Chatbotter skal også forstå sprogets danske sjæl

Chatbotter er i rivende fart blevet i stand til at svare på alt. Men de har næsten al deres viden fra engelsk. Nyt dansk projekt skal bidrage til, at det danske sprogs sjæl kommer med

Morten Mikkelsen
mikkelsen@k.dk

Hvis en dansker fortæller, at hun er i gang med at læse en ordentlig mursten, er de fleste her i landet klar over, at det ikke skal tages alt for bogstaveligt. Men kommunikerer man med ChatGPT eller en anden af de moderne chatbotter, går det ofte galt.

De er ellers på rekordtid blevet dygtige til at svare på næsten alt på et forståeligt dansk. Men de har langt det meste af deres viden fra engelsk, og på engelsk kan en *brick*, altså en mursten, ikke også betyde en tyk bog. Det er kun en mursten.

Eksemplet kommer fra Bolette Sandford Pedersen, professor ved center for sprogteknologi på Københavns Universitet.

"Der er en del metaforiske udtryk i dansk, som ChatGPT slet ikke forstår. For eksempel når vi kalder en tandbøjle for togskinner eller omtaler tykke briller som hinkesten," fortæller professoren.

Hun tilføjer, at også udtryk at tale om, at alting sejler, når det roder, eller at omtale en person som højpandet tager ChatGPT fejl af. Det sidste måske fordi janteloven ikke er så tydeligt indkodet i robotten som i det danske sprog.

Professoren arbejder i øjeblikket på et projekt, der gør det muligt systematisk at evaluere, hvor godt chatbotterne mestrer dansk, og de første resultater er lige offentliggjort i en større dansk forskningsartikel. Formålet er at undersøge, hvor meget de sprogmodeller, der bruges i Danmark, har med af det ordbogsautentiske danske sprog.

Af samme grund indgår Det Danske Sprog- og Litteraturselskab (DSL) også i forskningsprojektet, som er støttet af Carlsbergfondet. Og såvel Bolette Sandford Pedersen som ordbogsredaktør Nathalie Hau Sørensen fra DSL betoner, at selvom de kan finde morsomme eksempler på, hvad robotten ikke forstår, er de grundlæggende imponerede over, hvor meget bedre versionen ChatGPT 4.0, der kom i år, er i forhold til sidste års ChatGPT 3.5.

"Det har overrasket mig, hvor stort et spring det er lykkedes at foretage i de nyeste sprogmodeller i forhold til kendskab til dansk," fortæller Nathalie Hau Sørensen.

Også ordbogsredaktøren har dog haft held til at opdage punkter, hvor robotten bliver svag i koderne. For eksempel udsatte hun den for ældre danske stillingsbetegnelser som for eksempel en bødker, der fremstiller tønder. Stillet over for det ord er ChatGPT blank, men går ud fra, at en bødker er det samme som en snedker.

"Et andet område, chatbotten er dårlig til, er politisk ukorrekte ord som seksuelt ladet slang eller etniske betegnelser, som vi ikke bruger længere. De ord kan eller vil ChatGPT ikke forholde sig til. Eller også belærer de os om, at det er ord, man ikke må bruge," siger Nathalie Hau Sørensen.

Forskerne understreger dog, at projektet handler om meget mere end pudsige misforståelser. Overalt i Europa er der en stor optagethed af at evaluere og tilpasse chatbotternes sprogmodeller til de enkelte landes sprog, og der er behov for at udvikle systematiske måder at gøre dette på som tager udgangspunkt i de individuelle sprogsamfund.

Ifølge Bolette Sandford Pedersen er der en stor risiko for, at det danske sprog i al sin mangfoldighed vil svinde ind i en mulig fremtid, hvor en stor del af vores tekstproduktion måske kommer fra chatbotter. Og hvor vi mennesker lærer flere engelske ord- og betydningslån af chatbotten, end den lærer danske ord som murstensroman og bødker af os.

"Vi er optagede af, at arbejdet med at kommunikere med kunstig intelligens indbefatter, at man får mest muligt med af sprogets nordiske og danske sjæl. Ellers skævvrider vi sproget og mister noget dyrebart," siger Bolette Sandford Pedersen. ■