Tidsskriftet Tidsskrift for Den norske legeforening

Kunstig intelligens kan generere helseskadelig informasjon

FRA ANDRE TIDSSKRIFTER

SIMON ANDRUP
Tidsskriftet

Generativ kunstig intelligens basert på store språkmodeller kan produsere falsk og helseskadelig informasjon.

Illustrasjon: Tidsskriftet

De fleste av oss bruker internett som kilde til helseinformasjon, der feilinformasjon sprer seg opp til seks ganger raskere enn fakta. Store språkmodeller kan produsere ukorrekt og/eller misvisende informasjon om helse.

I en studie publisert i tidsskriftet BMJ undersøkte forfatterne muligheten for å generere feilinformasjon i de fire språkmodellene GPT-4, PaLM 2, Claude 2 og Llama 2 (1). De prøvde å generere falsk informasjon om at bruk av solkrem gir

hudkreft, og om at alkalisk kost helbreder kreft. Av de nevnte modellene var det bare Claude 2 som nektet å generere feilinformasjon, de andre modellene genererte store mengder overbevisende, men falsk og helseskadelig informasjon på kort tid.

– Det skal ikke mange forsøk til før man forstår at det er svært viktig hvordan en forespørsel til store språkmodeller blir formulert, sier Ragnhild Undseth, som er seksjonsleder for radiologisk forskning på Innovasjonssenteret, Oslo universitetssykehus, Rikshospitalet.

– Spørsmålene som ble brukt i denne studien, er dessverre ikke gjengitt, men det fremgår at språkmodellene ble bedt om å generere en tekst med en bestemt vinkling. Studien dreier seg derfor om hvor stor slagside man kan få til i en tekst som er generert av en stor språkmodell.

– At en av språkmodellene avslo forespørselen, er ikke uten videre bra, mener Undseth. Hva om formålet var godt, for eksempel ved å studere hvordan propaganda generert av kunstig intelligens kan se ut? Regulering av hva slags tekst som tillates, er ingen åpenbar fordel og innebærer også ulemper, enten de er tilsiktet eller ei. En forfatter som er kritisk til bruk av kunstig intelligens, ville nok lagt opp studien på en annen måte, sier Undseth.

REFERENCES

1. Menz BD, Kuderer NM, Bacchi S et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ 2024; 384: e078538. [PubMed][CrossRef]

Publisert: 5. august 2024. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.24.0276