SkoleGPT Teknisk Dokumentation

Følgende dokument giver et teknisk overblik over struktur og opbygning af SkoleGPT.

SkoleGPT består af to seperate komponenter:

- En webserver der håndterer UI, og står for kald til de bagvedliggende sprogmodeller.
- En webserver der indeholder sprogmodellerne. Den modtager kald og genererer svar.

I skrivende stund er sprogmodellen Gemma4, men den kan skiftes ud med en anden model hvis det ønskes.

Løsningen er hostet hos DBC Digital i deres datacentre i Danmark. Der kræves ingen login, og hverken spørgsmål eller svar bliver gemt. Denne tilgang er valgt for at prioritere brugernes sikkerhed og sikre fuld overholdelse af GDPR. Som det eneste logger vi antal kald og svartider for at kunne overvåge systemets tilstand.

Systemisk overblik

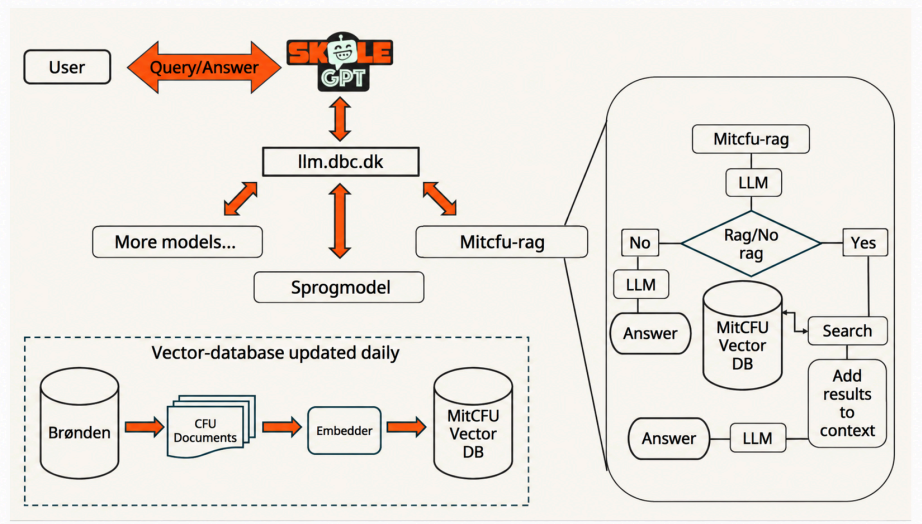

Figure 1: Arkitektur

Både LLM-proxy endpointet og UI endpointet er deployet i flere instancer for at maksimere throughput og gøre det stabilt i fohold til eventuelle udfald.

Webserver

Applikationen er skrevet i NodeJS+typescript og anvender frameworket NextJS. Applikationen er en videreudvikling af opensource projektet NextChat og er

valgt for at videreføre den første udgave af SkoleGPT projektet. NextChat er udgivet under MIT Licens.

Udover prompt har brugeren også mulighed for at specificere 3 yderligere model parametre:

- Systemprompt: Dette er den grundlæggende instruktion, som modellen modtager, inden noget andet behandles. Den definerer modellens funda mentale adfærd, såsom hvilket sprog den skal besvare spørgsmål på.
- Temperatur: Bestemmer modellens grad af kreativitet (lav temperatur = mere præcis og fokuseret, høj temperatur = mere kreativ og varieret). • Top-P: Angiver, hvor stor en del af det samlede sandsynlighedsfelt modellen vælger sine svar fra. For eksempel betyder en Top P på 0,9, at modellen ignorerer ord, som samlet udgør mindre end 10% af sandsynligheden.

LLM-Proxy

De tilgængelige sprogmodeller kører i docker med vLLM og tilgåes gennem vores proxy-lag. Det kræver en API nøgle at bruge endpointet. Dokumentationen kan ses her https://llm.dbc.dk/docs.

For at se en komplet liste over hvilke modeller ens API nøgle giver adgang til kan man tilgå https://llm.dbc.dk/v1/models.

I øjeblikket kører gemma4 modellen på dedikerede RTX6000PRO_96GB grafikkort.