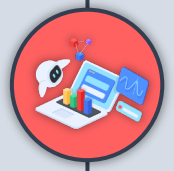
# HelseSvar

Bruk av kunstig intelligens til publikumsrettet informasjon

## Hvorfor startet vi HelseSvar?

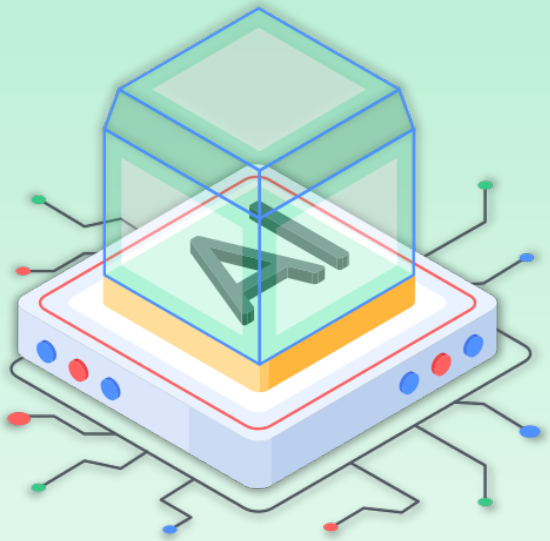
## Hva er HelseSvar?



## Hvordan utreder vi HelseSvar?

## Erfaringer og funn så langt

# Hvorfor startet vi HelseSvar?




Tidligere har kunstig intelligens vært for dårlig på språk. Nå som språkmodeller (også kalt LLM – Large Language Model) finnes har bruken eksplodert. Utfordringen er at språkmodeller har vist seg å være upålitelig når det gjelder kvalitetssikring og etikk.

Befolkningen eksponeres for helseinformasjon som er generert av kunstig intelligens på basis av kilder man ikke har oversikt over. Informasjonen er ikke kvalitetssikret, og kan være både feil, misvisende eller potentielt skadelig.

Kan (og bør?) Hdir forsøke å skape faglig kvalitetssikret motvekt i KI verden?

# Hva er HelseSvar?

Løsningen er en KI-assistent som benytter RAG (Retrieval Augmented Generation).

Poenget er å dele opp løsningen slik at:

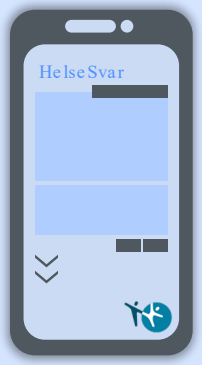
- språkmodellen tar seg av de språklige formuleringene og sikrer at innholdet kan forstås av grupper med lav helsekompetanse
- mens det faglige innholdet kommer utelukkende fra kvalitetssikrede kilder som Helsedirektoratet har kontroll over.

Småskala eksperiment på slutten av 2023.

# Hva er HelseSvar?

## Front-end

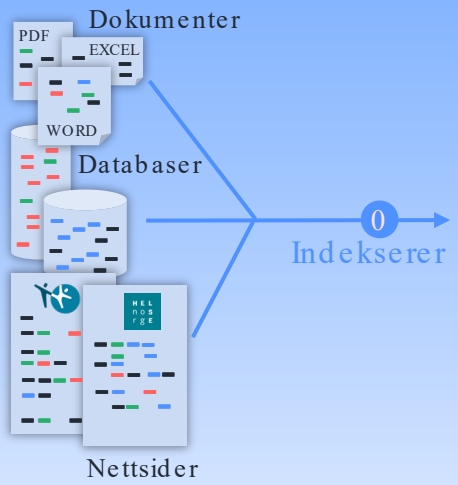
Enkel web-applikasjon ("klient") laget via React




## Back-end

Server som kjører løsningen via Hdir sin Azure infrastruktur

# Hva er HelseSvar?

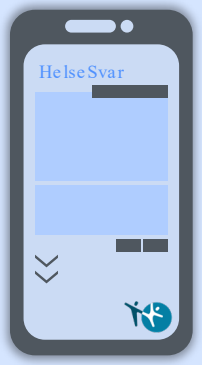


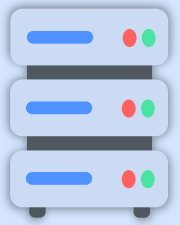

## Kunnskapsbase (Llama vektorindeks)

Innholdet i utvalgte kilder deles opp i brikker ("chunks") og kategorises basert på blant annet metadata og relasjoner til andre brikker ("vektorembedding"). Prosessen ender med et sett med filer eller indeks som vil bli datarammeverket for språkmodellen.

## Front-end

Enkel web-applikasjon ("klient") laget via React

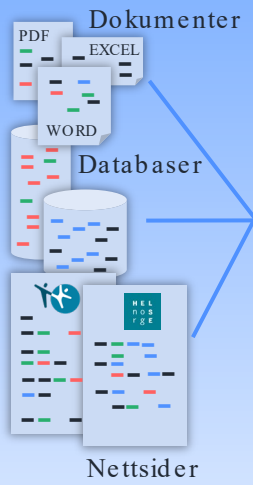


## Back-end

Server som kjører løsningen via Hdir sin Azure infrastruktur

# Hva er HelseSvar?



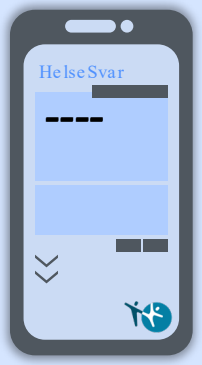

0 Indekserer

## Kunnskapsbase (Llama vektorindeks)

Innholdet i utvalgte kilder deles opp i brikker ("chunks") og kategorises basert på blant annet metadata og relasjoner til andre brikker ("vektorembedding"). Prosessen ender med et sett med filer eller indeks som vil bli datarammeverket for språkmodellen.

## Front-end

Enkel web-applikasjon ("klient") laget via React



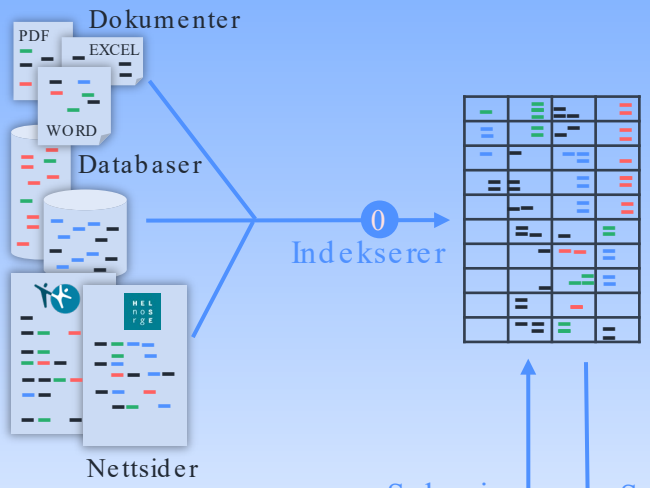

Sender spørsmål og instruksjoner

1

## Back-end

Server som kjører løsningen via Hdir sin Azure infrastruktur

# Hva er HelseSvar?




## Kunnskapsbase (Llama vektorindeks)

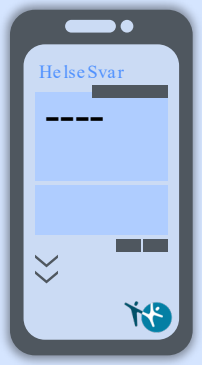
Innholdet i utvalgte kilder deles opp i brikker ("chunks") og kategorises basert på blant annet metadata og relasjoner til andre brikker ("vektorembedding"). Prosessen ender med et sett med filer eller indeks som vil bli datarammeverket for språkmodellen.

Søker i indeksen 2

3 Sender relevante tekstelementer

## Front-end

Enkel web-applikasjon ("klient") laget via React

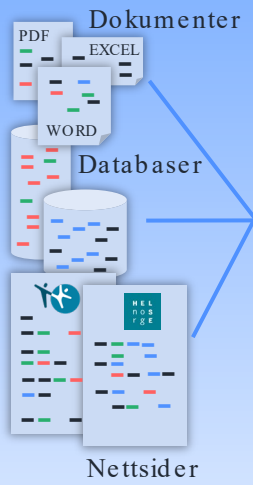



Sender spørsmål og instruksjoner 1

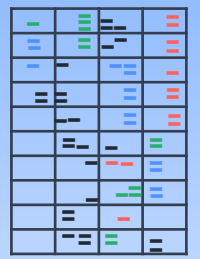
## Back-end

Server som kjører løsningen via Hdir sin Azure infrastruktur

# Hva er HelseSvar?




0 Indekserer



## Kunnskapsbase (Llama vektorindeks)

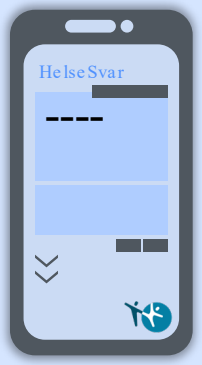
Innholdet i utvalgte kilder deles opp i brikker ("chunks") og kategorises basert på blant annet metadata og relasjoner til andre brikker ("vektorembedding"). Prosessen ender med et sett med filer eller indeks som vil bli datarammeverket for språkmodellen.

Søker i indeksen 2

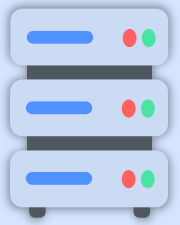
3 Sender relevante tekstelementer

## Front-end

Enkel web-applikasjon ("klient") laget via React




Sender spørsmål og instruksjoner 1

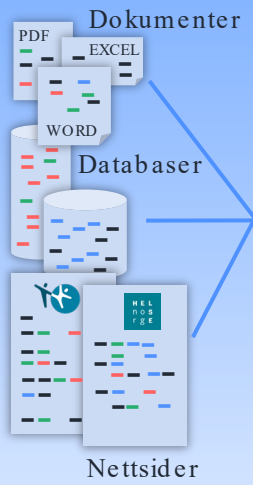


## Back-end

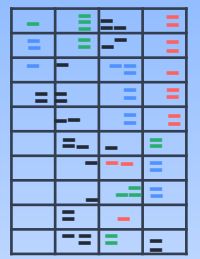
Server som kjører løsningen via Hdir sin Azure infrastruktur

4 Sender spørsmål, instruksjoner og tekstelementer

# Hva er HelseSvar?




0 Indekserer



## Kunnskapsbase (Llama vektorindeks)

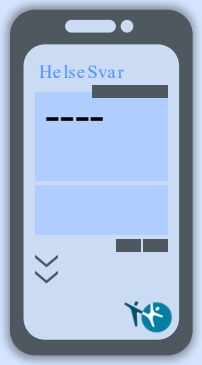
Innholdet i utvalgte kilder deles opp i brikker ("chunks") og kategorises basert på blant annet metadata og relasjoner til andre brikker ("vektorembedding"). Prosessen ender med et sett med filer eller indeks som vil bli datarammeverket for språkmodellen.

Søker i indeksen 2

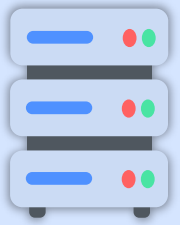
3 Sender relevante tekstelementer

## Front-end

Enkel web-applikasjon ("klient") laget via React



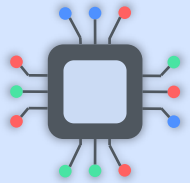

Sender spørsmål og instruksjoner 1



## Back-end

Server som kjører løsningen via Hdir sin Azure infrastruktur

4 Sender spørsmål, instruksjoner og tekstelementer

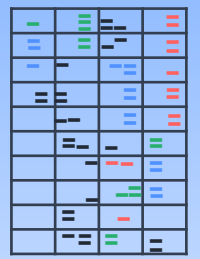


## Språkmodell (Open AI/GPT3.5 og 4)

Språkmodellen matcher embedding fra spørring til embedding fra vektorindeksen og setter sammen de meste relevante "chuncks" til å bygge et svar.

# Hva er HelseSvar?

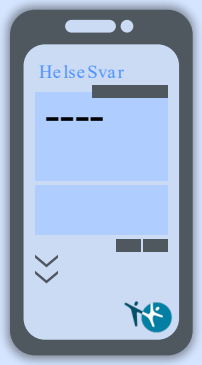

0 Indekserer



## Kunnskapsbase (Llama vektorindeks)

Innholdet i utvalgte kilder deles opp i brikker ("chunks") og kategorises basert på blant annet metadata og relasjoner til andre brikker ("vektorembedding"). Prosessen ender med et sett med filer eller indeks som vil bli datarammeverket for språkmodellen.

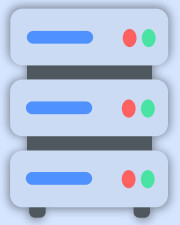
Søker i indeksen 2 | 3 Sender relevante tekstelementer




## Front-end

Enkel web-applikasjon ("klient") laget via React

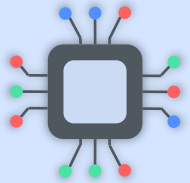
Sender spørsmål og instruksjoner 1



## Back-end

Server som kjører løsningen via Hdir sin Azure infrastruktur

Sender svar 5 | 4 Sender spørsmål, instruksjoner og tekstelementer

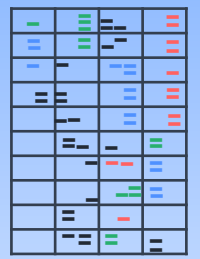


## Språkmodell (Open AI/GPT3.5 og 4)

Språkmodellen matcher embedding fra spørring til embedding fra vektorindeksen og setter sammen de meste relevante "chuncks" til å bygge et svar.

# Hva er HelseSvar?

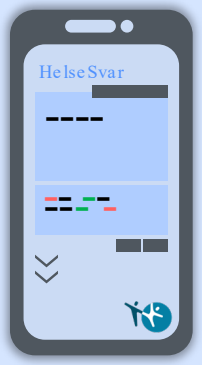

0 Indekserer



## Kunnskapsbase (Llama vektorindeks)

Innholdet i utvalgte kilder deles opp i brikker ("chunks") og kategorises basert på blant annet metadata og relasjoner til andre brikker ("vektorembedding"). Prosessen ender med et sett med filer eller indeks som vil bli datarammeverket for språkmodellen.

Søker i indeksen 2 | 3 Sender relevante tekstelementer

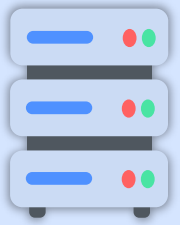



## Front-end

Enkel web-applikasjon ("klient") laget via React

Sender spørsmål og instruksjoner 1

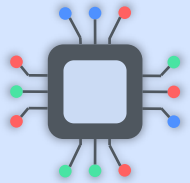
6 Sender svar



## Back-end

Server som kjører løsningen via Hdir sin Azure infrastruktur

Sender svar 5 | 4 Sender spørsmål, instruksjoner og tekstelementer



## Språkmodell (Open AI/GPT3.5 og 4)

Språkmodellen matcher embedding fra spørring til embedding fra vektorindeksen og setter sammen de meste relevante "chuncks" til å bygge et svar.

# Hvordan utreder vi HelseSvar?

Innledende testing så ut til å vise at prototypen:

- svarer med bra kvalitet og viser empati
- håndterer stavefeil, dialekter og upresist språk
- er lett å forvalte teknisk og relativt lite kostnadsdrivende
- sier ifra når den ikke finner svaret istedenfor å improvisere
- Kan vise til de kildene som er brukt

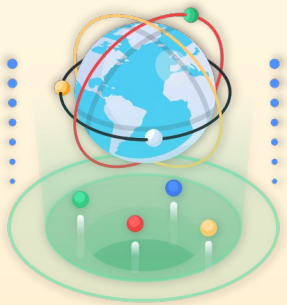
Prototypen ble derfor testet ytterligere under flere bruksområder og vurdert opp mot en rekke relevante problemstillinger. Eksperimentet gikk over til et forprosjekt i våren 2024 og ble meldt inn i Datatilsynets sandkasse. Målet med forprosjektet er å levere en konseptutredning innen årets slutt.

# Hvordan utreder vi HelseSvar?

- Validitetstesting av prototypen med ulike språksmodeller, både med og uten RAG, med og uten instruksjoner og ved bruk av ulike vektorindekser. Testene er gjennomført for oppgaver i fem bruksområder:
  - Veiledningsstøtte til å besvare spørsmål om tobakk
  - Veiledningsstøtte til å besvare spørsmål om prevensjon
  - Veiledningsstøtte til å besvare spørsmål om psykisk helse
  - Svare spørsmål i flere språk (Polsk, Arabisk og Somali)
  - Koding av dødsårsaker med ICD-11

- Utredning av nødvendige forhold til utarbeidelse av et konseptforslag (faglig kvalitet, brukeropplevelse, utvikling, forvaltning, arkitektur, sikkerhet, mm.)
- Utredning av personvernrettslige problemstillinger gjennom Datatilsynets sandkasse (4 workshop og noen ekstra drøftingsmøter)

# Erfaringer og funn så langt



## Kvalitet

RAG gir høyere faglig kvalitet enn det å kun bruke grunnmodeller, forutsatt at indeksen er satt opp med tilstrekkelig kompetanse. RAG viser stor variasjon i språklig kvalitet avhengig av språk, høy for noen, lav for andre, må testes før bruk. RAG åpner også for å vise kilder og sikre transparens, som øker tillit.

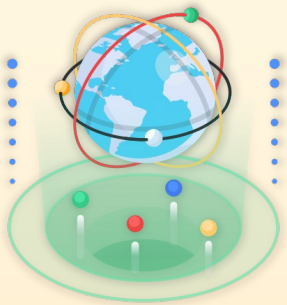
## Effektivitet

RAG er kostnadseffektiv å forvalte og testene som veiledningsverktøy en reduksjon av svartiden på 30 til 50%. Utredningen peker også på flere områder hvor løsningen har overføringsverdi.

## Teknologi

Testene har lært oss mye om hvordan vi kan styre KI-teknologien, fremfor å bli styrt. Læringspunkter inneholder blant annet fallgruver å unngå når det gjelder instruksjoner, bruk av vektorindekser vs. grafindekser, teknikker for å forebygge hallusinasjoner og feil, samt pre og post-prosesseringsteknikker for å sikre at indeksene brukes fremfor grunnmodellens treningsdata. Det er også utredet ulike alternativer for å avidentifisere spørsmål.

# Erfaringer og funn så langt



## Sandkasseprosessen

En unik drøftingsarena: annen arena enn tilsyn eller vanlig veiledning, hvor mindre tidspress åpner for dybdediskusjoner som skaper mye læring (for alle parter). Prosessen åpner også for en mer løsningsorientert tilnærming hvor målet for alle er å bygge en personvernvennlig løsning, ikke å unngå behandling av personopplysninger til enhver pris. Spesielt relevant for oss som jobber med helsespørsmål og ungdommer 13-20 år.

Status: fire workshoper er gjennomført med diskusjoner rundt relevant lovverk både eksisterende og kommende, sentrale forutsetninger, teknikker for anonymisering/avidentifisering, bruksvilkår hos LLM leverandører, behov for behandling av personopplysninger eller ikke (avhengig av type løsning). Pågående diskusjon om flere aktuelle behandlingsgrunnlag, prosjektet avventer endelig tilbakemelding fra Datatilsynet. Ulike tolkningsansvar avhengig av lovverk, og Hdir er behandlingsansvarlig til slutt, men svært nyttig med innspill.

# HelseSvar

Takk for oppmerksomheten