en forventning om, at modeller, særligt GPT-3.5-Turbo, er svage til tal og matematik i et omfang der

kompromitterer dens evne til at bruge numeriske svarformater frem for tekst-baserede. Denne forventning er primært bygget ud fra manglende matematik egenskaber ved ældre modeller, og at tidligere forskning på området satte modellerne til at vælge et tal mellem 0 og 100 for at udtrykke holdning (Bisbee, et al., 2024). Det har muligvis være simpelt nok for selv de ældre modeller at forholde sig til numerisk Likert fordeling, hvorimod det tekst-baserede svarformat har krævet flere tokens til at skrive svaret. Muligvis har det også fået modellen til at bruge flere tokens på at ræsonnere, hvilket kan have medført den øgede fejlrate, hvis de er opstået grundet token begrænsninger. Muligvis kan tekst også i højere grad være påvirket af dens alignment, hvis modellerne er trænet til at enighed er godt, så kunne det forklare den højere medianafstand ved den tekst-baserede prompt, grundet modellen i endnu højere grad overestimerer tillid.

6.4. Reproducerbarheden af LLM studier

Dette projekt har forholdt sig snævert til anvendelsen af LLM til generation af syntetisk data af spørgeskemaundersøgelser, men som den anden artikel af Argyle et al. (2025) kommer ind på i teoriafsnittet, så er der også gjort forsøg med LLM inden for andre dele af forskningsprocessen. Det er derfor værd at overveje mere bredt, hvordan LLM påvirker den forskning, den er anvendt i. I denne kontekst har jeg valgt at have et særligt fokus på reproducerbarheden af forskning som anvender modeller, givet den grundlæggende vigtighed af reproducerbarhed af videnskabeligt arbejde, samt at reproducerbarheden af LLM output har været særligt omtalt som problematisk (Barrie et al., 2025).

Som beskrevet i starten af projektet er der flere mulige og reelle fordele ved brugen af LLM inden for forskning, ikke bare inden for generation af syntetisk spørgeskema data, men også til evalueringsopgaver og kodning af tekst og transskriptioner. Modellernes evne til at processere store mængder data samt behandle forskellige typer af data gør, at de potentielt kan tillade studier at anvende større mængder data samt appliceres på mange forskellige felter. Dette er, foruden deres hjælp til opgaver som oversættelse og opsummering, noget som kan gøre forsknings mere tilgængelige (Argyle et al., 2025, Gu et al., 2025).

En af de mest grundlæggende udfordringer til reproducerbarhed, som opstår i forbindelse med brugen af LLM inden for forskning er, at de ikke er deterministiske, selv ved lavest mulig temperatur og samme seed. Inden for socialvidenskaben og politologien er denne mangel på determinisme muligvis et mindre problem, mennesker er ikke deterministiske og ved nogle metoder, som kodning af tekst og transskriptioner, er der en tolerance for inter-koder variabilitet. Udfordringen ved LLM er, at det ofte kan være uklart, hvorfor der opstår forskel, særligt hvis modellerne der bliver anvendt er en del af omskiftelige infrastrukturer, som dem der her er tilgået via OpenAIs API. Disse modeller kan blive opdateret eller justeret uden varsel, og modeller kan blive udfaset, hvilket begge kan ændre deres svar. Noget af dette kan løses ved at være omhyggelig i dokumentationen, men dette øger stadig sandsynligheden for fejl ved dannelsen eller applikationen af et forsøgs dokumentation (OpenAI, n.d.-a, n.d.-f).

En anden udfordring ved replicerbarheden er, at selv små semantiske ændringer i prompten kan potentielt medføre substantielle forskelle. Dette betyder at hvis et studie vil forsøge genskabe et andet studies resultater med samme metode,

men med nye data, for eksempel i et andet land eller sprog, kan modellen svare markant anderledes, selvom der reelt set måske ikke er nogen betydningsmæssig forskel i data. Dette åbner også for at forskere kan lave iterative ændringer i prompt indtil den giver resultater, der peger i en given retning, altså en form for p-hacking (Kosch & Feger, 2024, Seleznyov et al., 2025).

En anden udfordring som påvirker både replicerbarheden, men potentielt også andre kvalitetskriterier inden for forskning, er, at den data modellerne er trænede på bliver typisk ikke oplyst. Dette er problematisk i forbindelse med benchmarking af modeller og anvendelse af modeller til evaluering, fordi hvis træningsdata indeholder svar på spørgsmål, som bruges til at kontrollere eller teste om modellerne kan løse, og dermed er blevet bedre til bestemte opgaver, så kan sådan resultater blive misvisende. Med andre ord, fordi træningsdata holdes hemmeligt, åbner det muligheden for forurening af testdata (Mirzadeh et al., 2024).

Så på trods af de fordele LLM kan bringe, betyder deres anvendelse også, at noget af kontrollen over de forsøg som anvender LLM, enten som instrument eller subjekt, overlades til udbyderne af modellerne og den data, de er trænet på. Derudover påtvinger LLM også en hvis stochasticitet, og i den forbindelse også et øget krav til dokumentation, både for at sikre et studies eksterne validitet, men netop også reproducerbarhed.

6.5. Externaliteter

Dette projekt har primært fokuseret på brugen af LLM inden for forskning, men LLM bliver brugt mange andre steder, af mange forskellige mennesker til meget varierende formål (Ofcom, 2024, pp. 32–38; Ouyang et al., 2023; YouGov, 2025). Det er derfor vigtigt at forholde sig til hvordan denne endnu relativt nye teknologi påvirker andre dele af samfundet. Selv hvis LLM er et perfekt forskningsværktøj, så eksisterer det ikke i et vakuum. Så hvilke andre påvirkninger og eksternaliteter kan LLM påføre forskellige dele af samfundet i en bredere forstand?

Dette afsnit har et mere forsigtigt syn på LLM som teknologi, men forholder sig primært til problematikker og omkostninger, som enten er opstået, eller der er stærkt belæg for at opstå i den nærmere fremtid. Disse omkostninger bliver også sat i kontekst af dette studie, og hvad det kan betyde for anvendelsen af LLM-genereret syntetisk data. Fordi dette afsnit fokuserer på mulige yderligere omkostninger forbundet med LLM, er der ikke en videre afvejning af de mulige gevinster det kan medføre uden for forskningen.

6.5.1. Information og Tillid

LLM modeller kræver store mængder af højkvalitets menneskeskabt tekstdata for at trænes. Mængden af højkvalitetsdata er dog begrænset, hvilket kombineret med den hastige udvikling inden for LLM området medfører, at prisen for træningsdata stiger, og der vil være incitament at kræve licensbetaling for tekst, der måske ellers ville være offentligt tilgængeligt. Dette kan tvinge de firmaer som arbejder på at lave LLM ud i nogle ubehagelige valg (Tong et al., 2024; Villalobos et al., 2022).

De kunne forsøge at anvende data af lavere kvalitet eller syntetisk data, men i det omfang dette overhovedet tillader videreudvikling af modellerne, er der tegn på at det også vil medføre forværringer på nogle områder, særligt inden for variationen i modellernes output (Li et al., 2025; Seddik et al., 2024).

Alternativt kunne de forsøge at anskaffe højkvalitetsdata uden samtykke fra rettighedshaverne, hvilket kunne øge risikoen for retslige anklager og dermed omkostninger. Der eksisterer i forvejen retssager mod OpenAI på netop dette område, dog endnu uden udfald. Uanset udfaldet i de amerikanske retssager vil den nye EU regulering på området forøge dokumentations-omkostningerne ved blandt andet at kræve en ophavsretspolitik og et tilstrækkeligt detaljeret resumé af

træningsdata (European Union, 2024; *The New York Times Company v. Microsoft Corporation*, 2025).

Endelig kunne de også forsøge at betale licensprisen for den data, de vil bruge. Dette ville sandsynligvis stille firmaerne i en meget svag forhandlingsposition og ville derfor gøre købet meget dyrt. Givet offentlige rapporter indikerer, at førende LLM firmaer som OpenAI fortsat lider økonomisk underskud, kan de øgede omkostninger potentielt være eksistenstruende for industrien, med mulig undtagelse af firmaer som Google og Meta der kan betale for udviklingsomkostningerne med andre forretninger (Alphabet Inc., 2025; Hammond, 2025; Meta Platforms, Inc., 2025).

Dette har også en relevans i forhold til projektet, idet at løsningen på variansproblemet sandsynligvis vil kræve videreudvikling af LLM. Samtidig korroborerer det også begrænsningerne ved syntetisk data, også uden for spørgeskemakonteksten.

En anden problematik ved LLM er, hvordan de tillader svindelmetoder som phishing at blive mere sofistikerede og målrettede, samtidigt med at de også åbner for skaleringsmuligheder og kan gøre verificering sværere ved at kunne skabe både video, billeder og optagelser af tale (Bethany et al., 2025).

Derudover er der mulighed for anvendels af LLM til at manipulere den offentlige debat inden for et område ved at efterligne online aktører. Dette er et fænomen OpenAI selv erkender har fundet sted. Særligt har der været en misinformations kampange som brugte ChatGPT til at spræde kommentarer på X i forbindelse med valget i Rwanda i 2024 (Nimmo & Flossman, 2024; Sadeghi et al., 2025).

I forhold til projektet viser dette, at selvom variansen i modellernes output ikke indeholder menneskelig varians set i aggregat, betyder det ikke at de individuelle svar ikke godt kan bruges i stor skala og til menneskelignende adfærd. Samtidig kan det også indikere, at kvalitetstærsklen for ondsindet brug af syntetisk data potentielt er lavere end saglig, videnskabeligt brug.

6.5.2. Ressourcer

En anden omkostning ved LLM er det øgede ressourceforbrug, der kommer med udviklingen og anvendelsen af LLM, særligt strømforbrug. En rapport fra IEA

estimerer, at strømforbruget af samtlige datacentre i 2030 vil være lidt mere end Japans samlede strømforbrug i dag. Dette inkluderer datacentre, som bliver anvendt til andet end LLM, men de er den største faktor i væksten på området. Dette sker samtidigt med at elpriserne I USA, hvor flere af de nye datacentre bygges, stiger, og grønne energiinitiativer nedlægges af føderale myndigheder (International Energy Agency, 2025, pp. 62–70; U.S. Energy Information Administration, 2025; Volcovici & Groom, 2025).

Dette er relevant for projektet idet et grundlæggende argument for anvendelsen af syntetisk data er de lavere omkostninger, sammenlignet med menneskelige respondenter, men ved syntetisk data som kræver videreudvikling af LLM kan det medføre utilsigtede omkostninger for tredjeparter og samfundet som helhed.

6.5.3. Læring og Sundhed

En anden mulig omkostning ved brugen af LLM ligger i det cognitive. LLM anvendes i høj grad af studerende til at hjælpe med skolearbejdet, men også af voksne til at løse opgaver på jobbet. Udover udfordringerne, dette kan skabe for diverse uddannelsessystemer, peget et pre-print af et MIT studie på, at det også kan være et problem for brugeren, idet anvendelsen af LLM medfører kognitiv gæld, forstået som manglende indlæring grundet anvendelsen af hurtigere og nemmere løsninger, som ikke opbygger den samme viden og færdighed. Det skal hertil pointeres, at LLM også godt kan hjælpe med læringsprocesser, som et opfølgende led i læringen. Dette er understøttet af tidligere forskning uden for LLM som indikerer, at automatisering kan medføre øget uopmærksomhed (Kosmyna et al., 2025; Parasuraman & Manzey, 2010).

Relevansen af denne externalitet ligger i, at selv hvis udfordringerne som forhindrer LLM i at generere brugbart syntetisk data bliver løst, så er der potentiale for at forskere som anvender syntetisk data vil have en svagere forståelse af det datamateriale de anvender, og forholder sig mindre kritisk til det, så de svagheder der eventuelt stadig kunne være i data ville blive overset.

Et andet sted, hvor LLM kan medføre udfordringer, er mental sundhed og psykiatrien. Trods det er et mindretal, der rapporterer de bruger modellerne til dette formål, anvendes modellerne af nogle som personlig psykiater eller terapeut. Noget

af årsagen til denne anvendelse kan komme af at modellerne er let tilgængelige, kan virke anonyme, samt deres generelt bekræftende opførsel. En yderligere forklaring er den forværrede mental sundhed, særligt blandt unge, kombineret med den typisk lange ventetid på psykiatrisk udredning i Danmark. Trods evidensen på dette område er sparsomt, peger studier i retning af at det kan være en hjælp for nogle brugere, men der er også seriøse risici ved at bruge modellerne som terapeut, særligt ved alvorlige tilstande som spiseforstyrrelser og skizofreni. Omfanget af anvendelsen af LLM som personlig terapeut i danmark er endnu udokumenteret (Fitzsimmons-Craft & Jacobson, 2024; Jensen et al., 2024, pp. 20–27; Ofcom, 2024, pp. 32–38; Sundhedsstyrelsen, 2025).

Relateret til dette er også en række cases, hvor spirituelle eller personlige spørgsmål til modellen kan, uagtsomt, medføre psykose hos brugeren idet de bliver overbevist om at de har gjort et metafysisk fund, at modellen er guddommelig, eller forelsket i dem. Særligt kan det være antropomorfisering af modellernes menneskelignende opførsel, som kan fodre disse vrangforestillinger i sårbare brugere. Som beskrevet i det citerede preprint studie af Morrin et al. (2025), er det endnu uklart om det kan opstå i brugere som ellers ikke er prædisponerede for psykotiske episoder (Li et al., 2025; Morrin et al., 2025; Østergaard, 2023).

Relevansen af denne externalitet for projektets problemfelt er at hvis eventuelle løsninger på de problemer og svagheder som medføre modellernes manglende evne til at kunne generere syntetisk data kræver, at de i højere grad kan udtrykke menneskelignende adfærd, kan det blive en fare for en sårbare gruppe af brugere.

6.6. Afrunding

Det evindelige spørgsmål ved LLM, som så meget andet ny teknologi, er hvor dette vil føre hen, og som nærmest hver gang det spørgsmål stilles, er svaret svært at give med sikkerhed. Om LLM teknologien er den første gnist i en ny eksplosion af teknologisk udvikling, eller det første søm i menneskehedens ligkiste, kan afhænge af hvem man spørger.

Som Argyle, et al. (2025) beskriver i deres artikel, så er modellerne uransagelige, og det kan være svært på forhånd at sige, hvor godt de vil klare en given opgave. Denne uransagelighed gør dem svære at studere og svære at anvende. Men gemt bag uransageligheden er der også et enormt potentiale.

Personligt, efter at have udarbejdet dette projekt og forholdt mig til noget af forskningen på området, ser jeg, på godt og ondt, at LLM er kommet for at blive. Om udviklingen fortsætter, eller stagnerer, er jeg mindre vis på. Men i min optik er det også mindre vigtigt, for selv som det er nu, er der fordele og ulemper nok til at man bør forholde sig til modellerne, personligt såvel som på samfundsmæssigt plan.

7. Litteraturliste

Alphabet Inc. (2025, July 23). *Alphabet announces second quarter 2025 results* [Press release]. https://abc.xyz/assets/cc/27/3ada14014efbadd7a58472f1f3f4/2025q2-alphabet-earnings-release.pdf

Andersen, J., Andersen, J. G., Hede, A., Danneris, S., & Madsen, P. G. H. (2024). *Tillid i Danmark 2024: TrygFondens tryghedsmåling 2024.* TrygFonden smba. https://www.tryghed.dk

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). *Out of One, Many: Using Language Models to Simulate Human Samples.* Political Analysis, 31(3), 337–351. https://doi:10.1017/pan.2023.2

Argyle, L. P., Busby, E. C., Gubler, J. R., Hepner, B., Lyman, A., & Wingate, D. (2025). *Arti-"fickle" Intelligence: Using LLMs as a Tool for Inference in the Political and Social Sciences.* arXiv. https://arxiv.org/abs/2504.03822v1

Bail, C. A. (2024). *Can generative AI improve social science?* Proceedings of the National Academy of Sciences, 121(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Barrie, C., Palmer, A., & Spirling, A. (2025, May 1). *Replication for language models: Problems, principles, and best practices for political science* [Working paper]. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Bethany, M., Galiopoulos, A., Bethany, E., Bahrami Karkevandi, M., Beebe, N., Vishwamitra, N., & Najafirad, P. (2025, April 15). *Lateral phishing with large language models: A large organization comparative study* (arXiv preprint arXiv:2401.09727). arXiv. https://doi.org/10.48550/arXiv.2401.09727

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). *Synthetic replacements for human survey data? The perils of large language models.* Political Analysis, 32(4), 401–416. https://doi.org/10.1017/pan.2024.5

California Health Interview Survey. (2024). *CHIS 2023 methodology series: Report 4 – Response rates (Table 7-2)*. UCLA Center for Health Policy Research. https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/chis-design-and-methods/chis-methodology-reports-repository

Cao, Y. T., Sotnikova, A., Zhao, J., Zou, L. X., Rudinger, R., & Daumé III, H. (2025, July 31). *Multilingual large language models leak human stereotypes across language boundaries*. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)* (pp. 175–188). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.nlp4pi-1.15

Digitaliseringsministeriet. (2024, December). *Strategisk indsats for kunstig intelligens: Et styrket fundament for ansvarlig udvikling og anvendelse af kunstig intelligens i Danmark*. https://www.digmin.dk/Media/638687214351712933/Strategisk%20indsats%20for%20kunstig%20intelligens.pdf

Digitaliseringsstyrelsen. (n.d.). *Generativ kunstig intelligens skriver referater af høringssvar*. Tilgået d.17/08/2025. https://digst.dk/kunstig-intelligens/inspirationskatalog-til-generativ-kunstig-intelligens/generativ-kunstig-intelligens-hjaelper-roskilde-med-referater-af-hoeringssvar/

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2024). *Questioning the survey responses of large language models* (arXiv:2306.07951v4). arXiv. https://arxiv.org/abs/2306.07951v4

Eggleston, J. (2024). *Frequent survey requests and declining response rates: Evidence from the 2020 Census and household surveys*. *Journal of Survey Statistics and Methodology, 12*(5), 1138–1156. https://doi.org/10.1093/jssam/smae022

European Union. (2024, June 13). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union,* L 2024/1689. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

Fitzsimmons-Craft, E. E., & Jacobson, N. C. (2024). *Eating disorders care and the promises and pitfalls of artificial intelligence*. *Missouri Medicine, 121*(5), 345–349. https://pmc.ncbi.nlm.nih.gov/articles/PMC11482850/

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2025, March 9). *A survey on LLM-as-a-judge* (arXiv preprint arXiv:2411.15594). arXiv. https://doi.org/10.48550/arXiv.2411.15594

Hammond, G. (2025, June 10). *OpenAI expects subscription revenue to nearly double to $10bn*. *Financial Times*. Tilgået d.28/08/2025. https://www.ft.com/content/1ffc5fe7-6872-42a0-8b98-dc685f9c33c6

IBM. (2023, November 2). *What are large language models (LLMs)?* Tilgået August 28, fra https://www.ibm.com/think/topics/large-language-models

International Energy Agency. (2025, April). *Energy and AI* (World Energy Outlook special report). https://iea.blob.core.windows.net/assets/601eaec9-ba91-4623-819b-4ded331ec9e8/EnergyandAI.pdf

Jensen, H. A. R., Møller, S. R., Jezek, A. H., Davidsen, M., Ekholm, O., & Christensen, A. I. (2024). *Danskernes sundhed 2023*. Statens Institut for Folkesundhed, Syddansk Universitet. https://www.sdu.dk/sif/-/media/images/sif/udgivelser/2024/danskernes_sundhed_2023.pdf

Kosch, T., & Feger, S. (2024, April 24). *Risk or Chance? Large Language Models and Reproducibility in HCI Research* (arXiv preprint arXiv:2404.15782). arXiv. https://doi.org/10.48550/arXiv.2404.15782

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025, June 10). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay-writing task* (arXiv preprint arXiv:2506.08872). arXiv. https://doi.org/10.48550/arXiv.2506.08872

Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F., & Oudeyer, P.-Y. (2023, November 7). *Large language models as superpositions of cultural perspectives* (arXiv preprint arXiv:2307.07870). arXiv. https://doi.org/10.48550/arXiv.2307.07870

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M., Gururangan, S., Wortsman, M., Albalak, A., … Shankar, V. (2025, April 21). *DataComp-LM: In search of the next generation of training sets for language models* (arXiv preprint arXiv:2406.11794). arXiv. https://doi.org/10.48550/arXiv.2406.11794

Li, J., Li, Y., Hu, Y., Ma, D. C. F., Mei, X., Chan, E. A., & Yorke, J. L. (2025). *Chatbot-delivered interventions for improving mental health among young people: A systematic review and meta-analysis*. Worldviews on Evidence-Based Nursing, *22*(4), e70059. https://doi.org/10.1111/wvn.70059

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., & Halgamuge, M. N. (2024). *Inadequacies of large language model benchmarks in the era of generative artificial intelligence* (arXiv:2402.09880v2). arXiv. https://arxiv.org/abs/2402.09880v2

Meta Platforms, Inc. (2025, July 30). *Meta reports second quarter 2025 results* [Press release]. https://s21.q4cdn.com/399680738/files/doc_news/Meta-Reports-Second-Quarter-2025-Results-2025.pdf

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024, October). *GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models*. Apple Machine Learning Research. https://machinelearning.apple.com/research/gsm-symbolic

Nimmo, B., & Flossman, M. (2024, October). *Influence and cyber operations: An update* [Threat-intelligence report]. OpenAI. https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf

Mahendra, R., Spina, D., Cavedon, L., & Verspoor, K. (2025). *Evaluating numeracy of language models as a natural language inference task.* Findings of the Association for Computational Linguistics: NAACL 2025, 8336–8361.

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B., & Pollak, T. A. (2025, July 11). *Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it)* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/cmy7n.v5

Ofcom. (2024, November 28). *Online Nation 2024 report*. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2024/online-nation-2024-report.pdf

OpenAI. (2023a, January 31). *New AI classifier for indicating AI-written text*. OpenAI. Tilgået d.28/08/2025 https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

OpenAI. (2023b, March 1). *Introducing ChatGPT and Whisper APIs*. OpenAI. Tilgået d.28/08/2025 https://openai.com/index/introducing-chatgpt-and-whisper-apis/

OpenAI. (2025a, February 27). *Introducing GPT-4.5*. OpenAI. https://openai.com/index/introducing-gpt-4-5/

OpenAI. (2025b, April 14). *Introducing GPT-4.1 in the API*. OpenAI. https://openai.com/index/gpt-4-1/

OpenAI. (2025c, April 16). *Introducing OpenAI o3 and o4-mini*. OpenAI. https://openai.com/index/introducing-o3-and-o4-mini/

OpenAI. (2025d, August 7). *Introducing GPT-5*. OpenAI. https://openai.com/index/introducing-gpt-5/

OpenAI. (n.d.-a). *Backward compatibility*. OpenAI Platform. Tilgået d.26/05/2025. https://platform.openai.com/docs/api-reference/backward-compatibility

OpenAI. (n.d.-b). *GPT-3.5 Turbo*. OpenAI Platform. Tilgået d.25/05/2025. https://platform.openai.com/docs/models/gpt-3.5-turbo

OpenAI. (n.d.-c). *GPT-4.1.* OpenAI Platform. Tilgået d.26/05/2025. https://platform.openai.com/docs/models/gpt-4.1

OpenAI. (n.d.-d). *o3*. OpenAI Platform. Tilgået d.26/05/2025. https://platform.openai.com/docs/models/o3

OpenAI. (n.d.-e). *Reasoning guide*. OpenAI Platform. Tilgået d.26/05/2025. https://platform.openai.com/docs/guides/reasoning?api-mode=responses

OpenAI. (n.d.-f). *Reproducible outputs*. OpenAI Platform. Tilgået d.25/05/2025 https://platform.openai.com/docs/advanced-usage/reproducible-outputs

Ouyang, S., Wang, S., Liu, Y., Zhong, M., Jiao, Y., Iter, D., Pryzant, R., Zhu, C., Ji, H., & Han, J. (2023, October 19). *The shifted and the overlooked: A task-oriented investigation of user-GPT interactions* (arXiv preprint arXiv:2310.12418). arXiv. https://doi.org/10.48550/arXiv.2310.12418

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Rasmussen, J. (2025, June 10). *Ledere efterlyser massiv afprøvning af AI i gymnasiet – også som censors højre hånd*. *Gymnasieskolen*. Tilgået d.27/08/2025 https://gymnasieskolen.dk/articles/ledere-efterlyser-massiv-afproevning-af-ai-i-gymnasiet-ogsaa-som-censors-hoejre-haand/

Rupprecht, J., Ahnert, G., & Strohmaier, M. (2025, July 9). *Prompt perturbations reveal human-like biases in LLM survey responses* (arXiv preprint arXiv:2507.07188). arXiv. https://doi.org/10.48550/arXiv.2507.07188

Rystrøm, J., Kirk, H. R., & Hale, S. (2025). *Multilingual != Multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs* (arXiv:2502.16534v1). arXiv. https://arxiv.org/abs/2502.16534v1

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). *Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models.* In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 15295–15311). https://doi.org/10.18653/v1/2024.acl-long.816

Sadeghi, M., Dimitriadis, D., Arvanitis, L., Padovese, V., Pozzi, G., Badilini, S., Vercellone, C., Wang, M., Huet, N., Fishman, Z., Pfaller, L., Adams, N., & Wollen, M. (2025, May 5). *Tracking AI-enabled misinformation: 1,271 'unreliable AI-generated news' websites (and counting), plus the top false narratives generated by artificial intelligence tools*. NewsGuard. https://www.newsguardtech.com/special-reports/ai-tracking-center/

Sundhedsstyrelsen. (2025). *Analyse af ventetider til praktiserende psykiatere og børne- og ungdomspsykiatere: Analyseresultater*. https://www.sst.dk/-/media/Udgivelser/2025/Psykiatri/Analyse-af-ventetider-til-praktiserende-psykiatere-og-boerne--og-ungdomspsykiatere.ashx

Seddik, M. E. A., Chen, S.-W., Hayou, S., Youssef, P., & Debbah, M. (2024, April 7). *How bad is training on synthetic data? A statistical analysis of language model collapse* (arXiv preprint arXiv:2404.05090). arXiv. https://doi.org/10.48550/arXiv.2404.05090

Seleznyov, M., Chaichuk, M., Ershov, G., Panchenko, A., Tutubalina, E., & Somov, O. (2025, August 15). *When punctuation matters: A large-scale comparison of prompt robustness methods for LLMs* (arXiv preprint arXiv:2508.11383). arXiv. https://doi.org/10.48550/arXiv.2508.11383

Silver, L., Keeter, S., Kramer, S., Lippert, J., Hernandez Ramones, S., Cooperman, A., Baronavski, C., Webster, B., Nadeem, R., & Chavda, J. (2025, May 8). *Americans' trust in one another*. Pew Research Center. https://www.pewresearch.org/2025/05/08/americans-trust-in-one-another/

The New York Times Company v. Microsoft Corporation, No. 23-cv-11195 (S.D.N.Y. Apr. 4, 2025). https://www.nysd.uscourts.gov/sites/default/files/2025-04/yf%2023cv11195%20OpenAI%20MTD%20opinion%20april%204%202025.pdf

Tong, A., Wang, E., & Coulter, M. (2024, February 22). *Exclusive: Reddit in AI content licensing deal with Google*. Reuters. https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/

U.S. Energy Information Administration. (2025, July 24). *Electric Power Monthly: Table 6.6. Planned U.S. electric generating unit retirements* [Data set]. https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=table_6_06

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2022, October 26). *Will we run out of data? Limits of LLM scaling based on human-generated data* (arXiv preprint arXiv:2211.04325). arXiv. https://doi.org/10.48550/arXiv.2211.04325

Volcovici, V., & Groom, N. (2025, May 22). *House budget bill effectively halts US clean energy boom*. *Reuters*. Tilgået d.27/08/2025 https://www.reuters.com/sustainability/climate-energy/house-budget-bill-effectively-kills-us-clean-energy-boom-2025-05-22/

Wright, D., Arora, A., Borenstein, N., Yadav, S., Belongie, S., & Augenstein, I. (2024, October 31). *Revealing fine-grained values and opinions in large language models* (arXiv preprint arXiv:2406.19238). arXiv. https://doi.org/10.48550/arXiv.2406.19238

YouGov. (2025, April 25). *YouGov survey: AI uses* [Survey results]. https://d3nkl3psvxxpe9.cloudfront.net/documents/AI_Uses_poll_results.pdf

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2025). *A survey of large language models* [arXiv preprint arXiv:2303.18223]. arXiv. https://arxiv.org/abs/2303.18223

Østergaard, S. D. (2023). Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? Some additional considerations for clinicians. *Schizophrenia Bulletin, 49*(6), 1418–1419. https://doi.org/10.1093/schbul/sbad128