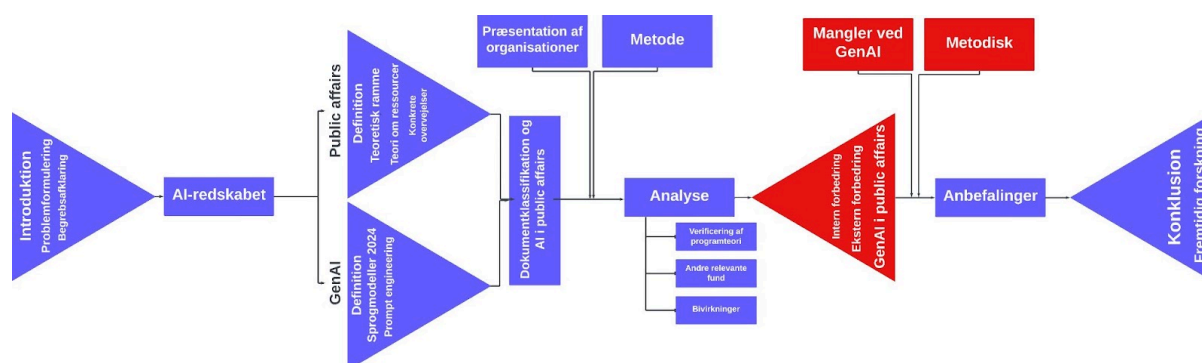


9.0 Implikationer

Først diskuteres, hvad fundene fra evalueringen medfører vedrørende AI-redskabet. Til dette kan man diskutere to ting: Hvordan man kan forbedre AI-redskabet til at lave flere korrekte klassifikationer af dokumenter, altså en forbedring af AI-redskabets ydeevne. Det kan man kalde *intern forbedring af AI-redskabet*. Alternativt kan man benytte GenAI helt anderledes og lave en form for database og bruge GenAI som opslagsværktøj til at gennemgå databaser, i dette tilfælde Folketingets procesdokumenter. Ved at lave en database kan organisationer efter egne behov og egne præmisser få information fra Folketingets procesdokumenter, hvilket afspejler, hvad SJ&K har benyttet AI-redskabet til (SJ&K1,5). Dette kan man kalde *ekstern forbedring* ved brugen af GenAI. Efter diskussionen om interne og eksterne forbedringer, diskuteres implikationerne for arbejdet med public affairs bredt set ved at have GenAI værktøjer til rådighed. Slutteligt diskuteres metodiske valg og hvad disse har af betydning for konklusionerne.



9.1 Intern forbedring af AI-redskabet:

For at udforske interne forbedringer er Simon Moe Sørensen, en data scientist specialiseret i GenAI, blevet inddraget. Han har lavet en vurdering af AI-redskabet og foreslået forbedringer, og påviser, at den underliggende sprogmodel i AI-redskabet har potentiale til at blive anvendt mere effektivt, end den er i AI-redskabet. Sørensen foreslår to primære forbedringer: Bedre prompt engineering og *fine-tuning* (bilag 28). Ud over hvad Sørensen har foreslået, er der to andre måder man kan lave interne forbedringer af AI-redskabet: Benytte en anden sprogmodel til at evaluere klassifikationen og en større tilpasning af AI-redskabet efter de enkelte organisationers behov.

9.1.1 Bedre prompt engineering

Sørensens påpeger, at AI-redskabet har for brede prompts, og at prompts skal deles mere op. Han foreslår: “Du separerer det (forskellige underemner inden for en organisation *red.*) ud i forskellige kategorier og beder sprogmodellen om for hvert emne at gå igennem hele din database flere gange, men med forskellige prompts” (bilag 28). Dette forslag til forbedring ligger i tråd med punkt 3 i prompt engineering fra afsnit 4.3., som var at dele komplekse opgaver ind i underopgaver. Promptet i AI-redskabet er allerede delt op efter organisationer og deres fokusområder, hvor for eksempel Dansk Standard har separate prompts for byggeri, energi og standarder, men den opdeling skal gøres endnu finere.

Fremadrettet kan prompts, på baggrund af Sørensens forslag, opdeles yderligere, eksempelvis med et prompt for Dansk Standard Energi, der kun klassificerer efter PtX, et prompt der kun klassificerer efter bæredygtige brændstoffer osv (bilag 28). Dette hjælper med at tilpasse klassifikationsprocessen og tager højde for sprogmodellens begrænsede evne til at behandle information på tværs af forskellige kontekster. Sørensens anbefaling er yderligere kvalificeret ved at redskabet allerede fungerer bedst i veldefinerede og specifikke kontekster som hos Danske Gymnasier, Techvirksomheden og Zero Carbon Shipping. Ved at gøre promptet endnu mere veldefineret og specifikt, kan AI-redskabet skabe endnu mere præcise og tilpassede klassifikationer.

9.1.2 Fine-tuning af modellen

Fine-tuning er, at man ændrer den underliggende sprogmodels struktur ved at ‘gentræne’ den med et mindre og klart afgrænset datasæt. I tilfældet med AI-redskabet kan det ske ved, at man tager Folketingets procesdokumenter og selv laver en klassifikation af mange dokumenter og bruger det som datasæt. På den måde bliver sprogmodellen trænet til at forstå, hvad en given organisation er interesseret i (bilag 28).

Mens bedre prompt engineering er en mere overkommelig og mindre omkostningsfuld løsning, er fine-tuning en mere grundlæggende og omkostningstung tilgang, da man skal lave flere end 50 eksempler for hver organisation, men det er ikke givet, at fine-tuning nødvendigvis medfører et bedre output (bilag 28). Litteraturen inden for AI engineering peger dog på, at fine-tuning kan løse mange problemer med hallucinationer, herunder potentielt problemet med de for mange falske positiver i AI-redskabet (Martino et al., 2023). Af de to er

bedre prompt engineering derfor den mest oplagte interne forbedring af AI-redskabet på kort sigt af de to.

9.1.3 Få en anden sprogmodel til at vurdere om klassifikationen er korrekt

En måde man kan øge AI-redskabets ydeevne på, ud over hvad Sørensen foreslår, er ved at benytte en anden sprogmodel til at evaluere AI-redskabets klassifikationer. Det kan skabe en kvalitetssikring af klassifikationerne af dokumenterne. Helt konkret kan man prompte en separat sprogmodel til at vurdere, om en klassifikation er korrekt. Et simplificeret prompt til den separate sprogmodel er: “En sprogmodel har vurderet om hvorvidt et politisk dokument er relevant for organisation x, som er interesseret i y. Du skal vurdere om sprogmodellen har ret. Sagen: **Information om sagen**. Klassificeringen fra sprogmodellen: **Klassificeringen**.” Man kan derefter designe AI-redskabet til at medtage den opdaterede klassifikation frem for den originale klassifikation, hvis der er uoverensstemmelser.

9.1.4 Større tilpasning af AI-redskabets brugerflade

En fjerde forbedring kunne være en større tilpasning af AI-redskabet. AI-redskabet har generelt været udformet som en one-size-fits-all løsning, men med en bias mod Danske Gymnasier, jf. afsnit 6.3.1, hvor det nævnes, at jeg har en tilknytning og viden om, hvordan Danske Gymnasiers arbejdsgange er. Givet at SJ&K allerede bruger AI-redskabet som en database, og at Dansk Standard ønsker opdateringer hver uge frem for hver dag, samt at alle fire organisationer ønsker forskellige forbedringer af brugerfladen, antyder det, at der kunne være potentiale for større succes, hvis AI-redskabet var mere tilpasset brugernes specifikke behov. En tilpasning kunne for eksempel involvere ændringer i monitoreringsrapporten, frekvensen af udsendelse, integrering i CRM-systemer eller lignende.

Disse ændringer er ikke blevet implementeret tidligere, primært fordi omkostningerne ved at udvikle og teste AI-redskabet har været for høje. Som nævnt, har hvert dokument, som AI-redskabet klassificerer, en tilknyttet omkostning, hvilket har lagt økonomiske begrænsninger på projektet. Fremadrettet ses tendenser til, at omkostningerne falder for brugen af sprogmodeller. Udviklingen af open-source sprogmodeller er i gang med at reducere omkostningerne betydeligt. For eksempel lancerede Meta i midten af april deres open-source LLAMA 3-model, som er på niveau med GPT3.5. I nærmeste fremtid har Meta annonceret, at de vil lancere LLAMA 3 400b-modellen, som forventes at være på niveau med

GPT-4, den førende model på markedet og som AI-redskabet benytter (Meta, 2024). Da disse modeller er open source, hvilket betyder, at de er offentligt tilgængelige, er omkostningerne ved at anvende dem markant lavere, hvilket sænker barriererne for at implementere flere forskellige prompts og udvide anvendelsen af GenAI. Derudover har OpenAI den 13. maj 2024 lanceret en ny og bedre model end GPT-4, som de kalder GPT-4o, som koster 50% mindre end GPT-4, som er anvendt i AI-redskabet. Det betyder, at man alt andet lige vil kunne lave dobbelt så mange prompts til samme pris i AI-redskabet (OpenAI, 2024).

9.2 Ekstern forbedring ved brugen af GenAI

AI-redskabet har fundamentalt fulgt den traditionelle tilgang til at få information fra Folketinget, som er via et nyhedsbrev, mens AI-redskabet blot har benyttet mere sofistikeret teknologi i form af GenAI. Men GenAI tillader helt nye måder at interagere med data på, hvilket rejser spørgsmålet om, hvorvidt den traditionelle tilgang stadig er den bedste måde at få informationer på.

GenAI har vist sig i stand til at klassificere relevante dokumenter for specifikke fagområder til en vis grad, som det ses i kasse to i programteorien. Selvom der er en del falske positive, opstår spørgsmålet: hvad nu hvis man kunne designe en GenAI-løsning, hvor falske positive ikke længere er et problem? Både SJ&K og Zero Carbon Shipping foreslår, at en database-baseret tilgang kunne være mere hensigtsmæssig for at håndtere parlamentariske procesdokumenter som type af data, hvor dataen bliver udtrukket af brugerne efter behov frem for at blive skubbet ud til dem af AI-redskabet i faste intervaller (SJ&K1; ZCS2). Danske Gymnasier og Dansk Standard blev præsenteret for samme ide, og de mener at det lyder som et nyttigt værktøj (DG3; DS2).

Det kunne man gøre via en teknologi som hedder en Retrieval-Augmented-Generator-model (RAG-model) (Merrit, 2023). En RAG-model kombinerer evnerne fra sprogmodeller med effektiv dataudvinding, hvilket tillader mere præcis og relevant informationsudtræk. En RAG-model fungerer ved at lave tekstoutputs via GenAI med indhold, der hentes direkte fra en relevant database eller et datasæt, hvilket også muliggør at få kilder på hvor information er hentet fra, hvilket øger tilliden. Det gør det muligt for en RAG-model at trække på en stor mængde af information, når den genererer svar, hvilket forbedrer både nøjagtighed og relevans af de genererede svar (ibid).

Helt konkret kunne RAG-modellen designes til at trække på præcis samme dokumenter som AI-redskabet gør og via et interface som ChatGPT vil man kunne spørge efter “hvad er seneste inden for elevfordeling?” eller “hvem har stillet spørgsmål om PtX?”. Det vil gøre GenAI-løsningen mere fleksibel og undgå problemet med falske positive, da der alligevel skal foregå en sortering og rangering af outputs fra RAG-modellen af brugerne.

Argumentet om at en RAG-model vil skabe værdi er ikke bare spekulativt og teoretisk. SJ&K har allerede anvendt det nuværende AI-redskab som en proxy for en RAG-Model jf afsnit 8.4.2 (SJ&K1,5). Siden at de på trods af designet af AI-modellen har benyttet den som en proxy RAG-model, er det et tegn på, at der er potentiale i at designe en RAG-model. Derudover viste jeg en demo af en RAG-model til Zero Carbon Shipping i anden runde af interviews for at vise hvordan en RAG-model fungerer og Zero Carbon Shipping så et stort potentiale i at benytte en løsning som den i deres public affairs arbejde (ZCS2).

Omvendt har en RAG-model en fundamental anderledes funktion end AI-redskabet, da brugeren aktivt skal opsøge information, og ikke får information leveret direkte til dem. Både Danske Gymnasier, Dansk Standard og Zero Carbon Shipping har sagt at en kombination af en løsning som AI-redskabet og en RAG-model er fordelagtig, men en RAG-model kan ikke erstatte en løsning som AI-redskabet (DG3, DS2, ZCS2). Det er vigtigt at have in mente når man skal konkludere hvordan GenAI kan forbedre organisationers arbejde med public affairs. Valget om at inkludere diskussionen om en RAG-model er blevet truffet fordi både SJ&K og Zero Carbon Shipping har omtalt en GenAI drevet database som nyttig uafhængig af hinanden og på egen drift, hvorfor det er relevant at undersøge og diskutere.

9.3 Implikationerne for public affairs ved at have GenAI værktøjer til rådighed

Uanset hvordan fremtidige GenAI-løsninger udvikles med henblik på at forbedre public affairs arbejdet, er der flere implikationer, der kan udledes baseret på den konceptuelle kontekst:

9.3.1 Bedre mulighed for politisk monitorering og forbedring af katalytiske kapaciteter

Hvis en GenAI-løsning, enten et forbedret AI-redskab eller en RAG-model, gør det lettere at være opdateret på relevante dokumenter fra Folketinget, betyder det, at en organisation lettere kan være ajour med udviklinger og tendenser. Som diskuteret i afsnit 3.3 fungerer denne tilgang som hjælp til politisk monitorering (Hegelund & Mose, 2013). Dette er afgørende i public affairs, hvor tidlig erkendelse af relevante informationer kan have betydning for public affairs-arbejdet (Esbensen, 2007). Implementeringen af GenAI i public affairs viser, hvordan moderne teknologi kan styrke organisationers evne til politisk monitorering og generelt øge deres katalytiske kapaciteter. Det understøttes af Obermans (2008) teori om, at styrkelse af organisationers strategiske kapaciteter gennem effektiv brug af værktøjer, i dette tilfælde GenAI, kan medføre mere effektiv public affairs.

GenAI kan designes til at forstærke de katalytiske kapaciteter ved at automatisere indsamling og klassifikation af politiske dokumenter, hvilket potentielt set placerer public affairs organisationer i en position for tidlig indgriben og reaktion, som de ellers ikke havde været i. Fra analysen af AI-redskabets ydeevne i praksis observeres det, at AI-redskabets kapacitet til at levere opdateret og relevant information kan påvirke organisationernes mulighed for at være mere ajour med information fra Folketinget og have et større vidensniveau, på trods af AI-redskabets begrænsede ydeevne.

Derudover er en implikation for public affairs, at når et GenAI redskab kan klassificere dokumenter fra Folketinget, kan GenAI med al sandsynlighed også klassificerer andre typer af tekstdokumenter, som yderligere vil forstærke organisationers katalytiske kapaciteter (Taubner & Weinhardt, 2023; Johansen, 2024). Et fremtidigt GenAI-redskab kunne altså udvides til også at omfatte rapporter, budgetter, omfattende love og reguleringer eller andre teksttunge datakilder. Hvis man kan få rettighederne til det, kan et GenAI-redskab også monitorere sociale medier og traditionelle medier. Denne tilgang ville øge mængden af information på tværs af forskellige kilder, hvilket kunne forbedre den måde, public affairs organisationer forstår og interagerer med information på og dermed øge de katalytiske kapaciteter.

Det er væsentligt at anerkende, at mens AI-redskabet har styrket visse katalytiske ressourcer, påpeger analysefundene også områder, der kræver forbedring, specielt i forhold til præcision og sensitivitet i klassifikationen. Denne afvejning mellem præcision og sensitivitet påvirker direkte kvaliteten af de katalytiske ressourcer ved at påvirke beslutningstagningens kvalitet. For højt fokus på sensitivitet fører til mange falske alarmer (falske positive), hvilket kan føre til ressourcespild, mens for højt fokus på præcision kan resultere i, at vigtige dokumenter overses (falske negative), hvilket potentielt kan koste strategiske muligheder. Der skal tages højde for disse udfordringer hvis public affairs organisationer ønsker at benytte et fremtidigt GenAI-redskab. Det vil sige, at det skal sikres, at de teknologiske forbedringer ikke blot løser eksisterende problemer, men også integreres effektivt i organisationens public affairs arbejde og strategi.

9.3.2 Pluralismen i public affairs

I overensstemmelse med det observerede skift fra korporatisme til pluralisme i dansk public affairs, kan GenAI værktøjer vise sig at være nyttige til at navigere i den øgede konkurrence om beslutningstagernes opmærksomhed. I en pluralistisk kontekst, hvor public affairs organisationer kæmper indbyrdes om at præge den politiske dagsorden, bliver adgang til rettidig og relevant information stadig vigtigere (Hegelund & Mose, 2013).

Anvendelsen af GenAI kan ses som en direkte reaktion på behovet for at styrke organisationers kapacitet til politisk monitorering for dermed at kunne levere bedre argumenter og dokumentation baseret på mere viden. Med sin evne til kontinuerligt at opdatere og klassificere politisk information kan GenAI-redskaber hjælpe organisationer med at identificere og agere på politiske sager, der matcher deres interesseområder, potentielt hurtigere og mere effektivt end før, og hurtigere end konkurrerende organisationer. Dette er vigtigt i den pluralistiske tradition, hvor tidsfølsomme og velinformede bidrag til den politiske debat potentielt kan afgøre, om en organisation formår at påvirke politiske beslutninger til egen fordel. Dette gør GenAI til et aktiv i pluralismen, hvor kapaciteten til hurtigt at optage og reagere på nye oplysninger kan have betydning for, hvilke organisationer lykkes med at sætte dagsordenen og forme politiske udfald.

9.3.3 Lavere barriere for at få information fra Folketinget

En række forskere finder, at ressourcer er en central faktor i public affairs (Christiansen & Nørgaard, 2003; Binderkrantz, Christiansen & Pedersen, 2014). For eksempel finder Crosson, Furnas og Lorentz (2020), at de public affairs organisationer, der besidder store finansielle ressourcer, ofte har bedre vilkår for at påvirke politikere. Dette rejser et ulighedsspørgsmål, siden en forestilling om public affairs er, at det er en mekanisme til at repræsentere diverse interesser inden for demokratiske samfund, idet public affairs muliggør deltagelse i demokratiet, hvor flere grupper kan påvirke politiske beslutninger inden for det politiske systems begrænsninger (Hegelund & Mose, 2013). Dog kan der rejses bekymringer om ulige repræsentation og potentialet for, at visse grupper dominerer politikudformningen, hvis de har flere ressourcer end mindre ressourcestærke grupper, hvilket er en mulighed i den pluralistiske public affairs tradition (Crosson, Furnas & Lorentz, 2020).

En implikation af at benytte GenAI i public affairs er, at GenAI kan reducere barriererne for at opnå relevant information fra Folketinget ved at benytte GenAI som ressource. Der er potentielt nogle organisationer uden stærke public affairs kapaciteter, som kunne drage fordel af et fremtidigt GenAI-redskab. Det kunne for eksempel være public affairs organisationer med små sekretariater, der ikke har ressourcer til at følge med i den store mængde af informationer fra Christiansborg.

Omvendt gælder det, at jo mere sofistikeret en GenAI løsning er, desto dyrere bliver den, ligesom med ethvert andet produkt eller løsning på kort sigt. Det kan derfor risikeres at når GenAI modnes, vil public affairs organisationer, der allerede har stor indflydelse og mange ressourcer, ende med at have adgang til de bedste og mest præcise GenAI-værktøjer. I mellemtiden må mindre public affairs organisationer nøjes med mindre præcise værktøjer. Det er noget som det datadrevne lobby-software-løsning firma Fiscal Note er kritiseret for: De er dyre og derfor mener nogle, at det er et demokratisk problem at ressourcestærke organisationer har bedre adgang til information frem for mindre ressourcestærke organisationer (Zaleski, 2018). Dette kan føre til yderligere polarisering af kapaciteter mellem public affairs organisationer, hvor større organisationer får præcis information hurtigt, mens mindre organisationer fortsat vil have en udfordring med at finde relevant information.

9.4 Mangler ved GenAI

De fordele, som tidligere er blevet nævnt, bygger på antagelsen om, at en fremtidig GenAI-løsning fungerer effektivt og er i stand til at identificere relevante dokumenter. Der er dog en risiko for, at GenAI og sprogmodellerne ikke lever op til denne præmis og måske aldrig vil gøre det, enten teknologimæssigt eller i en dansk kontekst. Set fra et ressourceperspektiv kunne det derfor muligvis være mere fordelagtigt at investere i de mere traditionelle aspekter af public affairs, såsom netværksopbygning og at sikre et effektivt flow af viden fra public affairs agenter til beslutningstagere (Esbensen, 2007). Der kunne være større værdi i at allokere katalytiske ressourcer til disse elementer af public affairs frem for at forsøge at finde potentielt relevante dokumenter, hvor kun en lille andel ender med at blive anvendt i public affairs arbejdet.

Det er også vigtigt at anerkende, at GenAI og sprogmodeller ikke nødvendigvis kan identificere de sjældne eller uventede pointer i tekstdataen, som findes i halerne af en normalfordeling som en sprogmodel fundamentalt er baseret på, da GenAI er typisk trænet til at outputte mod midten af en normalfordelingen af data (Young, 2024). Dette er ikke optimalt i public affairs, hvor det kan være afgørende at kunne isolere en specifik sætning eller detalje, som kan være afgørende for et argument. I public affairs er det ofte essentielt at 'finde djævelen i detaljen', som GenAI ikke er særligt god til, da GenAI er designet til at forstå bredere mønstre og tendenser i sproget (Young, 2024).

Selv hvis GenAI er i stand til at fange outlier-data og alt relevant information, er det ikke sikkert, at det vil forbedre effektiviteten i public affairs-arbejde markant. Der er en reel mulighed for, at der simpelthen ikke er nok relevante dokumenter tilgængelige, eller de dokumenter, der er tilgængelige ligger for sent i issue-cyklussen, da dokumenterne der kan blive klassificeret af et GenAI-redskab per definition allerede er online og bearbejdet af aktører (Esbensen, 2007).

På den anden side er der bare under denne specialeperiode blevet lanceret Meta's open source Llama3, OpenAI's nye sprogmodel GPTo, der er bedre end GPT4, og 28. maj kom det frem, at der er udviklet en DanskGPT, der er bedre end GPT4o i fire ud af syv danske sprogopgaver (Meta, 2024; OpenAI, 2024; Henrichsen, 2024), som peger i retningen af at GenAI med tiden vil blive bedre til at klassificere politiske dokumenter.

Derudover mener alle fire organisationer i dette speciale, at ideen bag AI-redskabet er nyttig og har potentiale til at skabe værdi. Dette tyder på, at der kunne være en fordel i enten at forbedre AI-redskabet eller teste de nye sprogmodeller, så GenAI kan være til nytte i public affairs. .

9.5 Metodiske diskussioner

9.5.1 Casetilgangen og valget af de fire organisationer

I dette speciale er der blevet anvendt en casetilgang, fordi formålet var at undersøge hvordan et specifikt AI-redskab kan bruges til at forbedre public affairs, hvor fire forskellige organisationer benyttede AI-redskabet med henblik på at der kunne foretages en evaluering. Man kunne i stedet have fokuseret på kun én organisation og tilpasset AI-redskabet præcist til deres specifikke behov. Et tættere samarbejde med en enkelt organisation omkring designet af AI-redskabet og tilhørende monitoreringsrapport kunne muligvis have øget chancen for at undgå implementeringsfejl og ført til bedre resultater. Man kunne ligeledes have fokuseret mere på prompt engineering og potentielt reduceret antallet af falske positive, hvilket ville have resulteret i en monitoreringsrapport uden meget støj fra falske positive. Det er vigtigt at huske, at det har været et bevidst valg at designe AI-redskabet som en ‘one-size-fits-all’ løsning med justeringer i enten prompt og frekvens, idet jeg har forsøgt at udforske en løsning, der teknisk er let at implementere bredt og med lave tekniske barrierer. Med andre ord er valget om at inddrage fire organisationer frem for blot én organisation en afvejning mellem intern og ekstern validitet, hvor stræben efter ekstern validitet er kommet på bekostning af intern validitet (Andersen, 2012).

9.5.2 Alternative teoretiske tilgange

En alternativ tilgang til dette speciale kunne have været ved at konstruere et lignende GenAI drevet værktøj, der udelukkende lavede klassificeringer af politiske dokumenter, men uden at teste det i organisationer. Dette ville betyde, at undersøgelsen alene ville være en skrivebordsundersøgelse af GenAIs potentiale. På den måde havde der været mere tid til at teste forskellige versioner af prompts og finjustere prompts yderligere. Det ville dog være en uhensigtsmæssig tilgang til at besvare problemformuleringen af en række grunde: For det første vil man ikke have adgang til viden fra organisationerne om hvad der reelt er relevant

for dem, ud over hvad man kan læse sig til. For det andet er det kun hvis organisationerne mener at en GenAI løsning er nyttig, at den er nyttig. Ideen med at bruge GenAI til at klassificere parlamentariske procesdokumenter er ikke verificeret i forskningen, selvom en række firmaer bruger andre teknologier til at løse samme problem. Derfor er det kun ved at inddrage organisationerne, at man kan validere ideen om at bruge GenAI i public affairs. For det tredje skal det huskes, at i standardtilgangen ved at benytte nyhedsbreve fra Folketinget er der også mange ‘falske positive’. For eksempel, hvis man følger Børne- og Undervisningsudvalget og er interesseret i gymnasiesektoren, er der stadig mange dokumenter i nyhedsbrevet, der omhandler folkeskolen. Hvis mængden af falske positive kan forblive den samme ved at gå fra nyhedsbrevstilgangen til en GenAI-tilgang, men der er flere sande positive ved GenAI-tilgangen, vil public affairs organisationerne – alt andet lige – finde mere nytte ved et GenAI-værktøj. Denne sontring vil man ikke kunne finde ved at lave en skrivebordsundersøgelse af GenAI.

9.5.3 Andre metodiske kritikpunkter

9.5.3.1 *Selektionsbias*

I dette speciale er der en fundamental selektionsbias, da alle de organisationer, der har hjulpet med evalueringen, har jeg på en eller anden måde været tilknyttet til og alle organisationerne har alle vist en interesse for GenAI. Det kunne have været interessant at udvide undersøgelsen til at omfatte organisationer, som oprindeligt er skeptiske over for brugen af GenAI, for at få et mere nuanceret billede af AI-redskabets påvirkning og modtagelse samt at udvide undersøgelsen til organisationer, som jeg ikke har en tilknytning til. Dette kunne have mitigeret selektionsbias og potentielt have fremskaffet mere robuste resultater.

9.5.3.2 *Manglede robuste evalueringsværktøjer af sprogmodeller*

Evaluerings af sprogmodeller er en udfordrende opgave, fordi der mangler robuste og standardiserede evalueringsværktøjer af sprogmodellers performance og svar (Huang, 2024). Dette medfører at der manuelt skulle evalueres hvert enkelt klassifikation, om end i stikprøver, for at sikre, at AI-redskabet klassificerede alle relevante dokumenter korrekt. Denne manuelle proces er en potentiel fejlkilde. Når jeg ikke har fuld indsigt i alle organisationernes arbejde, er det en udfordring at vurdere, hvad der er relevant for de enkelte

organisationer. Dette er forsøgt mitigeret ved at inddrage kategorier for edgecases og have dialog med organisationerne om, hvad der er relevant for dem.

9.5.3.3 *Kort periode for evaluering*

AI-redskabet har ikke været implementeret i en længerevarende periode, hvilket begrænser mængden af data tilgængelig for evaluering og den reelle nytteværdi, AI-redskabet har skabt. Det er på trods af, at Danske Gymnasier har sagt, at de tror at AI-redskabet vil identificere dokumenter, de ikke selv havde identificeret, over tid.

9.5.3.4 *Manglende kvantitative benchmarks og metrics*

Den manuelle klassificering af dokumenter er ikke sat op imod foruddefinerede benchmarks. Et benchmark kunne have været at sammenligne AI-redskabets klassificeringer med, hvad de forskellige organisationer ville have modtaget i deres respektive nyhedsbreve fra Folketinget. Ved ikke at benytte benchmarks, er kvantitativt datadrevet konklusioner per definition udelukket til at vurdere AI-redskabets evne til at klassificere dokumenter, hvilket går ud over den metodiske robusthed af specialet. Omvendt argumenteres det, at foruddefinerede benchmarks vil være arbitrært satte og ikke bidrage til vurderingen af AI-redskabet, da det handler om organisationernes syn på, hvorvidt AI-redskabet er god til at klassificere.

9.5.4 *Udvikler-bias*

Den mest markante metodiske udfordring i dette speciale har været udvikler-bias (Chattopadhyay, 2022). Som udvikler af AI-redskabet har jeg visse forudindtagelser, og selvom jeg har bestræbt mig på at være så transparent som muligt ved at dybdegående beskrive mit forskningsdesign, metode og data, vil der altid være en vis bias til stede (ibid). ‘Heldigvis’ har jeg fundet negative resultater, hvilket - ironisk nok - kan bidrage til specialets troværdighed. Hvis kun positive resultater var rapporteret, kunne det rejse tvivl om objektiviteten. Derfor er inklusionen af negative resultater ikke kun en ærlig fremstilling, men også et skridt mod at styrke troværdigheden af dette speciale, idet formålet er at fremme viden inden for hvordan GenAI kan benyttes i public affairs arbejde snarere end at promovere AI-redskabet. Dette er vigtigt, da meningen med dette speciale er, at public affairs organisationer, og andre interesserede aktører kan få indsigt i, hvor moden GenAI-teknologien er, hvis de overvejer at implementere eller udvikle lignende værktøjer.

