Helsedirektoratet

Rapport om store språkmodeller i helse- og omsorgstjenesten: risikoer og tilpasninger til norske forhold

Først publisert: 16. april 2025
Siste faglige endring: 16. april 2025

Innhold

Forord

Helse og omsorgsdepartementet har gitt Helsedirektoratet i oppdrag å utarbeide en felles KI-plan for helse- og omsorgstjenesten, i samarbeid med Direktoratet for medisinske produkter (DMP), Folkehelseinstituttet (FHI), Direktoratet for Strålevern og Atomsikkerhet (DSA), Helsetilsynet, de regionale helseforetakene og Kommunesektorens organisasjon (KS). Som en del av arbeidet i planen har Helsedirektoratet fått i oppdrag å gjøre en vurdering «av hvilke risikoer store språkmodeller introduserer, og identifisere tiltak for å sikre at helse- og omsorgstjenesten har tilgang til språkmodell(er) som er tilpasset norske forhold». Denne rapporten svarer på det oppdraget.

Arbeidet er gjennomført i form av intervjuer med eksperter og litteraturstudier av relevante forskningsartikler og rapporter knyttet til forskning på, utvikling av og bruk av store språkmodeller i Norge. Følgende personer og institusjoner har vært involvert i arbeidet: Benjamin Kille (NTNU/NorwAI), Svein Arne Brygfjeld (Nasjonalbiblioteket), Geir Thore Berge og Tor Oddbjørn Tveit (Sørlandet sykehus), Troels Sune Mønsted (Universitetet i Oslo), Christian Autenried (Helse vest IKT), Damoun Nassehi (Fastlege/UiB), Michael A. Riegler (Simula og OsloMet), Jens Osberg (Digitaliseringsdirektoratet), Karl Øyvind Mikalsen (Senter for pasientnær kunstig intelligens (SPKI)), Kristine Eide, Matias Jentoft og Bjarte Engelsland (Språkrådet). Vi vil takke for at de velvillig har stilt opp til intervjuer gjennom prosjektet. Vi vil rette en særlig stor takk til Michael Riegler for tett oppfølging, rådgivning i arbeidet og kvalitetssikring av tekster.

Kapittel 3 om risikoer ved bruk av store språkmodeller i helse- og omsorgstjenesten og kapittel 4 om tilpasninger til norske forhold er kunnskapsgrunnlag og bakgrunn for anbefalte tiltak.

Rapporten reflekterer kunnskap i starten av 2025. KI er et fagområde i rask utvikling og det kommer ny lovgivning fra EU som vil påvirke bruk av KI i Norge. Denne rapporten anbefaler tiltak som bør igangsettes allerede nå for å legge grunnlag for trygg og forsvarlig bruk av KI fremover.

Prioritering av anbefalte tiltak vil gjøres i dialog med HOD, KI-rådet og ledelsen i Helsedirektoratet. Prioriterte tiltak vi inngå løpende i Felles KI-plan for helse- og omsorgstjenesten. Et flertall av tiltakene vil strekke seg utover planperioden til felles KI-plan, som varer ut 2025.

Oslo, 1. april 2025

Sammendrag

Potensialet og forventningene til KI er høye i helse- og omsorgstjenesten, noe som blant annet er reflektert i den nasjonale digitaliseringsstrategien og Nasjonal helse- og samhandlingsplan [1][2]. Sistnevnte peker på at KI kan bidra til raskere og mer presis diagnostikk, bedre beslutningsstøtte til personell, forenklet logistikk, automatisering av administrative oppgaver, og å sette innbyggerne bedre i stand til å følge opp sin egen helse. KI vil også kunne utgjøre et betydelig bidrag til en bærekraftig utvikling av vår felles helsetjeneste.

Utvikling og bruk av kunstig intelligens (KI) og spesielt store språkmodeller har hatt en stor vekst de siste årene. Bruk av store språkmodeller har potensiale for å forbedre både effektiviteten og kvaliteten i helse- og omsorgstjenesten, eksempelvis gjennom automatisk tekstproduksjon, strukturering av fritekst i pasientjournalen, maskinassistert helsefaglig koding, hjelp til klarspråk, oversettelse, talegjenkjenning, kunnskapsstøtte, helseforskning, sortering og triagering av pasienthenvendelser, klinisk samtalerobot (spørsmål og svar), innbyggerrettet samtalerobot, helseutdanning og beslutningsstøtte i behandling [3].

På den andre siden er det store risikoer knyttet til å ta i bruk generativ KI. Denne rapporten beskriver viktige risikoer som den norske helse- og omsorgs-tjenesten må ta hensyn til for å kunne bruke KI-systemer som baserer seg på store språkmodeller på forsvarlig vis. Risikoene knytter seg både til iboende risiko forbundet med denne type KI-modeller, og til bruken av dem i helse- og omsorgstjenesten.

Foreløpig egner språkmodeller seg best til språklige og administrative oppgaver. Kvalitetssikring er viktig uansett bruk. KI-systemer med språkmodeller som skal brukes til å yte helsehjelp er mest sannsynlig medisinsk utstyr, og dermed regulert gjennom lov om medisinsk utstyr.

Tilpasning til norske forhold er én måte å redusere risikoer på knyttet til bruk av store språk-modeller i helse- og omsorgstjenesten. Rapporten beskriver *hva* slik tilpassing innebærer, *hvordan* tilpasninger kan gjøres, *hva som trengs*, og *hvem* som kan gjøre tilpassing.

Til slutt anbefaler rapporten tiltak som kan redusere risikoer og legge til rette for bruk som er tilpasset norske forhold i helse- og omsorgstjenesten:

- etablere kvalitetsrammeverk for store språkmodeller for norsk helse- og omsorgstjeneste
- bygge og dele kompetanse om utvikling og bruk av store språkmodeller
- etablere felles datagrunnlag
- utrede behov for infrastruktur med regnekraft tilpasset helsesektorens behov
- utrede forvaltningsstruktur for store språkmodeller i helse- og omsorgssektoren

De fleste tiltak vil måtte løses i samarbeid mellom en rekke aktører, eksempelvis Helsedirektoratet, Norske helsenett (NHN), Forskningsrådet, Nasjonalbiblioteket, UH-sektoren, regionale helseforetak, regionale IKT-selskaper og KS. Det foreslås at Helsedirektoratet tar et særlig ansvar for å etablere et kvalitetsrammeverk.

[1] https://www.regjeringen.no/no/tema/statlig-forvaltning/it-politikk/ny-nasjonal-digitaliseringsstrategi

[2] Nasjonal helse- og samhandlingsplan 2024–2027: Kortere ventetider og en felles helsetjeneste - regjeringen.no

[3] Helsedirektoratet har blant annet skrevet om muligheter med store språkmodeller i denne rapporten: Store språkmodellar i helse- og omsorgstenesta

Innledning

Potensialet og forventningene til store språkmodeller i helse- og omsorgstjenesten er høye, samtidig som risikoene er store. Det er avgjørende at KI-systemer med store språkmodeller som brukes i helse- og omsorgstjenesten er tillitsverdige og rapporten har identifisert fem tiltak som vil bidra det.

Det amerikanske National Institute for Standardization and Technology (NIST) karakteriserer tillitsverdig KI slik: *Gyldig og pålitelig, trygg, sikker og motstandsdyktig, ansvarlig og gjennomsiktig, forklarbar og tolkbar, har godt personvern og er rettferdig med kontroll på skadelig skjev*heter [4]. Den europeiske unionen publiserte i 2019 etiske retningslinjer hvor det beskrives at et tillitsverdig KI-system har tre komponenter som bør oppfylles gjennom hele systemets livssyklus: *(1) det bør være lovlig, i samsvar med alle gjeldende lover og forskrifter, (2) det bør være etisk, og sikre etterlevelse av etiske prinsipper og verdier, og (3) det bør være robust, både fra et teknisk og sosialt perspektiv, da KI-systemer, selv med gode intensjoner, kan forårsake utilsiktet skade* [5].

[4] Artificial Intelligence Risk Management Framework (AI RMF 1.0) (nist.gov), side 12 og Trustworthy and Responsible AI | NIST: Characteristics of trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.

[5] European Commission: Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI, Publications Office, 2019, https://data.europa.eu/doi/10.2759/346720

Mål

> Målet med denne rapporten er tredelt:
>
> 1. å beskrive og tydeliggjøre risikoer forbundet med bruk av generativ KI og store språkmodeller i helse- og omsorgstjenesten
> 2. å beskrive hvordan store språkmodeller kan tilpasses den norske helse- og omsorgstjenesten, som ett av flere mulige risikoreduserende tiltak
> 3. å foreslå tiltak som vil bidra til at store språkmodeller blir godt tilpasset forholdene i den norske helse- og omsorgstjenesten
>
> Rapporten er å anse som et underlag for diskusjoner om prioritering av og ansvar for aktuelle tiltak som vil bidra til dette.

Målgruppe

> Hovedmålgruppen for rapporten er Helse- og omsorgsdepartementet (HOD), KI-rådet, helse- og omsorgssektoren, samarbeidspartnere (eksempelvis Nasjonalbiblioteket og FoU-institusjoner), og andre myndigheter som er aktuelle for å igangsette de anbefalte tiltakene.
>
> Rapporten vil også være interessant for de som jobber med generativ KI og helse.

Avgrensninger

Denne rapporten omhandler generativ kunstig intelligens, hovedsakelig store språkmodeller. Store multimodale modeller er også omhandlet kort.

I mai 2024 publiserte Helsedirektoratet kunnskapsgrunnlaget *Store språkmodellar i helse- og omsorgstenesta* [6]. Dokumentet beskriver hvordan slike modeller kan bidra til å forbedre helsetjenesten, potensielle fordeler og utfordringer, og hvordan store språkmodeller kan tilpasses bruk i helse- og omsorgstjenesten. Denne rapporten omhandler ikke mulighetsrommet og utdyper utfordringer og hvordan språkmodeller kan tilpasses helsetjenesten.

I januar 2025 publiserte Helsedirektoratet *Rapport om kvalitetssikring: bruk av KI i helse- og omsorgstjenesten* [7]. Rapporten omhandler blant annet risikohåndtering og kvalitetssikring knyttet til innføring og bruk av KI generelt. Denne rapporten går mer i dybden på risikoer og kvalitetssikring knyttet til generative KI-modeller og spesielt store språkmodeller.

Helsedirektoratet utvikler KI-tjenester som bruker store språkmodeller, blant annet i Helsesvar og ETI. Personvernspørsmål knyttet til disse prosjektene blir adressert i Datatilsynets regulatoriske sandkasse og påfølgende rapport (som ikke er publisert når denne rapporten ble ferdigstilt) [8]. Denne rapporten adresserer også risikoer knyttet til personvern, men går ikke i dybden på spesifikke bruksområder.

[6] Store språkmodellar i helse- og omsorgstenesta

[7] Rapport om kvalitetssikring: Bruk av kunstig intelligens i helse- og omsorgstjenesten

[8] https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/pagaende-prosjekter2/he

Metode

Arbeidet med rapporten har blitt gjort gjennom en kombinasjon av intervjuer og litteraturstudier av norske og internasjonale forskningsartikler og relevante rapporter.

Vi har intervjuet fagpersoner med erfaring eller dyp kunnskap knyttet til forskning på, utvikling og bruk av store språkmodeller i Norge. Intervjuene er gjennomført i perioden oktober 2024 til februar 2025. Følgende institusjoner har vært involvert i arbeidet: NTNU, NorwAI, Nasjonalbiblioteket, Sørlandet sykehus, Universitetet i Oslo, Helse Vest IKT, et fastlegekontor, Universitetet i Bergen, Simulasenteret, OsloMet, Digitaliseringsdirektoratet, Senter for Pasientnær Kunstig Intelligens (SPKI) og Språkrådet.

Sentrale begreper

Definisjonen på et *KI-system* er etter KI-forordningen, og slik vi bruker betydningen i denne rapporten: ” …et maskinbasert system som er designet for å operere med varierende nivåer av autonomi og som kan vise tilpasningsevne etter utplassering, og som, for eksplisitte eller implisitte mål, utleder, fra inndata det mottar, hvordan det skal generere utdata som prediksjoner, innhold, anbefalinger eller beslutninger som kan påvirke fysiske eller virtuelle miljøer [9].

Et KI-system kan ha ulike former: som en selvstendig applikasjon på en PC eller mobil-applikasjon, en webtjeneste, en del av et fagsystem eller et journalsystem.

Et KI-system er medisinsk utstyr etter lov om medisinsk utstyr dersom det er ment å skulle brukes på mennesker i den hensikt å bidra til diagnostisering, forebygging, overvåking, prediksjon, prognostisering, behandling eller lindring av sykdom [10][11]

Et KI-system kan bestå av én eller flere *KI-modeller,* i tillegg til forbedringforslag som kunnskapsforankring.

En algoritme er en oppskrift som forteller hvordan noe gjøres, og består enkelt forklart av et sett instrukser som forvandler inndata til utdata. En maskinlæringsalgoritme er oppskriften systemet bruker for å bygge en *KI-modell* av virkeligheten, basert på dataene som blir matet inn (treningsdataene). Denne modellen kan så brukes for å ta nye beslutninger om nye data som kommer inn [12].

Grunnmodeller er store nevrale nettverk som er trent på store generelle datasett som kan bestå av tekst, bilder, lyd m.m. Slike modeller kan danne utgangspunkt for en rekke ulike løsninger [13].

Generativ KI omfatter løsninger som først og fremst genererer, eller frembringer, nytt materiale, for eksempel tekst eller bilder [14]. Generative KI-modeller skiller seg fra ikke-generative KI-modeller på blant annet følgende måter:

Generative KI-modeller skaper nytt innhold, som tekst, bilder eller musikk, basert på innlærte data. For eksempel kan en generativ modell lage nye bilder av katter basert på tidligere kattebilder.

Ikke-generative, også kalt klassifiserende, KI-modeller skiller mellom ulike klasser eller forutsier sannsynligheter for spesifikke utfall. De fokuserer på å finne grensene mellom kategorier i dataene. En klassifiserende KI-modell kan avgjøre om et gitt bilde inneholder en katt eller ikke.

Store språkmodeller (*Large Language Models - LLM*) er generative KI-modeller. Store språkmodeller, som de kjente GPT-modellene, er en form for grunnmodeller som er trent på enorme mengder tekst for å kunne forutsi hva som kommer etter et ord eller en stavelse i en gitt kontekst. Språkmodeller lagrer ikke teksten de er trent på. De «vet» ingenting om verden, og de går ikke gjennom nettsider for å finne fakta. De kan bare språk. Det er likevel lett å tro at modellene tenker, eller kan noe, fordi de er så gode til å skrive tekst som vi mennesker opplever som meningsfull [15].

I senere tid har store språkmodeller blitt utvidet til å håndtere andre modaliteter, som bilder, lyd og video. Disse kalles *multimodale KI-modeller*, men likevel brukes ofte betegnelsen "LLM". Dette skyldes flere forhold:

- *Kjernen i modellen er språklig*: Selv om de kan behandle flere modaliteter, er tekstgrensesnittet fortsatt den primære måten brukere interagere med modellen på.
- *Underliggende teknologi*: Modellene bygger på arkitekturer utviklet for språkprosessering (som transformere), og mye av treningen skjer på store tekstdatasett.
- *Generell faglig terminologi*: Innen KI-feltet brukes begrepet "LLM" fortsatt bredt, da de fleste multimodale modeller springer ut fra LLM-er som er utvidet til å tolke andre datatyper.

I denne rapporten brukes samme definisjon av *risiko* som i KI-forordningen: "Risiko betyr kombinasjonen av sannsynligheten for at en skade oppstår og alvorlighetsgraden av denne skaden" [16].

For øvrig er begreper som brukes i rapporten beskrevet i KI-faktaark KI-begreper [17].

[9] Artikkel 3 i KI-forordningen (oversatt):
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

[10] https://lovdata.no/dokument/NL/lov/1995-01-12-6

[11] Se for øvrig EØS-tillegget om medisinsk utstyr her. https://lovdata.no/static/NLX3/32017r0745.pdf

[12] Teknologirådet, side 16:
https://media.wpd.digital/teknologiradet/uploads/2022/02/Kunstig-intelligens-i-klinikken.pdf

[13]
https://www.regjeringen.no/contentassets/c499c3b6c93740bd989c43d886f65924/no/pdfs/nasjonal-digitalise

[14]
https://www.regjeringen.no/contentassets/c499c3b6c93740bd989c43d886f65924/no/pdfs/nasjonal-digitalise

[15]
https://www.regjeringen.no/contentassets/c499c3b6c93740bd989c43d886f65924/no/pdfs/nasjonal-digitalise

[16] Se definisjonen av "risiko" i KI-forordningen artikkel 3 (2),
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689#cpt_I

[17] Faktaark om KI-begreper

Risiko ved bruk av store språkmodeller i helse- og omsorgstjenesten

Bruk av store språkmodeller har stort potensiale for å forbedre både effektiviteten og kvaliteten i helse- og omsorgstjenesten, eksempelvis gjennom strukturering av fritekst i pasientjournalen, maskinassistert helsefaglig koding, beslutningsstøtte i behandling, kunnskapsstøtte, prediksjon, sortering og triagering av pasienthenvendelser, automatisk tekstproduksjon, klinisk samtalerobot (spørsmål og svar), innbyggerrettet samtalerobot (spørsmål og svar), hjelp til klarspråk, talegjenkjenning, helseutdanning, helseforskning, oversettelse og logistikk [18].

Samtidig har språkmodeller noen iboende utfordringer, som kan påvirke både pasient-sikkerheten og kvaliteten på helsetjenestene, noe som igjen kan påvirke den overordnede tilliten til helse- og omsorgstjenesten. Bruk av store språkmodeller er i en tidlig fase, og det er fremdeles knyttet usikkerheter til nytte og gevinster på kort og lang sikt. Forskere på Simula-instituttet påpeker at "kunstig intelligens (KI), og spesielt generativ KI, er en samling besnærende teknologier som, dersom man ikke har god teknologisk innsikt, kan forlede beslutningstakere i alle ledd i organisasjoner til å ta uinformerte vurderinger" [19].

Det er også usikkerheter knyttet til på hvilke områder store språkmodeller kan brukes i helse- og omsorgstjenesten på forsvarlig og hensiktsmessig vis. Foreløpig egner de seg best til språklige og administrative oppgaver med lav risiko. Kvalitetssikring er viktig uansett bruk.

KI-systemer med store språkmodeller som skal brukes til å yte helsehjelp har en høyere risiko. De er mest sannsynlig medisinsk utstyr, og dermed regulert gjennom lov om medisinsk utstyr [20]. Formålet med denne loven er å forhindre skadevirkninger, uhell og ulykker, samt sikre at medisinsk utstyr utprøves og anvendes på en faglig og etisk forsvarlig måte. Per 25. mars kjenner vi ikke til KI-systemer med generative språkmodeller som er CE-merket etter samsvarsvurderingen utført av meldt organ.

KI-forordningen har trådt i kraft i EU med mål om å etablere tillit til bruk av KI i EU og samtidig legge til rette for innovasjon. Forordningen har en risikobasert tilnærming, der kravene til et KI-system bestemmes av risikonivået ved tiltenkt bruk.

For øvrig må brukere så vel som utviklere av KI-systemer som skal anvendes i helse- og omsorgstjenesten i Norge, forholde seg til både generelle og sektorspesifikke regelverk. For eksempel gjelder sektorspesifikke regelverk som norsk helselovgivning, EUs personvernforordning, åndsverksloven og likestilling- og diskrimineringsloven.

Dette kapitlet beskriver både underliggende utfordringer ved store språkmodeller, og risikoer knyttet til bruk i helse- og omsorgstjenesten. Inkluderte risikoer er ikke uttømmende.

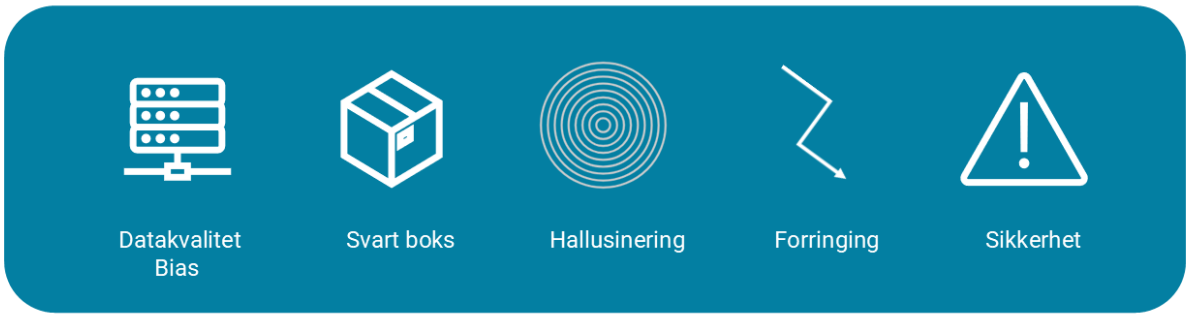
[18] Helsedirektoratet har blant annet skrevet om muligheter med store språkmodeller i denne rapporten: Store språkmodellar i helse- og omsorgstenesta

[19] https://www.digi.no/artikler/debatt-ki-feberen-nar-middelet-blir-malet/554186

[20] https://lovdata.no/dokument/NL/lov/1995-01-12-6

Risiko i store språkmodeller

Store språkmodeller er svært komplekse, og de genererte svarene er basert på statistiske mønstre i treningsdata fremfor faktisk forståelse av innhold. Feil, skjevheter og mangel på transparens kan få konsekvenser for sikkerheten til pasienter og kvaliteten på helsetjenester. Dette kapitlet beskriver noen av de underliggende årsakene som kommer fra trening av og egenskaper ved store språkmodeller (Figur 1).



Figur 1 Underliggende usikkerheter og egenskaper ved store språkmodeller som bidrar til risiko ved bruk i helse- og omsorgstjenesten inkluderer datakvalitet og bias i datagrunnlaget, "svart boks" problematikk, hallusinering, forringing av modellen over tid og sårbarheter ved cybersikkerhet og informasjonssikkerhet.

3.1.1 Dårlig datakvalitet og skjevheter i datagrunnlaget

Trening av store språkmodeller krever enorme mengder data hentet fra kilder som bøker, tidsskrifter, internettsider og sosiale medier. Gode kilder kan også være journalnotater og medisinske bilder.

Kvaliteten på dataene vil påvirke ytelsen til modellene. Noen faktorer som påvirker datakvalitet, er:

- *Ufullstendige eller mangelfulle data.* Frafall av innrullerte pasienter i kliniske studier kan føre til ufullstendige data. Manglende registrering av informasjon fører til ufullstendige og mangelfulle data [21]. Det kan være fordi dataene ikke er hentet inn systematisk. Pasientjournaler kan også mangle informasjon på grunn av varierende praksis i dataregistrering [22]. En annen årsak kan være at hensyn til personvern kan resultere i utelatelse av data.
- *Feilinformasjon* viser til informasjon som er feil eller misvisende [23]. Desinformasjon er en systematisk og bevisst form for feilinformasjon [24]. Små mengder feilinformasjon i treningsdata har vist seg å føre til misvisende svar fra medisinske KI-modeller [25]. Store språkmodeller kan være trent på data fra internett og/eller tredjeparts datakilder av varierende kvalitet. Disse kildene kan inneholde feilinformasjon som kan være vanskelig å oppdage.

KI-modeller kan videreføre og forsterke skjevheter som finnes i treningsdataene, samt bidra til å forsterke skjevheter som påvirker menneskelige vurderinger og dømmekraft [26][27]. Nedenfor er noen eksempler på typer skjevheter i data som er viktige å være klar over, spesielt på helseområdet:

- *Overvekt av historiske data* i forhold til nyere data*,* gjør at svarene fra modellene ikke er oppdatert i henhold til gjeldende praksiser og forskning*.* Innen medisin og helse bør ny kunnskap veies tung sammenlignet med historisk praksis.
- *Seleksjonsbias* i for eksempel publisering av vitenskapelige artikler. Det viser seg at positive funn og data oftere publiseres enn negative funn [28]. Kliniske studier med signifikante resultater rapporteres oftere enn negative studier [29]. Resultatet er at tekster og publikasjoner som brukes til trening av store språkmodeller i stor grad representerer de studiene som oppnår signifikans og bekrefter hypoteser. Ved rapportering av for eksempel intervensjonsstudier vektlegges ofte fordeler mer enn ulemper.
- *Manglende lokal kontekst.* Treningsdata for de store, generelle språkmodellene har sterk overvekt av tekst og informasjon på engelsk. I noen tilfeller trenes modellene kun på engelsk tekst. Norsk innhold fra internett vil uansett utgjøre en svært liten del av treningsdata [30][31]. Watson Oncology er et eksempel på at mangel på lokal kontekst kan påvirke ytelsen til en modell. Da modellen ble prøvd ut i Danmark hadde den betraktelig dårligere ytelse enn tidligere rapportert [32]. Danske klinikere pekte både på store forskjeller mellom danske og amerikanske retningslinjer for behandling av kreft, samt dårligere kvalitet på data fra kliniske studier utført i USA som mulige årsaker.
- *Skjeve eller mangelfulle data for bestemte grupper og minoriteter*. Ikke alle grupper og minoriteter er likt representert i datagrunnlaget som brukes for å trene store språkmodeller [33]. Historisk sett har for eksempel kvinner vært underrepresentert i kliniske studier [34]. Personer med nedsatt funksjonsevne eller minoritetsbakgrunn kan også være underrepresentert i kliniske studier. Dette kan skyldes bevisst ekskludering fra studiene, eller praktiske barrierer som transport, økonomi eller andre hindringer for deltakelse [35][36]. Dette er også relevant internt i Norge der det kan være mindre tilgjengelig data fra minoriteter og den samiske urbefolkning for eksempel [37].

3.1.2 Manglende transparens, forklarbarhet og tolkbarhet ("svart boks"-problemet)

Store språkmodeller kan fungere som "svarte bokser", der verken brukere eller utviklere kan forklare fullt ut hvordan et spesifikt svar (utdata) fra en KI-modell blir generert. Mangel på transparens, forklarbarhet og tolkbarhet kan bidra til denne "svart boks"-problematikken [38]. Dette er distinkte karakteristikker som bygger oppunder hverandre.

Transparens viser til *hva som skjedde* i modellen. Indikatorer som bidrar til transparens er tilgjengelig informasjon om:

- data brukt til trening
- kildekode
- metoder for trening
- evaluering
- ytelse

The Foundation Model Transparency index bygger på 100 ulike indikatorer på transparens relatert til trening, selve modellen, eller bruken av modellen [39]. Eksempelvis har Mistral 7B, som kan brukes av tredjeparter uten restriksjoner, en indeks på 55 %, det vil si at det ikke finnes noe tilgjengelig informasjon om 45 av de 100 indikatorene (deriblant datakilder) [40].

Forklarbarhet viser til hvordan et KI-system fungerer og hvilke mekanismer som styrer beslutningene det tar. Med mange millioner eller milliarder parametere og komplekse nevrale nettverk kan store språkmodeller operere som svarte bokser, selv med åpenhet rundt både data og kildekode [41].

Tolkbarhet viser til hvorfor modellen kom frem til et svar [42]. Det kan være å vise frem til brukere hvilke instruksjoner som er gitt for å komme frem til svaret, noe dagens KI-systemer sjelden viser. En metode som bidrar til bedre tolkbarhet, er resonnering (se KI-faktaark Intelligensforsterkning og kontroll [43]).

Store språkmodeller opererer sannsynlighetsbasert og genererer det mest sannsynlige svaret ut fra treningsdataene, men ofte uten en klar forståelse av når svarene er riktige. Manglende forklarbarhet og tolkbarhet kan for eksempel gjøre det vanskelig for klinikere å vurdere om anbefalinger fra språkmodeller er pålitelige, fordi de ikke kan evaluere hvilke mekanismer og data som ligger til grunn [44].

Metoder for å forankre modeller i fakta eller synliggjøre resonnementer eller hvilke parametere som er vektlagt (se KI-faktaark Intelligensforsterkning og kontroll [45]) kan gjøre det lettere for helsepersonell å vurdere kvaliteten på det genererte svaret [46]. Samtidig kan bruk av forankringsmetoder redusere oppmerksomheten rundt modellens begrensninger som fortsatt krever en kritisk evaluering av utdataene [47].

Mangelen på forklarbarhet og tolkbarhet gjør det også vanskelig å oppdage når modellen genererer feilaktige eller oppdiktede svar, også kjent som hallusinering.

3.1.3 Hallusinering

Fordi språkmodeller ikke har en innebygd forståelse av kunnskap og hva som er sannhet, men opererer basert på sannsynlighetsberegninger, kan de generere feilaktig, mangelfull, unøyaktig eller misvisende informasjon med stor overbevisning. Dette fenomenet, kjent som hallusinering, innebærer at modellen kan presentere informasjon som ser troverdig ut, men som enten er feil, tatt ut av kontekst eller ikke har noe grunnlag i treningsdataene.

Sannsynligheten for hallusinering øker dersom treningsdataene har lav kvalitet eller mangler representasjon, for eksempel dersom viktige medisinske perspektiver eller oppdaterte retningslinjer mangler. Selv i modeller trent på høykvalitetsdata kan hallusinering forekomme, da dette er en iboende begrensning ved store språkmodeller [48].

Hallusinering kan deles inn i to kategorier:

- fakta-hallusinering, der modellen genererer feil eller oppdiktede påstander.
- kontekst-hallusinering, der informasjon tolkes feil eller settes inn i en uriktig sammenheng.

I tillegg kan utelatelse av kritisk informasjon utgjøre en risiko, særlig i helsesammenheng, der det stilles strenge krav til nøyaktighet.

Temperaturinnstilling (se faktaboks om temperatur), ettertrening og metoder for å kunnskapsforankre modellen kan redusere forekomsten av hallusinering (se avsnitt 4.2).

Temperatur (grad av variasjon)

I fagområdet KI er temperatur en parameter som styrer hvor tilfeldig teksten som blir generert skal være. Temperaturen kan justeres i utvalgsfasen, der modellen velger det neste ordet basert på beregnede sannsynligheter. Lav temperatur (for eksempel 0.1-0.3) gjør at modellen oftere velger de mest sannsynlige ordene som gir mer forutsigbare og konservative svar. Høy temperatur (for eksempel 0.8-1.0) øker sjansen for at den velger mer uventede eller varierte ord som gir mer kreative og varierte svar. For medisinske anvendelser vil man ofte velge en lav temperatur for å sikre mer presise og konsekvente svar.

Eksempel (generert av ChatGPT-4o): Spørsmål: *"Hva er vanlige årsaker til hodepine?"*

- *Lav temperatur* (0.2) gir et kortfattet, nøyaktig og objektivt svar: *"Vanlige årsaker til hodepine inkluderer stress, dehydrering, spenninger i muskler, søvnmangel og migrene. Noen medisinske tilstander, som høyt blodtrykk eller bihulebetennelse, kan også forårsake hodepine."*
- *Høy temperatur* (0.8) gir et mer variert, uformelt og beskrivende svar: *"Hodepine kan skyldes alt fra stress og dehydrering til dårlig søvn eller spent nakkemuskulatur. Noen får hodepine av sterke lukter eller høye lyder, mens andre kan oppleve det som et symptom på underliggende sykdommer som bihulebetennelse eller migrene. I sjeldne tilfeller kan hodepine være et tegn på mer alvorlige tilstander."*

3.1.4 Forringing av modellens ytelse over tid

Store språkmodeller er dynamiske og påvirkes av både eksterne og interne faktorer over tid. Tre utfordringer knyttet til forringing av modellens ytelse er modelldrift, katastrofal glemsel og modellkollaps. Disse fenomenene kan redusere modellens evne til å gi nøyaktige, relevante og pålitelige svar, noe som er kritisk i helsesammenheng.

Modelldrift beskriver hvordan en modells ytelse kan svekkes over tid, enten fordi den møter nye data som skiller seg fra treningsgrunnlaget, eller fordi sammenhengen mellom inndata og utdata endrer seg [49]. Modelldrift kan føre til at KI-systemer gir utdaterte eller feilaktige anbefalinger eller informasjon. Dette er relevant i helsefeltet, der pasientpopulasjoner, behandlingsmetoder og medisinske retningslinjer stadig utvikler seg. Noen underliggende årsaker er:

- *Datadrift* som oppstår når distribusjonen av inndata modellen møter i daglig bruk, skiller seg fra dataene som ble brukt til å trene og optimalisere modellen. For eksempel kan en modell utviklet for en yngre pasientpopulasjon gi mindre presise anbefalinger dersom den i stedet brukes på en eldre pasientgruppe.
- *Konseptdrift som* skjer når forholdet mellom inndata og utdata endrer seg over tid. For eksempel, hvis en ny behandling eller endring i livsstil viser seg å redusere sammenhengen mellom alder og risiko for diabetes, vil modellen gi utdaterte eller misvisende anbefalinger fordi den fortsatt baserer seg på det gamle forholdet.
- *Ikke oppdatert informasjon*. Det er ressurskrevende å trene store språkmodeller, og de vil ikke til enhver tid være trent på den nyeste tilgjengelige informasjonen. Dette kan føre til at modellene gir utdaterte svar som påvirker modellens ytelse innen medisin og helse [50][51]

Katastrofal glemsel (*catastrophic forgetting eller catastrophic interference*) oppstår når en modell mister eller forringer tidligere lært kunnskap som følge av trening på nye data. For eksempel kan en stor språkmodell miste presisjonen for medisinsk diagnostikk hvis den ettertrenes på generell tekst.

Modellkollaps er en langsiktig utfordring som kan oppstå når en stadig større andel av treningsdataene er generert av språkmodeller fremfor mennesker. Dette kan føre til at modellene lærer fra tidligere modeller i stedet for fra ekte menneskelig kunnskap, noe som gradvis svekker kvaliteten og mangfoldet i treningsdataene [52]. Modellkollaps handler ikke bare om syntetiske data, men også om tap av modellens kompleksitet og variasjon. Når modeller lærer fra hverandre kan subtile mønstre, nyanser og ekstreme verdier i dataene forsvinne. Dersom det fører til en gradvis utarming av modellens evne til å fange opp komplekse sammenhenger og produsere variert innhold, kan det resultere i irreversible feil eller defekter. Denne risikoen er ikke akutt, men kan bli en utfordring på sikt dersom store mengder fremtidig data er syntetisk generert.

Katastrofal glemsel

Årsaker til katastrofal glemsel inkluderer:

- oppdatering av alle vektene i modellen ved ny trening
- sekvensiell læring der gamle data ikke lenger er tilgjengelige
- overtilpasning til nye data, slik at tidligere læring fortrenges

Det finnes mange strategier for løpende læring over tid (*continual learning*) som kan motvirke katastrofal glemsel. Disse inkluderer selektiv vektfrysing (*selective weight freezing*) for å bevare kritisk kunnskap, erfaringsrepetisjonsteknikker (*experience replay*) som blander historiske data med nye treningsdata eller arkitekturbaserte løsninger som progressive nevrale nettverk der nye oppgaver håndteres i separate lag. Disse metodene kan implementeres enkeltvis eller i kombinasjon for å oppnå mer robust læring over tid.

Kilde: https://link.springer.com/article/10.1007/s11063-024-11709-7

3.1.5 Sårbarheter ved cybersikkerhet og informasjonssikkerhet

Cybersikkerhet må være en del av vurderingen av leverandørkjeden til en språkmodell. Tradisjonelle cybersikkerhetstiltak som tilgangskontroll, logging og testing er fortsatt relevante, men må tilpasses språkmodellers unike sårbarheter. Risikovurderinger bør i tillegg inkludere modellens tilgangsnivåer, datahåndtering og beskyttelse mot manipulasjon av input, for å sikre trygge løsninger innen helse- og omsorgssektoren.

Store språkmodeller skiller seg fra tradisjonell programvare ved sin kompleksitet og manglende transparens (se 3.1.2), noe som skaper særlige cybersikkerhetsutfordringer. I motsetning til tradisjonell programvare, der kildekode kan gjennomgås for sikkerhets-vurdering, er mye av språkmodellers funksjonalitet basert på enorme datasett som kan være vanskelig å verifisere (se 3.1.1). Ofte er selve grunnmodellen en "svart boks", og leverandørkjeden kompleks med ulike aktører for modellutvikling, data og programvare. Dette gjør det krevende å oppdage sårbarheter knyttet til for eksempel forgiftede treningsdata eller bakdører i modellen.

Språkmodeller har brede bruksområder og kan brukes på uforutsette måter. Dersom de har tilgang til dokumenter eller systemer, kan de utilsiktet eksponere sensitiv informasjon. Et eksempel er Copilot, som har tilgang til alle data brukeren har tilgang til, noe som stiller store krav til kontrollrutiner [53]. I helsesektoren, hvor integritet og konfidensialitet er avgjørende, er dette en særlig utfordring.

En kjent risiko er *prompt injection*, hvor angripere manipulerer modellens svar ved hjelp av spesialutformede spørringer til systemet. Slik kan angriper for eksempel omgå sikkerhetsmekanismer eller avsløre treningsdata. Indirekte prompt injection, der skadelig input kommer fra eksterne dokumenter eller nettsider, er særlig bekymringsfullt da det kan påvirke svaret uten at brukeren fatter mistanke.

Store språkmodeller krever betydelige ressurser, noe som gjør dem sårbare for tjenestenektangrep (DDoS) ved mange eller ressurskrevende spørringer.

Open Worldwide Application Security Project [54] (OWASP) og MITRE [55] peker på en rekke spesifikke sikkerhetsrisikoer ved språkmodeller og oppdaterer dem jevnlig i takt med teknologiens utvikling.

[21] guideline-missing-data-confirmatory-clinical-trials_en.pdf

[22] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2776905

[23] https://tenk.faktisk.no/ordbok Feilinformasjon

[24] https://tenk.faktisk.no/ordbok Desinformasjon

[25] https://www.nature.com/articles/s41591-024-03445-1

[26] https://arxiv.org/abs/2201.11706

[27] https://www.nature.com/articles/s41562-024-02077-2#Sec2

[28] https://ebm.bmj.com/content/24/2/53.long

[29] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114023

[30] https://arxiv.org/abs/2101.00027

[31] URLer for domenet .no utgjorde 0.3% av alle URL'er i web-crawl for Common Crawl februar 2025: https://commoncrawl.github.io/cc-crawl-statistics/plots/tld/latestcrawl.html

[32] https://link.springer.com/article/10.1007/s00146-020-00945-9#ref-CR19

[33] https://www.who.int/publications/i/item/9789240084759 s. 10

[34] https://cordis.europa.eu/article/id/27270-exclusion-from-clinical-trials-harming-womens-health

[35] https://gh.bmj.com/content/8/11/e013473

[36] https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.23157

[37] https://uit.no/research/sshf-no?p_document_id=674134&Baseurl=/research/#region_752662 Senter for samisk helseforskning (SSHF) ble etablert i 2001 som følge av manglende kunnskap om helse og levekår til den samiske befolkningen.

[38] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf s16

[39] Center for Research on Foundation Models (CRFM) ved Stanford: https://crfm.stanford.edu/fmti/paper.pdf

[40] https://crfm.stanford.edu/fmti/May-2024/index.html 2024

[41] https://www.ibm.com/think/topics/black-box-ai

[42] https://www.ibm.com/think/topics/interpretability

[43] KI-faktaark publiseres her: KI-faktaark

[44] https://ai.nejm.org/doi/pdf/10.1056/AIra2400380

[45] KI-faktaark publiseres her: KI-faktaark

[46] https://ai.nejm.org/doi/pdf/10.1056/AIra2400380

[47] ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf

[48] https://arxiv.org/abs/2401.11817

[49] https://www.ibm.com/think/topics/model-drift

[50] https://www.who.int/publications/i/item/9789240084759 s.36

[51] https://arxiv.org/abs/2303.08774

[52] https://www.nature.com/articles/s41586-024-07566-y

[53] NTNU, sluttrapport: Copilot med personvernbriller på | Datatilsynet

[54] OWASP Top 10 for LLM Applications 2025 - OWASP Top 10 for LLM & Generative AI Security

[55] https://atlas.mitre.org/

Risiko ved bruk i helse- og omsorgstjenesten

Risiko er tett knyttet opp til bruksområder. Konsekvensene av de iboende utfordringene ved store språkmodeller, som hallusinering, bias og mangel på transparens (3.1), vil kunne variere fra ubetydelig til kritiske, avhengig av hvordan og hvor modellene brukes. Noen utfordringer oppstår i brukssituasjonen mens andre fremtrer over tid og med økt bruk.

Noen risikoområder kan medføre konsekvenser for pasientsikkerheten og kvaliteten i helsetjenesten (Figur 2): dialogen mellom menneske og maskin, avhengighet som kan oppstå ved utstrakt bruk av KI-verktøy og personvern ved håndtering av sensitive data. Andre risikoer er mer overordnet og knytter seg til kompleksitet i og samspill mellom ulike regelverk, utilsiktet ulovlig bruk av store språkmodeller, diskriminering av enkeltpersoner eller grupper på bakgrunn av skjevheter i modellene, miljø og bærekraft og hvem som har ansvar dersom feil oppstår.

Figur 2: Risikoer knyttet til bruken av store språkmodeller i helse- og omsorgstjenesten. Risikoer inkluderer mangelfull kompetanse om dialog med språkmodeller, reduserte faglige kunnskaper og ferdigheter, redusert personvern, komplekst regelverk, utilsiktet ulovlig bruk, diskriminering, miljø og bærekraft, vitenskapelige bevis og ansvar ved skade.

Arbeidet med rapporten viser at det særlig er usikkerheter forbundet med utvikling og bruk av generativ KI med høy risiko, som i diagnostikk og behandling. Blant annet så stiller WHO spørsmål ved om store språkmodeller vil kunne oppnå tilstrekkelig nøyaktighet til å rettferdiggjøre kostnader forbundet med utvikling, og sikker og effektiv implementering av slike verktøy i gitte bruksområder innen helsehjelpen [56].

De amerikanske myndighetene (*Food and Drug Administrasjon* (FDA)) har gjort en vurdering av medisinsk utstyr som benytter generativ kunstig intelligens (KI). Der peker de på flere utfordringer og risikoer som hallusinering, mangel på transparens, løpende læring, utfordringer med vitenskapelig dokumentasjon og behov for nye evalueringsmetoder (se faktaboks om FDA-rapporten under) [57]. Vi er ikke kjent med at EU har publisert tilsvarende.

Vurderinger knyttet til medisinsk utstyr med generativ KI utført av FDA

FDA har vurdert medisinsk utstyr som benytter generativ kunstig intelligens (KI) og peker på flere utfordringer og risikoer:

- *Hallusineringer*: Generativ KI kan produsere feilaktig innhold ("hallusineringer"). For eksempel kan utstyr som er laget for å oppsummere en samtale mellom pasient og helsepersonell utilsiktet generere en falsk diagnose som aldri ble diskutert.
- *Manglende transparens*: Utstyr som bruker grunnmodeller utviklet av tredjeparter har ofte begrenset tilgang til informasjon om modellenes arkitektur, treningsmetoder og datasett. Dette kan gjøre det vanskelig for produsenter å sikre kvaliteten og sikkerheten.
- *Løpende læring*: Generative modeller kan enten være statiske eller endre seg kontinuerlig ("løpende læring"). Løpende læring er forbundet med usikkerhet rundt modellens ytelse over tid, og pr. november 2024 hadde ikke FDA godkjent medisinsk utstyr som bruker løpende læring.
- *Utfordringer med vitenskapelig dokumentasjon*: Det kan være vanskelig å avgjøre hvilken type gyldig vitenskapelig dokumentasjon ("evidens") FDA bør be om for å kunne vurdere utstyrets sikkerhet og effektivitet gjennom hele livssyklusen.
- *Utfordringer med FDAs klassifiseringssystem:* Generativ KI kan introdusere nye eller annerledes risikoer som utfordrer FDAs nåværende klassifiseringssystem for medisinsk utstyr. Hvordan utstyr klassifiseres påvirker hvilke regulatoriske tiltak som er nødvendige for å sikre at utstyret blir trygt og effektivt.

FDA trekker også frem utfordringer knyttet til evaluering og testing av generative modeller:

- *Evaluering før markedslansering:* Store språkmodeller har komplekse parametere og kan gi ulike svar basert på små endringer i formulering eller prompter. Det er ikke mulig å teste alle mulige prompter før lansering. Dessuten kan manglende transparens og potensialet for uforutsette svar gjøre det spesielt vanskelig å evaluere slike systemer før de kommer på markedet. De (FDA) peker på viktigheten av planer og metoder for overvåking etter at utstyret er tatt i bruk.
- *Behov for nye evalueringsmetoder:* Dagens metoder for kvantitativ evaluering av ytelse kan være utilstrekkelige for å sikre trygg bruk av generativ KI. Nye, kvalitative evalueringsmetoder kan bidra til å karakterisere/beskrive modellens autonomi, transparens og forklarbarhet. Hvilke evalueringsmetoder som kreves, vil variere ut fra produktets spesifikke bruksområde og design.

FDA anbefaler produsenter av medisinsk utstyr å vurdere følgende sammenlignet med et ikke-generativt alternativ:

- Vil inkludering av generativ KI øke risikoklassen til et medisinsk utstyr?
- Vil et produkt med generativ KI kunne gi feilinformasjon og dermed utgjøre en risiko for folkehelsen?
- Er generativ KI hensiktsmessig for det tiltenkte bruksområdet?

Kilde: https://www.fda.gov/media/182871/download

3.2.1 Mangelfull kompetanse om dialogen med språkmodeller

Spørringer, også kalt prompting, ved hjelp av naturlig språk gjør det enkelt for brukere med ulik kompetanse og bakgrunn å benytte store språkmodeller. Svarene kan imidlertid variere basert på den eksakte ordlyden i spørsmålet som stilles og det er dermed ikke likegyldig hvordan spørsmålene formuleres [58]. Visse formuleringer, upresise spørringer og mangelfull beskrivelse av kontekst kan gi feil eller unøyaktig svar.

Kunnskap om dialogen med språkmodeller kan imidlertid forbedre kvaliteten til svarene fra språkmodeller. For eksempel har en studie vist at strukturerte prompter ga bedre resultater enn ustrukturerte [59] og at to-trinns resonnering (først spørre om sammendrag, så spørre om diagnose) ga bedre diagnostisk nøyaktighet enn enkel prompting [60]. En studie demonstrerte hvordan mer avanserte modeller (som GPT-4) håndterte komplekse prompter godt, mens svakere modeller (som GPT-3,5) presterte dårligere med slike prompter på kliniske resonneringsoppgaver [61]. En annen studie har vist at store språkmodeller har begrenset nytte for klinikere som beslutningsstøtteverktøy, men at de kan ha et større nyttepotensiale hvis brukerne får god opplæring i å formulere gode prompter [62].

Prompt og systemprompt

Prompt: En tekstbasert instruksjon som sendes til en språkmodell (som ChatGPT) for å få et svar. Et prompt er brukerens spørsmål eller forespørsler, og kvaliteten på et prompt påvirker direkte relevansen og kvaliteten på svaret.

Systemprompt: En overordnet instruksjon som definerer språkmodellens rolle, begrensninger og oppførsel. Et systemprompt er vanligvis ikke synlig for sluttbrukeren, men styrer hvordan modellen tolker og svarer på alle brukerensprompt. Et systemprompt kan for eksempel definere at modellen skal opptre som en helsefaglig assistent, svare kortfattet, og alltid henvise til faglige retningslinjer.

3.2.2 Reduserte faglige kunnskaper og ferdigheter

Økende bruk av KI og store språkmodeller i helse- og omsorgstjenesten kan føre til en gradvis avhengighet av verktøyene. Helsepersonell og pasienter kan komme til å støtte seg til teknologien, noe som på sikt kan påvirke både evnen til å gjøre egne vurderinger og innlærte ferdigheter.

Bekreftelsesbias (*confirmation bias, algorithmic appreciation*) innebærer at brukere ubevisst vektlegger informasjon som bekrefter deres eksisterende oppfatninger [63]. Både måten et prompt formuleres på, og tolkning av svaret gitt av en språkmodell kan påvirkes av bekreftelsestendens. Dette kan være en særlig utfordring ved bruk av KI-verktøy i helse- og omsorgstjenesten, da helsepersonell kan bli mer tilbøyelige til å akseptere svar fra en språkmodell dersom det stemmer overens med forventningene [64]. Det er dermed risiko for at feilaktige svar ikke blir oppdaget, spesielt dersom modellen presenterer informasjon med stor

selvsikkerhet, både med og uten kildehenvisning. Opplæring i kritisk bruk av KI-verktøy vil være avgjørende for å redusere denne risikoen.

Automatiseringsbias (*automation bias*) er en overdreven tiltro til KI-verktøy som kan forringe faglig integritet og kvalitet. Jo bedre KI-systemet blir, jo oftere vil det gi riktig svar og desto vanskeligere blir det for helsepersonell å fange opp de sjeldne tilfellene der de får feil svar fra KI-systemet. Dette kan føre til at brukere ikke stiller spørsmål ved eller vurderer andre kilder til informasjon, fordi de antar at KI-systemet er korrekt.

Tap av ferdigheter (deskilling). Flere kilder, deriblant WHO og National Health Service (NHS) i England, peker på tap av ferdigheter som en risiko ved utstrakt bruk av KI-systemer i helsetjenesten [65][66]. Tap av ferdigheter, kunnskap eller svekket beslutningsevne kan skje dersom ferdigheter ikke er i jevnlig bruk, som for eksempel hvis oppgaver overlates til maskiner, inkludert språkmodeller. Resultatet kan være at helsepersonell ikke overprøver eller utfordrer en beslutning foreslått av en modell. Det kan også føre til at de ikke kan utføre visse typer oppgaver eller prosedyrer i tilfeller der modellen er utilgjengelig, for eksempel ved nettverksfeil eller sikkerhetsbrudd.

En internasjonal undersøkelse blant helsepersonell viser at en stor andel er bekymret for at bruk av generativ KI kan svekke kritisk tenkning og føre til økt avhengighet av KI i kliniske beslutninger[67]. Tap av ferdigheter vil også gjelde neste generasjon helsepersonell som i økende grad vil møte KI-verktøy som en del av utdanning, opplæring og endringer i klinisk praksis [68]. Økt erfaring om bruk av store språkmodeller vil legge grunnlag for ny forskning som på sikt vil kunne endre vår forståelse av denne risikoen.

3.2.3 Redusert personvern og anonyme opplysninger

Det vil være risiko knyttet til etterlevelse av krav til personvern når store språkmodeller er i aktiv bruk. Hvis man legger inn personsensitive data i et prompt til en språkmodell som lagrer data og/eller lærer løpende, for eksempel i en webapplikasjon, er det risiko for at sensitive data brukes til å trene og oppdatere modellen, og at de kan komme på avveie.

En måte å sikre språkmodeller som håndterer personsensitiv informasjon på, kan være å lagre og kjøre modellen i private skytjenester eller lokalt (*on-prem*-løsninger). Innstillinger som sikrer kryptering av data (i bruk) og at data ikke skal lagres er andre metoder som skal sikre at inn- eller utdata ikke eksponeres.

En vanlig teknikk for å etterleve krav til personvern, er å kun dele og prosessere anonyme opplysninger. Risikoen for personvernet reduseres dersom personopplysninger blir anonymisert fordi de registrerte ikke lenger vil kunne gjenkjennes i datasettet. Anonyme opplysninger kan ikke knyttes til en enkeltperson, og er dermed ikke personopplysninger som reguleres av personvernforordningen.

Personopplysninger regnes som anonymiserte når de håndteres eller bearbeides slik at de ikke lenger kan knyttes til en identifisert eller identifiserbar fysisk person. I vurderingen av om opplysningene er anonyme eller ikke, må man se om det er mulig å spore opplysningene tilbake til de enkeltpersonene opplysningene knytter seg til.

Datasett som skal anonymiseres, må bearbeides for å unngå muligheter for reidentifisering (metode for å finne tilbake til en persons identitet fra anonymiserte data) eller bakveisidentifisering (metode for reidentifisering, ofte ved hjelp av kombinasjon med andre datakilder). Å vurdere hvorvidt opplysningene i et

datasett er å anse som reelt anonyme som resultat av anonymiseringsprosessen, vil bero på en helhetlig og risikobasert vurdering som påhviler dataansvarlig. Reell anonymisering kan være vanskelig å oppnå for enkelte datasett, eksempelvis for svært omfattende datasett med mange variabler.

Med dagens tilgang til store mengder data og kraftig analyseteknologi, vil det i større grad enn tidligere være mulig å reidentifisere enkeltpersoner gjennom opplysninger i datasettet [69]. Denne risikoen er det viktig å vurdere ved deling av opplysninger.

3.2.4 Komplekst regelverk

Utvikling og bruk av kunstig intelligens, herunder store språkmodeller i helsetjenesten, må skje innenfor rammene av gjeldende rett til enhver tid. Brukere (*deployere*) så vel som utviklere av KI-systemer som skal anvendes i helse- og omsorgstjenesten i Norge, må forholde seg til en rekke lover og forskrifter. Det gjelder både generelle og sektorspesifikke regelverk. For eksempel gjelder sektorspesifikke regelverk som norsk helselovgivningen og EUs forordning om medisinsk utstyr. Videre gjelder generelle regelverk som blant annet EUs personvernforordning, åndsverksloven og likestilling- og diskrimineringsloven. I tillegg kommer KI-forordningen, som har trådt i kraft i EU og så raskt som mulig skal innføres i norsk rett [70].

Sammenhengen mellom de ulike og noen ganger overlappende regelverkene er kompleks. Kompleksiteten kan føre til ulik regelverksforståelse og tolkning, at prosesser tar tid eller at handlingsrommet i regelverket ikke utnyttes full ut.

Uklarheter eller usikkerhet om det juridiske handlingsrommet innen gjeldende rett kan føre til manglende etterlevelse ved at virksomheter, helsepersonell eller innbyggere bruker KI-systemer med store språkmodeller i strid med krav oppstilt i regelverket, eller at de legger til grunn en for restriktiv tolkning i forhold til det handlingsrommet i regelverket. Det kan for eksempel være usikkerhet om et KI-system skal klassifiseres som medisinsk utstyr eller ikke, og i så fall hvilken risikoklasse det tilhører [71].

Risikoklasse i henhold til regelverket for medisinsk utstyr er førende for hvilken risikoklasse verktøyet får i henhold til KI-forordningen, og dermed avgjørende for hvilke krav som stilles til blant annet kvalitetssikring og dokumentasjon i henhold til begge lovene. KI-systemer som klassifiseres som medisinsk utstyr blir ofte [72] også klassifisert som høy-risikosystem i henhold til KI-forordningen [73]. For KI-verktøy hvor det er uklart om det er omfattet av definisjonen av medisinsk utstyr, eller hvilken klasse av medisinsk utstyr som gjelder, vil det derfor skape usikkerhet om hvorvidt det også er omfattet av flere krav knytet til KI-forordningen. Eksempler på dette kan være snakkeroboter som brukes i terapi, eller applikasjoner som bruker generativ KI for å lage journalutkast [74].

3.2.5 Utilsiktet ulovlig bruk

Store språkmodeller til allment bruk kan brukes til en rekke oppgaver, også av helsepersonell og pasienter. Fordi teknologien er lett tilgjengelig og enkel å bruke, er det en risiko for at KI-systemer tas i bruk til formål det ikke er utviklet eller regulert for. Noen eksempler er gitt nedenfor.

- *Ulovlig bruk av store språkmodeller i helsetjenesten.* Hvis helsepersonell skal bruke et KI-verktøy til medisinske formål må de i henhold til håndteringsforskriften bruke CE-merket utstyr [75]. Mangel på kvalitetssikrede KI-verktøy, gjennom for eksempel CE-merking eller annen kvalitetssikring, kan føre til at helsepersonell benytter store språkmodeller for bruksområder de ikke er tiltenkt, og dermed heller ikke er lovlig. Utålmodighet, uvitenhet eller mangel på tydelige retningslinjer kan føre til at helsepersonell bruker språkmodeller til oppgaver som ikke er i tråd med bruksområde(ne) definert av produsenter eller tilbydere av KI-verktøy.

- *Uforsvarlig bruk av KI-verktøy av innbyggere.* Innbyggere og pasienter har lett tilgang til KI-verktøy som markedsføres som livsstilsverktøy, og KI-modeller til allmenne formål som ikke har spesifikke medisinske formål. Innbyggerne kan likevel bruke verktøyene til for eksempel selvdiagnostisering eller til å få medisinske råd. Det er dermed en risiko for at de får feilaktig eller misvisende informasjon og/eller at informasjonen feiltolkes. Det kan påvirke helsebeslutninger og i verste fall true pasientsikkerheten dersom de ikke er i kontakt med helsetjenesten [76].

Det er også risiko knyttet til ulovlig bruk av helseopplysninger til trening av store språkmodeller. Dersom data som inneholder sensitiv informasjon skal benyttes til trening av en stor språkmodell åpner § 29 i helsepersonelloven for at det kan gis dispensasjon fra taushetsplikten for å bruke helseopplysninger innsamlet i helsetjenesten til gitte formål [77]. Hvis modellen på et senere tidspunkt tas i bruk til andre oppgaver, som medfører at man bruker helseopplysninger til et annet formål enn det dispensasjonsvedtaket omfatter, kreves det at man har lovlig grunnlag for bruk av opplysninger til det nye formålet. Hvis ikke man sørger for dette, kan det medføre bruk av helseopplysninger til nye formål uten tilstrekkelig hjemmelsgrunnlag.

3.2.6 Risiko for diskriminering

Manglende representativitet i treningsdata for store språkmodeller kan føre til skjevheter (*bias*) og dermed risiko for diskriminering i helse- og omsorgstjenesten, noe som også WHO påpeker i sin rapport [78].

Dersom skjevhetene påvirker noens rett til ytelser eller tilgang til helsetjenester, kan det utgjøre ulovlig diskriminering. KI-systemer som kan medføre denne risikoen faller inn under KI-forordningens definisjon av høy- risiko KI-systemer [79]. I KI-forordningen stilles blant annet krav til at både offentlige og private brukere (*deployers*) av høy-risiko KI-systemer i slike tilfeller må utføre en *Fundamental Rights Impact Assessment* (FRIA). Dette er en vurdering av hvordan et KI-system kan påvirke menneskerettighetene, på både individ- og gruppe-nivå[80]. Hvis det avdekkes risiko for diskriminering av en spesiell gruppe pasienter når et KI-system brukes kan et tiltak være å tilby dem et alternativ uten KI.

3.2.7 Miljø og bærekraft

Utvikling, testing og bruk av store språkmodeller krever ofte høyt energiforbruk og vil dermed kunne sette et betydelig miljøfotavtrykk. Energiforbruket knytter seg til forhåndstrening og ettertrening av modellene [81], og til bruk [82]. Den miljømessige påvirkningen omfatter også høyt forbruk av vann til nedkjøling av store datasentre og utvinning av sjeldne mineraler som er nødvendige for dagens maskinvare brukt til modelltrening[83][84].

Det anslås at helsesektoren står for 4-5 % av globale klimagassutslipp, og både innkjøp og bruk av utstyr og informasjonsteknologi bidrar til utslippene. Helsedirektoratet har publisert et veikart som inneholder konkrete tiltak for hvordan helse- og omsorgstjenesten kan bli mer miljøvennlig [85]. KI trekkes her frem som en teknologi som, brukt klokt, kan legge til rette for en mer bærekraftig helsetjeneste, samtidig som den er ressurskrevende.

Det er krav om at klima og miljø skal vektes med 30% i nye offentlige anskaffelser dersom relevant [86]. Samtidig kan det være vanskelig å få oversikt over klimaavtrykket ved anskaffelse av KI-løsninger. Leverandører oppgir ofte lite informasjon om energiforbruk og karbonutslipp, og det finnes ingen felles standard for hvordan slike beregninger skal gjennomføres. I tilknytning til krav om transparens i KI-forordningen vil det stilles spesifikke krav om dokumentasjon av energiforbruk til leverandører av store språkmodeller som skal brukes i EU [87].

Som et alternativ til store språkmodeller utvikles det også små modeller som krever betydelig mindre energi og ressurser både til trening og i bruk [88]. Ved hjelp av teknikker som modell-destillering [89] og målrettet finjustering yter disse modellene stadig bedre, og kan være et godt alternativ for å løse ulike oppgaver [90].

3.2.8 Vitenskapelige bevis og ansvar

Innføring av KI-verktøy med konsekvenser for pasienters medisinske behandling krever varsomhet. Konsekvensene av bruk av verktøyene kan være sammenlignbare med nye medikamenter eller annet medisinsk utstyr. Det vil være nødvendig med utvikling av nye tester og områder for evaluering og av store språkmodeller som skal brukes innen helse, samt studier og dokumentasjon av disse [91][92]. For eksempel anbefaler WHO å gjøre kliniske randomiserte studier for å bevise at et KI-system som skal brukes i klinisk praksis har bedre ytelse enn alternativer, og ikke kun i et laboratorium eller kontrollerte omgivelser [93]. Uten godt vitenskapelig grunnlag, rett opplæring og korrekt implementering vil ansvarsbyrden være uklar dersom feil har skjedd og helsepersonell har støttet seg på svar fra KI-verktøy. Slike situasjoner kan ha store etiske og juridiske konsekvenser. Merk at virksomhetsledere har ansvar for å legge til rette for korrekt implementering, opplæring og organisering.

Europakommisjonen anerkjente utfordringer knyttet til ansvarsspørsmål ved bruk av kunstig intelligens, og lanserte et forslag til et eget KI-erstatningsdirektiv (AI Liability Directive) i 2022 [94]. Målet til direktivet var å gi økt rettssikkerhet for både pasienter og helsepersonell dersom skade skulle oppstå som følge av bruk av KI-verktøy. I februar 2025 ble forslaget trukket tilbake på grunn av mangel på enighet blant medlemslandene, men problemstillingen er fremdeles like aktuell.

[56] https://www.who.int/publications/i/item/9789240084759

[57] https://www.fda.gov/media/182871/download

[58] https://arxiv.org/abs/2311.16452

[59] https://www.medrxiv.org/content/10.1101/2024.09.01.24312894v1.full

[60] https://www.medrxiv.org/content/10.1101/2024.09.01.24312894v1.full

[61] https://arxiv.org/abs/2308.06834

[62] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395#zoi241182r32

[63] https://journals.sagepub.com/doi/10.1037/1089-2680.2.2.175

[64] https://ai.nejm.org/doi/full/10.1056/AIe2400961

[65] https://www.who.int/publications/i/item/9789240084759 s.12

[66] https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/developing-he

[67] https://assets.ctfassets.net/o78em1y1w4i4/kWTSca6VXZ54DBhAIYxJU/386d36dc0c03c4fa8de0365bbb20

[68] https://pubmed.ncbi.nlm.nih.gov/38662419/

[69] https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/informasjonssikkerhet-internkontroll

[70] Regjeringen vil jobbe for at KI-forordningen blir innlemmet i EØS-avtalen så raskt som mulig: Fremtidens digitale Norge – nasjonal digitaliseringsstrategi 2024–2030, side 67.

[71] https://www.scup.com/doi/10.18261/olr.10.1.1

[72] Medisinsk utstyr av klasse IIa eller høyere etter loven om medisinsk utstyr

[73] Artikkel 6 i KI-forordningen: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689#cpt_III.sct_1

[74] https://www.scup.com/doi/10.18261/olr.10.1.1

[75] https://lovdata.no/dokument/SF/forskrift/2013-11-29-1373

[76] Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models s.15

[77] Reglene for utvikling og bruk av klinisk beslutningsstøtteverktøy - Helsedirektoratet

[78] Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models s. xi Tabell 2, s.21

[79] Punkt 5 a i vedlegg III i KI-forordningen

[80] Artikkel 27 i KI-forordningen

[81] Treningen av GPT-4 estimeres å ha forårsaket rundt 300 tonn CO-utslipp, til sammenligning er estimatet for et menneske 5 tonn årlig https://hbr.org/2023/07/how-to-make-generative-ai-greener

[82] Hugging Face sin ecological impact calculator estimerer at én spørring til GPT-4o med et relativt kort svar (400 tokens) forbruker 35,1 Wh strøm, ca. syv ganger mer enn å lade en mobiltelefon. EcoLogits Calculator - a Hugging Face Space by genai-impact besøkt februar 2025

[83] Reconciling the contrasting narratives on the environmental impact of large language models | Scientific Reports

[84] https://arxiv.org/pdf/2304.03271

[85] Veikart mot en bærekraftig, lavutslipps og klimatilpasset helse- og omsorgstjeneste - Helsedirektoratet

[86] Veileder til regler om klima- og miljøhensyn i offentlige anskaffelser | Anskaffelser.no

[87] https://code-of-practice.ai/?section=transparency#model-documentation-form endelig versjon av "Code of Conduct" ventet mai 2025

[88] Small language models (SLMs): A cheaper, greener route into AI | UNESCO

[89] Destillering er en teknikk der output fra en stor, mer kompleks modell blir brukt til å finjustere en mindre modell med mål om å oppnå lignende ytelse. What is Knowledge distillation? | IBM

[90] [2305.02301] Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes

[91] https://ai.nejm.org/doi/full/10.1056/AIe2401235

[92] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf s.28

[93] Ethics and governance of artificial intelligence for health

[94] https://www.ai-liability-directive.com/

Oppsummering

Bruk av generativ KI og store språkmodeller er foreløpig et umodent område. Selv om mulighetene er mange, er det også stor usikkerhet om risikoer knyttet til hvordan denne type KI kan brukes på en tillitsverdig måte i helse- og omsorgstjenesten. Vi mangler rett og slett innsikt i bivirkningene.

Fagområdet utvikler seg imidlertid raskt, med blant annet mekanismer for å forbedre, kunnskapsforankre og få bedre kontroll på kvaliteten til KI-systemer med generative språkmodeller.

Ett sentralt tiltak for å redusere risikoer og øke kvaliteten er å utvikle og gjøre tilgjengelig språkmodeller, det vil si grunnmodeller, som er tilpasset forholdene i den norske helse- og omsorgstjenesten, og som næringslivet og offentlig sektor kan bygge videre på til sitt bruk. Dette er i tråd med den nasjonale digitaliseringsstrategien som stadfester at Regjeringen vil jobbe for å videreutvikle en nasjonal infrastruktur for KI som skal gi tilgang til grunnmodeller tuftet på norske og samiske språk og samfunnsforhold [95].

De neste kapitlene omhandler hvordan store språkmodeller kan tilpasses den norske helse- og omsorgstjenesten (kapittel 4) og forslår tiltak som til sammen vil legge til rette for at bruken av generative KI-modeller kan baserer seg på grunnmodeller som er tilpasset forholdene i den norske helse- og omsorgstjenesten (kapittel 5).

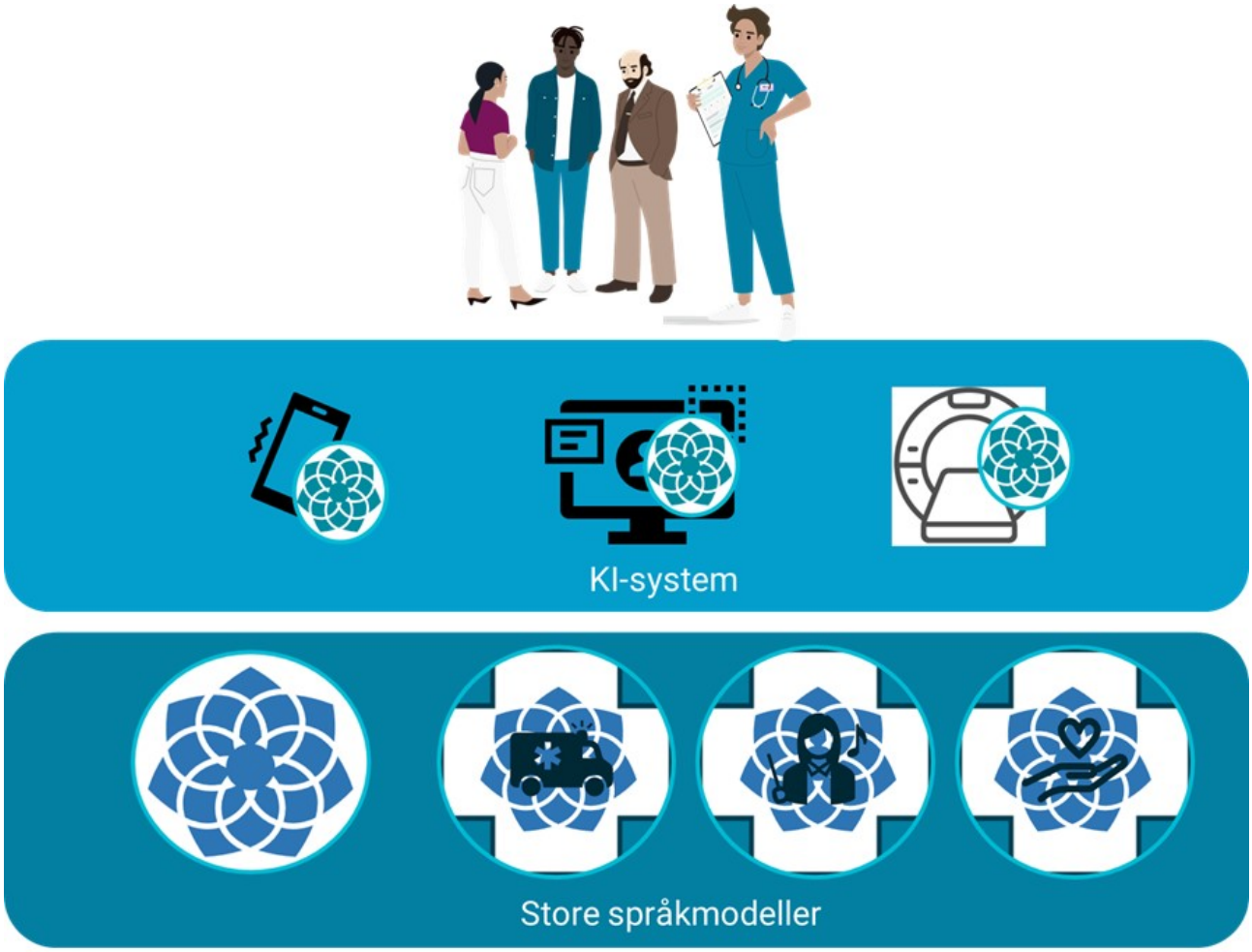
[95] I Statsbudsjettet for 2025 bevilges 20 millioner kroner til økt satsing på Nasjonalbibliotekets arbeid med kunstig intelligens og trening av generative språkmodeller. https://www.nb.no/pressemeldinger/regjeringen-vil-satse-stort-pa-digitalisering-og-kunstig-intelligens-ved-na

Tilpassing til den norske helse- og omsorgstjenesten

For at en generativ språkmodell skal kunne brukes på en trygg og effektiv måte i norsk helse- og omsorgstjeneste, bør den være godt tilpasset norske forhold. I dette kapitlet vil vi se på *hva* tilpassing av språkmodeller til norske forhold innebærer, *hvordan* språkmodeller kan tilpasses norske forhold, *hva som trengs* for å tilpasse språkmodeller norske forhold, og *hvem* som kan tilpasse språkmodellene.

I denne rapporten ser vi i liten grad på generell tilpassing til norske forhold, for eksempel språk i Norge eller generell offentlig forvaltning og kunnskap om samfunnsforhold og verdier, men først og fremt på tilpassing til den norske helse- og omsorgstjenesten. På dette området vil vi søke å samarbeide med arbeid som skjer tverrsektorielt.

Det å tilpasse til norske forhold utelukker ikke at språkmodellen samtidig bør være tilpasset internasjonale forhold. Helsetjenestene blir også utført i en internasjonal kontekst, og helsefaglig kunnskap og nyvinninger skjer internasjonalt. En språkmodell bør ikke være tilpasset norske forhold ene og alene.



Figur 3: Store språkmodeller kan være komponenter i KI-systemer. KI-systemer kan bygge på en eller flere ulike språkmodeller og har gjerne et tiltenkt bruksområde.

Hva er tilpassing til norske forhold?

Norske forhold er et bredt og vagt begrep som kan omfatte svært mye. Noe vil gjelde generelt for bruk i Norge, men mye vil gjelde spesifikt for helse- og omsorgssektoren i Norge. Vi vil se nærmere på det siste.

Basert på innhentet kunnskap kan tilpasning til norske forhold deles opp i fem områder: språk i Norge, allmennkunnskap i Norge, offentlig forvaltning i Norge, lovverk i Norge, verdier og etikk i Norge. På lignende vis kan tilpasning til den norske helse- og omsorgstjenesten deles opp i fem områder: helsefagspråk i Norge, helsefaglig kunnskap og praksis i Norge, helseadministrativ kunnskap og praksis i Norge, norsk helselovgivning og verdier og etikk som gjelder i helse- og omsorgstjenesten [96], se Figur 4.

Figur 4. Språkmodeller kan tilpasses ulike områder: til norske forhold generelt og for helse- og omsorgstjenesten spesielt. Områder for tilpassing gjelder språk, faglig kunnskap, administrativ kunnskap, lovverk og verdier og etikk.

4.1.1 Språk i Norge

4.1.1.1 Bokmål og nynorsk

Språkmodeller tilpasset norske forhold, må være i stand til å utføre oppgavene på norsk. Språkloven gir flere føringer for det offentliges språkbruk i Norge. Ifølge språklovens § 4 er norsk det nasjonale hovedspråket i Norge, og bokmål og nynorsk er likestilte skriftspråk i offentlige organ [97]. I tillegg skal offentlige organ følge den offisielle rettskrivningen for både bokmål og nynorsk. Språkmodeller som skal benyttes i den offentlige helse- og omsorgstjenesten, må derfor beherske bokmål og nynorsk innenfor den fastsatte rettskrivningen. Språkmodeller bør også kunne håndtere variasjonen innenfor bokmål og nynorsk.

Språkrådet slår fast i en språklig undersøkelse fra oktober 2024 at språket i ChatGPT ikke holdt tilstrekkelig høy nok kvalitet til å benyttes av offentlig forvaltning [98]. Mye taler for at språkmodeller fremdeles ikke er godt nok tilpasset norsk språk.

4.1.1.2 Samiske språk

Samiske språk, dvs. lulesamisk, sørsamisk og nordsamisk, har status som urspråk. Ifølge språklovens § 5 er samiske språk og norsk språk likeverdige [99].

Samelovens formål er å legge forholdene til rette for at den samiske folkegruppe i Norge kan sikre og utvikle sitt språk, sin kultur og sitt samfunnsliv. Den stiller krav til når offentlige organer skal bruke samiske språk. Samtidig gir loven enkeltpersoner språklige rettigheter i møte med offentlige organ. Sameloven § 3-5 gir personer utvidet rett til bruk av samisk i helse- og omsorgsinstitusjoner i forvaltningsområdet for samisk språk:[100][101][102].

Det bør arbeides for å utvikle og tilgjengeliggjøre språkmodeller som fungerer på samisk med både det språklige og kulturelle perspektivet.

4.1.1.3 Andre språk i Norge

Norge er et flerspråklig land, og vi finner en rekke andre språk i bruk i tillegg til norsk og samiske språk: de nasjonale minoritetsspråkene, skandinavisk, norsk tegnspråk og nyere minoritetsspråk.

Nordsamisk, sørsamisk, lulesamisk, kvensk, romanes og romani er alle definert som minoritetsspråk i Norge, og dermed beskyttet av minoritetsspråkpakten [103]. Språkbrukere av disse minoritetsspråkene har ikke egne, konkrete rettigheter knyttet til møte med helse- og omsorgstjenesten. Tilsvarende som for samisk bør det jobbes for at KI-verktøyene tar hensyn til det kulturelle perspektivet i møtet mellom brukere fra minoritetsgrupper og storsamfunnet, noe som vil være i tråd med pasient- og brukerrettighetsloven. I helse- og omsorgssektoren vil dette i stor grad gjelde de nyere minoritetsspråkene.

Skandinaviske språk kan benyttes i pasientjournalen. Ifølge pasientjournalforskriften § 10 kan dansk og svensk benyttes i den utstrekning det er forsvarlig [104].

Tolkelovens § 6 gir offentlige organer plikt til å "bruke tolk når det er nødvendig for å ivareta hensynet til rettssikkerhet eller for å yte forsvarlig hjelp og tjeneste. I vurderingen av om bruk av tolk er nødvendig, skal det blant annet legges vekt på om samtalepartene kan kommunisere forsvarlig uten tolk, og på sakens alvorlighet og karakter." Skal språkmodeller benyttes i slike tilfeller, for eksempel til automatisk tolking, vil tolkeloven trolig gjelde. Språkmodeller må i så fall tilpasses aktuelle språk.

Flerspråklighet i språkmodeller

Norsk helsevesen benytter en innfløkt blanding av norsk og internasjonal medisinsk terminologi. Dette krever avanserte embedding-rom som kan håndtere flere språk samtidig.

Et embedding-rom er et flerdimensjonalt matematisk rom der ord, setninger, bilder eller andre data representeres som vektorer, slik at lignende begreper plasseres nær hverandre basert på semantisk eller kontekstuell likhet.

Moderne språkmodeller implementerer dette gjennom delte vektorrom for medisinsk terminologi, samtidig som de opprettholder separate representasjoner for bokmål og nynorsk med et felles medisinsk vokabular som fundament. Omfattende testing har vist at denne arkitekturen, spesielt med dedikerte medisinske embedding-nivå, resulterer i en betydelig økning i presisjon ved krysslingvistisk medisinsk kommunikasjon.

Kilde: https://link.springer.com/article/10.1007/s10462-025-11162-5

4.1.1.4 Norsk helsefaglig språk

Skal språkmodeller benyttes i helsefaglige sammenhenger, bør de i tillegg til norsk beherske det norske helsefaglige fagspråket, inkludert fagterminologien. Fagterminologi omfatter spesialiserte faguttrykk, forkortelser og sjargong.

Fagspråket kjennetegnes av mange synonymer, der norske, greske, latinske og hybride (latinske termer tilpasset norsk skrivemåte) ofte er i bruk om hverandre. I tillegg varierer fagterminologien i stor grad mellom de ulike yrkesgruppene og spesialitetene.

Språkmodellen bør derfor være i stand til både å motta instrukser og produsere tekst med etablert og riktig norsk terminologi for det aktuelle bruksområdet. Det vil eksempelvis være uheldig om en språkmodell automatisk oversetter og lager helt nye direkteoversettelser der det finnes godt etablerte fagtermer som har hevd i norsk fagspråk. Dette har hittil vært en utfordring ved bruk av KI-verktøy for oversettelse.

4.1.1.5 Dialekter

Skal språkmodeller benyttes til å prosessere tale, for eksempel tale-til-tekst, bør språkmodellene også beherske dialektene som er i bruk i Norge samt andre muntlige talevariasjoner som for eksempel skandinaviske språk. Dialektbruk har vært en ikke ubetydelig utfordring med tidligere språkteknologi. Imidlertid har nyere generative språkmodeller som NB-Whisper vist lovende resultater (se faktaboks NB-Whisper). Dette blir særlig viktig for flere typer verktøy, for eksempel tale-til-sammendrag (*ambient scribe*) og muntlig språkbruk i samtaleroboter (*conversational AI*).

NB-Whisper

Nasjonalbiblioteket har tilpasset språkmodellen Whisper fra selskapet OpenAI til norsk. Det er en språkmodell som kan konvertere tale til tekst, for eksempel samtaler og monologer.

NB-Whisper har blitt trent på store mengder språkressurser fra Nasjonalbiblioteket, inkludert norsk talespråkkorpus. Det gjør at språkmodellen forstår norsk tale og norske dialekter langt bedre enn tidligere teknologi.

Som andre generative modeller kan NB-Whisper også oppleve hallusineringer, noe som er særlig viktig å være oppmerksom på i helse- og omsorgstjenesten.

NB-Whisper er fritt tilgjengelig til ibrukttagelse, også i helse- og omsorgssektoren.

NB-Whisper er et godt eksempel på hvordan en internasjonal språkmodell kan tilpasses norske forhold. Systematisk arbeid med å bygge gode språkressurser i bunn har lagt til rette for at det offentlige eller næringslivet kan utvikle løsninger for helsetjenesten. Flere teknologileverandører tilbyr i dag tale til tekst-teknologi basert på NB-Whisper.

NB-Whisper er tilgjengeliggjort av Nasjonalbiblioteket, som forvalter og videreutvikler modellen.

Kilde:
https://www.nb.no/pressemeldinger/nasjonalbiblioteket-deler-kunstig-intelligens-som-skjoner-norske-dialekt

4.1.1.6 Klarspråk

Språkloven stiller krav til klarspråk, dvs. klart og korrekt språk tilpasset målgruppen. I tillegg er pasient- og brukerrettighetsloven § 3-5 relevant: Når det skal gis informasjon om en pasient eller brukers helsetilstand og helsehjelp, skal informasjonen være tilpasset mottakerens individuelle forutsetninger, blant annet kultur og språkbakgrunn. Språkmodeller bør derfor være i stand til å differensiere språkbruk og valg av ord avhengig av hvem som er brukerne/leserne, enten det er innbyggere med ulike nivåer av helsekompetanse eller fagpersoner.

Forskjell mellom språk dreier seg ikke alltid om ulike ord og setningsstrukturer. Språk er også et kulturelt uttrykk. Normer for god språkbruk, som for eksempel høflighet, stil og tone, varerier mellom kulturer og språk. En tilpasset språkmodell bør derfor også være i stand til å uttrykke seg på en måte som er i tråd med normene for god språkbruk i Norge.

4.1.2 Helsefaglig kunnskap og praksis i Norge

Helsefaglig kunnskap kan være internasjonal og uavhengig av land og kulturer. Samtidig er mye helsefaglig kunnskap betinget av spesifikke kulturområder. Eksempelvis vil kunnskap om kosthold og ernæring og tilhørende råd kunne variere fra land til land, og man snakker gjerne om nordiske kostholdsråd [105]. En språkmodell tilpasset norske forhold må kunne bygge på norsk (eller nordisk) kunnskap og vår helsefaglige forståelse av verden, som kostholdeksemplet illustrerer.

Språkmodeller som brukes i helse- og omsorgstjenesten bør også virke i tråd med praksisen i tjenesten. Det finnes en rekke nasjonale faglige retningslinjer og pasientforløp som representerer normert klinisk praksis i Norge, for eksempel antibiotikaveiledere, pakkeforløp for kreft og veiledende planer for sykepleiepraksis. En forutsetning for trygg bruk av språkmodeller i Norge er nettopp at den blir trent på data som inneholder slike nasjonale retningslinjer.

Det er ikke gjort noen systematisk analyse av i hvilken grad tilgjengelige språkmodeller baserer seg på slik kunnskap.

Uten spesifikk trening eller veiledning, kan språkmodeller falle tilbake på generell eller internasjonal kunnskap fremfor å følge landsspesifikke retningslinjer og praksiser.

Praksiser og retningslinjer kan tydeliggjøres og defineres gjennom standarder, og de kan hjelpe språkmodeller med å generere innhold som er i tråd med norske forhold. En internasjonal undersøkelse kan tyde på at integrering av standarder i språkmodeller kan forbedre deres etterlevelse av slike standarder [106].

4.1.3 Helseadministrativ kunnskap og praksis i Norge

Den norske helse- og omsorgstjenesten er strukturert på en annen måte enn i andre land. Det gjelder blant annet fordelingen mellom privat og offentlig helsetjenestetilbud og den geografiske og administrative innretningen. Videre gjelder det også ordninger som for eksempel fastlegeordningen og skillet mellom primær- og spesialisthelsetjenesten for øvrig, samt den sentrale helseforvaltningen [107].

En språkmodell som er tilpasset norsk helseadministrativ kunnskap og praksis bør være i stand til å håndtere kunnskap om dette. Det er i dag usikkert i hvilken grad tilgjengelige språkmodeller er i stand til det [108].

Det er viktig å være ekstra oppmerksom på at helseadministrativ kunnskap og praksis følger en annen dynamikk enn den helsefaglige ved at den endres gjennom for eksempel politiske vedtak og juridiske endringer.

4.1.4 Helselovgivning i Norge

Rettigheter og plikter til pasienter, pårørende, helsepersonell og virksomheter varierer fra land til land. En språkmodell som er tilpasset norske forhold, innebærer også tilpasning til norsk lovverk.

En rekke lover og forskrifter regulerer helse- og omsorgssektoren, fra generelle lover som helse- og omsorgstjenesteloven til mer spesifikke lover som abortloven. Det er helt avgjørende at en språkmodell i bruk i Norge ikke gir informasjon som bryter med verken norsk lovverk eller norsk helselovverk. Det er ikke gjort noen systematisk analyse av hvorvidt tilgjengelige språkmodeller baserer seg på slik kunnskap.

4.1.5 Verdier og etikk i norsk helse- og omsorgstjeneste

Ingen språkmodell er helt nøytral når det gjelder kulturelle preferanser, ideologi og verdier. Forskning indikerer at store språkmodeller speiler kulturelle verdier fra treningsdataene. Hvilke ytringer som oppfattes som krenkende eller aggressive, vil variere fra kultur til kultur, og vil basere seg på språkmodellenes treningssett [109]. Nye store språkmodeller fra andre kulturområder enn de vestlige har også ført til større oppmerksomhet knyttet til sensur og sensitivitet, også i den offentlige debatten. Allment kjente, men politisk betente politiske hendelser i Kina synes å ha blitt sensurert i DeepSeeks språkmodell [110]. Syn på og kunnskap om kjønn vil eksempelvis variere mellom ulike land og kulturer.

Det samme vil også gjelde helse- og omsorgssektoren. I St.meld. nr. 26 (1999–2000) *Om verdiar for den norske helsetenesta* slås det fast at det ukrenkelige menneskeverdet er den grunnleggende verdien for tjenesten. Videre står likhet, rettferdighet, likeverdig tilgang til tjenester av god kvalitet, faglig forsvarlighet, menneskeverd og solidaritet med de svakest stilte sentralt.

Dette slås også fast i Helsetjenesten i Norge at "[d]et er et grunnleggende prinsipp at i Norge skal alle ha samme rett til helsetjenester, uavhengig av alder, kjønn, bosted og økonomi"[111].

I flere stortingsmeldinger de senere årene har det vært lagt særlig vekt på brukermedvirkning som en grunnleggende forutsetning for god helse og god omsorg [112]. Pasienter, brukere og pårørende skal bli sett og hørt. Brukermedvirkning er en grunnleggende forutsetning i pasientens helsetjeneste, der ingen beslutninger skal tas om deg uten deg. Det skal tas utgangspunkt i menneskets helhetlige behov, herunder fysiske, psykiske, sosiale, åndelige og eksistensielle behov.

Slike verdier er ikke nødvendigvis universelle, og det må sikres at språkmodeller i helse- og omsorgstjenesten er i tråd med dette verdigrunnlaget.

[96] https://tidsskriftet.no/2024/11/kronikk/kritisk-veiskille-integrering-av-kunstig-intelligens

[97] https://lovdata.no/dokument/NL/lov/2021-05-21-42

[98] https://sprakradet.no/aktuelt/ki-sprakets-fallgruver/

[99] https://lovdata.no/dokument/NL/lov/2021-05-21-42?q=spr%C3%A5kloven

[100] Forvaltningsområdet for samiske språk - Sametinget

[101] Den nye forskriften til samelovens språkregler vil føre til utvidelse av hvilke landsomfattende organer samiske språkbrukere skal ha rett til skriftlig svar fra.

[102] https://lovdata.no/dokument/NL/lov/1987-06-12-56/kap3#kap3

[103] Minoritetsspråkpakten - regjeringen.no

[104] https://lovdata.no/dokument/SF/forskrift/2019-03-01-168?q=pasientjournalforskriften

[105] https://pub.norden.org/nord2023-003/recommendations.html

[106] https://arxiv.org/abs/2402.12593

[107] Se for eksempel Nylenna, Magne (2024) "Helsetjenesten i Norge"

[108] ChatGPT 4o ble spurt i januar 2025 om hvilke sykehus som finnes i Møre og Romsdal. Spørsmålet på engelsk gav alle fire sykehusene, mens spørsmålet på norsk gav tre sykehus: Kristiansund sykehus manglet.

[109] https://arxiv.org/html/2402.10946v1 og
https://academic.oup.com/pnasnexus/article/3/9/pgae346/7756548

[110] https://www.kode24.no/artikkel/4-grunner-til-a-vaere-skeptisk-til-deepseek/82594605 ,
https://www.khrono.no/fastlaste-verdier-for-alltid/939158

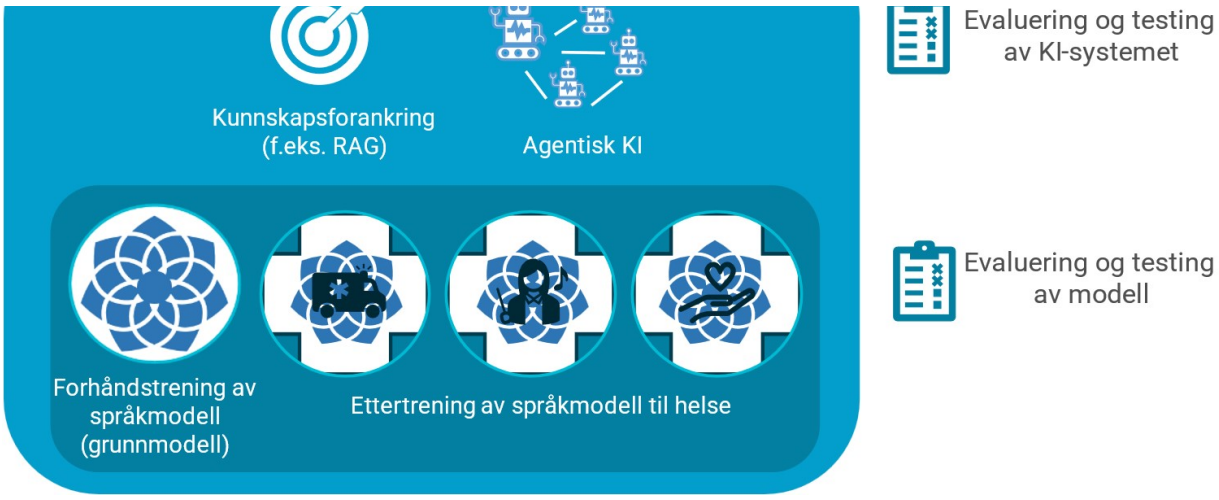
[111] Nylenna, Magne "Helsetjenesten i Norge" (2024, s.33).

[112] For eksempel Meld. St. 38 (2020–2021) Nytte ressurs og alvorlighet – Prioritering i helse- og omsorgstjenesten: https://www.regjeringen.no/no/dokumenter/meld.-st.-38-20202021/id2862026/

Hvordan tilpasse til norske forhold?

Hvordan språkmodeller blir utviklet og tilpasset, er blant annet behandlet i rapporten "Store språkmodellar i helse- og omsorgstenesta" [113].

Metodene for å tilpasse en språkmodell er ikke gjensidig utelukkende. *Forhåndstrening* innebærer å trene en modell fra grunnen av slik at den kan tilpasses norske forhold helt fra begynnelsen av. *Ettertrening* innebærer å bygge videre på en forhåndstrent generell modell med datakilder som er relevante for helsesektoren. Et KI-system som bruker språkmodeller kan også forbedres og tilpasses til den konkrete bruken, ved for eksempel kunnskaps-forankring eller ved agentisk KI. Dette kapitlet beskriver ulike måter å forhåndstrene, ettertrene, kunnskapforankre, evaluere og teste store språkmodeller på.



Figur 5: Tilpassing til norske forhold kan skje ved forhåndstrening av en stor språkmodell og/eller ved ettertrening av forhåndstrente språkmodeller. Et KI-system som bruker språkmodeller kan forbedres og tilpasses til den konkrete bruken, ved for eksempel kunnskapsforankring eller ved såkalt agentisk KI. Både KI-modellen og KI-systemet må evalueres og testes.

4.2.1 Forhåndstrening av språkmodeller

Forhåndstrening av store språkmodeller gir full kontroll over modellen og innsyn i modellens læringsprosess. Tilnærmingen er imidlertid meget kostbar, da den vil kreve omfattende mengder relevante data og betydelige regneressurser. GPT og Gemini er eksempler på store, markedsledende forhåndstrente språkmodeller som har blitt utviklet fra grunnen av. Disse kalles også grunnmodeller, og leveres ofte av store amerikanske teknologiselskaper. Imidlertid blir det pekt på utfordringer knyttet til blant annet forklarbarhet, ansvarlighet, suverenitet og språklig kvalitet i disse modellene [114]. Som et svar på slike utfordringer, har flere europeiske land tatt initiativ til å bygge opp egne forhåndstrente språkmodeller.

De store internasjonale grunnmodellene har vist seg å være kraftige og anvendbare på mange områder, og inneholder allerede en del norsk språk, og brukes i Norge. I Norge er det flere initiativer for å utvikle norske grunnmodeller, både gjennom BERT-baserte modeller som NorBERT og større modeller som NorGPT fra NorwAI og NORA.LLM-modeller fra NORA-konsortiet, for eksempel NorMistral.

Mange store internasjonale grunnmodeller er forhåndstrent på betydelig mengde generell helsefaglig kunnskap. Det finnes også internasjonale språkmodeller som er forhåndstrente fra grunnen av spesifikt for helse. Et eksempel er MedFound, som er trent på pasientjournaler og medisinske tekster [115]. Denne modellen er langt mindre enn de større internasjonale grunnmodellene nevnt ovenfor. Slike mindre modeller kan være mer energieffektive, men bruksområdene blir mer avgrenset siden de er trente på mindre data.

Den teknologiske utviklingen går raskt, og helt nye modellarkitekturer som DeepSeeks språkmodeller R1 og Open R1 representerer en ny og mer energieffektiv retning for hvordan informasjon kan hentes ut og prosesseres. Dette gjør at tidligere tilnærminger, særlig BERT-baserte modeller, i stor grad er blitt utdaterte for generative oppgaver, selv om de fortsatt kan være nyttige for spesifikke anvendelser.

Ved valg av hvilken forhåndstrent modell å bygge videre på, er det viktig å vurdere om den er åpen eller lukket. I åpne modeller kan enten kildekoden være tilgjengelige og/eller treningsdataene være kjente og dokumenterte. Omvendt, vil tilgang til kildekode og/eller treningsgrunnlag være begrenset eller manglende i lukkede modeller. Dette gjør det vanskelig, om ikke umulig, å forklare og kvalitetssikre modeller som baserer seg på dem.

4.2.2 Ettertrening av forhåndstrente språkmodeller (grunnmodeller)

Selv om mange grunnmodeller allerede behersker noe norsk språk og er trent på en del internasjonal helsefaglig kunnskap, bør de ettertrenes hvis de skal bli bedre tilpasset norske forhold. Ettertrening innebærer å bygge videre på en forhåndstrent generell modell med relevant data, og kan basere seg på en eller flere grunnmodeller.

Nasjonalbibliotekets NB-Whisper og Binerics NorskGPT er eksempler på norsk, generell (ikke-helsefaglig) ettertrening.

4.2.2.1 Finjustering med norske helsefaglige språkressurser

Grunnmodeller kan finjusteres ved hjelp av domenespesifikke helsefaglige ressurser som treningsdata. En finjustert språkmodell vil da i langt større grad kunne håndtere norsk fagspråk og terminologi samt helsefaglig og helseadministrativ kunnskap og praksis.

En grunnmodell kan finjusteres til:

- en relativt stor og generell helsefaglig språkmodell. Den vil legge til rette for en felles helsefaglig språkmodell for norsk helse- og omsorgssektor som kan brukes og videreutvikles av hele sektoren, inkludert offentlige og private aktører.
- flere mindre og spesialiserte språkmodeller, innrettet mot ulike formål (fag, spesialitet eller oppgave). De trenes på tekster tilhørende en gitt oppgave eller spesialitet.

Slik finjustering kan skje på ulike måter. En mulig tilnærming er å trene grunnmodellen videre på norskspråklige, helsefaglige tekster slik at algoritmene i modellen endres. Det krever at grunnmodellen er åpen (se ovenfor). Denne formen for finjustering trenger betydelig regnekraft [116]. En annen tilnærming innebærer at finjusteringer skjer som et lag utenfor den forhåndstrente modellen som ikke gjør endringer i den opprinnelige forhåndstrente modellen, og trenger dermed mindre regnekraft [117].

Uavhengig av tilnærming er det viktig å være klar over at en språkmodell aldri vil være utlært. Fagspråk, kunnskap og praksis i helse- og omsorgstjenesten er aldri konstante, men blir oppdatert i tråd med den faglige utviklingen på området. Det er derfor viktig å ta høyde for løpende læring (*continous learning*) ved finjustering av en språkmodell. Det forskes nå på ulike måter dette kan skje på, for eksempel ved federert læring [118]. En annen tilnærming er et oppdateringsregime der språkmodeller blir versjonert og utgitt periodisk med oppdatert kunnskap. Det vil kreve en aktiv forvaltning, se faktaboks om løpende oppdatering av språkmodeller under.

Løpende oppdatering av språkmodeller

Store språkmodeller kan oppdateres løpende ved å etablere en systematisk oppfølgings- og oppdateringsmekanisme slik at for eksempel nye kliniske retningslinjer, oppdaterte data og endringer i fagterminologi raskt kan innlemmes i modellen. Dette kan for eksempel innebære regelmessige oppdateringer gjennom kontinuerlig læring eller periodiske finjusteringer.

Det kan også etableres en feedback-loop med klinisk ekspertise, for eksempel ved å inkludere en mekanisme for kontinuerlig tilbakemelding fra helsepersonell som bruker modellen. Tilbakemeldingene brukes til å identifisere eventuelle feil eller utdaterte anbefalinger, slik at modellen kan revideres og forbedres i tråd med gjeldende fagkunnskap og praksis.

4.2.2.2 Instruksjonsjustering i tråd med helsefaglige og administrative oppgaver

Skal en språkmodell kunne håndtere helsefaglige oppgaver, må den trenes for det. Slike oppgaver kan være oversettelser, tolking, formidle informasjon til pasienter eller lage utkast til journalnotater. Det kan gjøres ved å finjustere språkmodellen med hjelp av datasett som inneholder eksempel på oppgavene som skal løses. Prosessen kan være iterativ og involvere mennesker som gir tilbakemeldinger underveis.

Behovet for instruksjonsjustering gjelder både grunnmodeller og finjusterte modeller. Selv om modeller er instruksjonsjustert på engelsk fra før, kan det likevel være behov for spesifikk norsk instruksjonsjustering. Dette er særlig relevant for domener der presis forståelse av norske instrukser er kritisk, eller der kulturelle og fagspesifikke aspekter er viktige.

4.2.1.3 Finjustering i tråd med norske lovverk, verdier og etikk

En språkmodell kan finjusteres for å fungere i tråd med norsk lovverk, verdier og etiske prinsipper. Slik tilpasning kan gjøres på flere måter:

- gjennom systematisk trening på eksempler som demonstrerer ønsket etisk atferd
- ved hjelp av forsterket læring med menneskelig tilbakemelding (*Reinforcement Learning from Human Feedback* - RLHF) der mennesker vurderer modellens svar opp mot definerte etiske retningslinjer
- ved å bygge inn etiske prinsipper i treningsdataene og evalueringskriteriene

Formålet er at modellens atferd og genererte tekst skal reflektere og være konsistent med definerte etiske prinsipper. Det bør sees i sammenheng med en nasjonal tilnærming til felles verdier og etikk. For helsetjenesten står verdier som konfidensialitet, selv-bestemmelse og ikke-skade sentralt. Samtidig må man ta høyde for at verdier og etiske prinsipper ikke er konstante, men utvikler seg.

Når man tar i bruk utenlandske modeller, er det viktig å undersøke hvilke verdier og prinsipper som allerede er innebygd i modellen og vurdere hvordan disse samspiller med eller eventuelt avviker fra norske verdier og helsetjenestens behov.

For å overholde lovverket kan tilpasningsprosessen inkludere integrerte sikkerhets- og personvernstrategier som eksplisitt inkluderer tiltak for å beskytte sensitive helsedata.

Kunnskapsforankring (*grounding*) er en teknikk for å forbedre kvaliteten og relevansen til svarene til en språkmodell uten å endre selve språkmodellen. Kvalitetssikrede og kuraterte datakilder som eksempelvis virksomhetsinterne eller eksterne kunnskapsbaser kan utgjøre et kunnskapsgrunnlag.

4.2.3 Kunnskapsforankring og agentisk KI

Såkalt RAG *(Retrieval Augmentet Generation)* [119] er én type teknikk for kunnskapsforankring som har vist seg å være lovende som et effektivt alternativ til helsefaglig finjustering. En forutsetning er at kunnskapsbasen er relevant for norske forhold. Samtidig må språkmodellen som benyttes, være i stand til å håndtere språk i Norge.

Det er fremdeles for tidlig å slå fast om kombinasjonen av RAG og store språkmodeller fungerer tilstrekkelig bra for de aktuelle bruksområdene i den norske helse- og omsorgstjenesten. Det er behov for mer erfaring og kunnskap på området for at RAG kan anbefales brukt i stor skala i helse- og omsorgstjenesten.

Når en slik RAG-løsning benyttes, hentes kunnskap fra forhåndsdefinerte kunnskapsbaser. Språkmodellens funksjon er først og fremst å tolke og formulere spørsmålet/instruksen for å finne riktig informasjon i kunnskapsbasen og deretter generere svaret etter at informasjon er hentet ut. Tilnærmingen gjør det mulig med direkte tilpassing til bruksformålet, krever lite regne- og dataressurser, og kan redusere feil og oppdiktede svar. Helsedirektoratet tester ut RAG for informasjonstjenesten Helsesvar (se boks under).

HelseSvar

Helsedirektoratet har testet RAG-teknologien for å generere forslag til svar på helsespørsmål fra unge knyttet til ulike områder, som for eksempel prevensjon, tobakk og psykisk helse. Målet er å effektivisere arbeidet til helsepersonell når de svarer ungdom som tar kontakt gjennom ung.no.

Flere ulike språkmodeller har blitt testet, sammen med en RAG-løsning som benytter en kunnskapsbank med tidligere spørsmål og svar, eller artikler publisert på nettet.

I sluttrapporten ble det konkludert med at "RAG og virksomhetsdata gir faglig sterke resultater. Men det er nødvendig å forbedre den språklige tilpasningen, spesielt for nynorsk og enklere språklige uttrykk, for å gjøre KI-assistentene mer tilgjengelige og brukervennlige for ungdom." Dette stiller krav til at man bearbeider tekster og formuleringer tilpasset målgruppen, i virksomhetsdataene som legges til grunn i RAG-løsningen.

Videre konkluderes det med følgende: "På kort sikt anbefaler vi at KI-løsningen HelseSvar (som bruker RAG), primært benyttes som støtteverktøy for svarere innen helse, fremfor som selvhjelpsverktøy til innbyggere. RAG-løsningen leverer korrekte svar i 80–90 % av tilfellene, noe som illustrerer en høy grad av presisjon og pålitelighet. Selv om løsningen svarer svært godt på helserelaterte spørsmål, forekommer det at den svarer feil grunnet misforståelser i konteksten på spørsmålet, eller uklare virksomhetsdata. Den menneskelige kvalitetssikringen er derfor nødvendig for spørsmål om helse."

Etter at rapporten er publisert jobber prosjektet videre med teknikker som hever svarkvaliteten. Det gjelder for eksempel teknikker for hvordan svarene bygges opp, spesielt gjennom bruk av agenter som tolker spørsmålet, utarbeider en plan for innhenting av relevant informasjon og komponerer svaret på en strukturert og kontrollert måte.

Kilde: "Rapport – HelseSvar. Konseptutredning for en KI-assistent for innbyggerrettet informasjon". Internrapport i Helsedirektoratet.

Agentisk KI innebærer å dele opp komplekse spørsmål i mindre oppgaver og kombinere flere verktøy som kan svare på oppgaver. Agenten tolker brukerens spørsmål og utarbeider en strukturert arbeidsflyt med spesifikke trinn for å svare ut spørsmålet (se KI-faktaark Intelligensforsterkning og kontroll) [120].

4.2.4 Evaluering og testing av språkmodeller for norske forhold

Vi mangler nasjonale prinsipper for testing, evaluering og kvalitetssikring av språkmodeller for norske forhold, noe som er en utfordring som gjelder hele den offentlige forvaltningen og ikke bare helse- og omsorgssektoren.

Evaluering og testing kan gjøres av både grunnmodeller og finjusterte modeller blant annet for å kunne sammenligne og eventuelt velge den språkmodellen som egner seg best å trene videre og tilpasse til bruksformålet. Evaluering og testing av et KI-system gjøres for å kunne vurdere hvor godt det yter i forhold til den tiltenkte bruken.

For å kunne vurdere en språkmodells ytelse kreves en strukturert tilnærming til evaluering av de ulike aspektene som beskrives i avsnitt 4.1. Et sentralt aspekt er språk, der modellen kan testes i hvordan den behersker norsk medisinsk fagterminologi, eller hvordan den behersker bokmål, nynorsk og i noen sammenhenger også samisk i helsefaglig kontekst. Presisjonsgraden i helsefaglig kommunikasjon kan også vurderes.

Den kontekstuelle forståelsen er like viktig, der tester kan vise hvordan modellen yter i samsvar med norsk helsetjenestes organisering og struktur. Dette kan gjelde norske behandlingsretningslinjer og prosedyrer, samt lokale administrative rutiner og dokumentasjonskrav.

For å sikre grundig evaluering er det mulig med en stegvis tilnærming til testing. Denne kan starte med grunnleggende evaluering av språk og sikkerhet, gå videre til domenespesifikk testing for helsesektoren, fortsette med bruksområdespesifikk testing, og kontekstspesifikk testing i implementeringsmiljøet, da gjerne som en del av et KI-system. Dersom språkmodellen inngår i et KI-system som faller innunder loven om medisinsk utstyr vil det stilles egne krav til testing og validering i henhold til denne loven [121].

[113] Store språkmodellar i helse- og omsorgstenesta

[114] Teknologirådet. 2024. *Generativ kunstig intelligens i Norge: https://teknologiradet.no/publication/generativ-kunstig-intelligens-i-norge/* og https://www.nora.ai/norsk/news/2024/vi-trenger-apne-norske-sprakmodeller.html

[115] https://pubmed.ncbi.nlm.nih.gov/39779927/

[116] Full parameter fine-tuning https://arxiv.org/abs/2306.09782

[117] Parameter efficient fine tune-tuning (PEFT) https://arxiv.org/abs/2410.21228

[118] https://medium.com/@bhat_aparna1/federated-learning-with-large-language-models-balancing-ai-innovatio

[119] RAG (Retrieval-Augmented Generation) er en KI-teknikk som kombinerer informasjonsgjenfinning (retrieval) med tekstgenerering (generation) for å forbedre svarenes kvalitet og relevans.

[120] KI-faktaark publiseres her: content://d16e0ae7-17a1-48f0-9809-7a8bd25eb943

[121] https://lovdata.no/dokument/NL/lov/1995-01-12-6

Hva trengs for å tilpasse til norske forhold?

Gjennom arbeidet med rapporten har vi fått innspill på hva som bør være på plass for å tilpasse store språkmodeller og bruken av dem til norske forhold i helse- og omsorgstjenesten. Innspillene omfatter

tilgang til data, regnekraft, kvalitetsrammeverk, kompetanse og et forvaltningsregime (se figur 6). Dette kapitlet går gjennom disse områdene, som også danner utgangspunkt for anbefalinger til tiltak i Felles KI-plan (se kapittel 5).

Figur 6: Sentrale områder som bør være på plass for å tilpasse store språkmodeller til norske forhold inkluderer data, regnekraft, kvalitetsrammeverk, kompetanse og et forvaltningsregime.

4.3.1 Tilstrekkelig datagrunnlag og datakvalitet

For å tilpasse språkmodeller til norsk helse- og omsorgssektor er det behov for betydelige mengder data av høy kvalitet til både trening og testing.

Digitaliseringsstrategien trekker frem viktigheten av bedre og enklere tilgang til data for helse- og omsorgssektoren: "Helse- og omsorgstjenesten har mye informasjon som kan være nyttig å bruke til å utvikle KI, slik som registerdata, medisinske bilder og pasientjournalnotater. Det må bli enklere for relevante aktører å få tilgang til helsedata for å bruke disse med KI. Bedre og enklere tilgang til helsedata er viktig både for videreutvikling av vår felles helsetjeneste, for forskning og for næringsutvikling, men forutsetter at hensynet til nasjonale sikkerhetsinteresser ivaretas" [122].

En nøkkelfaktor for bruk av data til tilpassing til norske forhold for er datakvaliteten. Vi har beskrevet risikoer knyttet til dårlig datakvalitet i store språkmodeller i avsnitt 3.1.1. Datakvalitet refererer til hvor pålitelig, nøyaktig og anvendelig data er for et gitt formål. Høy datakvalitet er avgjørende for gode analyser, beslutninger og modeller.

Nøkkelfaktorer for god datakvalitet er:

- nøyaktighet: dataene gjenspeiler virkeligheten korrekt.
- representativitet: dataene representerer hele befolkningen i Norge, også minoriteter og den samiske urbefolkning.
- fullstendighet: ingen viktige data mangler.
- konsistens: dataene er sammenhengende på tvers av kilder og systemer.
- pålitelighet: dataene kommer fra en troverdig kilde og er verifiserbare.
- aktualitet: dataene er oppdaterte og relevante.
- unikhet: ingen dupliserte eller motstridende oppføringer.
- relevans: dataene er nyttige for formålet de brukes til.
- balanse: de ulike helsefaglige områdene må være dekket

Få datakilder vil være perfekte, men ulike rammeverk for å beskrive datasett kan brukes for å øke transparensen rundt dataene som benyttes til utvikling av språkmodeller til bruk i helse- og omsorgsektoren [123]. Både *Model Documentation Form* [124] for store språkmodeller knyttet til KI-forordningen og model cards [125] fra Huggingface har dedikerte felter for beskrivelse av data.

Innsamling og bearbeiding av data er et betydelig arbeid. Det gjelder også arbeid med å avklare om det finnes juridiske, tekniske eller andre begrensninger.

Man skiller blant annet mellom merkede data (*labelled data*) og umerkede data (*unlabelled data*). Merkede data har ofte høyere kvalitet siden de eksempelvis kan inneholde semantisk eller kontekstuell informasjon som gir en merverdi.

Spørsmålet om tilgjengeligheten *av* nødvendige data har vært vesentlig for flere av informantene i denne rapporten.

4.3.1.1 Kategorier av data

Datagrunnlag for språkmodeller kan overordnet deles i tre hovedgrupper:

A. Fritt tilgjengelige data
B. Data med begrenset tilgang pga. rettighetshensyn
C. Data med begrenset tilgang pga. taushetsplikt og personvernhensyn

A Fritt tilgjengelige data

Regjeringens KI-strategi har som "mål at data som kan gjøres åpent tilgjengelig skal deles slik at de kan brukes av andre" [126]

Det finnes en rekke data som fritt kan benyttes til trening av språkmodeller, for eksempel nettsider fra offentlige helsemyndigheter:

- Helsefaglig kvalitetssikret kunnskap og informasjon, for eksempel fra FHI, spesialisthelsetjenesten, Helsenorge.no, Helsebiblioteket.
- Nasjonale krav og anbefalinger fra Helsedirektoratet
- Metodebøker fra spesialisthelsetjenesten
- Frie fagspråklige ressurser, for eksempel helsefaglige terminologier og klassifikasjoner uten krav til lisens, for eksempel fremtidig ICD-11 og termene i SNOMED CT.

B Data med avgrenset tilgang pga. rettighetshensyn

Noen data kan det være begrenset tilgang til pga. rettighetshensyn, for eksempel

- Fag- og lærebøker (se faktaboks om Mimir-prosjektet)
- Kunnskapsbaser og oppslagsverk, for eksempel Felleskatalogen og Store medisinske leksikon
- Fagspråklige ressurser, for eksempel helsefaglige terminologier og fagordbøker med krav om lisens

Mímir-prosjektet og trening på opphavsbeskyttet materiale

På oppdrag fra Regjeringen samarbeidet Nasjonalbiblioteket med NorwAI ved NTNU og Language Technology Group ved UiO i 2024 om å analysere hvilken betydning opphavsrettsbeskyttet materiale har på den språklige kvaliteten i språkmodeller.

Språket i språkmodeller med og uten opphavsrettsbeskyttet materiale som norske aviser og bøker ble sammenlignet. Resultatene viste at språkmodeller trent på innhold der rettighetsbelagt norsk materiale inngår, oppnår bedre kvalitet.

Målet med prosjektet var å samle inn empiri som kan legge grunnlaget for eventuelle avtaler mellom staten og rettighetshavere om bruk av innhold under opphavsrett for KI-formål. Det arbeides nå med å etablere prinsipper knyttet til slike avtaler (pr, 1, mars 2025).

Kilde: https://www.nb.no/content/uploads/2024/08/Mimirprosjektet_teknisk-rapport.pdf

Trening på fagspråkressurser

Det finnes flere fagspråklige ressurser som kan benyttes til å trene språkmodeller, blant annet terminologier og klassifikasjoner.

SNOMED CT er en internasjonal maskinlesbar terminologi med nærmere 370 000 helsefaglige begreper. Ca. 130 000 av begrepene er oversatt til norsk og dekker helsefaglige områder som anatomi, funn, symptomer, diagnoser, prosedyrer, substanser og legemiddel. SNOMED CT blir hovedsakelig brukt til dokumentasjon og samhandling av pasientopplysninger.

Oversettelsen er gjort til bokmål, mens en økende del også finnes på nynorsk. Ressursen er flerspråklig ved at det er en direkte kobling mellom engelske og norske termer.

Oversettelsene til norsk er fritt tilgjengelige fra Språkbanken hos Nasjonalbiblioteket. Internasjonalt har termene i SNOMED CT blitt brukt til trening av flere språkmodeller (pubmed.ncbi.nlm.nih.gov).

Helsedirektoratet forvalter den norske versjonen av SNOMED CT med jevnlige oppgraderinger.

SNOMED CT inneholder også en dypere maskinlesbar struktur (ontologi) med blant annet begrepsrelasjoner. SNOMED CTs ontologi har blitt brukt til RAG i språkmodeller, også i Norge, men det kreves lisens fra Helsedirektoratet.

En annen aktuell ressurs er Den internasjonale statistiske klassifikasjonen av sykdommer og beslektede helseproblemer (ICD), som eies og forvaltes av WHO. Helsetjenesten i Norge bruker i dag ICD-10 som kodeverk for sykdommer og dødsårsaker, og inneholder derfor termer og begreper som er i etablert bruk internasjonalt.

Helsedirektoratet forvalter den norske versjonen av ICD. WHO har nå oppdatert ICD-10 til ICD-11. ICD-11 har et oppdatert medisinsk innhold som er langt mer omfattende, og som også omfatter en terminologi. Ressursen blir nå oversatt til norsk, og vil bli fritt tilgjengelig for bruk.

Tilsvarende er ICPC-2 den internasjonale klassifikasjonen for helseproblemer, diagnoser og andre årsaker til kontakt med primærhelsetjenesten. ICPC-2 er i bruk i Norge og Helsedirektoratet vedlikeholder både denne, og nettsiden der den oppdaterte internasjonale utgaven av ICPC-2 (English version, ICPC-2e-English) publiseres på vegne av Wonca International Classification Committee (WICC).

Helsedirektoratet forvalter også en rekke andre relevante klassifikasjoner, for eksempel Norsk prosedyrekodeverk, Norsk laboratoriekodeverk med tilhørende kodeverk (Prøvemateriale, Anatomisk lokalisasjon, Tekstlige resultatverdier og Undersøkelsesmetode), Norsk patologikodeverk (NORPAT), og Aktivitetskodene for patologilaboratoriene (APAT). Utarbeidelsen av Norsk prosedyrekodeverk er initiert fra Nordisk råd og har fortsatt en nordisk- baltisk kjerne (NCSP). Dette inneholder eksempelvis termer for prosedyrer og prosedyregrupper i bruk i spesialisthelsetjenesten i de nordisk-baltiske landene.

C Data med avgrenset tilgang pga. taushetsplikt og personvernhensyn

Flere datakilder har begrenset tilgang pga. taushetsplikt og personvernhensyn. Dette gjelder for eksempel:

- journaltekster
- data fra kvalitetsregistre og andre helseregistre
- helseundersøkelser som for eksempel HUNT
- søknader og svar knyttet til norske helseadministrative virksomheter

Datakilder som for eksempel tekster fra pasientjournaler kan være særskilt relevante for å trene store språkmodeller fordi de speiler faktisk språkbruk og helsepersonells bruk av kunnskap. Imidlertid kan det være tidkrevende å få tilgang til slike data [127]. Det skyldes blant annet at helsepersonell er underlagt taushetsplikten, jf. helsepersonellovens kapittel 5. Det begrenser fri behandling av helseopplysninger fra pasientjournaler og andre helseregistre til trening og bruk av språkmodeller. Paragraf 29 i helsepersonelloven åpner likevel for at det kan søkes om dispensasjon fra taushetsplikten for bruk av helseopplysninger fra behandlingsrettede helseregistre (pasientjournaler) når visse vilkår er oppfylt [128]. Eksempel på at det er gitt en slik dispensasjon, er journaltekster til trening av språkmodellen Klinisk NorBERT hos Helse vest IKT (se faktaboksen Trening på journaltekster) og NorDeClin-BERT hos Nasjonalt senter for e-helseforskning.

Trening på journaltekster, Klinisk NorBERT

Helse Vest IKT har i samarbeid med helseforetakene på Vestlandet trent en språkmodell ved hjelp av blant annet journaltekster. For å ivareta personvernet har tekstene blitt anonymisert ved at stedsnavn blir erstattet med et annet stedsnavn, for- og etternavn med et annet for- og etternavn, datoer med annen dato osv. Som en del av forskningsprosjektet har man analysert kvaliteten til anonymiseringen.

For å få tilgang til journaltekstene til trening av modellen, har det blitt søkt om og blitt innvilget dispensasjon fra taushetsplikten for forskningsformål i helsepersonelloven hos REK.

Man ser for seg at Klinisk NorBERT kan benyttes til for eksempel automatiserte tekstanalyser og maskinstøttet koding. Språkmodellen er en såkalt BERT-modell som skiller seg fra generative språkmodeller. Derfor er det liten fare for at rådataene som språkmodellen opprinnelig ble trent på, kan bli gjenskapt.

Klinisk NorBERT kan bli gjort tilgjengelig til bruk hos andre aktører i helse- og omsorgstjenesten under visse vilkår. Skal språkmodellen finjusteres for et konkret bruksområde, kreves det ny godkjennelse fra REK eller Helsedirektoratet for å bruke helsedata til dette formålet.

Helse Vest IKT ser nå på mulighetene for å trene generative språkmodeller.

Kilde: Helse vest IKT

Syntetiske data

Avgrenset tilgang til personsensitive data har ført til en diskusjon om behovet for å lage syntetiske tekster [129]. Syntetiske data som ikke kan tilbakeføres til personer, er ikke personopplysninger og omfattes dermed ikke av taushetsplikten. Imidlertid finnes det også en rekke utfordringer knyttet til syntetiske data.

Forskning på såkalte ASR-systemer (Automatic Speech Recognition) og nedstrøms KI-modeller viser at syntetiske data kan ha vesentlige begrensninger. Forskning viser at syntetiske transkripsjoner kan inneholde feil, hallusinasjoner og unaturlige språkstrukturer som påvirker ytelsen til modeller som bruker disse dataene [130]. For eksempel viser en studie at ved bruk av simulerte ASR-utdata for å trene modeller (ved å benytte tekst-til-tale og deretter tale-til-tekst), kunne modellene bli mer robuste mot feil, men kvaliteten varierte fortsatt sammenlignet med autentiske data [131].

En særlig bekymring ved bruk av syntetiske helsedata er at hallusinasjoner i språkmodeller kan føre til at falsk informasjon blir innlemmet i datasettene [132]. Studien viser at selv avanserte systemer som Whisper kan "finne på eller 'hallusinere' hele fraser og setninger" i omtrent 1% av transkripsjonene [133]. I helsekontekst kan slike feilaktige innslag føre til at modeller trenes på medisinsk informasjon som aldri ble uttalt, noe som kan ha alvorlige konsekvenser for diagnostisering og behandlingsanbefalinger.

I tillegg er man, som nevnt, avhengig av autentiske data for å lage syntetiske data, så man unngår ikke nødvendigvis problemstillingen knyttet til personvern. Kvaliteten på syntetiske data avhenger sterkt av kvaliteten og representativiteten til kildedataene, og studien indikerer at visse feilmønstre i originaldataene kan forsterkes eller skape nye problemer i syntetiske datasett.

Det er derfor fremdeles behov for mer kunnskap for å kunne si om syntetiske datasett kan være tilstrekkelig egnet til å trene språkmodeller, særlig når det gjelder domener med høy risiko som helsesektoren. Forskere anbefaler en kombinasjon av forbedret ASR-teknologi, feilkorrigerende metoder, robust trening og tverrmodal validering for å redusere problemene med syntetiske data, men disse utfordringene er fortsatt ikke fullstendig løst [134].

4.3.2 Infrastruktur for regnekraft

Tilpassing av språkmodeller til norske forhold vil kreve omfattende regnekraft og tilhørende infrastruktur. Behovet vil være avhengig av mange faktorer, blant annet hvordan tilpassingen foregår.

Forskningsrådet har gjort en konseptvalgutredning om behovet for regnekraft og organisering av en nasjonal infrastruktur [135]. Rapporten peker på et investeringsbehov på 3,4 milliarder kroner de neste fem årene. Utredningen skal gjøre en kost-nytte-vurdering for hvert forvaltningsområde i steg 2. Imidlertid omfatter ikke dette trinnet helse- og omsorgssektoren og peker på at det kreves en særskilt utredning siden sektoren er omfattende og kompleks med særskilte krav til personvern.

Den teknologiske utviklingen i det siste antyder at det nå utvikles langt mer effektive måter å utvikle språkmodeller enn tidligere, noe som kan redusere tidligere antatt behov for regnekraft.

Imidlertid vil den store mengden data med sensitive data fortsatt gi utfordringer knyttet til infrastruktur for regnekraft. Det kan derfor være behov for en egnet infrastruktur som er i stand til å håndtere slike data til trening og finjustering av språkmodeller.

4.3.3 Kvalitetsrammeverk for store språkmodeller

Evaluering av språkmodeller er komplekst og vil omfatte flere typer tester (*benchmarks*). Det er gjort relevant forskning på tester for tilpassete språkmodeller til generelle norske forhold, for eksempel språk, ved Universitetet i Oslo [136].

Evalueringen bør omfatte både generell og bruksorientert testing. Den generelle kvalitetsmålingen kan baseres på standardiserte tester som evaluerer grunnleggende medisinsk kunnskap, for eksempel gjennom tilpassede versjoner av medisinske eksamensspørsmål. Den bruksorienterte og kontekstspesifikke kvalitetsmålingen kan være testing i reelle eller simulerte brukssituasjoner som er representative for modellens tiltenkte bruk i norsk helse- og omsorgstjeneste.

Testing ved hjelp av eksamensspørsmål

Internasjonal forskning benytter ofte medisinske eksamensspørsmål for å teste kvaliteten til språkmodeller. Et vanlig datasett er MedQA, som inkluderer over 60 000 spørsmål fra flere medisineksamener i USA og Kina.

Slike datasett kan være formålstjenlig for å teste generell medisinkompetanse, for eksempel anatomi og diagnoser. Imidlertid vil den kliniske hverdagen der KI-verktøy skal fungere, skille seg vesentlig fra en eksamenssituasjon. Testing ved hjelp av eksamensspørsmål vil derfor ikke nødvendigvis være tilstrekkelig for å måle kvaliteten til språkmodeller.

Et annet spørsmål er hvorvidt eksamensspørsmål fra USA og Kina vil fungere godt i norsk kontekst, og det er mulig at det bør lages egne datasett med eksamensspørsmål basert på norsk pensum.

Et tilleggsmoment er at flere internasjonale språkmodeller allerede kan være trent på datasett som MedQA. Datasettet vil da være forurenset og være uegnet for å teste språkmodeller.

For øyeblikket finnes det ikke et generelt akseptert kvalitetsrammeverk for evaluering av språkmodeller for helsesektoren, men flere peker på at det er viktig for å kunne bruke språkmodeller på en ansvarlig måte innen helse [137]. Et slikt rammeverk vil måtte utvikles stegvis og basere seg på velprøvde metoder, beste praksiser og standarder. En koalisjon som har som mål å fremme ansvarlig utvikling av KI for helse (Coalition for Health AI (CHAI™)) skisserer for eksempel et rammeverk og påpeker viktigheten av å utvikle gode tester for å kunne evaluere store språkmodeller med tanke på fem grunnleggende prinsipper: nytteverdi, rettferdighet og likebehandling, åpenhet, sikkerhet, og personvern og datasikkerhet [138].

Gjennom arbeidet med denne rapporten har det blant annet blitt foreslått å etablere en lagvis modell for å evaluere og teste ulike egenskaper som er viktige for den norske helse- og omsorgs-sektoren. Nedenfor beskrives overordnet mulige områder for testing og hva en lagvis modell for evaluering kan innebære.

4.3.1.1 Mulige områder for testing

Kvalitetsrammeverket kan omfatte tester av modeller på flere områder.

Helsefaglig språk:

- medisinsk terminologi, inkludert faguttrykk, forkortelser og sjargong
- bokmål, nynorsk og eventuelt samisk i helsefaglig kontekst
- språklig presisjon, språk tilpasset ulike målgrupper som helsepersonell og pasienter, inkludert personer med ulik kulturell bakgrunn

Helsefaglig kunnskap og praksis:

- nasjonale faglige retningslinjer
- etablerte medisinske prosedyrer og behandlingsprotokoller
- akutte versus ikke-akutte situasjoner
- identifisering og forklaring av medisinske sammenhenger

Helseadministrativ kunnskap og praksis:

- norsk helsesektors struktur og organisering
- administrative rutiner og prosedyrer
- henvisnings- og dokumentasjonspraksis
- samhandling mellom ulike nivåer i helsetjenesten

Verdier og etikk:

- respekt forpasientrettigheter
- personvern og taushetsplikt
- etisk forsvarlige anbefalinger
- komplekse etiske dilemmaer

Lovverk:

- lovgivning relevant for helse- og omsorgssektoren
- juridiske krav til pasientbehandling

- helsepersonells plikter og ansvar
- dokumentasjonskrav og meldeplikt

For alle disse områdene er det essensielt å etablere både kvantitative og kvalitative metoder.

Evalueringen kan inkludere ulike dimensjoner: [139]

- systematisk evaluering av modellens nøyaktighet og pålitelighet
- vurdering av modellens evne til å erkjenne egen usikkerhet
- testing av modellens robusthet under ulike forhold
- kontinuerlig evaluering av modellens ytelse over tid for å oppdage eventuell ytelsesforringelse

4.3.1.2 Lagvis modell for testing av språkmodeller

For å sikre en grundig og systematisk evaluering, har det gjennom arbeidet med denne rapporten blitt foreslått en lagvis tilnærming som omfatter evaluering av grunnleggende, domenespesifikke, bruksområdespesifikke og kontekstspesifikke egenskaper (se figur 7).

Det kan være hensiktsmessig å begynne utviklingen av rammeverket for de nederste nivåene, samt bruksområder med lav risiko.

Der bruksområdet faller inn under regelverket for medisinsk utstyr og/eller KI-forordningen vil standarder, både eksisterende og under arbeid, utgjøre egne rammeverk som kan brukes for å ivareta krav i lovverket.

Figur 7: Lagvis tilnærming til evaluering og testing av språkmodeller, fra grunnleggende, domenespesifikk, bruksområdespesifikk til kontekstspesifikk

Beskrivelsen under er en skisse som illustrerer hva som typisk kan bli testet på hvert nivå, og kan danne et utgangspunkt for videre konkretisering.

1. *Grunnleggende evaluering* (nivå 1) omfatter grunnleggende egenskaper som er kritiske for anvendelser i helsesektoren. Det inkluderer testing av språklig kvalitet på både bokmål og nynorsk og samisk der det er relevant, og hvordan modellen håndterer generelt medisinsk språk og terminologi, grunnleggende sikkerhet og personvern, samt teknisk ytelse som responstid og stabilitet. På dette nivået kan det vurderes å etablere minimumskrav som alle modeller må oppfylle for å kunne brukes i helsesektoren.
2. *Domenespesifikk evaluering* (nivå 2) omfatter domenespesifikke egenskaper som er viktige for helsesektoren. Det kan inkludere testing av modellen med tanke på spesialisert medisinsk fagspråk, kliniske retningslinjer, helseadministrative prosesser og etiske rammer. På dette nivået vurderes også modellens evne til å håndtere regionale og lokale forhold i norsk helsesektor.
3. *Bruksområdespesifikk evaluering* (nivå 3) omfatter spesifikke bruksområder innen helsesektoren. For eksempel vil en modell som skal brukes i akuttmottak, testes spesifikt for evnen til å assistere med triage-vurderinger, akuttmedisinske prosedyrer og koordinering med andre avdelinger. En modell for å lage tale-til-sammendrag av pasientsamtaler, bør testes for akkurat dette bruksformålet. Andre bruksområder vil ha andre spesifikke krav som evalueres. Testing på dette nivået bidrar til å vurdere hvordan modellen er egnet for sitt tiltenkte formål.
4. *Kontekstspesifikk evaluering* (nivå 4) handler om modellens egnethet i den spesifikke implementeringskonteksten. Det omfatter testing av integrasjon med lokale systemer, tilpasning til etablerte arbeidsprosesser, og håndtering av særskilte dokumentasjons- og personvernkrav. Testing på dette nivået vurderer også modellens evne til å møte lokale kvalitetsindikatorer og spesifikke behov i implementeringsmiljøet.

Fordelene med et felles akseptert rammeverk for testing er at det:

- gir grunnlag for sammenligning mellom ulike modeller
- muliggjør systematisk evaluering fra det generelle til det spesifikke
- forenkler identifisering av svakheter og forbedringsområder
- legger til rette for målrettet optimalisering
- sikrer at kritiske aspekter blir evaluert
- effektiviserer testprosessen ved å avdekke fundamentale mangler tidlig

Testingen bør ikke bare gjøres én gang, men kontinuerlig over tid, for å fange opp endringer i ytelse, identifisere nye utfordringer og sikre vedvarende kvalitet i tjenesten.

Rammeverket bør jevnlig revideres og oppdateres i takt med teknologisk utvikling og nye erfaringer fra praktisk bruk.

4.3.4 Kompetanse om utvikling og bruk

Det er avgjørende at helse- og omsorgssektoren har tilstrekkelig tilgang til nødvendig kompetanse i hele livsløpet til språkmodeller, fra utvikling, tilpassing, testing og bruk av språkmodeller som er tilpasset norske forhold [140]. Flere av informantene har gitt tilbakemelding på at mer kompetanse i helse- og omsorgssektoren er nødvendig, blant annet innen kunstig intelligens (KI), informasjons- og kommunikasjonsteknologi (IKT), helsefag, juss, lingvistikk og økonomi.

For bruk av KI-verktøy som ledd i helsehjelp, gjelder som ellers kravene som følger av helselovgivningen. Helse- og omsorgstjenesten har plikt til å sørge for at helsehjelpen som gis er forsvarlig. Blant annet må den sørge for at de KI-verktøyene som brukes bidrar til at helsehjelpen som gis, er trygg og sikker. Ifølge KI-forordningen hviler det et ansvar på virksomheter som innfører KI-løsninger (*deployer*) i å lære opp brukerne. Virksomheten har også et ansvar for å delegere oppgaver knyttet til å sørge for menneskelig overblikk til personer som har den nødvendige kompetansen og gjennomgått den nødvendige opplæringen for å utføre denne funksjonen [141].

KI-kompetanse (*AI literacy*) er et sentralt begrep KI-forordningen. Det finnes også forsøk på å operasjonalisere og konkretisere begrepet på nivåene 'kunnskap', 'forståelse' og 'ferdigheter' [142].

Det forskes også på KI-kompetanse. En internasjonal metastudie peker på at helsepersonell og studenter har lav KI-kompetanse [143]. Vi kjenner ikke til slike kartlegginger for norske forhold, men det kan være rimelig å anta at det ikke finnes tilstrekkelig KI-kompetanse for trygg og effektiv bruk, utvikling og testing av språkmodeller i norsk helse- og omsorgstjeneste.

4.3.5 Forvaltning av språkmodeller

I dag finnes det ikke et nasjonalt forvaltingsregime, inkludert infrastruktur, for store språkmodeller. Teknologien er fremdeles i rask bevegelse, og nye internasjonale språkmodeller blir stadig introdusert. Språkmodeller blir også utviklet og tilpasset til norske forhold av aktører både i offentlig sektor og i næringslivet. Det finnes imidlertid ingen samlet nasjonal oversikt over hvilke modeller som finnes og hvilke som benyttes i helse- og omsorgssektoren.

Et stabilt forvaltningsregime kan legge til rette for trygg innføring av språkmodeller i helse- og omsorgstjenesten ved å sikre blant annet kvalitet, aktualitet og juridiske rammer.

Teknologirådet anbefaler i sin rapport å definere utvalgte norske språkmodeller som en nasjonal fellestjeneste, på linje med ID-porten, Altinn og Folkeregisteret, for å sikre god drift, forvaltning og tilgang. Både norske forhåndstrente språkmodeller og tilpassede språkmodeller kan inngå. Teknologirådet peker videre på at det bør etableres en ny funksjon for utvikling og drift av en slik fellestjeneste [144]. Dette vil også kunne gjelde for helse- og omsorgssektoren, slik som andre fellestjenester innenfor sektoren. En nasjonal forvaltning vil kunne adressere flere av risikoene beskrevet ovenfor gjennom for eksempel testing og evaluering av språkmodeller for sektoren.

En felles forvaltning vil kunne omfatte flere funksjoner, blant annet:

- forvaltning og videreutvikling av felles datagrunnlag for tilpassing til norske forhold
- forvaltning og videreutvikling av kvalitetsrammeverk
- forvaltning av felles språkmodeller
- veiledningstjeneste for helsesektoren
- samarbeide med forskningsinstitusjoner som deltaker i relevante forskningsprosjekter
- gjøre språkmodeller tilgjengelige for bruk i sektor eller videreutvikling av for eksempel leverandører eller offentlig sektor
- koordinere innsats i sektoren knyttet til språkmodeller

Felles forvaltning vil ikke erstatte næringslivets rolle som leverandør, men heller legge til rette for trygge rammer for innovasjon og bruk, for private og offentlige organisasjoner. Videreutvikling gjort av kommersielle aktører kan bidra til et bredere tilbud av løsninger. Eksempelvis kan en forvaltningsorganisasjon teste, tilpasse og tilgjengeliggjøre en eller flere forhåndstrente helsefaglige språkmodeller. Slike modeller kan videreutvikles videre til spesifikke bruksområder av for eksempel private eller offentlige aktører.

En forvaltningsorganisasjon kan sikre langsiktighet, kvalitet og koordinert utvikling av språkmodeller for sektoren. Organisasjonen bør ha ansvar for å følge den teknologiske utviklingen, vurdere nye muligheter, men også risikoer og etablere tydelige rammer for hvordan modeller kan anvendes, med særlig vekt på pasientsikkerhet og personvern. Slik kan det gi råd om implementering og bruk av språkmodeller i helsesektoren.

Forvaltning kan foregå i en eksisterende organisasjon eller det kan opprettes en egen organisasjon, for eksempel et eget senter. Det vil være behov for nærmere utredning av behov for og hvordan en slik forvaltning bør organiseres.

[122] Fremtidens digitale Norge. Nasjonal digitaliseringsstrategi 2024-2030, s. 66.

[123] Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations - The Lancet Digital Health

[124] https://code-of-practice.ai/?section=transparency#model-documentation-form

[125] https://huggingface.co/docs/hub/en/model-cards

[126] https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/

[127] https://ehealthresearch.no/nyheter/2024/en-revolusjon-for-helsesektoren-den-forste-norske-kliniske-sprakm

[128] Taushetsplikt og opplysningsrett - Helsedirektoratet

[129] https://tidsskriftet.no/2024/11/kronikk/syntetiske-datasett-kan-gi-bedre-ki-modeller-helsetjenesten

[130] https://arxiv.org/html/2408.14418v1, https://dl.acm.org/doi/abs/10.1145/3630106.3658996

[131] https://arxiv.org/abs/2108.13048

[132] https://arxiv.org/html/2401.01572v1

[133] https://dl.acm.org/doi/abs/10.1145/3630106.3658996

[134] High-Precision Medical Speech Recognition Through Synthetic Data and Semantic Correction: United-MedASR

[135] https://www.forskningsradet.no/nyheter/2025/investere-milliarder-infrastruktur-tungregning/

[136] https://arxiv.org/abs/2412.06484

[137] https://jamanetwork.com/journals/jama/fullarticle/2825147

[138] https://chai.org/generative-ai-work-group/

[139] Disse dimensjonene er blant annet beskrevet i disse artiklene: Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review | Digital Health | JAMA | JAMA Network og HELM: [2211.09110] Holistic Evaluation of Language Models.

[140] https://healthinformaticsjournal.com/downloads/files/524.pdf

[141] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[142] https://www.pwc.nl/en/services/audit-assurance/pwc-accountancy-insights/data-it-and-internal-control/ai-lite

[143] https://healthinformaticsjournal.com/downloads/files/524.pdf

[144] Teknologirådet (2024) Generativ kunstig intelligens i Norge: https://teknologiradet.no/publication/generativ-kunstig-intelligens-i-norge/

Hvem kan tilpasse til norske forhold?

Flere aktører er aktuelle for å tilpasse språkmodeller til norske forhold.

4.4.1 Universitets- og høyskolesektoren

I universitets- og høyskolesektoren finnes det flere initiativ som både lager forhåndstrente norske språkmodeller og finjusterer internasjonale modeller, for eksempel Norsk forskningssenter for KI-innovasjon (NorwAI) ved NTNU. NORA-konsortiet inkluderer mange samarbeidspartnere, blant annet UiO og Norsk Regnesentral.

Nasjonalt senter for e-helseforskning utviklet en helsefaglig språkmodell (NorDeClin-BERT).

Fremtidige forskningsprosjekt knyttet til språkmodeller vil spille en viktig rolle når det gjelder tilpassing til norske forhold. Det er viktig å at noen av disse initiativene inkluderer tilpassing til norsk helse- og omsorgssektor, altså helsefaglige språkmodeller.

4.4.2 Helse- og omsorgssektoren

I helse- og omsorgssektoren har både Helse vest IKT og Sørlandet sykehus vært involvert i tilpassing av språkmodeller til norske forhold i forskningsøyemed. Helse Vest IKT ser nå også på spesifikke bruksområder for språkmodellen Klinisk NorBERT.

Næringslivet spiller en svært viktig rolle ved å finjustere språkmodeller og tilby tjenester som baserer seg på språkmodeller, blant annet norske forhåndstrente modeller, men særlig internasjonale modeller. Leverandører tilbyr nå tjenester basert på språkmodeller til blant annet fastleger.

4.4.3 Nasjonalbiblioteket

Nasjonalbiblioteket er en sentral forskningsinstitusjon som utvikler og tilpasser språkmodeller for norske forhold. Nasjonalbiblioteket har utviklet og tilgjengeliggjort språkmodellen NB-Whisper for transkripsjon av lydopptak, og modellen har blitt videreutviklet av leverandører for helse- og omsorgssektoren.

Nasjonalbiblioteket vil etter alt å dømme få en særskilt posisjon i Norge når det gjelder tilpassing av språkmodeller til norske forhold. Regjeringen foreslår på statsbudsjettet for 2025 å bevilge til sammen 40

millioner kroner til trening av norske og samiske språkmodeller (grunnmodeller) til bruk i kunstig intelligens. Av disse vil 20 millioner kroner gå til Nasjonalbiblioteket, som får i oppdrag å trene og tilgjengeliggjøre språkmodeller, mens 20 millioner kroner skal gå til å finansiere regnekraft til trening av modellene. Regnekraften skal leveres av tungregneselskapet Sigma2 AS, som eies av Sikt – Kunnskapssektorens tjenesteleverandør [145]. Nasjonalbibliotekets posisjon vil kunne bli svært viktig for arbeidet med språkmodeller i helse- og omsorgssektoren også, og det vil kunne være hensiktsmessig med et tett samarbeid.

[145] https://www.regjeringen.no/contentassets/e628c0d8c3a44971bb48bd5a487cb52d/nn-no/pdfs/prp20242025

Anbefalinger

For at bruk av KI-verktøy i helse- og omsorgstjenesten skal være forsvarlig og sikker er det avgjørende at kvaliteten på KI-modellene og -systemene er av høy kvalitet. Det å kunne tilpasse store språkmodeller til norske helse- og omsorgstjenesten vil kunne bidra til å redusere flere risikoer knyttet til å bruk KI-verktøy i helsehjelpen.

Regjeringens digitaliseringsstrategi har KI som et eget innsatsområde og vil blant annet: [146]

- jobbe for å videreutvikle en nasjonal infrastruktur for KI som skal gi tilgang til grunnmodeller tuftet på norske og samiske språk og samfunnsforhold
- utrede hvordan man kan bruke data og tekst fra offentlige virksomheter til etisk og trygg trening av nasjonale språkmodeller
- klargjøre hva som er lovlig bruk av åndsverk og andre vernede arbeider i tekst- og datautvinningsprosesser

I Statsbudsjettet for 2025 bevilges 20 millioner kroner til økt satsing på Nasjonalbibliotekets arbeid med kunstig intelligens og trening av generative språkmodeller [147]. Norges forskningsråd har bevilget over 1 milliard kroner til å etablere fire til seks KI-forskingssentre på tvers av sektorer for å svare på utfordringer som samfunnsutfordringer ved den digitale teknologiutviklingen [148].

Regjeringen kunngjorde 21. mars 2025 at de skulle opprette "KI-Norge", som skal være en ny nasjonal arena for innovativ og ansvarlig KI og en del av det nasjonale forvaltningsapparatet som nå etableres. Samtidig publiserte regjeringen at Nasjonal kommunikasjonsmyndighet (Nkom) får tilsynsansvar for EUs KI-forordning [149].

Anbefalingene i denne rapporten som knytter seg til tilpassing til den norske helse- og omsorgstjenesten vil i størst mulig grad søke synergier og samarbeid med relevant arbeid som foregår på nasjonalt nivå.

For å kunne tilrettelegge for at bruken av generative KI-modeller kan baserer seg på grunnmodeller som er tilpasset forholdene i den norske helse- og omsorgstjenesten, anbefaler Helsedirektoratet følgende tiltak:

1. etablere kvalitetsrammeverk for store språkmodeller som brukes i norsk helse- og omsorgstjeneste
2. bygge og dele kompetanse om utvikling og bruk av store språkmodeller
3. etablere felles datagrunnlag
4. utrede infrastruktur med regnekraft tilpasset helsesektorens behov
5. utrede forvaltningsstruktur for store språkmodeller for helsesektoren

De fleste tiltak vil måtte løses i samarbeid mellom en rekke aktører, både i helsesektoren og utenfor som Nasjonalbiblioteket, U&H-sektoren, Digitaliseringsdirektoratet og Forskningsrådet.

[146] https://www.regjeringen.no/no/dokumenter/fremtidens-digitale-norge/id3054645/?ch=5

[147] https://www.nb.no/pressemeldinger/regjeringen-vil-satse-stort-pa-digitalisering-og-kunstig-intelligens-ved-nasjon

[148] https://www.forskningsradet.no/forskningspolitikk-strategi/ltp/kunstig-intelligens/

[149] https://www.regjeringen.no/no/aktuelt/gjor-norge-klar-for-trygg-og-innovativ-ki-bruk/id3093081/

Etablere kvalitetsrammeverk for store språkmodeller for norsk helse- og omsorgstjeneste

Foreslått tiltakseier: Helsedirektoratet i første omgang

I samarbeid med: Helse- og omsorgstjenesten**,** Folkehelseinstituttet, Helsetilsynet, DSA, FoU-miljøer, Nasjonalbiblioteket, Digitaliseringsdirektoratet, Norges forskningsråd, Innovasjon Norge, mm.

Relevant for: Helse- og omsorgssektoren

Problem som skal løses: Det er knyttet stor usikkerhet til hva som er tilstrekkelig god kvalitet for å bruke store språkmodeller i helsesektoren, og hvordan dette måles. Evaluering av språkmodeller er komplekst og det finnes for øyeblikket få allment aksepterte metoder for evaluering og testing (*benchmarks*) av store språkmodeller for helse- og omsorgssektoren. Det gjør det også vanskelig å velge riktig språkmodell å videreutvikle til KI-systemer av god kvalitet.

Forslag: Sektoren utvikler og etablerer felles kvalitetsrammeverk for testing og evaluering av språkmodeller for å bidra til trygg bruk av generative KI-modeller i helse- og omsorgssektor.

Etableringen av kvalitetsrammeverket vil måtte skje i tett samarbeid med flere aktører i helse- og omsorgssektoren, offentlig og privat sektor, forskningssektoren og næringslivet.

Det kan være hensiktsmessig å starte med administrative bruksområder med lav risiko og med grunnleggende evaluering, som er det nederste nivået i Figur 7. Rammeverket kan stegvis utvides til bruksområder av høyere risiko etter hvert som området modnes og EU klargjør både regelverk og standarder knyttet til KI og medisinsk utstyr.

Arbeidet bør omfatte følgende:

1. Identifisere organisasjoner som bør være med i arbeidet og hvem som bør ha ansvar for hva
2. Kartlegge relevante eksisterende tester (benchmarks), rammeverk, beste praksiser og standarder for testing og evaluering av store språkmodeller
3. Systematisere erfaringer fra tilsvarende arbeid i sektoren, inkludert Helsedirektoratets arbeid med KI-tjenester som Helsesvar og Enklere Tilgang til Informasjon (ETI).
4. Stegvis oppbygging og pilotering av rammeverket basert på konkrete bruksområder.

Rammeverket kan omfatte både kvantitative og kvalitative evalueringsmetoder.

For øvrig vises det særlig til avsnitt 4.3.3 og kapittel 4.4

Bygge og dele kompetanse om utvikling og bruk av store språkmodeller

Foreslått tiltakseier: Flere tiltak, med ulike tiltakseiere

I samarbeid med: Helsedirektoratet, helsesektoren, andre relevante miljøer, som Nasjonalbiblioteket, Digitaliseringsdirektoratet, FoU-institusjoner og UoH-institusjoner.

Relevant for: Helse- og omsorgssektoren

Problem som skal løses: Det mangler nødvendig kompetanse for å tilpasse, evaluere og ta i bruk språkmodeller i den norske helse- og omsorgstjenesten.

Forslag: Helsesektoren jobber systematisk med å bygge og dele kompetanse, spesielt gjennom erfaringer fra praktisk utprøving og bruk.

Arbeidet bør omfatte:

1. Denne rapporten kan brukes for å heve kompetanse og øke bevissthet om risikoer og risikoreduserende tiltak, i helsetjenesten, internt i Helsedirektoratet, i helse-myndighetene og i hele sektoren. Spredning kan gjøres aktivt gjennom presentasjoner, webinarer mm.
2. Videreutvikle tverretatlig regulatorisk veiledningstjeneste med temaer knyttet til bruk av store språkmodeller i helse- og omsorgstjenesten
3. Følge opp seminar med fastlegene, knyttet til bruk av KI-verktøy i allmennlegetjenesten, i samarbeidet med Norsk forening for allmennlegetjenesten.
4. Arbeide for at pågående og framtidige forskning- og utviklingsprosjekter på KI også omfatter språkmodeller tilpasset norsk helse- og omsorgsektor.
5. Følge med på nordiske, og internasjonale tilnærminger til utvikling, testing og bruk av språkmodeller i helse- og omsorgssektoren.
6. Dele erfaringer og kompetanse gjennom for eksempel seminar, rapporter og veiledningsmateriell.
7. Pilotere tilpasninger av store språkmodeller til den norske helse- og omsorgstjenesten i samarbeid med en eller flere nøkkelaktører, forslagsvis Nasjonalbiblioteket.
8. Pilotere rammeverk for testing/kvalitetsmåling av språkmodeller i reelle brukssituasjoner i tett samarbeid med fagmiljøer som kan bidra til tilpasninger til norsk helsetjeneste.

For øvrig vises det særlig til avsnitt 4.3.4 og kapittel 4.4

Etablere felles datagrunnlag

Foreslått tiltakseier: Flere tiltak, med ulike tiltakseiere

I samarbeid med: Helsedirektoratet, Helsesektoren, de regionale helseforetakene, Kommunenes organisasjon (KS), FHI og Nasjonalbiblioteket, DSA

Relevant for: Helse- og omsorgssektoren

Problem som skal løses: Helse- og omsorgssektoren mangler et tilstrekkelig felles datagrunnlag for å utvikle, tilpasse og teste språkmodeller enklere og mer kostnadseffektivt, og dermed bidra til at det utvikles KI-verktøy av høy kvalitet som er tilpasset norske forhold.

Forslag: Helse- og omsorgssektoren bør samarbeide for å gjøre mer data av god kvalitet tilgjengelig til både trening (forhåndstrening og ettertrening), kunnskapsforankring (for eksempel RAG) og testing av språkmodeller som skal brukes i helse- og omsorgssektoren.

Arbeidet bør omfatte:

1. å kartlegge åpne og lett tilgjengelige datakilder, som for eksempel retningslinjer og metodebøker, for utvikling av store språkmodeller for helse- og omsorgstjenesten (Hdir, i samarbeid med FHI). Oversikt over fritt tilgjengelige tekster og andre språkressurser med helsefaglig kunnskap og praksis kan legges på informasjonssiden om KI hos Helsedirektoratet.
2. å gjøre fagspråklige datakilder som for eksempel terminologier og klassifikasjoner tilgjengelige til gjenbruk sammen med andre relevante data, for eksempel på helsedata.no, helsedirektoratet.no eller i språkbanken i Nasjonalbiblioteket (Hdir), også nynorsk og samisk.
3. å etablere datagrunnlag som utgjør felles nasjonale prinsipper for verdier og etikk i språkmodeller i norsk helse- og omsorgstjeneste.
4. vurdere hvordan sensitive data som for eksempel journalnotater, epikriser og skjemaer kan brukes til å utvikle store språkmodeller (klargjøre regelverk gjennom veiledning, tekniske løsninger mm.) (Hdir, i samarbeid med relevante (forsknings-) miljøer)
5. å gjøre rettighetsbelagte datakilder mer tilgjengelig, for eksempel ved å vurdere å etablere en kompensasjonsordning for rettighetseiere av helsefaglig innhold i tråd med nasjonale prinsipper (se Mimir-prosjektet til Nasjonalbiblioteket [150]) å utrede behov for og etablering av en felles infrastruktur for deling av språkdata, for eksempel gjennom helsedata.no (FHI, Hdir, helseforetakene). Se også anbefalt tiltak under om etablering av infrastruktur om regnekraft (kap. 5.2)

For øvrig vises det særlig til avsnitt 4.3.1 og kapittel 4.4.

[150] Første fase av Mimir-prosjektet er dokumentert her: https://www.nb.no/content/uploads/2024/08/Mimirprosjektet_teknisk-rapport.pdf

Utrede infrastruktur med regnekraft tilpasset helsesektorens behov

Foreslått tiltakseier: Flere tiltak, med ulike tiltakseiere

I samarbeid med: Helsedirektoratet, Norske helsenett (NHN), Forskningsrådet, Nasjonalbiblioteket, U&H-sektoren, de regionale helseforetak, regionale IKT-selskaper og Kommunenes organisasjon (KS), helsesektoren forøvrig

Relevant for: Helse- og omsorgssektoren

Problem som skal løses: Helsesektoren er en omfattende og kompleks sektor med særskilte krav til personvern. Eksisterende infrastruktur for regnekraft i sektoren er i liten grad egnet til å håndtere store mengder og sensitive data. Det er imidlertid ingen planer om å utrede behov for regnekraft for utvikling og bruk av store språkmodeller spesifikt for helse- og omsorgssektoren.

Forslag:

Det settes i gang en utredning for å vurdere behov, kost og nytte knyttet til regnekraft og infrastruktur for å kunne trene, kunnskapsforankre, teste og bruke store mengder og sensitive data i helse- og omsorgssektoren og foreslå ulike tilnærminger for å løse behovet.

For øvrig vises det særlig til avsnitt 4.3.2 og kapittel 4.4. Noe av behovet for slik infrastruktur er beskrevet i kapittel 4.3.4 i felles KI-plan for helse- og omsorgstjenesten.

Utrede forvaltningsstruktur for store språkmodeller i helse- og omsorgssektoren

Foreslått tiltakseier: Flere tiltak, med ulike tiltakseiere

I samarbeid med: Helsedirektoratet, de regionale helseforetakene, Kommunenes organisasjon (KS), de regionale IKT-selskapene, Nasjonalbiblioteket, Nasjonal kommunikasjonsmyndighet, Digitaliseringsdirektoratet, Digitaliserings- og forvaltningsdepartementet, Kunnskapsdepartementet

Relevant for: Helse- og omsorgssektoren

Problem som skal løses: Mangel på felles forvaltning av språkmodeller er et hinder for effektiv bruk av ressurser og deling av kompetanse på området. Lite samordning kan gi varierende og dårlig kvalitet på språkmodeller i sektor. Språkmodeller blir i dag i liten grad delt mellom aktører i sektor.

Forslag: Utrede organiseringen av en egen organisasjon/senter med ansvar for forvaltning av språkmodeller i helse- og omsorgssenteret.

En felles forvaltning vil kunne omfatte flere funksjoner, blant annet:

- forvaltning og videreutvikling av felles datagrunnlag for tilpassing til norske forhold
- forvaltning og videreutvikling av kvalitetsrammeverk
- forvaltning av felles språkmodeller
- veiledningstjeneste for helsesektoren
- samarbeide med forskningsinstitusjoner som deltaker i relevante forskningsprosjekter
- gjøre språkmodeller tilgjengelige for bruk i sektor eller videreutvikling av for eksempel leverandører eller offentlig sektor
- koordinere innsats i sektoren knyttet til språkmodeller

For øvrig vises det særlig til avsnitt 4.3.5 og kapittel 4.4