

problematiske» svar: 14 prosent – og flest «ikke-problematiske» svar: 60 prosent.

Generelt klarte chatbotene seg dårligst innen kategoriene ernæring, idrettsprestasjoner og stamceller og best innen vaksiner og kreft.

Ingen av botene var i stand til å lage feilfrie kildelister. Gjennomsnittsskåren for hvor fullstendige kildene var, lå på bare 40 prosent.

For å forstå utfordringene, har Groth selv prøvd å bygge digitale prototyper som kunne hente informasjon fra sikre nettsider og gi klare kildehenvisninger.

Men det er utrolig vanskelig å gjøre dette konsistent.

– Det er virkelig vanskelig å få disse språkmodellene til å gjenbruke kilder uten å hallusinere og finne opp svar. Så jeg kan godt forstå hvorfor det blir slik, sier han.

Dette unnskylder likevel ikke selskaper som Google, Meta og OpenAI. Groth peker på at tech-gigantene prøver å få så mange som mulig til å bruke KI-en deres. Et godt eksempel er Google, som har innført KI-svar øverst i søkeresultatene sine.

– Da Google lanserte sine KI-svar i fjor, valgte de jo selv å levere svar av tvilsom kvalitet, også på søk etter helseinformasjon, sier han.

Da Google introduserte KI-svar øverst i søk, falt trafikken markant på offentlige nettsider i Danmark, for eksempel Sundhed.dk, viser en rapport fra Center for Digital Psykiatri.

Fra å ha 1,6 millioner brukere i april 2025, har tallet falt til én million ved utgangen av året.

Samtidig har en norsk studie vist at unge heller vil ha råd fra KI enn fra legen.

Til tross for at vi, ifølge Groth, godt vet at KI kan ta feil.

– Det skyldes i stor grad brukervennligheten. Det er så enkelt for meg å få svar på akkurat det som interesserer meg nå, på det tidspunktet og stedet jeg er interessert i det, sier han.

Derfor er vi også villige til å gå på kompromiss med kvaliteten på svarene.

Slik klarte de fem chatbotene seg

Gemini (Google): Høyest nøyaktighet i studien med 60 prosent ikke-problematiske svar. Var best til å ta med advarsler. Skrev de korteste og mest leselige svarene, men var dårlig til å sitere kilder: 30,2 prosent fullstendighet.

DeepSeek (High-Flyer): I midtsjiktet med 52 prosent ikke-problematiske svar. Sammen med Grok best til å angi kilder, selv om det var mangler med en fullstendighets-score på 62 prosent.

Meta AI (Meta): Gjennomsnittlig resultat med 50 prosent ikke-problematiske svar. Skiller seg ut som den eneste chatboten som blankt nektet å svare på potensielt farlige spørsmål.

ChatGPT (OpenAI): Høy feilrate med 22 prosent svært problematiske svar. Dårligst til å sitere kilder, bare 22 prosent fullstendighet. Ga færrest advarsler til brukeren (56 prosent) og brukte det vanskeligste språket.

Grok (xAI): Studiens dårligste på nøyaktighet med flest kritiske feil (30 prosent svært problematiske svar). Skrev de lengste svarene, og selv om den var blant de beste på kil-

desitering (61,9 prosent), inneholdt kildene ofte feil.

Kanskje må chatbotene bare «lære» litt mer?

Groth er ikke så sikker. Faktisk kan det gå motsatt vei, mener han.

De nyeste modellene er ikke nødvendigvis mer sannferdige enn dem de erstatter.

– Mange av dem hallusinerer faktisk litt oftere enn dem som fantes i 2025. Det ligger i kjernen av teknologien at det er en kreativ sannsynlighetsberegner. Så jeg kan godt forestille meg at det blir verre – ikke bedre, sier han.

Myndighetene bør se på om chatboter som gir helsefaglige råd, skal regnes som medisinsk utstyr, mener han. Da må det stilles samme strenge krav til disse som til annet utstyr innen helse.

– Vi vet at et av de primære bruksområdene for disse chatbotene er helse. Og vi vet at de svarer villig på helsespørsmål. Jeg kan ikke se hvorfor disse produktene ikke er medisinsk utstyr, sier Groth.

Fram til det skjer, trenger du ikke å slette chatbotene fra app-samlingen din, sier han. Men du bør være «svært kritisk» når det gjelder helsen din og søke andre steder enn hos chatbotene.

KILDE:

1. Generative artificial intelligence-driven chatbots and medical misinformation: an accuracy, referencing and readability audit, BMJ-journals (2025), DOI: 10.1136/bmjopen-2025-112695
2. ©Videnskab.dk. Oversatt av Trine Andreassen for forskning.no. Les originalartikkelen på videnskab.dk her.

ET NYTT VERKTØY I BEHANDLINGEN AV PERIODONTITT

Klinisk dokumentert antibakteriell lysbehandling med Lumoral®
Scan QR-koden for dokumentasjon

