Tidsskriftet Tidsskrift for Den norske legeforening

AI-generated and doctors' answers to health-related questions

SHORT REPORT

TIRIL EGSET MORK

Faculty of Medicine
University of Bergen
Author contribution: development of the project concept, the text for the web application and the instructions for GPT-4. Collection and analysis of data, preparing the first draft, and revising the final manuscript. Mork and Mjøs share first authorship.
Tiril Egset Mork, medical student.
The author has completed the ICMJE form and declares no conflicts of interest.

HÅKON GARNES MJØS

Faculty of Medicine
University of Bergen
Author contribution: development of the project concept, the text for the web application and the instructions for GPT-4. Collection and analysis of data, preparing the first draft, and revising the final manuscript. Mork and Mjøs share first authorship.
Håkon Garnes Mjøs, medical student.
The author has completed the ICMJE form and declares no conflicts of interest.

HARALD GISKEGJERDE NILSEN

Western Norway University of Applied Sciences
Author contribution: development of the web application used for data collection, all software development for the data collection and retrieval of questions from studenterspør.no.
Harald Giskegjerde Nilsen, BSc in IT.
The author has completed the ICMJE form and declares no conflicts of interest.

SINDRE KJELSRUD

Western Norway University of Applied Sciences
Author contribution: development of the web application used for data collection, all software development for the data collection and retrieval of questions from studenterspør.no.
Sindre Kjelsrud, BSc in IT.
The author has completed the ICMJE form and declares no conflicts of interest.

ALEXANDER SELVIKVÅG LUNDERVOLD

Department of Computer science, Electrical engineering and Mathematical sciences
Western Norway University of Applied Sciences
Author contribution: supervising the development of the instructions for GPT-4, the generated responses from GPT-4, facilitating data for analysis, and acting as supervisor for Nilsen and Kjelsrud.
Alexander Selvikvåg Lundervold, professor.
The author has completed the ICMJE form and declares no conflicts of interest.

ARVID LUNDERVOLD

Department of Biomedicine
University of Bergen
Author contribution: conducting statistical analyses, and development and establishment of this as a collaborative project with the Western Norway University of Applied Sciences.
Arvid Lundervold, doctor and professor emeritus in medical information technology with more than 30 years of experience in AI.
The author has completed the ICMJE form and declares no conflicts of interest.

IB JAMMER

ib.jammer@helse-bergen.no
Department of Anaesthesia and Intensive Care
Haukeland University Hospital
Author contribution: supervision, development of the project, interpretation of data, and preparation of the manuscript.
Ib Jammer, PhD, anaesthetist.
The author has completed the ICMJE form and declares no conflicts of interest.

* Tiril Egset Mork and Håkon Garnes Mjøs have contributed equally to this article.

** Ib Jammer and Arvid Lundervold have contributed equally to this article.

Background

Several studies have investigated how large language models answer health-related questions. In a study from 2023, responses to health-related questions in English generated by the language model GPT-3.5 were perceived as more empathetic and informative than responses from doctors. We wanted to apply the newer language model GPT-4 in Norwegian to investigate how respondents with a healthcare background rated responses to health-related questions from doctors and those generated by the language model.

Material and method

A total of 192 health-related questions with corresponding answers from doctors were sourced from the website Studenterspør.no. The language model GPT-4 was used to generate a new set of answers to the same questions. Both sets of answers were evaluated by 344 respondents with a background in health care. The respondents, who were blinded to whether the answer was generated by a doctor or the language model, were asked to rate the empathy, quality of information and helpfulness of the answers.

Results

The survey consisted of 344 respondents and 192 questions. The average number of evaluations per answer was 5.7. There was a significant difference between doctors' answers and those generated by GPT-4 in terms of perceived empathy ($p < 0.001$), quality of information ($p < 0.001$) and helpfulness ($p < 0.001$).

Interpretation

The answers generated by GPT-4 were rated as more empathetic, informative and helpful than the answers from doctors. This suggests that AI could serve as an aid to healthcare personnel by drafting good responses to health-related questions.

Main findings

Medical responses from an AI language model were perceived as more empathetic, informative and helpful compared to responses from doctors.

Several studies have examined how artificial intelligence (AI) responds to health-related questions. *Generative Pre-training Transformer* (GPT) is an AI model that can understand and generate human language. A US study published in 2023 found that answers generated by the GPT-3.5 language model to health-related questions in English were perceived as more empathetic and informative than answers from doctors (1). How AI responses are perceived could have significant implications and great potential value for the healthcare sector.

Since language, culture and medical guidelines vary between countries, we wanted to investigate how those with a background in health care in Norway perceive responses from large language models to health-related questions, compared to answers from doctors. We also examined whether the responses were evaluated differently by doctors and licensed medical students, compared to those with other backgrounds in health care.

Material and method

A total of 192 health-related questions and corresponding answers from doctors, sourced from the website Studenterspør.no, were included in the study. Studenterspør.no is a platform where students can submit questions and receive answers from healthcare personnel. The responses are published anonymously. We developed a script to retrieve questions and answers from the category 'Body, sex and identity' and the subcategory 'Illness and symptoms'. This category was chosen because it contains a wide range of health-related questions and a high proportion of answers provided by doctors.

We developed a set of instructions for GPT-4 to ensure that the model's responses adhered to the desired format, length, content and language. It was emphasised that the responses from GPT-4, like those from Studenterspør.no, should not be regarded as medical assistance, in accordance with the Health Personnel Act (2). Instead, they were to serve as health-related guidance and advice, as opposed to replacing medical advice from healthcare personnel. The instructions were developed iteratively until GPT-4 produced satisfactory answers to a set of test questions. The instruction set was then locked, and the same instructions were applied to all the questions in the study. The results were analysed using Python.

Respondents were recruited from email lists for emergency care, nursing homes and hospital departments, at stands and via posters at Haukeland University Hospital, Facebook groups for healthcare personnel and directly through contacts within the health service. Doctors, licensed medical students, and those working, studying or with a background in health care were included in the study. Data were collected from the participating 344 respondents in the period 15 January to 18 February 2024.

The survey was distributed via a customised web application where respondents could read one question with two corresponding answers at a time, and provide their rating for each of the different dimensions. The application included information about data protection and definitions of the evaluation criteria. Participants were informed that one answer was generated by GPT-4 and one was written by a doctor, but they were not told which was which. The questions were assigned randomly. Respondents evaluated the empathy, quality of information and helpfulness of the answers based on a five-point

Likert scale. For quality of information, it was also possible to answer 'Don't know'. Respondents could skip questions, and the survey could be completed after evaluating five questions, or earlier if desired. The survey could be taken multiple times.

Detailed information about the inclusion of questions, answer generation, definitions of the evaluation terms, analysis and results, as well as complete instructions and examples of questions and answers, is available here: https://github.com/MMIV-ML/helseveileder.

Results

A total of 344 respondents evaluated the 192 questions, providing a total of 1109 ratings of question-answer sets. The average number of ratings per answer was 5.7 (standard deviation 6.7), with a median of 5. Nineteen respondents (5.4 %) participated in the study more than once. Among the respondents, 44 (12.8 %) were doctors or licensed medical students, while 300 (87.2 %) were not doctors or licensed medical students but were studying, working or had a background in health care.

Figure 1 shows respondents' ratings of empathy, quality of information and helpfulness. Note the shift toward higher scores for GPT-4 responses across all three dimensions. Empathy: $\chi^2 = 571.26$, df =4, $p < 0.001$, quality of information: $\chi^2 = 204.24$, df =4, $p < 0.001$ and helpfulness: $\chi^2 = 258.49$, df =4, $p < 0.001$.

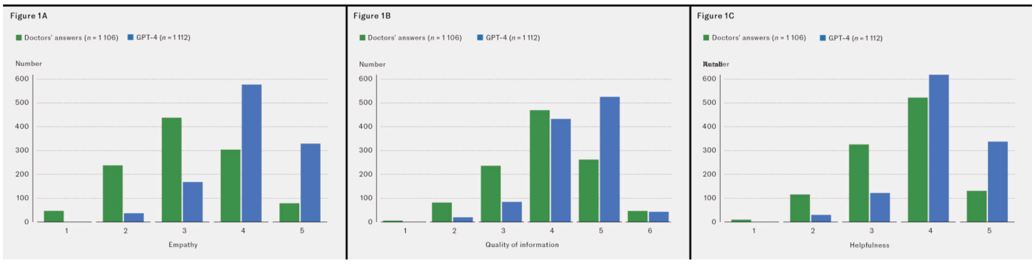

Figure 1 Evaluation of responses to 192 health-related questions by 344 respondents. The figures show answers generated by the GPT-4 language model (blue) and by doctors (green) for the dimensions of empathy (a), quality of information (b) and helpfulness (c). Higher scores indicate more positive ratings.

Discussion

GPT-4-generated responses to health-related questions were considered more empathetic, informative and helpful than those from doctors. Our findings indicate that the doctors and licensed medical students did not assess quality of information differently to other respondents who work, study or have a background in health care.

The findings of our study align with the results of a previously published study (1), and the value of large language models has also been demonstrated in other studies. For example, preliminary, non-peer-reviewed findings suggest that responses generated by language models may offer improved diagnostic accuracy and conversational quality (3), or that language model responses to questions about anaesthesia care are of the same quality as the content in academic sources (4). These examples demonstrate that AI can provide answers that are as good as, and sometimes better than, those of doctors, indicating that AI can be a valuable tool.

However, other studies report conflicting findings. One study found that doctors who answered electronic patient enquiries using GPT-4-generated draft replies spent more time reading and editing the drafts and did not save any time completing their responses (5). The study also showed that the doctors' responses became longer. This highlights the importance of further investigation into how integrating this form of AI can actually improve health care and aid healthcare personnel.

Unlike the 2023 study (1), we used GPT-4 instead of the older GPT-3.5 and developed customised instructions for the model. In addition to empathy and quality of information, this study also examined perceptions of the helpfulness of doctors' and GPT-4s' answers. All respondents were blinded to whether the responses they evaluated were written by doctors or generated by the language model. Our instructions were designed to make it difficult to identify whether a response was AI-generated. Unlike previous studies, none of the respondents in this study were involved in its design or publication.

One limitation of our study is that respondents may have recognised the language model's responses, which could introduce confirmation bias based on their attitudes to AI. We chose not to ask respondents to identify the source of the responses to avoid drawing attention to this. However, a limitation of our study is that we cannot assess the extent to which they recognised the source or how this may have affected the results.

There may also be selection bias if individuals with strong positive or negative views are overrepresented while those with more neutral attitudes are underrepresented.

Respondents self-reported whether they were doctors or licensed medical students, without verification against the health personnel registry. Collecting additional information from respondents would have allowed us to examine the impact of factors such as work experience and occupation.

Conclusion

The study shows that responses to health-related questions generated by the language model GPT-4 were rated as more empathetic, informative and helpful than those from doctors. This suggests that AI could serve as an aid for healthcare personnel by generating high-quality draft responses to health-related questions.

We would like to thank Studenterspør.no for their cooperation and for allowing us to use questions and answers from their website.

The article has been peer-reviewed.

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 Accessed 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 Accessed 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. doi: 10.1001/jamanetworkopen.2024.6565. [PubMed][CrossRef]

Publisert: 10 February 2025. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.24.0402
Received 28.7.2024, first revision submitted 16.10.2024, accepted 14.12.2024.