



DATA OG MODELLER

Oversigt og offentliggørelse af data og modeller

Dato: 02-02-2026

Status: Final

Version: 1.3

1. Formål

At dokumentere hvilke data og AI-modeller der er anvendt og udviklet i projektet, hvordan de er håndteret, og hvordan de – i relevant og lovlig form – kan **offentliggøres og deles** med andre aktører. Dokumentet skal sikre transparens, genbrug og reproducérbarhed, samt danne grundlag for videre udvikling og samarbejde.

Resume

Dataindsamlingsaktiviteter og resulterende data opdelt på projektets use cases: Sygeplejefaglig udredning, opsummering af journaldata, og indtaling af besøgsbeskrivelser er beskrevet og overordnede opgørelser for data kan findes i tabel 2, 3 og 4.

Til projektets forskellige formål er forskellige anvendte og trænede modeller benyttet.

Talegenkendelse er primært understøttet ved brug af en whisper-large-v3 model hostet lokalt, indledningsvis via Transformers og senere via Faster-whisper udstillet i en FastAPI applikation udviklet i Python.

Kategorisering af tekst til sygeplejefaglig udredning er understøttet ved brug af en fine-tunet embedding model via SetFit modulet. Modellen er også udstillet i en FastAPI applikation udviklet i Python.

Opsummering af tekst er understøttet ved brug af en lokal Llama3.3-70B-Instrukt model udstillet via Ollama og forskellige eksternt hostede modeller fra OpenAI udstillet via Azure AI. Modellerne GPT-4o og GPT-3o er de primære. De eksternt hostede modeller er også benyttet til syntetisk datagenerering.

Relevant data og modeller vil blive offentliggjort på Hugging Face til benyttelse under MIT-license.

Indholdsfortegnelse

1. Formål.....	2
Resume	2
2. Introduktion.....	4
3. Data.....	5
3.1 Syntetisk data.....	5
3.2 Overblik.....	5
Sygeplejefaglige udredninger (UC1)	5
Journal data.....	9
Indtaling af besøgsbeskrivelser.....	10
Offentliggørelse af data.....	11
Ægte data.....	11
Supplerende materiale.....	11
4. Modeller	12
4.1 Overblik.....	12
4.2 Anvendte modeller.....	12
Whisper	12
Llama	12
GPT	13
Embedding modeller.....	13
Andre	13
4.3 Trænede modeller	13
5. Bilag	14
Bilag A: Mapning fra helbredstilstande til sygeplejetilstande.....	14

Figur-oversigt

Figur 1: Sygeplejefaglig udredning - Komponent oversigt.....	6
Figur 2: Data og indsamlingsaktiviteter ifm. UC1 Afprøvningsklient og efterfølgende behandling	7
Figur 3: Data og indsamlingsaktiviteter ifm. syntetisk datagenerering og efterfølgende behandling.....	8
Figur 4: Data og indsamlingsaktiviteter ifm. UC3 Afprøvningsklient	11

Tabel-oversigt

Tabel 1: Sygeplejefaglig udredning - Datasæt overblik med efterbehandlingsstatus	8
Tabel 2: Sygeplejefaglig udredning – Syntetisk data valideringsstatus.....	9
Tabel 3: Indtaling af besøgsbeskrivelser - Data oversigt.....	10

2. Introduktion

Vi har i projektet arbejdet med 3 områder for værdiskabelse i sygeplejen, kaldet use cases, og har til hver af dem udviklet en løsning baseret på talegenkendelse, tekstklassificering og/eller generative tekst modeller. Løsningerne har det til fælles at de forsøger at afhjælpe byrden forbundet med at generere eller orientere sig i journaldata. Disse løsninger har vi gennem afprøvninger vist frem til sygeplejefagligt personale fra et bredt udsnit af kommuner.

Projektet har arbejdet med forskellige muligheder for værdiskabelse i syge- og hjemmeplejen samlet i 3 use cases (UC). Følgende er use cases overskrift, efterfulgt af en kort beskrivelse:

- **UC1: Sygeplejefaglig udredning**

Støtte til oprettelse af sygeplejetilstandsnotater på baggrund af en sygeplejefaglig udredningssamtale (SFU-samtale) mellem borger og sygeplejerske. Støtten ydes gennem udkast til notater baseret på samtals indhold.

- **UC2: Opsummering af borgerjournal**

Opsummering af borgerjournal til støtte ved borgerbesøg udført af hhv. Sygeplejersker (SPL) og sundheds- og omsorgsassisterter og -hjælpere (SSA/SSH).

- **UC3: Indtaling af besøgsbeskrivelse**

Støtte til oprettelse og opdatering af besøgsbeskrivelser til hjælp med hjemmebesøg udført af SSA/SSH. Besøgsbeskrivelserne oprettelse på baggrund af en indtaling.

Du kan læse mere om arbejdet med at definere, afgrænse og afprøve projektets use cases i det supplerende dokument "Talt: Uses cases og afprøvning".

Dette dokument indeholder en oversigt over data og modeller benyttet i TALT-projektet, samt en status på deres offentliggørelse. Dokumentet er derfor delt i to overordnede kapitler, data og modeller.

3. Data

I forbindelse med udvikling, afprøvning og evaluering af projektets løsninger og analyser er en række data indsamlet og genereret. Nedenstående er en opgørelse af denne data fordelt på de forskellige typer af data brugt.

3.1 Syntetisk data

Vi har i projektet benyttet syntetiske data til alle formål. Dette skyldes primært juridiske udfordringer, hvor den nødvendige hjemmel til at benytte ægte data ikke har været til stede.

Alle objektiver, fra modeltræning og evaluering til afprøvning af projektets løsninger foregår ved syntetiske data. I den forbindelse har vi etableret følgende ramme for distinktion af syntetisk data:

- Manuelt syntetisk data

Data genereret "manuelt", primært af fagligt personale i projektkommunerne og af afprøvningskommunerne i forbindelse med afprøvning af løsningerne. Dette data forventes at være det tætteste vi kommer på "ægte" data og bliver derfor primært benyttet til evaluering af modeller og samlede løsninger.

- Automatisk syntetisk data

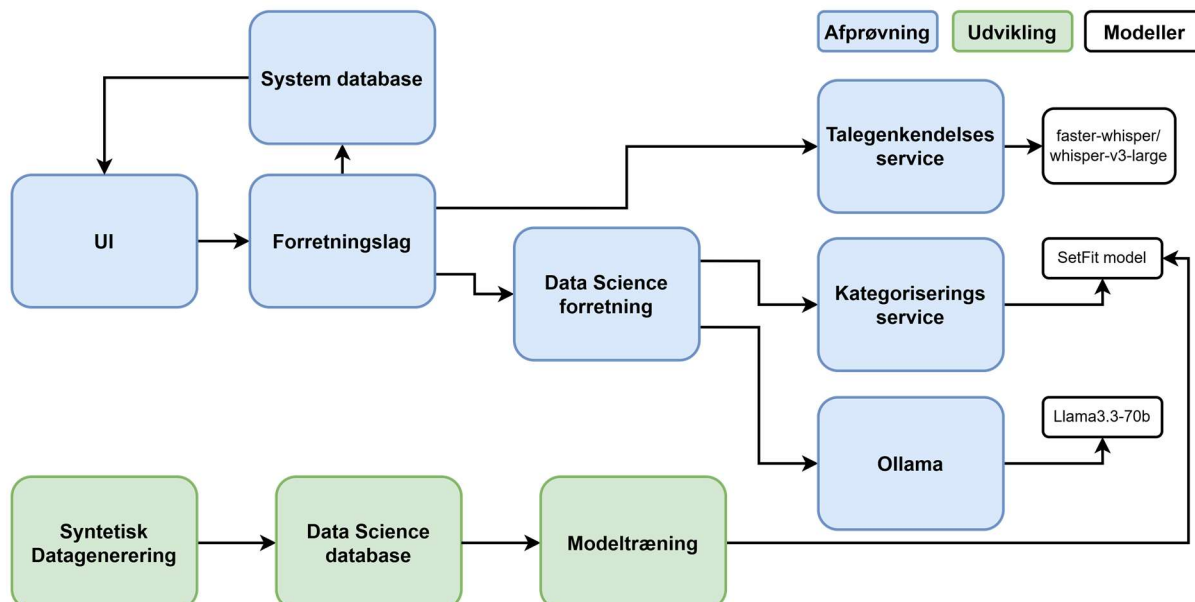
Data genereret "automatisk", primært gennem den syntetiske datagenerering ved prompt af store sprogmodeller. Denne data beriges med validering og feedback fra fagligt personale. Denne data benyttes udelukkende til træning af modeller og udvikling af løsningerne.

3.2 Overblik

Sygeplejefaglige udredninger (UC1)

I forbindelse med udvikling, afprøvning og evaluering af løsningen til understøttelse af sygeplejefaglige udredning er en række data indsamlet. Se figur 1 for et samlet overblik over komponentarkitekturen for understøttelsen af sygeplejefaglige udredninger.

Figur 1: Sygeplejefaglig udredning - Komponent oversigt



Komponenter markeret i blå udgør løsninger. Grønne komponenter er brugt til udvikling af løsningen. Hvide komponenter er modeller, både anvendte og træned.

Se figur 2 for en skematisk oversigt over konkrete dataindsamlingsaktiviteter og data indsamlet i forbindelse med brug af løsningen til sygeplejefaglige udredning (SFU-samtale) og efterbehandlingen heraf.

I forbindelse med brug af løsningen og efterbehandling indsamles følgende data:

1. Lydoptagelse af SFU-samtale (input fra bruger)
2. Udkast til transskription af samtale (via talegenkendelse service)
3. *Rettet transskription af samtale (input fra bruger)*
4. Udkast til kategorisering af samtale (via kategoriserings service)
5. *Valideret kategorisering af samtale (input fra bruger)*
6. Udkast til helbredstilstandsnotater (via Ollama)
7. Rettede helbredstilstandsnotater (input fra bruger)

Bemærk at punkt 3 og 5 ikke understøttes af løsningen, men er brugerinput der indhentes via efterbehandling¹ af den enkelte SFU-samtale optaget i løsningen ved at transskriptionen af samtalen rettes (punkt 3) og kategorisering af samtalen valideres (punkt 5).

Foruden brug af løsningen til understøttelse af SFU-samtaler har vi i projektet generet syntetisk data. Se figur 3 for en skematisk oversigt over konkrete dataindsamlingsaktiviteter og data indsamlet i forbindelse med den syntetiske datagenerering og efterbehandling heraf.

I forbindelse med syntetisk datagenerering og efterfølgende behandling indsamles følgende data:

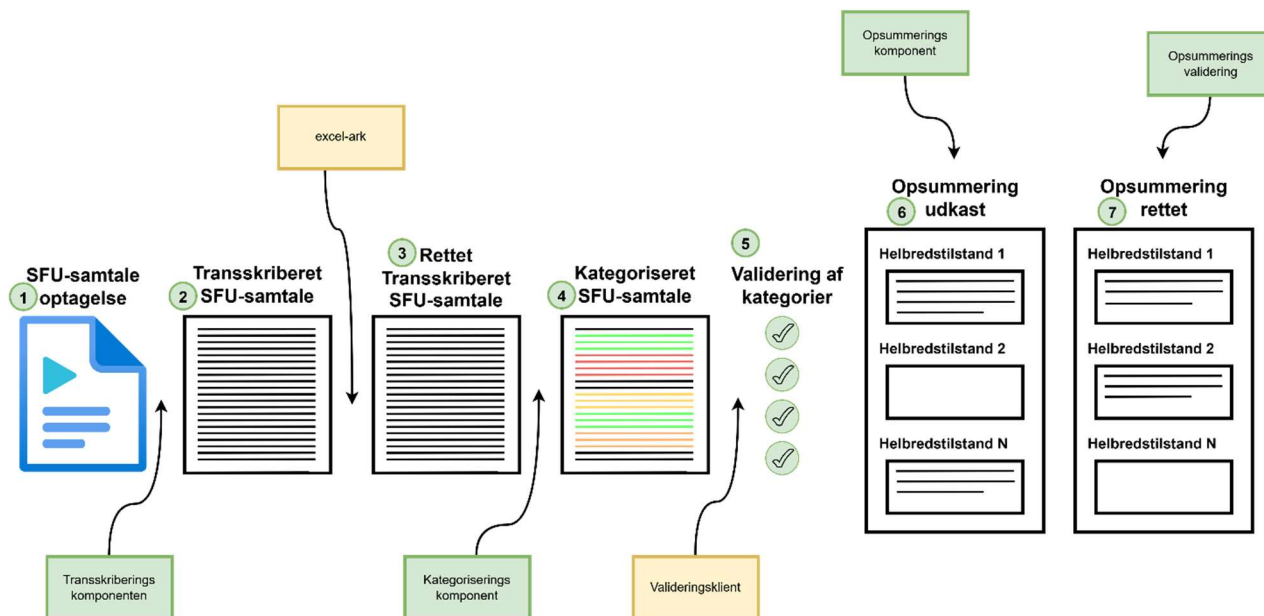
1. Syntetisk SFU-samtale med udkast til kategorisering (via syntetisk datagenereringskomponenten)

¹ Se 'Evaluerings og syntetiske data'-dokumentet for yderligere beskrivelse af processen for efterbehandling

2. Validering af kategorisering af samtale (via valideringsklient)
3. Udkast til helbredstilstandsnotater (via opsummeringskomponenten)
4. Rettede helbredstilstandsnotater (via valideringsklient)

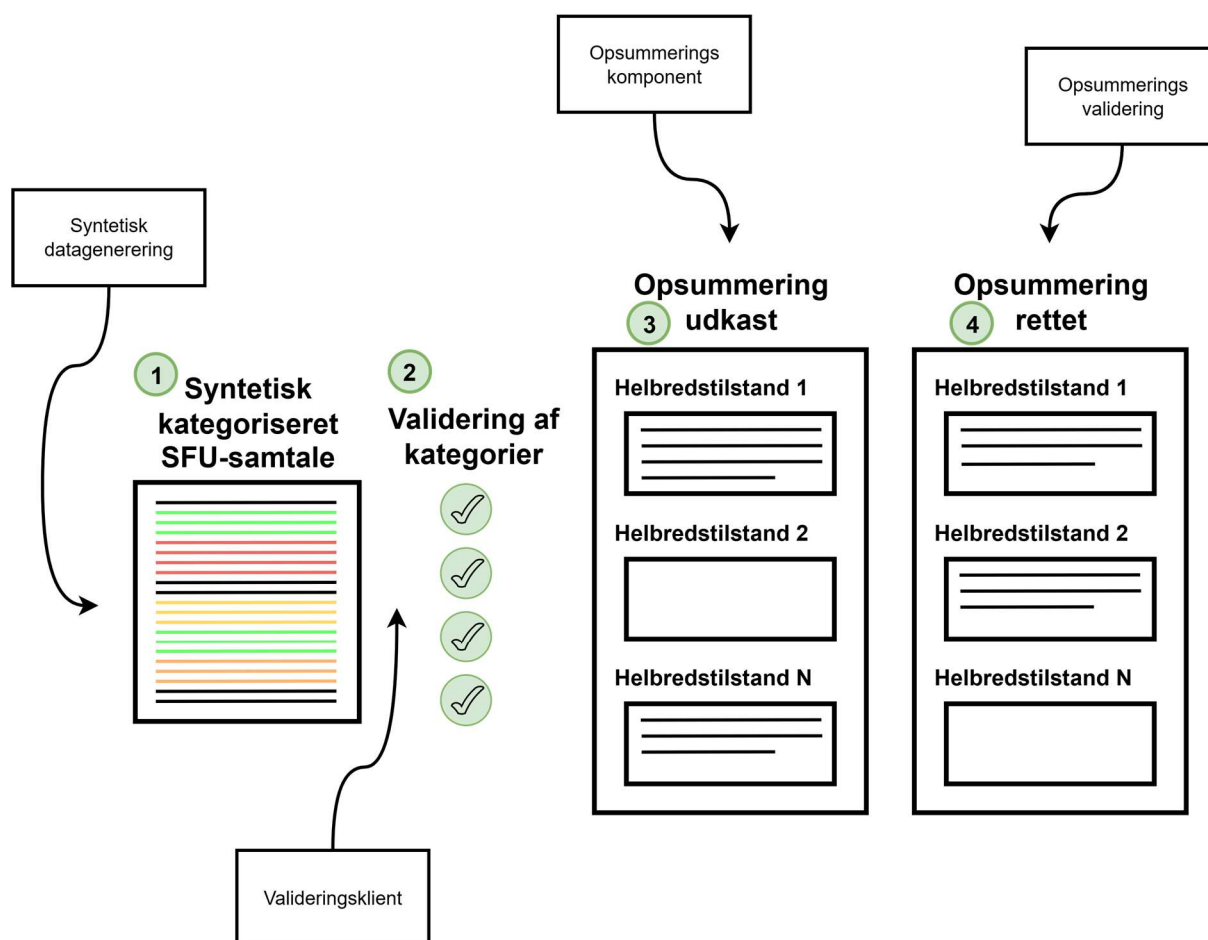
Alle konkrete dataindsamlingsaktiviteter i ovenstående grupper er udført i forskellige tempi. Alle datatyper er derfor meget forskelligt repræsenteret på tværs af den samlede portefølje af data på sygeplejefaglige udredninger.

Figur 2: Data og indsamlingsaktiviteter ifm. UC1 Afprøvningsklient og efterfølgende behandling



Aktiviteter markeret med grøn er understøttet ifm. afprøvning af løsningen. Aktiviteter markeret med gult kræver efterbehandling.

Figur 3: Data og indsamlingsaktiviteter ifm. syntetisk datagenerering og efterfølgende behandling



Se tabel 2 for et overblik over de forskellige typer datasæt og deres efterbehandlingsstatus. Afprøvningsdata refererer til data indsamlet ifm. afprøvnings af systemet. Evalueringsdata er et indledende datasæt skabt af manuelle syntetiske samtaler og blev brugt til at etablere en baseline for løsningens performance. Syntetisk data er et datasæt bestående af flere iterationer af syntetiske SFU-samtaler skabt via syntetisk datagenereringskomponenten², se tabel 3 for en oversigt over den faglige valideringsstatus indhentet via valideringsklienten.

Tabel 1: Sygeplejefaglig udredning - Datasæt overblik med efterbehandlingsstatus

Datasæt	Antal samtaler	Transskription rettet	Kategorisering valideret	Opsummeringer rettet
Afprøvningsdata	99	23	0	99
Evalueringsdata	12	12	12	12
Syntetisk data	240	-	0	0

² Se 'Evaluering og syntetiske data-dokumentet for beskrivelse af processen for syntetisk datagenerering

Tabel 2: Sygeplejefaglig udredning – Syntetisk data valideringsstatus

Antal samtaler	Valideret positivt	Valideret negativt	Ikke valideret
240	25	40	175

Journal data

Projektets use case 2 og 3 afhænger af realistisk og repræsentativt data fra omsorgsjournaler. Til dette har vi i projektet defineret en journalrepræsentation, der har til formål at repræsentere det mest relevante data fra omsorgsjournalen, uafhængigt af det elektroniske omsorgsjournal system (EOJ). Hvilke felter der indgår i journalrepræsentationen, er besluttet af projektets faglige spor med fokus på de use cases, der bygger på data fra omsorgsjournalerne. Journalrepræsentationen indeholder følgende information om borgeren:

- Aktuelle sygeplejetilstande
- Generelle oplysninger om borgeren
 - Overordnede helbredsoplysninger
 - Motivation
 - Vaner
 - Mestring
 - Ressourcer
- Udvalgte kliniske observationer
- Udvalgte vigtige observationer
- Besøgsbeskrivelser (fordelt på 4 tidsrum: morgen, middag, aften, nat)

Den samlede repræsentation er primært etableret for at understøtte projektets use case 2 og indeholder derfor felter til at understøtte opsummering af omsorgsjournaler til sygeplejersker og social- og sundhedsassistenter og -hjælpere. Information om borgers besøgsbeskrivelse understøtter også udvikling af projektets use case 3, hvor genererede beskrivelser bruges som grundlag til indtallingsløsningen.

Til generering af manuelt syntetiske journalrepræsentationer blev et excel-ark oprettet, hvor det faglige spor kunne indskrives samlede repræsentationer. De blev brugt til både udvikling og afprøvning af løsningen til opsummering af journal data. Til hjælp med udvikling af løsningerne blev et bud på opsummeringer genereret til en række af repræsentationerne. To typer af opsummeringer blev genereret, en rettet mod sygeplejersker og en rettet mod social- og sundhedshjælpere og –assistenter (SSH/SSA).

I forløbet med use case 2 blev der genereret 64 syntetiske journalrepræsentationer, hvoraf ca. 44 af dem blev brugt til løbende udvikling og interne tests af løsningen. De resterende blev brugt til en ekstern afprøvning, hvor løsningen blev vist frem til og evalueret af repræsentanter fra de 11 afprøvningskommuner.

Indtaling af besøgsbeskrivelser

I forbindelse med udvikling, afprøvning og evaluering af løsningen til understøttelse af indtaling af besøgsbeskrivelser er flere typer data indsamlet. Se figur 4 for en skematisk oversigt over konkrete dataindsamlings aktiviteter og data indsamlet i forbindelse med brug af løsningen til indtaling af besøgsbeskrivelser.

I forbindelse med brug af løsningen indsamles følgende data:

1. Lydoptagelse af indtaling (input fra bruger)
2. Transskribering af indtaling (via transskriberingskomponent)
3. Udkast til besøgsbeskrivelse (via opsummeringskomponent)
4. Rettet besøgsbeskrivelse (input fra bruger)

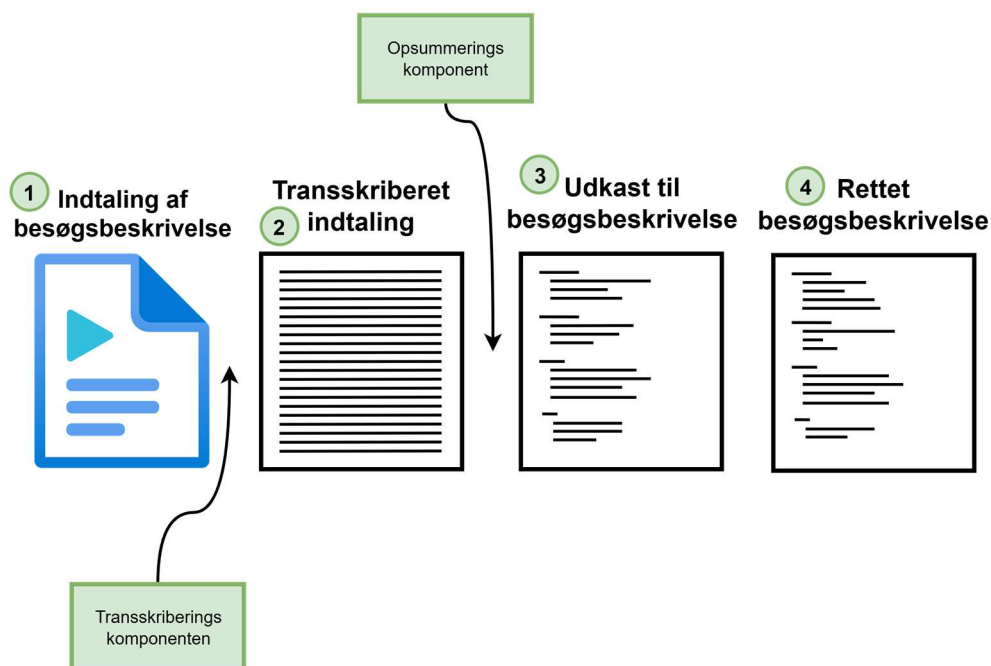
Indsamling af rettede transskriptioner er hverken understøttet af løsningen eller efterbehandling af data.

Se tabel 4 for en oversigt over data genereret til udvikling og indsamlet i forbindelse med afprøvning af løsningen. Tabellen viser data genereret og indsamlet i forskellige stadier af løsningen, først data genereret til den indledende udvikling 'Udvikling af løsning', dernæst data indsamlet ifm. den første afprøvning afholdt internt i projektkommuner og senest data genereret til udviklingen af feature 'Opdatering af besøgsbeskrivelser'.

Tabel 3: Indtaling af besøgsbeskrivelser - Data oversigt

Formål/Indhold	Dikteringer af Besøgsbeskrivelser (lydfiler)	Besøgsbeskrivelser
Udvikling af løsning	21 fra KK 22 fra Aarhus 27 fra Aalborg	24 fra Aarhus 46 fra KK 27 fra Aalborg
Intern afprøvning	15 fra KK 13 fra Aarhus 24 fra Aalborg	-
Udvikling af feature "Opdatering af besøgsplaner"	7 fra KK 13 fra Aalborg	-

Figur 4: Data og indsamlingsaktiviteter ifm. UC3 Afprøvningsklient



Offentliggørelse af data

Ovenstående data vil blive offentliggjort på projektets hjemmeside. Brug af data er indbefattet af MIT-licens³

Ægte data

Da projektets overordnede rammer ikke tillod at løsningerne blev udviklet og afprøvet direkte ved brug af ægte data, var vi i projektet begrænset til statistiske udtræk fra EOJ-systemet til validering af det syntetiske data. Denne validering er derfor sket i ret begrænset omfang, hvorfor den primære validering af det syntetiske data (både manuelt og automatisk) sker via faglig validering af personale fra projektkommunerne.

Supplerende materiale

Vi har brugt sygeplejefaglige materiale til at informere flere af de tekniske processer. Materialet er typisk information produceret og samlet til at etablere journaliseringspraksis og instruere medarbejdere i de projektkommunerne, samt en del materiale fra Fælles Sprog III (FSIII) og dens forenkling.

I forbindelsen med forenklingen af FSIII skete overgangen fra 43 helbredstilstande til 23 sygeplejetilstande. Denne overgang betød at projektets data og løsninger skulle migreres, se bilag A for nøglen, der blev brugt til denne migrering.

³ [MIT License](#)

4. Modeller

I forbindelse med udvikling og evaluering af projektets løsninger har vi gjort brug af en række forskellige modeller til flere forskellige formål. Størstedelen af modeller er præ-trænede modeller hentet fra nettet eller tilgået gennem tilkøbte services, her omtalt som 'anvendte modeller', men vi har også trænet modeller selv. Denne afsnit er et overblik over hvilke modeller vi har benyttet og til hvilke formål.

4.1 Overblik

I løbet af projektet har vi gjort brug af en række forskellige modeller og modelarkitekturer. Herunder er en oversigt over modellerne, med en kort intro, samt en beskrivelse af, hvad vi har brugt den til. Oversigten er opdelt i 'anvendte' og 'trænede' modeller. Forskellen mellem de to kategorier afhænger af hvorvidt vi blot har taget en modellen fra hylden og brugt den, eller om vi har trænet eller fine-tunet modelarkitekturen selv.

4.2 Anvendte modeller

Whisper

Whisper er en talegenkendelsesmodel udgivet af OpenAI⁴. Vi har i projektet brugt modelarkitekturen som vores primære talegenkendelsesmodel implementeret i Python. Vi har gjort brug af to forskellige implementationer i Python, først transformers modulets 'automatic-speech-recognition' pipeline udviklet af Hugging Face⁵ og senest Faster-whisper implementeret via CTranslate2⁶ udviklet af SYSTRAN⁷. I begge implementeringer har vi brugt large-v3 modellen. Mere information om den konkrete implementering vil kunne findes i kodebasen, når den er offentliggjort, samt i dokumentet 'Vej til drift'.

Ud over den lokalt hostet implementering af Whisper, har vi også benyttet en eksternt hostet løsning. Denne model var whisper large-v3-turbo udstillet via Cloudflare⁸.

Llama

Llama er en familie af store sprogmodeller udgivet af Meta AI⁹. Vi har i projektet benyttet modellen Llama3.1-70B¹⁰, men skiftede undervejs til den instruks trænede version Llama3.3-70B-Instruct¹¹. Disse modeller har vi hostet lokalt på projektets GPU via Ollama¹².

Ud over de lokalt hostede implementeringer via Ollama, har vi også benyttet en eksternt hostet løsning. Denne model var Llama3.3-70B hostet via Azure AI¹³.

⁴ [GitHub - openai/whisper: Robust Speech Recognition via Large-Scale Weak Supervision](https://github.com/openai/whisper)

⁵ [Automatic speech recognition with a pipeline - Hugging Face Audio Course](https://huggingface.co/docs/transformers/main_classes/automatic-speech-recognition)

⁶ [GitHub - OpenNMT/CTranslate2: Fast inference engine for Transformer models](https://github.com/OpenNMT/CTranslate2)

⁷ [GitHub - SYSTRAN/faster-whisper: Faster Whisper transcription with CTranslate2](https://github.com/SYSTRAN/faster-whisper)

⁸ [whisper · Cloudflare Workers AI docs](https://www.cloudflare.com/ai-workers/docs/whisper/)

⁹ [Docs & Resources | Llama AI](https://llama.meta.com/docs/)

¹⁰ [meta-llama/Llama-3.1-70B · Hugging Face](https://huggingface.co/meta-llama/Llama-3.1-70B)

¹¹ [meta-llama/Llama-3.1-70B-Instruct · Hugging Face](https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct)

¹² [Ollama's documentation - Ollama](https://ollama.com/docs/)

¹³ [Azure AI Foundry](https://azure.microsoft.com/en-gb/solutions/ai-factory/)

GPT

GPT er en familie af store sprogmodeller udgivet af Open AI¹⁴. Vi har i projektet brugt forskellige modeller fra familien: GPT-4o, GPT-4.1, GPT-3o, GPT-5. Modellerne er brugt gennem et API udstillet via ¹⁵ ressourcen, der udstiller modeller fra Azure AI.

Embedding modeller

Vi har i projektet, primært til kategorisering af tekst, brugt forskellige embedding modeller. Modellerne er downloadet via Hugging Face og implementeret i løsningerne via python.

Vi har brugt følgende embedding modeller: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2¹⁶ fra sentence-transformers¹⁷ og intfloat/multilingual-e5-large-instruct¹⁸.

Andre

I projektets begyndelse eksperimenterede vi med forskellige open-weight modeller, før vi landede på at bruge Llama som vores primære lokale model. Heriblandt følgende modeller: Llama3-pro-8b-instruct, mistral-7b-instruct¹⁹, phi-3-mini-128k-instruct²⁰, phi-3-medium-128k-instruct

Modeller blev udstillet lokalt via Ollama og llama.cpp²¹

4.3 Trænede modeller

Vi har i projektet trænet modeller til tekstklassificering til understøttelsen af sygeplejefaglige udredninger. Til dette har vi eksperimenteret med forskellige model-strukturer baseret på tekst embedding modeller. Motiveret af vores data begrænsninger begyndte vi tidligt i projektet at benytte finetuning af embedding modeller via python modulet SetFit²² til opgaven.

Modeller brugt i forbindelse med løsningen til sygeplejefaglig udredning er offentliggjort på [Github](#).

¹⁴ [Models - OpenAI API](#)

¹⁵ [Azure OpenAI in Foundry Models | Microsoft Azure](#)

¹⁶ [sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 · Hugging Face](#)

¹⁷ [SentenceTransformers Documentation — Sentence Transformers documentation](#)

¹⁸ [intfloat/multilingual-e5-large-instruct · Hugging Face](#)

¹⁹ [Models Overview | Mistral AI](#)

²⁰ [Phi Open Models - Small Language Models | Microsoft Azure](#)

²¹ [GitHub - ggml-org/llama.cpp: LLM inference in C/C++](#)

²² [GitHub - huggingface/setfit: Efficient few-shot learning with Sentence Transformers](#)

5. Bilag

Bilag A: Mapning fra helbredstilstande til sygeplejetilstande

FUNTIONSEVNE	
Problemer med personlig pleje	
Problemer med daglige aktiviteter	Problemer med daglige aktiviteter
BEVÆGAPPERAT	
Problemer med mobilitet og bevægelse	Problemer med mobilitet og bevægelse
ERNÆRING	
Problemer med væskeindtag	Problemer med ernæring
Problemer med fødeindtag	
Problemer med vægtændring	Problemer med vægt
Problemer med overvægt	
Problemer med undervægt	
HUD OG SLIMHINDER	
Problemer med diabetisk sår	Problemer med diabetisk sår
Problemer med tryksår	Problemer med tryksår
Problemer med cancersår	Problemer med <i>andre</i> sår
Problemer med kirurgisk sår	
Problemer med arterielt sår	
Problemer med venøst sår	
Problemer med blandingssår	
Problemer med traumesår	
Andre problemer med hud og slimhinder	Problemer med hud og slimhinder
KOMMUNIKATION	
Problemer med kommunikation	Problemer med kommunikation
PSYKOSOCIALE FORHOLD	
Problemer med socialt samvær	Problemer med socialt samvær
Problemer med misbrug	Problemer med misbrug
Mentale problemer	

Emotionelle problemer	Problemer med Trivsel
RESPIRATION OG CIRKULATION	
Respirationsproblemer	Respirationsproblemer
Cirkulationsproblemer	Cirkulationsproblemer
SEKSUALITET, KØN OG KROPSOPFATTELSE	
Problemer med seksualitet, køn og kropsopfattelse	Problemer med seksualitet, køn og kropsopfattelse
SMERTER OG SANSEINDTRYK	
Aktuelle smerter	Problemer med smerter
Periodevise smerter	
Kroniske smerter	
Problemer med synssans	Problemer med sanser
Problemer med lugtesans	
Problemer med hørelse	
Problemer med smagssans	
Problemer med følesans	
SØVN OG HVILE	
Døgnrytmeproblemer	Problemer med søvn og hvile
Søvnproblemer	
VIDEN OG UDVIKLING	
Problemer med hukommelse	Problemer med hukommelse
Problemer med sygdomsindsigt	Problemer med <i>sundhedskompetence</i>
Problemer med indsigt i behandlingsformål	
Kognitive problemer	Kognitive problemer
UDSKILLELSE AF AFFALDSSTOFFER	
Problemer med vandladning	Problemer med vandladning
Problemer med urininkontinens	
Problemer med afføringsinkontinens	Problemer med mave og tarm
Problemer med mave og tarm	