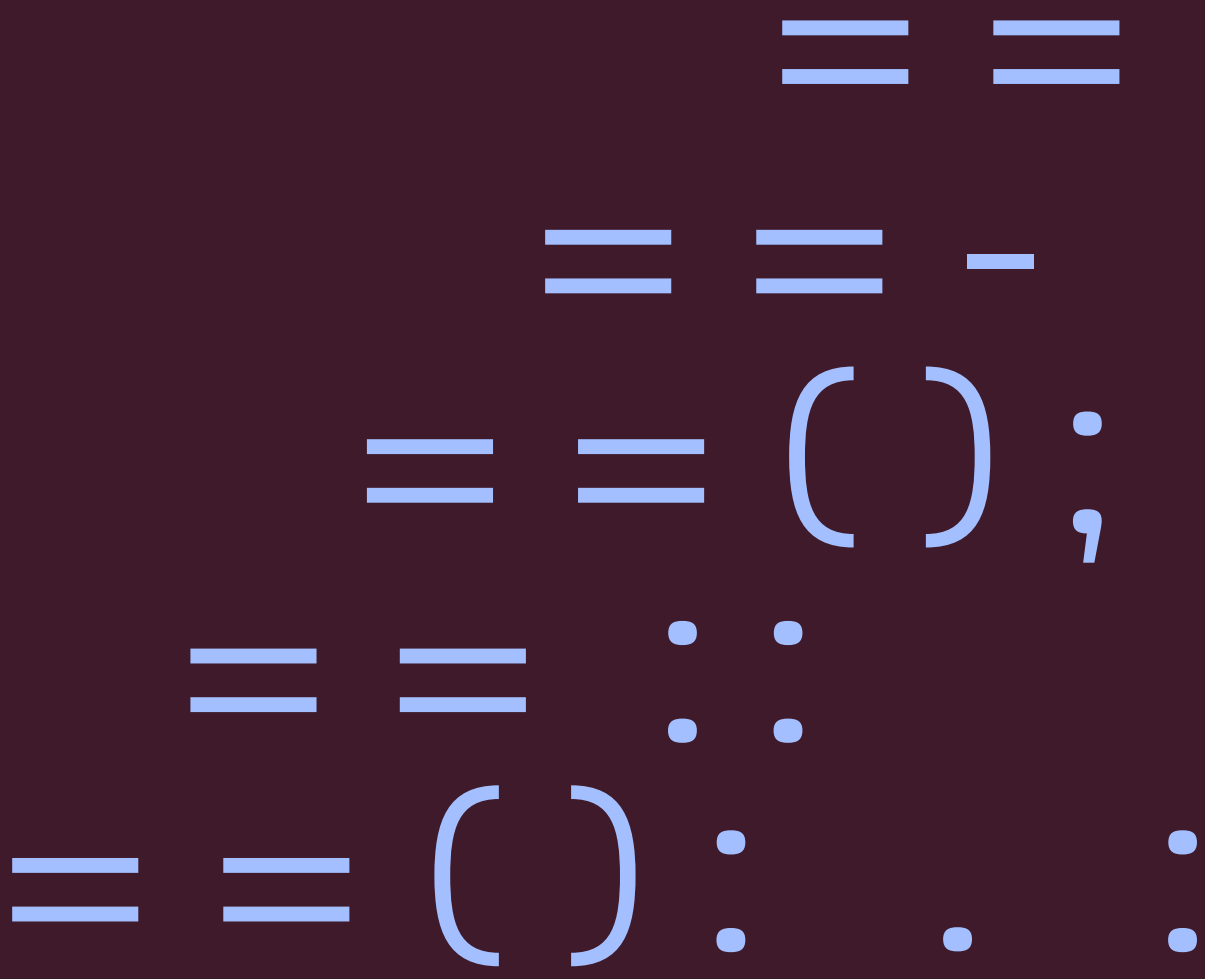
Digitaliserings-
styrelsen

Opsamling på workshop om

Evaluering og benchmarking af sprogmodeller på dansk



November 2024

Indhold

1. Introduktion og baggrund

Digitaliseringsstyrelsen varetager det fællesoffentlige initiativ *sprogteknologi.dk*, der har til formål at fremme udviklingen af dansksproget kunstig intelligens gennem tre områder:

- Videndeling om og udstilling af eksisterende danske sprogressourcer.
- Koordination omkring og udvikling af nye danske højkvalitets sprogressourcer.
- Fællesoffentligt, tværsektorielt og europæisk samarbejde om sprogteknologi.

Det politiske afsæt for sprogteknologi.dk findes i den tidligere Nationale strategi for kunstig intelligens fra 2019, Digitaliseringspagten og Økonomiaftalerne for 2020 mellem Regeringen, KL og Danske Regioner, hvori etableringen af en fælles dansk sprogressource blev vedtaget. Sidenhen har indsatsen for styrket dansk sprogteknologi også været indskrevet i nyere politiske aftaler.

Den politiske forankring af sprogteknologi.dk er i 'Styregruppen for digital innovation og grøn omstilling', der består af en række statslige myndigheder, KL og Danske Regioner. Digitaliseringsstyrelsen agerer sekretariat for sprogteknologi.dk.

Arbejdet for sprogteknologi.dk tager bl.a. udgangspunkt i nogle af de anbefalinger, som Sprogteknologiudvalget, nedsat af Kulturministeriet, fremsatte i rapporten "Dansk sprogteknologi i verdensklasse" i april 2019.

I 2023 arrangerede Digitaliseringsstyrelsen to workshops, hvor indsatsens fokusområder og målsætning blev diskuteret med interessentlandskabet for dansk sprogteknologi. På disse workshops blev det bl.a. efterspurgt, at Digitaliseringsstyrelsen hjælper med at skabe opmærksomhed omkring manglen på danske evalueringsdatasæt, som er nødvendige i arbejdet med at kvalitetssikre og benchmarke sprogmodellers præstationer på dansk. Evaluering og benchmarking af sprogmodeller er en forudsætning for dels at finde de bedst egnede løsninger til en given opgave, og dels for at målrette det videre arbejde med at udvikle konkurrencedygtige sprogmodeller.

Siden da har Digitaliseringsstyrelsen været i dialog med nogle af de danske aktører, som aktivt arbejder med evaluering og benchmarking af sprogmodeller på dansk for at komme behovene på området nærmere. Dette har ledt til, at Digitaliseringsstyrelsen i september 2024 afholdte en workshop om evaluering og benchmarking af sprogmodeller på dansk. Workshoppen blev afholdt med henblik på at invitere det bredere danske interessentlandskab for sprogteknologi med til en fælles afdækning af de videre behov på området.

2. Om workshoppen

Workshop om evaluering og benchmarking af sprogmodeller på dansk fandt sted fredag d. 20. september 2024 hos Digitaliseringsstyrelsen fra kl. 09:30 til 16:00. På workshoppen deltog omkring 40 personer fordelt over 30 forskellige virksomheder, myndigheder og forskningsinstitutioner. Invitationer var sendt ud til Digitaliseringsstyrelsens liste over interessenter på det sprogteknologiske område[1], og der var derudover delt en åben invitation på sprogteknologi.dk's hjemmeside samt LinkedIn.

Dagen var delt i to hovedsektioner. Første del havde til formål **at gøre status på større igangværende danske indsatser med evaluering og benchmarking af sprogmodeller**. Sessionen skulle dels bidrage til videndeling om de eksisterende indsatser, men også give mulighed for at deltagerne kunne stille spørgsmål og skabe et fælles forståelsesgrundlag til de videre drøftelser senere på dagen. Sessionen indeholdte følgende oplæg:

- Generel introduktion til evaluering og benchmarking af sprogmodeller og ScandEval (ved Dan Sattrup Nielsen, Alexandra Instituttet).
- Scandinavian Embedding Benchmark (ved Kenneth Enevoldsen, Aarhus Universitet).
- Danish Semantic Reasoning Benchmark (ved Bolette Sandford Pedersen, Københavns Universitet og Nathalie Hau Sørensen, Det Danske Sprog- og Litteraturselskab).
- Danoliterate (ved Søren Vejlgaard Holm, Danmarks Tekniske Universitet).
- Evaluering af sprogmodeller i Norge (ved Hans Christian Farsethås, Universitetet i Oslo).
- Evalueringsdatasæt for 1000 danske talemåder og faste udtryk (ved Nathalie Hau Sørensen, Det Danske Sprog- og Litteraturselskab).

Anden del, som var diskussionsbaseret, havde to formål. Med udgangspunkt i oplæggene skulle deltagerne først **kortlægge yderligere behov på området**. Kortlægningsopgaven blev vendt i plenum og dannede grundlag for et fælles billede af potentielle indsatser. Dette blev efterfølgende udgangspunktet for anden opgave, hvor deltagerne i fællesskab skulle **specificere en "backlog" med opgaver**, som bør løses for at hjælpe Danmark bedre på vej i arbejdet med at evaluere og benchmarke sprogmodeller.

Dagen blev rundet af med tid til networking, hvor deltagerne også fik mulighed for at snakke videre om eksisterende og potentielle projekter inden for evaluering og benchmarking samt dansk sprogteknologi mere generelt.

[1] Ønskes man tilføjet på denne liste, så man kan få lignende invitationer i fremtiden, er man velkommen til at rette henvendelse på info@sprogteknologi.dk.

3. Metode

I dette afsnit redegøres for designet af de opgaver, som deltagerne skulle løse på workshoppen. Udfaldet af opgaverne har dannet grundlag for afsnit 4 og 5 i denne opsamling.

Opsætning

Efter oplæggene i dagens første session blev deltagerne inddelt i grupper på ca. otte personer hver. Der var opstillet fem gruppeborde, og deltagerne fik mulighed for selv at danne grupperne. Oplægsholderne blev dog opfordret til at fordele sig, så der var én oplægsholder i hver gruppe. Opfordringen var med henblik på at sikre, at der var mindst én person til stede i hver gruppe, som havde praktisk erfaring med evaluering og benchmarking af sprogmodeller, og som dermed kunne hjælpe drøftelserne på vej, hvis der opstod behov for dette. Grupperne forblev de samme på tværs af opgaverne.

Opgave 1: Kortlægning af yderligere behov på området

Første opgave havde til formål at skabe overblik over de ressourcer, Danmark har og mangler i arbejdet med at evaluere og benchmarke sprogmodeller på dansk. Til formålet blev hver gruppe udstyret med tre A3-ark, hvorpå der hhv. stod "Det har vi", "Det skal vi have mere af" og "Det skal vi ikke have mere af". Grupperne havde til opgave at diskutere disse tre overskrifter og skrive stikord fra deres drøftelser på post-it-notes og sætte på A3-arkene. Det blev understreget, at grupperne gerne måtte tænke i konkrete datasæt, domæner, projekter og evalueringsopgaver/-scenarier. Der blev sat en halv time af til dette, hvorefter hver gruppe med udgangspunkt i A3-arkene på skift præsenterede deres væsentligste diskussionspunkter. Under præsentationerne blev der taget noter på storskærm, som opsummerede de forskellige inputs. Der blev desuden diskuteret på tværs af disse pointer i plenum. De noter, som relaterede til overskriften "Det skal vi have mere af" blev efterfølgende printet og delt ud til alle grupperne, så hver gruppe havde en liste med det fulde overblik over inputs til brug for opgave 2.

Opgave 2: Backlog og prioritering

Anden opgave havde til formål at præcisere og prioritere opgaver, som bør løses for at hjælpe Danmark videre i arbejdet med at evaluere og benchmarke sprogmodeller på dansk. "Backlog"-konceptet blev brugt for at lede tankerne hen på udarbejdelsen af specifikke og håndgribelige opgaver, som kan igangsættes ved lejlighed, frem for at formulere primært overordnede og tematiske behov, som er sværere at afgrænse og igangsætte konkrete indsatser omkring. Hver gruppe blev udstyret med et A3-ark, som var inddelt i tre kolonner med en række instruktioner, som listet nedenfor:

- **Kolonne 1: "To do"**
    - Definér opgaven så specifikt som muligt.
    - Tænk gerne her i specifikke datasæt, evalueringsopgaver/-scenarier, projekter.
- **Kolonne 2: "Hvem?"**
    - Angiv hvilken aktør, der bør løse opgaven.
    - Fx Digitaliseringsstyrelsen, men også gerne andre aktører.
- **Kolonne 3: "Afhængigheder og behov fra andre"**
    - List forudsætninger for, at aktøren i "Hvem?" kan løse opgaven og giv eventuelt et bud på, hvem der kan hjælpe med dette.

Der blev sat en halv time af til at diskutere og udfylde arket i grupperne, hvorefter hver gruppe igen på skift præsenterede deres A3-ark. Her blev der igen taget noter på storskærm i en tabel, der var magen til den, som fremgik af A3-arket. Der var en tendens til, at tredje kolonne, "Afhængigheder og behov fra andre", kom mindre i spil end de øvrige to kolonner. På storskærmen var der endvidere tilføjet en fjerde kolonne, som skulle repræsentere en rangorden, som kunne bruges til at prioritere i opgaverne. Det var hensigten, at alle i plenum skulle være med til at rangordne de opgaver, som blev noteret. Dette blev dog uhensigtsmæssigt grundet uoverskueligheden i de noter, der blev genereret på storskærmen, og med enighed fra deltagerne udeblev prioriteringsopgaven. I stedet blev deltagerne bedt om at forholde sig til, om der var andre relevante spørgsmål, der burde vendes. Efter endnu en drøftelse i plenum blev dagen rundet af.

Efter workshoppen

Efter workshoppen blev de skriftlige artefakter, som deltagerne udarbejdede under opgaverne, digitaliseret og brugt som understøttende empiri til at udarbejde et forslag til en "backlog" på området, som præsenteres i afsnit 4 i denne opsamling. Forud for publicering har opsamlingen desuden været sendt til kommentering blandt deltagerne med henblik på at indsamle yderligere inputs eller bemærkninger, som burde fremgå af dokumentet.

4. Forslag til "backlog" med opgaver

I dette afsnit opsummeres de specifikke opgaver, som deltagerne på workshoppen har defineret. Opgaverne er listet i vilkårlig rækkefølge og er udfærdiget af Digitaliseringsstyrelsen med udgangspunkt i de skriftlige og mundtlige inputs, som deltagerne bidrog med på workshoppen. Opgaverne er dermed ikke nødvendigvis politisk prioriterede men kan bruges som inspiration til nye indsatser for evaluering og benchmarking af sprogmodeller på dansk.

Menneskelig baseline / Dansk "arena"-tilgang	
To do	Der skal samles en gruppe af mennesker, som manuelt skal evaluere sprogmodellers svar og danne et datasæt med menneskelige præferencer. Sammensætningen skal være divers og repræsentativ for Danmarks befolkning for at mindske bias. Resultaterne kan bruges til at tilpasse automatisk evaluering og i forskningsprojekter og vil være værdifulde for allerede eksisterende evalueringsframeworks. Opgaven kan evt. indebære opsætning af infrastruktur til brugerpanel, som fremadrettet kan bruges som fundament for menneskelige baselines.
Hvem?	De aktører, der driver de eksisterende evalueringsframeworks, kan stå for at implementere de menneskelige baselines i deres frameworks. Selve arbejdet med at etablere et brugerpanel kan fx udføres af Digitaliseringsstyrelsen i samarbejde med CBS, som arbejder med *citizen data science*, mens en platform for indsamling af besvarelser fx kan bygge videre på Danoliterate.
Afhængigheder	Det vil kræve både finansiering og juridisk grundlag, evt. med hjælp fra Danmarks Evalueringsinstitut.

Domænespecifikke evalueringsopgaver	
To do	Der skal udarbejdes domænespecifikke evalueringsopgaver for domæner som fx jura og sundhed (disse to er uddybet i separate opgaver). Arbejdet skal i samarbejde med domæneeksperter indebære en kortlægning af, hvad evalueringsscenarierne skal indeholde for de respektive domæner. Inden for sundhed findes der fx allerede gode data, som kræver relativt lidt manipulation til formålet.
Hvem?	Det bliver ikke de aktører, der driver de eksisterende evalueringsframeworks, men disse aktører optager gerne de færdige evalueringsopgaver i disse frameworks. Evalueringsopgaverne må udarbejdes af nogen, der har en interesse i de respektive domæner og forstår sig på behovene.
Afhængigheder	Det vil kræve både domænespecifikke kompetencer og kompetencer med benchmarking af sprogmodeller.

Datasæt for resumering af tekst	
To do	Der skal indsamles datasæt, som kan danne grundlag for evalueringsscenarier for sprogmodellers præstationer for resumering af tekst. Datasættene bør bestå af tekster parret med dertilhørende resuméer.
Hvem?	Selve indsamlingen af data kan fx foretages af Digitaliseringsstyrelsen. Udarbejdelse af en decideret evalueringsopgave kan fx laves af universiteterne.
Afhængigheder	Der findes en række relevante dataindehavere, som kan hjælpe til med arbejdet. Disse er fx Folketinget, tidsskrift.dk, Danske Taler, Danske Medier, videnskabelige tidsskrifter, offentlige rapporter.

Sundhedsdatasæt	
To do	Der skal indsamles datasæt, som kan danne grundlag for evalueringsscenarier for sprogmodellers præstationer inden for sundhedsdomænet.
Hvem?	Selve indsamlingen af data kan fx foretages af Digitaliseringsstyrelsen. Udarbejdelse af en decideret evalueringsopgave skal ske i samarbejde med domæneeksperter.
Afhængigheder	Der findes en række relevante dataindehavere, som kan hjælpe til med arbejdet. Disse er fx kommuner og regioner, sundhed.dk, netdoktor.dk, Sundhedsdatastyrelsen, RAIN.

Juridiske datasæt	
To do	Der skal indsamles datasæt, som kan danne grundlag for evalueringsscenarier for sprogmodellers præstationer inden for det juridiske domæne.
Hvem?	Selve indsamlingen af data kan fx foretages af Digitaliseringsstyrelsen. Udarbejdelse af en decideret evalueringsopgave skal ske i samarbejde med domæneeksperter, fx folkene bag Markup Legal eller Ailex.
Afhængigheder	Der findes en række relevante dataindehavere, som kan hjælpe til med arbejdet. Disse er fx Folketinget, Retsinformation, Domstolsstyrelsen, Karnov Group.

Question-Answering datasæt	
To do	Der skal indsamles datasæt, som kan danne grundlag for evalueringsscenarier for sprogmodellers præstationer i forbindelse question-answering-opgaver.
Hvem?	Selve indsamlingen af data kan fx foretages af Digitaliseringsstyrelsen. Udarbejdelse af en decideret evalueringsopgave kan fx laves af universiteterne.
Afhængigheder	Der findes en række relevante dataindehavere, som kan hjælpe til med arbejdet. Disse er fx borger.dk, SKAT, sundhed.dk, Nationale Tests og øvrige uddannelsesdata hos STIL og STUK, søgemaskineresultater.

Datasæt for metaforer	
To do	Der skal indsamles datasæt, som kan danne grundlag for evalueringsscenarier for danske sprogmodellers præstationer i forbindelse med metaforiske udtryk og overførte betydninger. Datasættet skal bl.a. afdække, om der opstår kulturelle bias i forhold til kulturspecifikt metaforisk sprogbrug, og om den rette danske 'sprogtone' i den forbindelse bibeholdes i modellernes output. Datasættet kan med fordel tage afsæt i danske ordbogsdata.
Hvem?	Arbejdet kan bygge videre på Danish Semantic Reasoning Benchmark og kan dermed fx bestå i et samarbejde mellem Det Danske Sprog- og Litteraturselskab og Center for Sprogteknologi ved Københavns Universitet.
Afhængigheder	Folketinget kan måske hjælpe med at skaffe spontant genererede talemåder fra Folketingssalen under forudsætning af, at nogen står for at indsamle disse.

Kortlægning af behov	
To do	Der skal laves en kortlægning af, hvilke behov industri og myndigheder har i forbindelse med sprogmodeller. Dette indebærer at identificere use cases hos anvendere og potentielle anvendere af sprogmodeller, som kan give bedre indsigt i, hvor der er målrettet behov for yderligere evalueringsopgaver. Der kan fx tages inspiration fra Kommunernes AI-landkort.
Hvem?	Kortlægningen kan fx udføres af Digitaliseringsstyrelsen, KL og OpenDataDK, Danske Regioner og brancheorganisationerne.
Afhængigheder	Kortlægningen skal komme anvenderne af sprogmodeller til gode, men forudsætter samtidig, at anvenderne bidrager aktivt med at få indblik i specifikke use cases og evt. hjælper med at forsyne relevante evalueringsdata.

Evalueringsopgave for RAG	
To do	Der skal udarbejdes en evalueringsopgave, som kan hjælpe med at evaluere sprogmodellers præstationer i forbindelse med Retrieval Augmented Generation (RAG). Evalueringsopgaven skal både kombinere performance for "retrieval"-delen og "generation"-delen samt afdække, om sprogmodeller er i stand til at svare "ved ikke" eller lignende erkendelser i stedet for at hallucinere.
Hvem?	Arbejdet kan fx løftes af Alexandra Instituttet og/eller Aarhus Universitet.
Afhængigheder	Der skal identificeres relevante data, som kan danne grundlag for evalueringsopgaven.

Whitepaper for RAG	
To do	Der skal udarbejdes et whitepaper, der definerer en evalueringsprocedure for RAG-løsninger. Whitepaperet skal have fokus på, hvordan man evaluerer en RAG-løsning med henblik på at maksimere værdien for brugerne i et organisatorisk setup, hvor man har en videndatabase med en redaktør og et antal brugere.
Hvem?	Arbejdet kan fx løftes af Alexandra Instituttet og/eller Aarhus Universitet i samarbejde med Aarhus Kommune.
Afhængigheder	Lignende arbejde er muligvis på vej på IT Universitet i København.

Governance model for evalueringsframeworks	
To do	Der skal udarbejdes en langsigtet governancemodel for danske evalueringsframeworks, som sikrer, at opgaven med at evaluere og benchmarke sprogmodeller på dansk er varigt forankret hos en eller flere aktører. Det skal bl.a. indebære en ensrettet procedure for optagelse af nye evalueringsopgaver i eksisterende evalueringsframeworks med henblik på at sikre transparens og målrettet udvikling af nye evalueringsopgaver. Endelig skal det være med til at sikre, at nye evalueringsopgaver er baseret på et lovligt datagrundlag.
Hvem?	Arbejdet kan fx med hjælp fra Digitaliseringsstyrelsen koordineres på tværs af de aktører, som allerede varetager danske evalueringsframeworks.
Afhængigheder	Arbejdet forudsætter en varig finansiering til vedligehold af evalueringsframeworks, som er forankret hos en eller flere aktører.

Multimodale datasæt	
To do	Der skal tilvejebringes multimodale datasæt, som indeholder fx tekst, tale/lyd, billede og video, der er annoterede. Formålet er at bidrage til, at sprogmodeller også repræsenterer danske kulturelle normer og værdier på tværs af modaliteter, så at et AI-genereret billede af fx en pølsevogn i højere grad også kan forestille en traditionel dansk pølsevogn.
Hvem?	Indsatsen kan fx koordineres af Digitaliseringsstyrelsen.
Afhængigheder	Der findes en række relevante dataindehavere, som kan hjælpe til med arbejdet. Disse er fx Det Kongelige Bibliotek og Danske Medier.

Dansk standardisering om bias og fairness	
To do	Der skal tilvejebringes dansk standardisering om bias og fairness i sprogmodeller, fx særligt i forhold til køns- og etnicitetsbias i generative sprogmodeller. Arbejdet skal danne grundlag for udarbejdelsen af evalueringsopgaver, som kan skabe overblik over bias og fairness i sprogmodeller.
Hvem?	Arbejdet kan fx drives af universiteterne og Alexandra Instituttet.
Afhængigheder	Det er en forudsætning, at fx Dansk Standard hjælper med processen for tilvejebringelse af dansk standardisering på området, og at fx Digitaliseringsstyrelsen hjælper med at udarbejde relevant vejledning om emnet.

Undersøgelser af prompting-teknikker	
To do	Der skal laves undersøgelser, som belyser effekter af prompting-teknikker på sprogmodellers præstationer, herunder hvilke tiltag, man kan foretage for at forbedre sprogmodellers præstationer via prompting. Indsigterne fra sådanne undersøgelser kan danne grundlag for guidelines og værktøjer, som kan hjælpe anvendere med at få bedre outputs fra sprogmodeller. Arbejdet kan med fordel undersøge, om der er perspektiver i prompting-teknikker, som har relevans for specifikt danske interaktioner med sprogmodeller,
Hvem?	Arbejdet kan fx drives af universiteterne eller øvrige aktører, der arbejder med at eksperimentere med prompting-teknikker som led i opgaveløsningen.
Afhængigheder	Det vil være nødvendigt at koordinere indsatsen, således at forskellige sprogmodeller bliver dækket af undersøgelserne.

Adgang til beregningskapacitet	
To do	Det skal sikres, at der findes en fast og transparent måde at tildele adgang til beregningskapacitet på tværs af danske og europæiske HPC-faciliteter, herunder klare rammer for anvendelsesvilkår.
Hvem?	Aktører som DeiC og EuroHPC kan være behjælpelige her.
Afhængigheder	Det er en forudsætning, at fx NNF Gefion kan tages i brug.

5. Fremtidsperspektiver

Ud over de konkrete outputs var der flere drøftelser på workshoppen, som ikke kan opsummeres i specifikke og afgrænsede formater, men som også er vigtige at tage med videre i det fremadrettede arbejde med at evaluere og benchmarke sprogmodeller på dansk. Disse drøftelser opsummeres i dette afsnit.

Det blev på workshoppen pointeret, at drøftelserne fra dagen og den backlog, som blev genereret, med fordel kan danne grundlag for en såkaldt BLARK (Basic Language Ressource Kit) for evaluering af sprogmodeller på dansk. En BLARK er en samling af grundlæggende sprogressourcer for et givent sprog, som er nødvendige for at udvikle sprogteknologi. Deltagerne foreslog derfor, at det kan være relevant at afdække, om der er behov for en mere formaliseret BLARK for evaluering og benchmarking på dansk. I så fald kan man fx bygge videre på indeværende backlog. I et sådan arbejde kan man også kigge mod andre lande og se på hvilke ressourcer, der bliver udarbejdet for andre sprog, og om der er inspiration at hente i disse.

Der var bred konsensus på workshoppen om, at der er behov for flere originale danske domænespecifikke datasæt til evaluering af sprogmodeller. Erfaringerne peger på, at oprindeligt engelske evalueringsopgaver og -datasæt, som oversættes, ikke giver det samme gode evalueringsgrundlag som oprindeligt danske datasæt. De nævnte domæner i den foreslåede backlog bør derfor ikke tænkes som udtømmende for de danske behov på området.

I forlængelse af dette var der også en efterspørgsel fra deltagerne efter generelle afklaringer for anonymisering af data, som kan give fx industriaktører mulighed for lettere at dele domænespecifikke data til evalueringsformål. Det blev foreslået at inkludere Datatilsynet i et sådan afklaringsarbejde, så der er opmærksomhed på tilstrækkelig sikring i forbindelse med data om borgere. Samtidig var der også et ønske fra deltagerne om at få afklaret, hvorvidt syntetiske data er tilstrækkelige til formålet, eller om denne type data er uegnet til arbejdet med at evaluere sprogmodeller.

Foruden de igangværende indsatser med evaluering og benchmarking af sprogmodeller, som blev præsenteret på workshoppen, blev der også efterspurgt vejledning til anvendere af dansk sprogteknologi til at hjælpe med evaluering af løsninger til specifikke behov. Deltagerne foreslog muligheden for at udarbejde et white paper med fokus på, hvordan man evaluerer løsninger til brug i praksis. I forlængelse heraf efterspurgte deltagerne også et fokus på stabilitet og sikkerhed, fx i forhold til jailbreaking, så anvendere kan vide sig sikre i de løsninger, de benytter sig af. Evaluering af bæredygtighedsaspekter i løsninger var ligeså et vigtigt element, som deltagerne efterspurgte øget fokus på.

Endelig var der konsensus på workshoppen om, at danske indsatser med evaluering og benchmarking af sprogmodeller ikke skal ske isoleret fra eller uden hensyntagen til europæiske initiativer på området. Det europæiske arbejde med TrustLLM har fx et evalueringsspor, som Alexandra Instituttet leder, og som ScandEval lige nu er koblet op på. Center for Sprogteknologi på Københavns Universitet deltager endvidere i en EU Cost Action, UNIDIVE (Universality, Diversity, and Idiosyncrasy in Language Technology), hvor der arbejdes med at understøtte og evaluere sproglig og kulturel diversitet i store sprogmodeller. Derudover repræsenterer Digitaliseringsstyrelsen Danmark i Alliancen for sprogteknologi, der er et europæisk infrastrukturkonsortium, som arbejder mod at skabe bedre forudsætninger for udviklingen af sprogteknologi på europæiske sprog, og hvor produktionen af evalueringsopgaver også løbende er på tale. Endelig er der med den nye forordning for kunstig intelligens også ved at blive udarbejdet nye europæiske standarder inden for kunstig intelligens i regi af CEN-CENELEC, hvor man

bl.a. også har øje for evaluering af sprogmodeller, bias og bæredygtighed. Dansk indflydelse på disse standarder koordineres af Dansk Standard.

Til sidst var et fælles budskab fra workshoppens deltagere, at de nuværende indsatser med evaluering og benchmarking af sprogmodeller på dansk ikke nødvendigvis er varige, men derimod lige nu i høj grad udvikles, drives og vedligeholdes af ildsjæle. Hvis Danmark skal sikre sig, at arbejdet fortsætter, så vi også har gode forudsætninger for at evaluere sprogmodeller i fremtiden, er der behov for et fokus på forankring af denne form for indsatser, så vedligehold og ejerskab sikres. Workshoppen og den foreslåede backlog illustrerer, at der stadig er meget arbejde, der bør tages hånd om på området. Digitaliseringsstyrelsen vil kigge nærmere på de opgaver, hvor deltagerne har peget på styrelsen som en potential bidragsyder, og opfordrer desuden deltagerne og øvrige aktører til at gøre det samme for de opgaver, hvor de selv står nævnt som potentielle bidragsydere.

Til slut vil Digitaliseringsstyrelsen gerne takke deltagerne og oplægsholderne for en spændende dag og for mange gode inputs!

Digitaliserings-
styrelsen