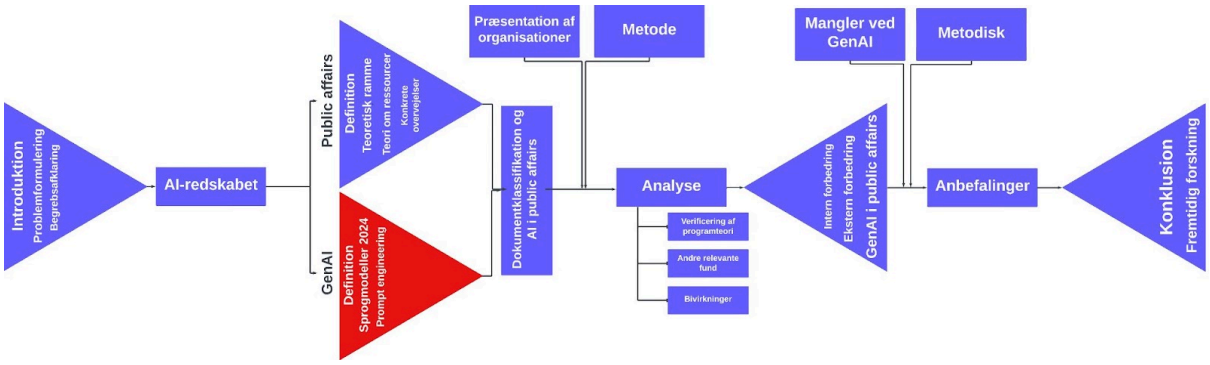
4.0 Konceptuel kontekst for GenAI og sprogmodeller

Den næste del af den konceptuelle kontekst handler om GenAI. Først defineres GenAI og sprogmodeller, hvorefter der sker en kortlægning af landskabet for sprogmodeller for foråret 2024. Det leder hen til et teknisk afsnit omhandlende en teknik i brugen af sprogmodeller, der hedder *prompt engineering*.



4.1 Definition og afgrænsning af GenAI

GenAI er en teknologi, hvor computermodeller kan skabe og outputte nyt indhold via træning på eksisterende datasæt (Johansen, 2024; Taubner & Weinhardt, 2023). Det betyder at GenAI kan outputte tekst, billeder, lyde, videoer eller lignende baseret på brugernes prompts, som er den tekst, man i menneskeligt sprog 'beder' eller 'spørger' computeren om at gøre. Det, der derfor er interessant og nyt ved GenAI, er, at man kan have en interaktion med AI på et menneskeligt sprog, hvor outputtet også er i menneskeligt sprog eller et format, som mennesker kan forstå (Johansen, 2024). Det vil sige, at man ikke længere skal kunne programmere og kode for at interagere med sofistikeret teknologi, hvilket gør det lettere tilgængeligt. Det står i modsætning til *prædiktiv AI*, som er den 'traditionelle' brug af AI, hvor outputtet oftest er mindre intuitivt at forstå, da outputtet fra prædiktiv AI modeller oftest er sandsynligheder eller sekvenser af tal. Det er den type af AI, der er baseret på træningsdatasæt og valideringsdataset, som kræver matematiske beregninger (Taubner & Weinhardt, 2023). Det er for eksempel prædiktiv AI som sociale medier bruger til at personalisere indhold baseret på forudsigelser af brugeres adfærd.

GenAI er et felt i hurtig udvikling, hvor nye fremskridt og anvendelser opstår løbende. Den voksende kapacitet af GenAI-teknologier til at forstå og generere menneskeligt sprog åbner for nye muligheder for at arbejde med teksttung data, optimere arbejdsprocesser, og tilvejebringe nye indsigter gennem dataanalyse, herunder dokumentklassifikation (ibid).

På trods af en række forskellige teknologier inden for GenAI koncentrerer dette speciale sig om teknologien sprogmodeller. Dette valg motiveres primært af, at sprogmodeller er den teknologi, der kan forstå og outputte tekst. Derudover er sprogmodeller avancerede, brugbare, billige, let tilgængelige og i høj grad en teknologi som mange erhverv kommer til at benytte fremadrettet (ibid). ChatGPT er for eksempel en sprogmodel.

4.2 Sprogmodeller

En sprogmodel er en type af GenAI, der har til formål at generere og forstå menneskeligt sprog på en måde, der efterligner menneskelig kommunikation. Sprogmodeller er baseret på komplekse algoritmer og store mængder data, som de er trænet på. Grundlæggende fungerer en sprogmodel ved at analysere og behandle tekstdata. Denne evne opnår den ved at blive præsenteret for store mængder tekstdata. Herfra lærer den sandsynlighederne for, at ord fremtræder efter hinanden eller i samme tekst givet en bestemt kontekst. Dette gør det muligt for sprogmodellen at generere tekst, der er både sammenhængende, meningsfuldt og kontekstuelt relevant via sprogmodellens evne til at forstå semantikker (Teubner & Weinhardt, 2023).

To centrale egenskaber ved sprogmodeller er, at de er *autoregressive* og *stokastiske*. Autoregressiv betyder, at sprogmodellerne beregner sandsynligheden for det næste ord i en sætning baseret på de foregående ord, mens stokastisk betyder, at deres output ikke er deterministisk, men sandsynlighedsbaseret (Murgia, 2023). For eksempel, i sætningen "Danmarks statsminister i 2023 var", kan en sprogmodel forudsige, at det mest sandsynlige næste ord er "Mette" og det mest sandsynlige ord efter dette er “Frederiksen” baseret på den læring, sprogmodellen har opnået under sin træning. Denne egenskab gør sprogmodeller effektive til opgaver som tekstgenerering og klassifikation. Det er vigtigt at forstå at sprogmodeller er autoregressive og stokastiske, da dette ligger til grund for deres evne til at skabe sammenhængende og logisk sekventielt output. Samtidigt forklarer det, hvorfor modellerne kan *hallucinere,* som er et fagbegreb, der betyder, at en sprogmodel kan generere

noget, der er fejlagtigt eller faktuelt forkert (Johansen, 2024). Hallucinationen er et resultat af sprogmodellers stokastiske opbygning, hvor sandsynlighedsbaseret valg kan lede til uventede og upræcise resultater. Dette kan mitigeres via en teknik der hedder *prompt engineering*, som introduceres i afsnit 4.3.

Et yderligere lag af kompleksitet i konteksten af sprogmodeller vedrørende klassifikation af dokumenter på dansk er, at sprogmodeller ikke er ligeligt trænede på alle sprog. Sprogmodellerne har en tendens til at præstere bedre på engelsk, grundet at deres træningsdata primært stammer fra internettet, som i overvejende grad er på engelsk, og at de fleste benyttede sprogmodeller er designet af amerikanske firmaer. Dette kan medføre en begrænset forståelse for mindre sprog som dansk og danske kontekster (Kehlet, 2024).

4.2.1 Versioner af sprogmodeller

Der er mange forskellige sprogmodeller. Den mest kendte er ChatGPT som er udviklet af firmaet OpenAI, men der findes også Gemini som er udviklet af Google, Llama3 som er udviklet af Meta, Opus som er udviklet af Anthropic, Le Chat som er udviklet af Mistral og mange flere. Sprogmodeller kan rangeres på forskellige benchmarks alt efter hvor gode de er til forskellige opgaver, herunder *reasoning*, som er sprogmodellers evne til at forstå komplekse emner og give meningsfulde og korrekte svar (Basal, 2023). Inden for hver sprogmodel kan der findes forskellige versioner, som for eksempel ChatGPT3.5 og ChatGPT4, hvor 3.5 versionen er billigere, men også performer dårligere på en række benchmarks ift ChatGPT4. I starten af dette speciale var den bedste sprogmodel GPT4, som også er den der bliver brugt i AI-redskabet (Vellum.ai, 2024).

Generelt bliver der ofte lanceret nye sprogmodeller, eller eksisterende sprogmodeller bliver forbedret. Trenden indtil videre er, at kvaliteten stiger, og at sprogmodellerne bliver billigere at bruge. Denne trend vil sandsynligvis fortsætte, hvilket blandt andet ses i den seneste udmelding om en alliance af offentlige og private aktører, der vil skabe en dansk sprogmodel. Desuden blev der undervejs i specialeperioden lanceret en ny sprogmodel, OpenAIs GPT-4o, som anses for at være den bedste sprogmodel i maj 2024. (Hays & Rafieyan, 2024; Kehlet, 2024; OpenAI, 2024; Vellum.ai, 2024).

4.3 Prompt engineering

Siden det er et menneske, der skriver promptet, altså kommandoerne eller instruktionerne til sprogmodellen, er der blevet udviklet en teknik til at maksimere sandsynligheden for at en sprogmodel giver et godt og præcist resultat. Denne teknik kaldes prompt engineering (platform.openai, 2023). Med andre ord er formålet med prompt engineering at styre sprogmodellens svar i den ønskede retning og forbedre både korrektheden og relevansen af de genererede svar (ibid).

Effektiv prompt engineering kræver en forståelse af, hvordan forskellige formuleringer af prompts påvirker sprogmodellens svar. Dette inkluderer valget af nøgleord, strukturering af spørgsmål og præcisering af kontekst, som kan hjælpe med at minimere risikoen for uønskede eller irrelevante svar. For eksempel kan et prompt, der er for åbent eller vagt, føre til, at modellen genererer bredere eller mindre specifikke svar, mens en mere detaljeret og specifikt formuleret prompt kan guide modellen til at producere mere fokuserede og detaljerede svar (ibid).

Der findes fem strategier for at optimere prompt engineering der er relevante for AI-redskabet:

1. *Skrive klare instruktioner:* Det kunne for eksempel være at bede sprogmodellen om at svare i et bestemt format som "Svar ja eller nej *og* begrund hvorfor".
2. *Few-shot learning:* Man skal ikke blot bede en sprogmodel om at klassificere dokumenter ud fra, hvad der er relevant for en given public affairs organisation. For eksempel specificeres det, at Danske Gymnasier er interesserede i ungdomsuddannelser, SU-forhold, elevtrivsel eller lignende. Alternativet er zero-shot learning, hvor man blot prompter sprogmodellen til at svare på om et dokument er relevant for Danske Gymnasier uden at specificere, hvad Danske Gymnasier er interesseret i, og lader det være op til sprogmodellen at vurdere hvad Danske Gymnasier er interesseret i baseret på sit træningsdata.
3. *Test ændringerne systematisk*: Siden en sprogmodel er autoregressiv og stokastisk, er det ikke altid tydeligt, hvordan et svar fra sprogmodellen ser ud. Derfor er det nødvendigt at teste og ændre promptet iterativt.
4. *Giv sprogmodellen roller:* Man kan for eksempel bede en sprogmodel om at påtage sig en rolle for at kvalificere svarene.Dette ses i prompten i AI-redskabet, hvor det specificeres, at sprogmodellen er ekspert i politik og har til opgave at klassificere dokumenter.

5. *Inkluder meget kontekst i prompt:* Jo mere kontekst og information man kan inkludere i sit prompt, jo større er sandsynligheden for at sprogmodellen ikke leverer hallucinerende svar, hvorfor alt data om dokumenterne bliver inkluderet i promptet (ibid).

Disse fem strategier bør benyttes, hvis man ønsker at få det bedste resultat fra en sprogmodel.