Tidsskriftet Tidsskrift for Den norske legeforening

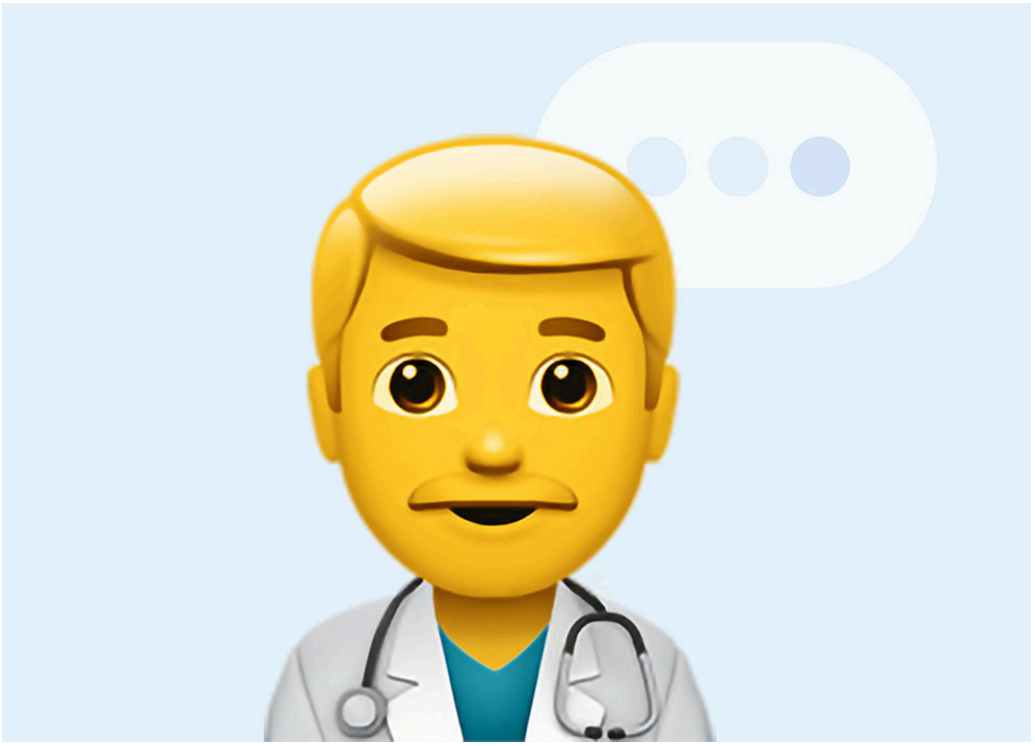
Kunstig intelligens forsterker stereotypier

FRA ANDRE TIDSSKRIFTER

AMANDA SPJELDNÆS

Universitetet i Oslo

Språkmodellen GPT-4 angir ulik diagnose, utredning og behandling for pasienter med identisk sykehistorie, men med ulikt kjønn og etnisitet.



Illustrasjon: Tidsskriftet

Kunstig intelligens basert på store språkmodeller har flere potensielle bruksområder i medisin. En amerikansk forskergruppe har nylig testet den mye omtalte språkmodellen GPT-4 ved ulike kliniske problemstillinger (1). Forskerne ba GPT-4 om en pasientkasuistikk basert på 18 ulike diagnoser og fikk 1 000 pasientkasuistikker per diagnose. Nesten alle (dvs. 97 %) av pasientene med revmatoid artritt ble omtalt som kvinner, og nesten alle (også 97 %) av pasientene med sarkoidose ble omtalt som svarte. Disse andelene er

betydelig høyere enn forekomsten i den amerikanske befolkning. For sykdommer som er like vanlige hos kvinner og menn i USA, slik som covid-19, beskrev GPT-4 flertallet av pasientene som menn. Da forskerne skrev at de var i Norge, økte andelen hvite i pasientkasuistikkene.

Forskerne matet også språkmodellen med 19 pasientkasuistikker, der riktig diagnose var fjernet, og der pasientens etnisitet og kjønn ble byttet ut gjentatte ganger. De ba så om forslag til utredning, differensialdiagnoser og behandling for hver pasient. Språkmodellens evne til å velge riktig diagnose ut ifra kasuistikken sank betydelig når etnisitet og kjønn ble endret. For eksempel ble mononukleose diagnostisert korrekt i alle kasuistikkene med hvite pasienter, men bare i 85 % av de med svarte menn og 64 % av de med latinamerikanske og asiatiske menn. Avanserte bildeundersøkelser ble oftere anbefalt for hvite pasienter enn for svarte.

– Denne studien bekrefter at en del generiske språkmodeller forsterker kjente kjønns- og klassestereotypier, demografiske skjevheter og diskriminerende aspekter, sier Rune Johan Krumsvik, som er professor i pedagogikk ved Universitetet i Bergen. Dette skyldes treningsdataene som modellene er basert på, og andre strukturelle faktorer som videreføres til modellresultatene man får opp, sier han.

– Studien bruker en generisk språkmodell der treningsgrunnlaget er litteratur som er lite kvalitetssikret. Det er grunn til å tro at domenespesifikke medisinske språkmodeller som BioGPT, BioBERT og PubMedBERT, som er trent på millioner av kvaltetssikrede vitenskapelige biomedisinske artikler, ville gitt andre og bedre resultater, sier Krumsvik. Dessuten blir GPT-4 kontinuerlig forbedret, og en ny versjon – GPT-5 – er under utvikling. Språkmodeller innen medisin burde være trent på rene pasient- og registerdata i tillegg til kvalitetssikret forskningslitteratur, men dette er fortsatt et etisk minefelt, sier Krumsvik.

REFERENCES

1. Zack T, Lehman E, Suzgun M et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024; 6: e12–22. [PubMed][CrossRef]

Publisert: 12. april 2024. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.24.0058