

## Projektbeskrivning: Träning av LLaMa2-modellen med YouTube-undertexter

### Bakgrund:

Det finns ett ökande intresse för att utforska hur förtränade språkmodeller presterar med olika typer av data. Ett område som hittills varit relativt outforskat är data i form av undertexter från videor på YouTube.

### Mål:

A: Kartlägga kostnader och resursåtgång associerade med träning av en språkmodell som LLaMa2.

B: Förstå modellens prestation med data som endast kommer från YouTube-undertexter.

C: Skapa en proof-of-concept webbapplikation där användare kan interagera med den tränade modellen.

### Projektspecifikationer:

#### Datainsamling:

Utveckla en motor för att extrahera undertexter från specificerade YouTube-kanaler eller spellistor.

Verifiera att all datainsamling överensstämmer med "fair use" för akademiska syften.

#### Datatillredning och förträning:

Rensa och förbereda data för träning.

Träna eller finjustera LLaMa2-modellen med den insamlade datan.

#### Evaluerings:

Utvärdera modellens prestation genom korsvalidering eller andra lämpliga tekniker.

Mäta modellens svars kvalitet genom att jämföra den med befintliga modeller eller genom att använda en standardiserad uppsättning frågor och svar.

#### Webbapplikation:

Implementera en webbserver som kan köra LLaMa2-modellen.

Skapa en användarvänlig frontend där användare kan skicka frågor till modellen och få svar.

Resurser:

Team: 4-6 dataingenjörsstudenter från Chalmers högskoleingenjörsprogram.

Hårdvara: PC med AMD Ryzen 9 7950X, 64GB RAM, Nvidia RTX 4090.

Budget: Begränsad budget tillgänglig från Chalmers vid behov.

Tidsram:

Projektet ska vara slutfört senast vid jul.

Risken:

Eftersom detta är ett utforskande projekt finns det alltid en risk att den tränade modellen inte uppfyller förväntningarna. Det är också viktigt att säkerställa att alla aktiviteter, särskilt datainsamling, utförs inom ramen för lagliga och etiska riktlinjer.

Förslag för Evaluering:

För att verkligen förstå hur väl modellen presterar kan man överväga följande:

Använd en uppsättning standardfrågor för att jämföra den nytränade modellen med andra kända modeller.

Genomföra användartester där frivilliga interagerar med modellen och sedan ge feedback om svars kvaliteten.

Mät svarstiden för att se om den tränade modellen kan svara snabbt nog för realtidsinteraktioner.

Säg till om något är oklart!

Med vänlig hälsning,  
Alixander Ansari

0704743171  
[alixansari@gmail.com](mailto:alixansari@gmail.com)