

Applied Clinical Informatics

A Survey of Clinicians' Views of the Utility of Large Language Models

Matthew Spotnitz, Betina Idnay, Emily R Gordon, Rebecca Shyu, Gongbo Zhang, Cong Liu, James J Cimino, Chunhua Weng.

Affiliations below.

DOI: 10.1055/a-2281-7092

Please cite this article as: Spotnitz M, Idnay B, Gordon E R et al. A Survey of Clinicians' Views of the Utility of Large Language Models. ACI 2024. doi: 10.1055/a-2281-7092

Conflict of Interest: The authors declare that they have no conflict of interest.

This study was supported by National Center for Advancing Translational Sciences (<http://dx.doi.org/10.13039/100006108>), UL1TR001873, U.S. National Library of Medicine (<http://dx.doi.org/10.13039/100000092>), R01LM009886, R01LM014344, T15LM007079, National Human Genome Research Institute (<http://dx.doi.org/10.13039/100000051>), R01HG012655

Abstract:

Objective: Large language models (LLMs) like ChatGPT are powerful algorithms that have been shown to produce human-like text from input data. A number of potential clinical applications of this technology have been proposed and evaluated by biomedical informatics experts. However, few have surveyed healthcare providers for their opinions about whether the technology is fit for use.

Materials and Methods: We distributed a validated mixed-methods survey to gauge practicing clinicians' comfort with LLMs for a breadth of tasks in clinical practice, research and education, which were selected from the literature.

Results: A total of 30 clinicians fully completed the survey. Of the 23 tasks, 16 were rated positively by more than 50% of the respondents. Based on our qualitative analysis, healthcare providers considered LLMs to have excellent synthesis skills and efficiency. However, our respondents had concerns that LLMs could generate false information and propagate training data bias.

Discussion: Our survey respondents were most comfortable with scenarios that allow LLMs to function in an assistive role, like a physician extender or trainee.

Conclusion: In a uniquely rigorous and comprehensive mixed-methods survey of clinicians about LLM use, healthcare providers were encouraging of having LLMs in healthcare for many tasks, and especially in assistive roles. There is a need for continued human-centered development of both LLMs and artificial intelligence (AI) in general.

Corresponding Author:

Dr. Matthew Spotnitz, Columbia University Irving Medical Center, Biomedical Informatics, 630 W 168th Street, 10032-3784 New York, United States, mes698@mail.harvard.edu

Affiliations:

Matthew Spotnitz, Columbia University Irving Medical Center, Biomedical Informatics, New York, United States
Betina Idnay, Columbia University Irving Medical Center, Biomedical Informatics, New York, United States
Emily R Gordon, Columbia University Irving Medical Center, Biomedical Informatics, New York, United States

[...]

Chunhua Weng, Columbia University, Department of Biomedical Informatics, New York, United States



This article is protected by copyright. All rights reserved.

Accepted Manuscript

A Survey of Clinicians' Views of the Utility of Large Language Models

Matthew Spotnitz, MD, MPH¹, Betina Idnay, PhD, RN¹, Emily R. Gordon, BA^{1,2}, Rebecca Shyu, BS¹, Gongbo Zhang, PhD¹, Cong Liu, PhD¹, James J. Cimino, MD^{1,3}, Chunhua Weng, PhD¹

1. Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY
2. Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, NY
3. Informatics Institute, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL

Corresponding authors:

Matthew Spotnitz, MD, MPH, 630 West 168th Street New York, NY 10032, United States (e-mail: mes698@mail.harvard.edu).

Chunhua Weng, PhD, 630 West 168th Street New York, NY 10032, United States (e-mail: cw2384@cumc.columbia.edu).

Abstract

Objective Large language models (LLMs) like ChatGPT are powerful algorithms that have been shown to produce human-like text from input data. A number of potential

clinical applications of this technology have been proposed and evaluated by biomedical informatics experts. However, few have surveyed healthcare providers for their opinions about whether the technology is fit for use.

Materials and Methods We distributed a validated mixed-methods survey to gauge practicing clinicians' comfort with LLMs for a breadth of tasks in clinical practice, research and education, which were selected from the literature.

Results A total of 30 clinicians fully completed the survey. Of the 23 tasks, 16 were rated positively by more than 50% of the respondents. Based on our qualitative analysis, healthcare providers considered LLMs to have excellent synthesis skills and efficiency. However, our respondents had concerns that LLMs could generate false information and propagate training data bias.

Discussion: Our survey respondents were most comfortable with scenarios that allow LLMs to function in an assistive role, like a physician extender or trainee.

Conclusion: In a mixed-methods survey of clinicians about LLM use, healthcare providers were encouraging of having LLMs in healthcare for many tasks, and especially in assistive roles. There is a need for continued human-centered development of both LLMs and artificial intelligence (AI) in general.

Keywords

Artificial Intelligence

Clinical decision support

Implementation and deployment

Clinical informatics

Clinical information systems

Introduction:

Large language models (LLMs), which are a type of artificial intelligence (AI), are designed to process and understand human language. They are usually trained on massive amounts of text. For example, ChatGPT is a very efficient LLM that has garnered a great deal of public attention for its capabilities since its recent introduction in late 2022¹⁻³. The possible healthcare applications of LLMs are numerous. Representative examples include generating clinical documentation, personalized educational materials, and original scientific manuscripts^{4,5}.

One well-known limitation of ChatGPT is its tendency for “hallucination,” the generation of text that is perceived as convincing but is not accurate⁶. A second limitation is that ChatGPT can propagate bias that is intrinsic in the training data. These issues have raised concerns about the safety of LLM use in healthcare. Specifically, some researchers envision scenarios where ChatGPT could provide clinical care advice that is outdated, inaccurate or incomplete⁷⁻¹⁰.

Determining the best uses of LLMs in healthcare has been the focus of recent studies. In a prior publication, clinicians with informatics expertise evaluated LLMs for clinical decision support and concluded that they may provide valuable assistance (Liu

et. al., *JAMIA* 2023)¹¹. However, that study did not address the experience of novice LLM users. Furthermore, few studies have investigated healthcare provider comfort with LLMs or used both quantitative and qualitative methods. Those studies either asked general questions about the suitability of LLMs in different healthcare domains^{12,13} or compared it to human performance on one healthcare delivery service¹⁴.

Presently, applications of LLMs are being developed at a rapid pace and could have widespread adoption within healthcare by novice and expert users alike. The most ethical and effective implementation of the technology must consider user requirements and concerns from representative stakeholders of this technology in the clinical setting. In this study, we surveyed diverse practicing clinicians about using LLMs for tasks in clinical practice, research and education and summarized their perceptions of the potential and limitations of LLMs, in order to inform the development of clinically meaningful evaluation standards for LLMs to ensure their appropriate and ethical implementation in clinical settings.

Methods:

Study Design and Sampling

The survey instrument, which is shown in Supplementary Appendix 1, was developed by two authors with both clinical and informatics experience (M.S., B.I.) and refined based on feedback from a third author (E.R.G.). It was implemented through *Qualtrics* (Qualtrics, Provo, UT), took about 15 minutes to complete, and gauged clinicians' perceptions on the appropriateness of using LLMs in clinical practice, research and education.

The opening questions quantified participants' experience in clinical medicine and informatics with multiple-choice answers. Next, there were questions that asked if the amount of LLM experience in healthcare within the past year had exceeded 50 hours. Then, there were questions that prompted rating the appropriateness of LLM use for 23 different tasks in clinical practice, research and education on a 5-point Likert scale (i.e., highly appropriate to highly inappropriate). Those tasks represented a sample of proposed LLM uses that were synthesized from the literature and included, but were not limited to, optimizing alerts for clinical decision support, providing a differential diagnosis, writing a discharge summary, recommending treatment options, translating radiology reports into layperson language, writing scientific manuscripts, and generating personalized study plans for students or trainees among others^{2,6,7,8,15-25}.

We distributed the proposed clinical practice tasks over two questions, and had one question for research tasks and one question for education tasks. The purpose of this section was to measure the appropriateness of these tasks by category and to determine if any individual tasks were negative or positive outliers. We hypothesized that perceptions about the strengths, limitations and ethical concerns about LLMs could contribute to the ratings. Therefore, we had open-ended questions about each of those as well as an open-ended question about other possible uses of LLMs.

Data Collection

We recruited participants with an email invitation that was sent to a listserv of clinicians at Columbia University Irving Medical Center and by word of mouth. To be eligible, participants needed to be practicing clinicians affiliated with Columbia

University within the past 12 months and were able to comprehend and communicate fluently in English. Respondents were compensated with a \$20 Amazon Gift Card for completing the survey.

Data Analysis

We calculated descriptive statistics on the participants and tabulated their ratings for each question. Two independent reviewers performed an inductive thematic analysis with the narrative comments. Both of them performed independent coding of free text using NVivo (Version 14) with generation of themes. They met regularly for a total of 3 times and developed themes iteratively. Once a consensus was reached, the reviewers determined a final list of themes and applied them to the narrative comments. A third reviewer was available to resolve any discrepancies.

Results:

We recruited practicing clinicians from internal medicine, otolaryngology, ophthalmology, pediatrics, urology, anesthesiology, neurosurgery and general surgery. We distributed a prescreening survey to 350 clinicians, among whom 108 responded, and 30 were eligible and enrolled. All completed the survey. Their demographics are shown in Table 1.

Survey Ratings

Heat maps of the ratings for clinical, research, and education tasks are shown in Tables 2a-c. Of the 23 tasks, 16 (69.6%) had positive ratings by at least half of the participants. The highest rated tasks were “assist with vaccine development by predicting the antigenicity of different proteins from genomic data” (25 positive ratings from 30 participants), “model the spread and transmission of an infectious disease” (25 positive ratings), “generate case studies for training purposes” (24 positive ratings), “monitor data for an emerging disease cluster” (24 positive ratings) and “generate alerts to improve compliance with clinical guidelines” (24 positive ratings). In contrast, 7 out of 23 tasks had positive ratings by fewer than half of the participants. Two of the tasks with the lowest number of positive ratings also had the highest number of negative ratings, which were “Respond to patient questions about a radiology report” (7 positive ratings, 16 negative ratings), and “Write an original scientific manuscript” (5 positive ratings, 20 negative ratings).

Thematic Analysis

We received 20 open-ended responses about LLM strengths, limitations, and ethical concerns, respectively. The responses about the limitations and ethical concerns of LLMs were very similar; therefore, we combined them for a total of 40 responses. There were 19 responses about additional uses of LLMs. The themes and corresponding examples are shown in Table 3. Some respondent answers addressed multiple themes and were mapped to each of them. The full responses to our open-ended questions are shown in Supplementary Appendix 2.

Discussion:

Large language models (LLMs) promise to transform healthcare. A human-centered approach is critical to ensure ethical and effective implementation of this powerful technology in clinical settings. This was the first study of clinical practitioners, who were mostly novice LLM users and from diverse healthcare domains, to rate tasks that may be improved by LLMs.

The fundamental theorem of biomedical informatics is user augmentation so that “a person working in partnership with an information resource is better than that same person unassisted²⁶.” Similarly, the clinicians who we studied were encouraging of having LLMs as their assistants. The tasks that leveraged LLMs for supportive roles were rated the highest. In the qualitative analysis, emerging themes were that LLMs were highly skilled at different tasks; however, there were ethical concerns about using the technology. Supportive LLM roles may have been more popular, because in those scenarios clinicians could correct for false information that the algorithms might generate.

Therefore, we expect that clinicians would prefer to have LLMs function more like trainees or physician extenders than attending physicians. LLMs could assist clinicians by drafting notes and reports, making suggestions for patient triage, extracting patient information from charts, and identifying discrepancies from standard patient care. Since LLMs are very skilled at processing large amounts of data, they could help monitor patients in critical care and perioperative settings. Also, they could help translate

medical information between languages, or from technical jargon into layperson language. The contributions of LLMs to these tasks could be reviewed by a clinician.

However, the notion of allowing LLMs to function without supervision in clinical practice raises ethical concerns. They have a propensity to produce false information and propagate data bias, which could lead to incorrect medical decisions. Furthermore, LLMs lack human empathy, which could be a source of mistrust with patients. Instead, patients are more likely to trust medical advice from a clinician because of the human connection. Overall, we believe that clinicians would prefer to have LLMs assist them instead of replace their practice.

Our study participants were encouraging of LLM assistance in the research and education domains as well. In research, the processing power of LLMs would allow them to help with a range of statistical analyses. Also, their linguistic capabilities could translate ideas across human and programming languages. Those skills could be especially useful in large research networks, which consist of individuals from different countries and who have different programming skills. However, having LLMs author an original manuscript instead of a researcher would raise similar ethical concerns to allowing LLMs to function as an autonomous clinician. The education tasks raised the fewest ethical concerns, perhaps because students have regular supervision and a smaller role in direct patient care than clinicians or researchers.

Our sampling method followed a defined set of recruitment criteria and enrolled a total of 30 practicing clinicians who completed the survey for this study. While a larger number of respondents would have been desirable, ours covered a variety of clinical domains and provided valuable, original insights regarding the ethical and reliable uses

of LLMs in clinical settings. Given the unusually rapid evolution of LLMs technology, this early study is timely and makes meaningful contributions by including the voices of key stakeholders of implementing LLMs for clinical tasks.

A limitation of our study, and a potential source of sampling bias, is that only a relatively small number of participants from a single medical center were recruited by convenience sampling. Also, we used self-reported data as key elements of our analysis. These data may introduce biases due to varying accuracy in self reports and varying awareness of the problems by reporting individuals. Despite these limitations, we have developed an instrument that is capable of discerning different opinions about LLM use. We hope our findings can be taken into consideration by developers as the field continues its rapid evolution. As further progress is made, and clinicians have more significant experience with this technology, subsequent studies can build on our methods and experience to study larger sample sizes at multiple institutions to gain additional insights for future directions.

Future studies with a larger and more diverse sample will be warranted to ensure the generalizability of the results and allow for stratification by variables that could affect perceptions of LLM use, such as age, duration of clinical training, provider specialty, and experience with the technology. Those perceptions could be tracked longitudinally to gauge how they change over time. A more robust study about participants' general knowledge of LLMs and AI would strengthen future studies. Specifically, gauging to what extent participants understand how AI algorithm are able to work, predict, learn, and generate responses, would be valuable part of an analysis. Furthermore,

comparing the perceptions regarding different LLMs, and how LLM generated errors compare to human errors, may provide a more balanced view of the technology.

Our study found that healthcare providers would prefer to have LLMs assist than replace them. That finding has implications for future development and implementation of LLMs in clinical practice, research, and education. Studying active clinicians with novice LLM experience helped identify that preference. Therefore, for optimal development and implementation of LLMs in healthcare, continued human centered development is critical.

Conclusion:

Clinicians are generally supportive of the use of large language models (LLMs) for many tasks in clinical practice, research and education, especially where LLMs play a supportive role to humans. Continued human centered development of the technology is critical.

Clinical Relevance Statement:

We studied healthcare providers about the best uses of large language models (LLMs) in healthcare. The clinicians who we studied were encouraging of having LLMs assist them for a range of tasks. The results of our work have implications for implementation of LLMs in healthcare.

Multiple Choice Questions

Which of the following are ethical concerns about Large Language Model (LLM) use?

- A. Efficiency
- B. Confabulation or Hallucination
- C. Ability to Synthesize Information
- D. Capacity to Make Technical Language Accessible.

Answer B). Confabulation or hallucination can cause the LLMs to generate false information, which can lead to incorrect medical decisions. The other answer choices are advantages of the technology.

What is the fundamental theorem of biomedical informatics?

- A. An Information Resource is Better Without Assistance from a Person
- B. An Information Resource Working in Partnership with a Person is Better than an Information Resource Unassisted
- C. A Person Working in Partnership with an Information Resource is Better than that Same Person Unassisted
- D. A Person is Better Without an Information Resource.

Answer C). The fundamental theorem of biomedical informatics states that people are more effective when partnered with an information resource. The alternatives, which are to have no cooperation with information resources and people, or to have people assist information resources, are less effective.

Protection of Human Subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human

Subjects, and was reviewed by Columbia University Irving Medical Center Institutional Review Board (AAAU7954).

Funding/Acknowledgements

This work was supported by National Library of Medicine (NLM) grants R01LM014344 and R01LM009886 to Dr. Weng, National Human Genome Institute (NHGRI) grant R01HG012655 to Dr. Liu, and by National Center for Advancing Clinical and Translational Science (NCATS) grant UL1TR001873 to Columbia University Irving Medical Center. Dr. Idnay and Rebecca Shyu acknowledge support from NLM grant T15LM007079.

Conflict of Interest

All authors have no relevant disclosures.

References

1. Sezgin E, Sirrianni J and Linwood SL. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. JMIR Med Inform. 2022 Feb; 10(2): e32875
2. Elkassem AA and Smith AD. Potential Use Cases for ChatGPT in Radiology Reporting. AJR 2023; 221(3):373-376
3. Rao A, Pang M, Kim J et. al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow. JMIR 2023 25:e58659

4. Cascella, M., J. Montomoli, and V. Bellini. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *Journal of Medical Systems*: 47(33) (2023)
5. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *HealthcarePapers*; 11(6): 867 (2023).
6. Athaluri S, Manthena S, Kesapragada V, et al. Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus* 15(4): e37432
7. Peng Y, Rousseau JF, Shortliffe EH et. al. AI-generated text may have a role in evidence-based medicine. *Nature Medicine* 29: 1593–1594 (2023)
8. Tang L, Sun Z, Idnay B et. al. Evaluating large language models on medical evidence Summarization. *npj Digital Medicine* 6:158 (2023)
9. Deik A. Potential Benefits and Perils of Incorporating ChatGPT to the Movement Disorders Clinic. *J Mov Disord* 2023;16(2):158-162
10. Shashavar Y and Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Hum Factors* 2023 e47564
11. Liu S, Wright AP, Patterson BL et. al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *JAMIA* 2023, 30(7), 1237–1245
12. Hosseini M, Gao CA, Leibovitz DM et. al. An exploratory survey about using ChatGPT in education, healthcare, and research. *medRxiv* 2023.03.31.23287979
13. Choudhury A, Shamszare H. Investigating the Impact of User Trust on the Adoption

and Use of ChatGPT: Survey Analysis. JMIR 2023;25:e47184

14. Dash D, Rahul T, Banda JM et. al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. Arxiv:2304.13714

15. Hirose T, Harada Y, Yokose M, et. al. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int. J. Environ. Res. Public Health 2023, 20, 3378

16. Cheng K, Li Z, He Y et. al. Potential Use of Artificial Intelligence in Infectious Disease: Take ChatGPT as an Example. Annals of Biomedical Engineering 2023; 51:1130–1135

17. Patel SB, Lam K. ChatGPT: The Future of Discharge Summaries? The Lancet Digital Health 2023; 5(3):E107-E108

18. Galido PV, Butala S, Chakerian M, et al. A Case Study Demonstrating Applications of ChatGPT in the Clinical Management of Treatment-Resistant Schizophrenia. Cureus 2023; 15(4): e38166

19. Sharma S, Pajai S, Prasad R, et al. A Critical Review of ChatGPT as a Potential Substitute for Diabetes Educators. Cureus 2023; 15(5): e38380

20. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can Chat-GPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health 2023;13:01003

21. Májovský M, Černý M, Kasal M, et. al. Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. JMIR 2023; e46924

22. Abd-alrazaq A, AlSaad R, Alhuwail D et. al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ 2023;9:e48291
23. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ 2023;9:e46885
24. Karaback M, Ozkara BB, Margetis K et. al. The Advent of Generative Language Models in Medical Education. JMIR Med Educ 2023;9:e48163
25. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA 2023;330(1):78-80
26. Friedman CP. A “Fundamental Theorem” of Biomedical Informatics. JAMIA 2009; 16(2):169-170

Table 1 Participant Information. N, Number; %, Percent

Survey Characteristic	N (%)
<i>Clinical Training</i>	
1-2 years	12 (40)
3-5 years	3 (10)
More than 5 years	15 (50)
<i>Informatics Training</i>	
None	26 (86.7)
1-2 years	3 (10)
3+ years	1 (3.3)

<i>LLM use within the past 12 months</i>	
Less than 50 hours	28 (93.3)
50+ hours	2 (6.7)

Table 2a : Heatmap of Ratings for Large Language Model (LLM) Uses in Clinical Practice Tasks. Orange = Lowest; Yellow =Highest.

Task	Highly Inappropriate	Inappropriate	Neutral	Appropriate	Highly Appropriate
Generate alerts to improve compliance with clinical guidelines	1	2	3	19	5
Provide a differential diagnosis	0	2	9	16	3
Describe how to perform a procedure	0	7	6	16	1
Translate radiology reports into layperson language	2	2	5	15	6
Synthesize and present patient data from the electronic health record for clinical decision support	2	5	4	15	4
Write discharge summaries	2	6	2	13	7
Suggest patient management or treatment options	2	5	8	13	2
Check for inaccuracies in a radiology report, and notify providers of them	2	1	10	11	6
Report current information on a topic for clinical decision support	3	1	14	10	2
Write radiology reports	3	8	12	6	1
Respond to patient questions	2	14	7	6	1

about a radiology report					
--------------------------	--	--	--	--	--

Table 2b: Heatmap of Ratings for Large Language Model (LLM) Uses in Research

Tasks. Orange = Lowest; Yellow = Highest.

Task	Highly Inappropriate	Inappropriate	Neutral	Appropriate	Highly Appropriate
Suggest how to interpret a data set	1	2	7	18	2
Assist with vaccine development by predicting the antigenicity of different proteins from genomic data	1	0	4	16	9
Model the spread and transmission of an infectious disease	1	1	3	16	9
Monitor data for an emerging disease cluster	1	2	3	15	9
Generate programming code	0	0	10	14	6
Write a literature review for a research publication	5	8	8	8	1
Write an original scientific manuscript	10	10	5	3	2

Table 2c: Heatmap of Ratings for Large Language Model (LLM) Uses in Education

Tasks. Orange = Lowest; Yellow = Highest.

Task	Highly Inappropriate	Inappropriate	Neutral	Appropriate	Highly Appropriate
Generate interactive simulations for training purposes	0	1	6	17	6
Generate case studies for training purposes	0	1	5	15	9
Write quizzes and self-	0	2	6	15	7

assessments for students or trainees					
Generate a personalized study plan for a medical student	0	1	8	15	6
Write personalized patient education texts for students or trainees	3	3	12	8	4

Table 3 Summary of Narrative Comments about Perceived Advantages, Ethical Concerns, and Clinical Applications of Large Language Models (LLMs) with Representative Examples

Advantages (n=20)	Ethical Concerns (n=40)	Recommended Clinical Applications for Using LLMs (n=19)
<p><u>Aptitude for specific tasks (n=10)</u></p> <p><i>Ability to generate first drafts with low effort</i></p> <p><i>It can also help students and providers come up with differential diagnosis</i></p>	<p><u>False information (n=15)</u></p> <p><i>Hallucination, fabrication, reinforcement of assumptions and biases</i></p> <p><i>With the confabulation/hallucination issue, does not allow for the uncertainty that is almost</i></p>	<p><u>Drafting documentation (n=8)</u></p> <p><i>Note templates/drafts, especially for routine and predictable things like procedures</i></p>

<p><i>Theoretically could reduce paperwork / administrative work</i></p> <p><i>Ability to write code for novice programmers</i></p>	<p><i>always present in medicine</i></p> <p><i>Its propensity to make up information</i></p>	
<p><u>Synthesis ability (n=9)</u></p> <p><i>Synthesize large amounts of data quickly</i></p> <p><i>Good at synthesizing information in a clear concise fashion</i></p>	<p><u>Worsens patient care (n=14)</u></p> <p><i>This technology if unchecked at a patient care level may have serious implications of harm to patients</i></p> <p><i>Major concern about inappropriate use by lay public to self-diagnose</i></p> <p><i>Also worry about who gets care from healthcare workers vs from direct-to-patient LLM which could be less personalized, and initially less</i></p>	<p><u>Decision support (n=5)</u></p> <p><i>Anything providing recommendations to patients or providers</i></p> <p><i>Flagging concerning trends (VS, lab values) earlier; providing guidance in managing chronic conditions</i></p>

	<i>validated and trustworthy</i>	
<u>Efficiency (n=8)</u> <i>Saves time and improves efficiency</i>	<u>Data bias (n=12)</u> <i>Results are only as good as the data sets that are fed into the LLM</i> <i>Given the fast pace of evidence in healthcare, can be trained on old evidence</i> <i>Poor data quality leads to poor answers</i> <i>Replicates existing biases</i>	<u>Patient communication (n=4)</u> <i>Drafting replies to patient messages in the outpatient inbox modeled off of the provider's communication style</i>
<u>Accuracy (n=5)</u> <i>Fairly accurate and provide higher quality, more-personalized information than most patient-facing information available on the</i>	<u>Human oversight critical (n=7)</u> <i>They should not replace informed decision making for patients or clinical decision making for doctors completely</i>	

<i>internet</i>		
<u>Accessibility (n=4)</u> <i>Translating medical documents into plain English</i>	<u>Impersonal (n=6)</u> <i>Worry about who gets care from healthcare workers vs from direct-to-patient LLM which could be less personalized</i>	
	<u>Legal concerns (n=5)</u> <i>Grey area of ethical/legal limitations</i>	
	<u>Privacy (n=3)</u> <i>Worries of patient confidentiality</i>	
	<u>Worsens clinicians (n=2)</u> <i>If we become reliant on LLMs, we may lose opportunities to practice interpreting/synthesizing data</i>	

	<i>ourselves</i>	
--	------------------	--

