Tidsskriftet Tidsskrift for Den norske legeforening

Kunstig intelligens og legers svar på helsespørsmål

KORT RAPPORT

TIRIL EGSET MORK*

Det medisinske fakultet
Universitetet i Bergen
Forfatterbidrag: utvikling av prosjektetideen, teksten til web-applikasjonen og instruksjonene til GPT-4. Innsamling og analyse av data, skriving av første utkast og revisjon av endelig manus. Mork og Mjøs deler førsteforfatterskap.
Tiril Egset Mork er medisinstudent.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

HÅKON GARNES MJØS*

Det medisinske fakultet
Universitetet i Bergen
Forfatterbidrag: utvikling av prosjektetideen, teksten til web-applikasjonen og instruksjonene til GPT-4. Innsamling og analyse av data, skriving av første utkast og revisjon av endelig manus. Mork og Mjøs deler førsteforfatterskap.
Håkon Garnes Mjøs er medisinstudent.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

HARALD GISKEGJERDE NILSEN

Høgskulen på Vestlandet
Forfatterbidrag: utvikling av web-applikasjonen brukt til datainnsamling, all programutvikling i forbindelse med datainnsamlingen og innhenting av spørsmål fra studenterspør.no.
Harald Giskegjerde Nilsen har fullført bachelorprogram i informasjonsteknologi.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

SINDRE KJELSRUD

Høgskulen på Vestlandet

Forfatterbidrag: utvikling av web-applikasjonen brukt til datainnsamling, all programutvikling i forbindelse med datainnsamlingen og innhenting av spørsmål fra studenterspør.no.
Sindre Kjelsrud har fullført bachelorprogram i informasjonsteknologi.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

ALEXANDER SELVIKVÅG LUNDERVOLD

Institutt for datateknologi, elektroteknologi og realfag
Høgskulen på Vestlandet

Forfatterbidrag: veiledning av utviklingen av instruksjonene gitt til GPT-4, generert responser fra GPT-4, tilrettelagt data for analyse og vært veileder for Nilsen og Kjelsrud.
Alexander Selvikvåg Lundervold er professor.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

ARVID LUNDERVOLD**

Institutt for Biomedisin
Universitetet i Bergen

Forfatterbidrag: gjennomføring av statistiske analyser, og utvikling og forankring av dette som et samarbeidsprosjekt med Høgskulen på Vestlandet.
Arvid Lundervold er lege og professor emeritus i medisinsk informasjonsteknologi med mer enn 30 års erfaring innen kunstig intelligens.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

IB JAMMER**

ib.jammer@helse-bergen.no
Kirurgisk serviceklinikk
Haukeland universitetssjukehus

Forfatterbidrag: veiledning, utvikling av prosjektet, tolkning av data, samt utarbeidelse av manuskriptet.
Ib Jammer er ph.d. og anestesilege.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

* Tiril Egset Mork og Håkon Garnes Mjøs har bidratt i like stor grad til denne artikkelen.
** Ib Jammer og Arvid Lundervold har bidratt i like stor grad til denne artikkelen.

Bakgrunn

Flere studier har undersøkt hvordan store språkmodeller besvarer helsespørsmål. I en studie fra 2023 ble svar på engelskspråklige helserelaterte spørsmål generert av språkmodellen GPT-3.5, oppfattet som mer empatiske og kunnskapsrike enn svar fra

leger. Vi ønsket å anvende den nyere språkmodellen GPT-4 på norsk for å undersøke hvordan svar på helserelaterte spørsmål fra leger og svar generert av språkmodellen ble vurdert av respondenter med helsefaglig bakgrunn.

Materiale og metode

Vi benyttet 192 helserelaterte spørsmål med tilhørende svar fra leger, hentet fra nettstedet Studenterspør.no. Språkmodellen GPT-4 ble benyttet til å generere et nytt sett med svar på de samme spørsmålene. Begge settene med svar ble vurdert av 344 respondenter med helsefaglig bakgrunn. Respondentene, som var blindet for hvilket svar som var generert av lege eller språkmodellen, ble bedt om å rangere svarenes grad av empati, kunnskap og hjelpsomhet.

Resultater

Det var 344 respondenter og 192 spørsmål i undersøkelsen. Gjennomsnittlig antall vurderinger per svar var 5,7. Det var signifikant forskjell mellom legesvar og svar fra GPT-4 i oppfatningen av empati ($p < 0,001$), kunnskap ($p < 0,001$) og hjelpsomhet ($p < 0,001$).

Fortolkning

Svarene generert av GPT-4 ble vurdert som mer empatiske, kunnskapsrike og hjelpsomme enn svarene fra leger. Det antyder at kunstig intelligens kan avlaste helsepersonell ved å formulere gode svarutkast på helsespørsmål.

Hovedfunn

Medisinske svar fra en kunstig intelligent språkmodell ble oppfattet som mer empatiske, kunnskapsrike og hjelpsomme sammenlignet med svar fra leger.

Flere studier har undersøkt hvordan kunstig intelligens besvarer helserelaterte spørsmål. *Generative Pre-training Transformer* (GPT) er en kunstig intelligensmodell som kan forstå og generere menneskelig språk. En amerikansk studie publisert i 2023 fant at språkmodellen GPT-3.5 sine svar på engelskspråklige helserelaterte spørsmål ble oppfattet som mer empatiske og kunnskapsrike enn svar fra leger (1). Hvordan svar fra kunstig intelligens oppfattes, kan ha betydelige implikasjoner og stor nytteverdi for helsesektoren.

Siden språk, kultur og medisinske retningslinjer varierer mellom ulike land, ønsket vi å undersøke hvordan personer med helsefaglig bakgrunn i Norge oppfatter svar fra store språkmodeller på helserelaterte spørsmål, sammenlignet med svar fra leger. I tillegg undersøkte vi om svarene ble vurdert ulikt av leger og medisinstudenter med lisens, sammenlignet med personer med annen helsefaglig bakgrunn.

Materiale og metode

192 helserelaterte spørsmål og tilhørende svar fra leger fra nettstedet Studenterspør.no ble inkludert i studien. Studenterspør.no er et nettsted der studenter kan sende inn spørsmål og få svar fra helsepersonell. Svarene publiseres anonymisert. Vi utviklet et script for å samle inn spørsmål og svar fra kategorien «Kropp, sex og identitet» og underkategorien «Sykdom og symptomer». Denne kategorien ble valgt fordi den inneholder varierte helserelaterte spørsmål og har en stor andel spørsmål besvart av lege.

Vi utviklet et sett med instruksjoner for GPT-4 for å sikre at modellens svar hadde ønsket form, lengde, innhold og språk. Vi vektla at svarene fra GPT-4, i likhet med svarene fra Studenterspør.no, ikke skulle oppfattes som helsehjelp, i tråd med helsepersonelloven (2). I stedet skulle de gi helseveiledning og rådgivning, uten å erstatte medisinske råd fra helsepersonell. Instruksjonene ble utviklet iterativt til GPT-4 ga tilfredsstillende svar på et sett testspørsmål. Deretter ble instruksjonssettet låst, og de samme instruksjonene ble brukt på alle spørsmålene i studien. Analyse av resultatene ble utført i Python.

Rekruttering av respondenter skjedde gjennom e-postlister for legevakt, sykehjem og sykehusavdelinger, på stand og postere på Haukeland universitetssjukehus, Facebook-grupper for helsepersonell og direkte kontakt med bekjente innen helsevesenet. Respondentene som oppga å være lege eller medisinstudent med lisens, eller som oppga å jobbe, studere eller ha bakgrunn innen helsevesenet, ble inkludert i studien. Innsamling av data fra de 344 inkluderte respondentene foregikk fra 15. januar til 18. februar 2024.

Spørreundersøkelsen ble distribuert via en egenutviklet webapplikasjon der respondentene kunne lese ett spørsmål med to tilhørende svar om gangen, samt angi sin vurdering for hver av de ulike dimensjonene. I applikasjonen fantes informasjon om personvern og definisjoner av evalueringskriteriene. Deltakerne fikk vite at ett svar var generert av GPT-4, og ett svar var skrevet av lege, men de fikk ikke spesifisert hvilket som var hvilket. Spørsmålene ble tildelt tilfeldig. Respondentene evaluerte dimensjonene empati, kunnskap og hjelpsomhet i svarene ved hjelp av en femdelt Likert-skala. For dimensjonen kunnskap var det i tillegg mulig å svare «vet ikke». Det var mulig å hoppe over spørsmål, og undersøkelsen be avsluttet etter å ha vurdert fem spørsmål, eller tidligere om ønskelig. Det var mulig å gjennomføre undersøkelsen flere ganger.

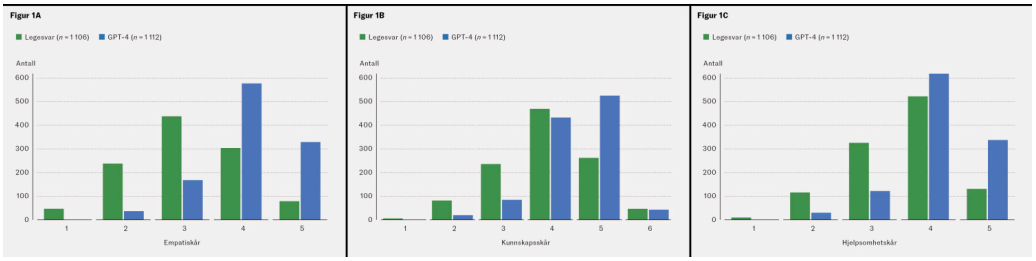
Utfyllende informasjon om inklusjon av spørsmål, generering av svar, definisjoner av evalueringstermene, analyser og resultater, samt komplette instruksjoner og eksempler på spørsmål og svar, er tilgjengelige her: https://github.com/MMIV-ML/helseveileder.

Resultater

Til sammen 344 inkluderte respondenter vurderte de 192 spørsmålene, og det ble avgitt totalt 1 109 vurderinger av sett med spørsmål og svar. Gjennomsnittlig antall vurderinger per svar var 5,7 (standardavvik 6,7), med medianverdi 5. Nitten

respondenter (5,4 %) deltok i studien mer enn én gang. Blant respondentene oppga 44 (12,8 %) å være lege eller medisinstudent med lisens, mens 300 (87,2 %) ikke var lege eller medisinstudent med lisens, men studerte, arbeidet eller hadde bakgrunn innen helsevesenet.

Figur 1 viser respondentenes vurdering av empati, kunnskap og hjelpsomhet. Merk forskyvningen mot høyere skår for GPT-4-svar for alle de tre dimensjonene. Empati: χ^2 = 571,26, df =4, p < 0,001, kunnskap: χ^2 = 204,24, df =4, p < 0,001 og hjelpsomhet: χ^2 = 258,49, df =4, p < 0,001.



Figur 1 Vurdering av svar på 192 helserelaterte spørsmål fra 344 respondenter. Figurene viser svar generert av språkmodellen GPT-4 (blå) og av leger (grønn) for dimensjonene empati (a), kunnskap (b) og hjelpsomhet (c). Høyere skår er bedre.

Diskusjon

Svar generert av GPT-4 på helserelaterte spørsmål, ble vurdert som mer empatiske, kunnskapsrike og hjelpsomme enn svar fra leger. Våre funn indikerer at respondenter som var leger eller medisinstudenter med lisens, ikke ga en annen vurdering av kunnskap sammenlignet med andre respondenter som jobber, studerer eller har bakgrunn innen helsevesenet.

Funnene i vår studie samsvarer med resultatene fra en tidligere publisert studie (1). Nytteverdien av store språkmodeller er også vist i andre studier. For eksempel viser foreløpige, ikke-fagfellevurderte resultater at svar generert av språkmodeller kan gi høyere diagnostisk nøyaktighet og bedre samtalekvalitet (3), eller at språkmodellers svar på anestesiologiske spørsmål er likeverdige med akademiske ressurser (4). Eksemplene demonstrerer at kunstig intelligens kan gi like gode og noen ganger bedre svar enn leger, og at kunstig intelligens således kan være et nyttig hjelpemiddel.

Imidlertid rapporterer andre studier motstridende funn. En studie viste at leger som besvarte elektroniske pasientspørsmål ved hjelp av svarutkast generert av GPT-4, brukte mer tid på å lese og redigere utkastene, og de sparte ikke tid på å ferdigstille svarene (5). Studien viste også at legenes svar ble lengre. Dette understreker viktigheten av videre utforskning av hvordan integrasjon av denne formen for kunstig intelligens faktisk kan forbedre helsehjelp og avlaste helsepersonell.

I motsetning til studien fra 2023 (1) brukte vi GPT-4 fremfor den eldre GPT-3.5, og vi utviklet spesialtilpassede instruksjoner til modellen. I tillegg til empati og kunnskap undersøkte vi i denne studien også hvor hjelpsomme svarene fra leger og GPT-4 ble oppfattet. Alle respondenter i studien var blindet for om svarene de vurderte var skrevet av leger eller generert av språkmodellen. Våre instruksjoner var tilpasset for å gjøre det

vanskelig å identifisere om et svar var generert av kunstig intelligens. I motsetning til tidligere studier var ingen av respondentene i denne studien involvert i utforming eller publikasjon.

En svakhet ved vår studie er muligheten for at svar skrevet av språkmodellen kan gjenkjennes, noe som kan føre til en bekreftelsesbias basert på respondentens holdninger til kunstig intelligens. Vi ba ikke respondentene gjette avsenderen, for å unngå et fokus på dette, men en begrensning er at vi ikke kan vite i hvilken grad de faktisk gjenkjente avsenderen og hvordan dette påvirket resultatene.

Det kan også foreligge en seleksjonsbias dersom de med sterke positive eller negative holdninger er overrepresentert, og de med mer nøytrale holdninger er underrepresentert.

Respondentene oppga selv om de var lege eller medisinstudent med lisens, uten at dette ble kontrollert mot helsepersonellregisteret. Innhenting av flere opplysninger fra respondentene ville gjort det mulig å undersøke betydningen av faktorer som arbeidserfaring og yrke.

Konklusjon

Studien viser at svar på helserelaterte spørsmål generert av språkmodellen GPT-4 ble vurdert som mer empatiske, kunnskapsrike og hjelpsomme enn svar fra leger. Dette indikerer at kunstig intelligens kan avlaste helsepersonell ved å formulere gode svarutkast på helserelaterte spørsmål.

Vi takker Studenterspør.no for samarbeidet og for at vi har fått bruke spørsmål og svar fra deres nettsted.

Artikkelen er fagfellevurdert.

REFERENCES

1. Ayers JW, Poliak A, Dredze M et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023; 183: 589–96. [PubMed][CrossRef]

2. Helse- og omsorgsdepartementet. LOV-1999-07-02-64. Lov om helsepersonell m.v. (helsepersonelloven). https://lovdata.no/dokument/NL/lov/1999-07-02-64 Lest 10.10.2024.

3. Tu T, Palepu A, Schaekermann M et al. Towards conversational diagnostic AI. arXivorg. Preprint 11.1.2024. https://arxiv.org/abs/2401.05654 Lest 13.10.2024.

4. Segal S, Saha AK, Khanna AK. Appropriateness of answers to common preanesthesia patient questions composed by the large language model GPT-4 compared to human authors. Anesthesiology 2024; 140: 333–5. [PubMed][CrossRef]

5. Tai-Seale M, Baxter SL, Vaida F et al. AI-Generated draft replies integrated into health records and physicians' electronic communication. JAMA Netw Open 2024; 7: e246565. [PubMed][CrossRef]

Publisert: 10. februar 2025. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.24.0402
Mottatt 28.7.2024, første revisjon innsendt 16.10.2024, godkjent 14.12.2024.