

Original Paper

# Human-Machine Agreement in Medical Ethics: Patient Autonomy Case-Based Evaluation of Large Language Models

Vamshi Mugu<sup>1</sup>, MS, MD; Brendan Carr<sup>2</sup>, MD, MBA; Ashish Khandelwal<sup>1</sup>, MBBS, MD; Mike Olson<sup>1</sup>, MD; John Schupbach<sup>2</sup>, MD, MBA; John Zietlow<sup>3</sup>, MD; T N Diem Vu<sup>3</sup>, MD; Alex Chan<sup>1</sup>, DO; Christopher Collura<sup>4</sup>, MD; John Schmitz<sup>1</sup>, MD

<sup>1</sup>Department of Radiology, Mayo Clinic, Rochester, MN, United States

<sup>2</sup>Department of Emergency Medicine, Mayo Clinic, Rochester, MN, United States

<sup>3</sup>Department of Trauma, Critical Care, and General Surgery, Mayo Clinic, Rochester, MN, United States

<sup>4</sup>Department of Neonatal Medicine, Mayo Clinic, Rochester, MN, United States

**Corresponding Author:**

Vamshi Mugu, MS, MD

Department of Radiology

Mayo Clinic

200 1st street SW

Rochester, MN, 55905

United States

Phone: 1 507 284 2511

Email: [vkugu@gmail.com](mailto:vkugu@gmail.com)

## Abstract

**Background:** Medical ethics provides a moral framework for the practice of clinical medicine. Four principles, that is, beneficence, nonmaleficence, patient autonomy, and justice, form the cornerstones of medical ethics as it is practiced today. Of these 4 principles, patient autonomy holds a pivotal position and often takes precedence in ethical dilemmas that result from conflicts among the 4 principles. Its importance serves as a constant reminder to the clinician that the “needs of the patient come first.” With their remarkable ability to process natural language, large language models (LLMs) have recently pervaded nearly every aspect of human life, including medicine and medical ethics. Reliance on tools such as LLMs, however, poses fundamental questions in medical ethics, where human-like reasoning, emotional intelligence, and an understanding of local context and values are of utmost importance.

**Objective:** While emphasizing the central role of the human factor, we undertake a bold venture to establish some confidence in LLMs, as it pertains to medical ethics by not only evaluating the status quo of foundational LLMs but also exploring ways to improve the LLMs by using patient autonomy-based hypothetical cases. Although literature today is certainly lacking in such ventures, we also believe projects such as ours must be frequently revisited in the field of LLMs, which is evolving at a pace that is both rapid and unprecedented.

**Methods:** We evaluated 3 foundational LLMs (ChatGPT, LLaMA, and Gemini) on hypothetical cases in patient autonomy. We used Cohen  $\kappa$  to compare LLM responses to the consensus from a physician panel. McNemar test was used during the improvement phase and to report the final significance of improved agreement of each LLM with physician consensus. *P* values less than .05 were considered significant. An agreement with  $\kappa < 0$  was designated as poor, 0-0.2 as slight, 0.2-0.4 as fair, 0.41-0.6 as moderate, 0.61-0.8 as substantial, and 0.81-1 as almost perfect.

**Results:** There was slight to fair agreement between the foundational LLMs and the physician consensus. With iterative improvement techniques, this agreement evolved to be substantial or higher (Cohen  $\kappa$  of 0.73-0.82). The degree of improvement was statistically significant (*P*=.006 for ChatGPT, *P*<.001 for Gemini, and *P*<.001 for LLaMA).

**Conclusions:** Although LLMs hold great potential for use in medicine, there needs to be an abundance of caution in using foundational LLMs in domains such as medical ethics. With adequate human oversight in testing and utilizing established techniques, LLM responses can be better aligned to human responses, even in the domain of medical ethics.

(*JMIR Med Inform* 2025;13:e77061) doi: [10.2196/77061](https://doi.org/10.2196/77061)

**KEYWORDS**

large language models; medical ethics; patient autonomy; generative AI; human in the loop

**Introduction**

The American Medical Association defines medical ethics as a moral framework for the practice of clinical medicine [1]. Proposed by Beauchamp and Childress [2], the 4 principles, that is, beneficence, nonmaleficence, patient autonomy, and justice, are often regarded as the cornerstones of medical ethics as the field is taught and practiced today. Among these 4 principles, patient autonomy holds a particularly important place due to its frequent precedence over the other 3 principles when an ethical dilemma arises due to conflict among the 4 principles [3].

Large language models (LLMs) represent a new paradigm in artificial intelligence with remarkable abilities to process and generate text in natural language [4]. Foundational LLMs can perform many different tasks but sometimes lack domain-specific capabilities [5]. Although the abilities of foundational LLMs have translated into numerous applications in medical education, research, and practice [6,7], concerns remain about their accuracy, biases, and potential misuse [8]. A recurrent theme in proposed approaches to address these concerns is the importance of human expert participation in the evaluation and governance of LLMs. This importance of the human factor, which is also emphasized in this study, has prompted the development of guidelines from various governing bodies across continents [9-11].

There is a notable scarcity of literature on the role of LLMs in medical ethics, particularly on ways to improve the reliability of foundational LLMs on tasks specific to medical ethics. The

**Textbox 1.** Sample hypothetical case. Note that the hypothetical case deals with the capacity to consent and treatment refusal.

**Sample hypothetical case**

David is a 79-year-old male with a history of myocardial infarction whose wife, Mary called 911 about David's chest pain. The pain started 30 minutes ago and is similar to some of his prior episodes which needed admission to a hospital. Upon arrival of the emergency medical services (EMS) team, David refuses any intervention. David seems distressed from pain but is orientedx3.

Should the EMS team take David to the nearest hospital?

Three foundational LLMs were chosen for this study: ChatGPT version 4o (ChatGPT), LLaMA 3.1 70b BF16 (LLaMA), and Gemini 1.5 (Gemini). Although the open-source LLaMA was used locally with Ollama version 0.1.33 (Ollama, Inc), the 2 closed-source models, ChatGPT and Gemini, were accessed using their respective web interfaces. Where applicable, the default parameters (eg, temperature of 0.8, top\_p of 0.95 for Ollama) were used. There were no concerns for Health Insurance Portability and Accountability Act (HIPAA) noncompliance due to the hypothetical nature of the cases. Five physicians (P1-P5) with respective board certifications from emergency medicine, surgery, and radiology comprised the physician panel.

**Study Design**

Our study was designed in 2 phases (Figure 1): evaluation phase and improvement phase. The evaluation phase was designed to compare the responses from foundational LLMs to those from

comparative evaluation of LLMs in medical ethics is nearly nonexistent, with only ChatGPT being evaluated to any considerable extent [12]. Although the body of knowledge, particularly prompt engineering [13], to elicit improved responses from LLMs in general is enlarging, the discussion of these techniques in the context of medical ethics is lacking. We venture on a bold undertaking to not only evaluate the foundational LLMs in their ability to analyze cases in patient autonomy but also explore techniques to improve the reliability of LLM responses, thereby attempting to increase trust in the technology while emphasizing the indispensable role of the human factor.

**Methods****Ethical Considerations**

After obtaining approval from the Mayo Clinic Institutional Review Board (which also governs research ethics per approval 25-001042), 44 hypothetical cases in patient autonomy requiring yes or no responses were composed. No real patient information was used for this research. The first author adopted these cases from the literature [14] and from personal and collective clinical experiences, with a focus on capacity to consent, occupational exposure, confidentiality, informed consent for a minor patient, patient preferences, treatment refusal, and training needs.

**Hypothetical Case**

A hypothetical case is presented in Textbox 1. The LLMs and the physicians on the panel (described subsequently) were blinded to the sources of the cases and others' responses.

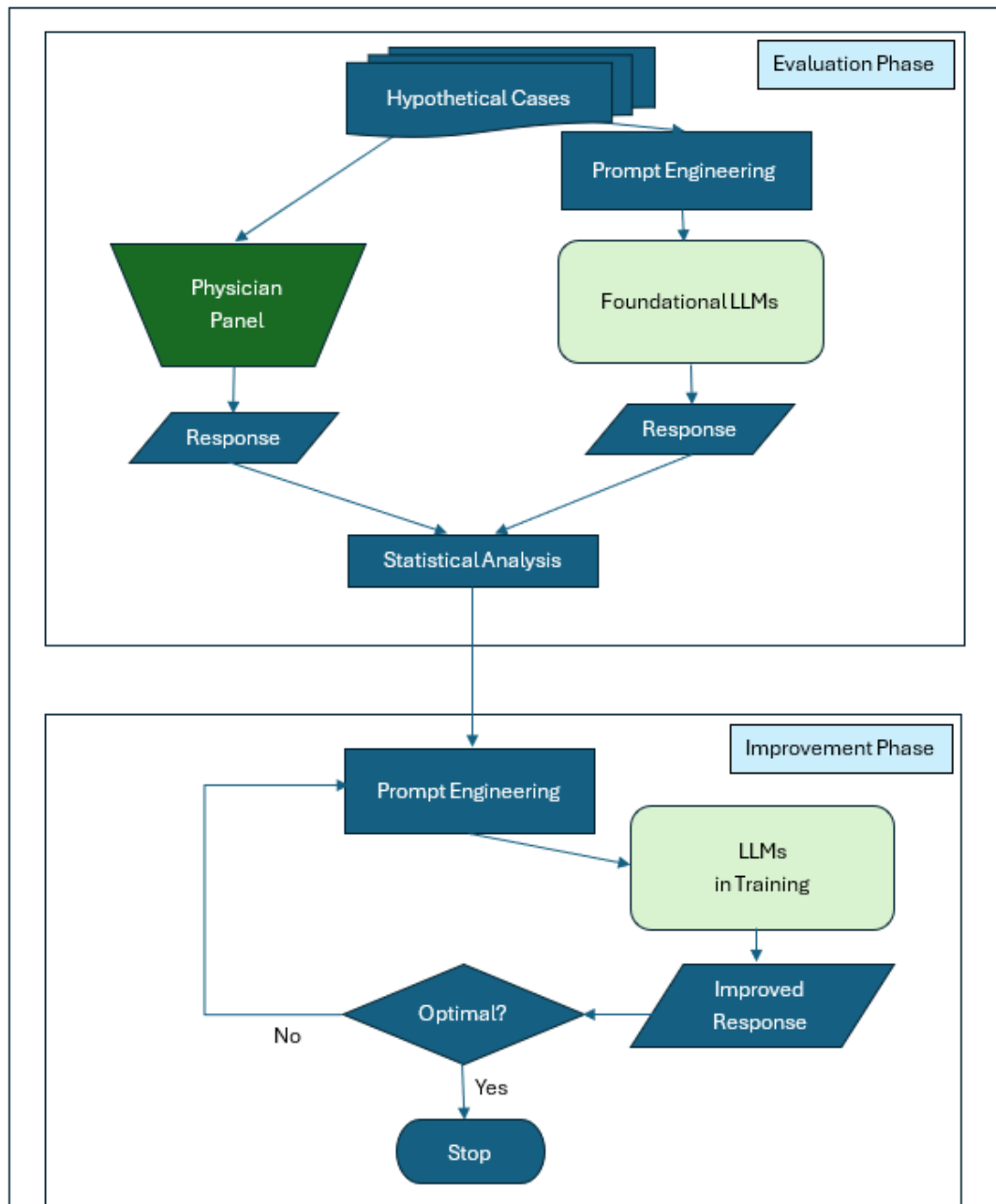
the physician panel, while the improvement phase was designed to optimize the responses from LLMs to improve agreement with the physician panel. The LLMs and the physicians on the panel were blinded to each other's responses. Interobserver agreement among the physicians on the panel and among the foundational LLMs was calculated using Fleiss  $\kappa$ . Cohen  $\kappa$  was used to compare the responses of each LLM with the physician consensus, which was defined as the majority response from the physicians in the panel. A missing value from an LLM was considered "No" for statistical analysis purposes. For illustration, pairwise agreements were depicted as a heatmap and listed as proportions in a tabular format. Python statsmodels (version 0.15.0) was used for statistical testing. An agreement with  $\kappa < 0$  was considered to be poor, 0 to 0.2 was considered slight, 0.21 to 0.4 as fair, 0.41 to 0.6 as moderate, 0.61 to 0.8 as substantial, and 0.81 to 1 as almost perfect. Although we reserved several prompt engineering techniques to employ during the

improvement phase, we nevertheless used some techniques such as role playing, forcefulness, and chain-of-thought during the evaluation phase [13].

During the improvement phase, prompt engineering techniques such as chain-of-thought, N-shot prompting, directional stimulus, versioning, rephrase-and-respond, and long context prompting were used to improve LLM responses [13,15-18].

The goal of this iterative process was to improve LLMs to produce responses that were better aligned to the physician consensus, reducing any statistically significant differences. The stopping point of the iterative process was determined to be when no more reduction in the number of differing responses could be achieved. The degree of improvement was assessed using McNemar test. A *P* value less than .05 was considered significant.

**Figure 1.** Two-phase design with evaluation phase and improvement phase. The evaluation phase compares the responses of foundational LLMs to physician panel consensus. The improvement phase is designed to improve the agreement between the LLMs and the physician panel consensus. LLM: large language model.



**Sample Human-Model Interaction**

A sample human-model interaction during the evaluation phase and the improvement phase is provided in [Textbox 2](#). This hypothetical case deals with consenting for a minor patient and

demonstrates human-model interaction during the evaluation and improvement phases. The use of 2 prompt engineering techniques, that is, question-refinement and chain-of-thought, is illustrated with this sample.

**Textbox 2.** Sample hypothetical case with human-model interaction during the evaluation and improvement phases. Two prompt engineering techniques, that is, question-refinement and chain-of-thought, are demonstrated here.

#### Sample interaction with a model demonstrating evaluation and improvement phases

**Hypothetical case:** Rachel is a 9-year-old girl from Virginia. She fell at the school basketball court and was taken to the emergency department by the teacher. She is found to have an elbow fracture that needs nonemergent surgical fixation. Can the surgeon proceed with the surgery?

#### Interaction snippet during the evaluation phase

Gemini's response

Answer: Yes

Thought Process:

1. The surgeon has a duty to act in the best interest of the patient. In this case, the surgery is necessary to treat the elbow fracture and prevent complications.
2. Delaying the surgery could lead to worsening of the condition, pain, and potential long-term disability.
3. Therefore, proceeding with the surgery is in the patient's best interest.

#### Interaction snippet during the improvement phase

Refined question (provided by ChatGPT upon human request): Given that Rachel, a 9-year-old girl from Virginia, fell at the school basketball court and was taken to the emergency department by her teacher, where she was diagnosed with an elbow fracture that requires nonemergent surgical fixation, does the surgeon have the necessary consent and authority to proceed with the surgery?

Gemini's response

Answer: No

Thought Process:

1. Rachel is a minor and therefore lacks the capacity to provide informed consent for the surgery.
2. Her legal guardian, likely her parents, must provide consent.

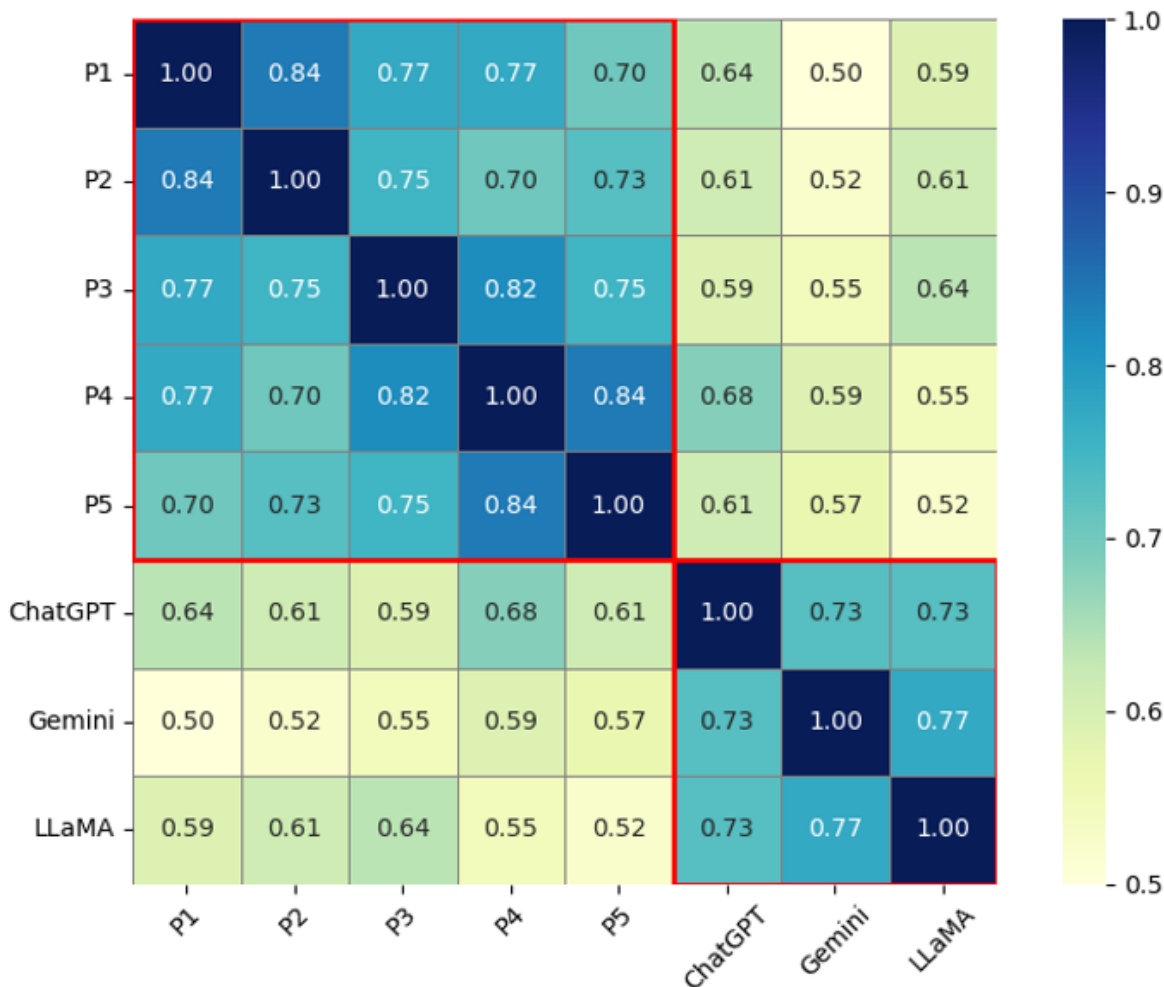
## Results

Of the 44 cases, 6 focused on informed consent, 6 on capacity to consent, 6 on occupational exposure, 6 on training needs, 7 on treatment refusal, 6 on patient preferences, and 7 on confidentiality. A sample hypothetical case is presented in [Textbox 1](#). There was substantial agreement among the physicians in the panel ( $\kappa=0.77$ ). There was also substantial agreement among the foundational LLMs ( $\kappa=0.73$ ). There was slight agreement between the foundational Gemini and the physician consensus ( $\kappa=0$ ), slight agreement between the foundational LLaMA and the physician consensus ( $\kappa=0.14$ ), and fair agreement between the foundational ChatGPT and the

physician consensus ( $\kappa=0.32$ ). Pairwise agreement is depicted in [Figure 2](#) as a heatmap for illustration purposes. However, note that the Improvement phase was designed to improve the agreement between LLMs and the physician panel consensus rather than individual physician responses.

At the end of the improvement phase, agreement between all the improved LLMs and the physician consensus was substantial or higher ( $\kappa=0.77$ , 95% CI 0.59-0.96 for ChatGPT;  $\kappa=0.73$ , 95% CI 0.52-0.93 for Gemini; and  $\kappa=0.82$ , 95% CI 0.65-0.99 for LLaMA). The degree of improvement for all the LLMs was statistically significant ( $P=.006$  for ChatGPT,  $P<.001$  for Gemini, and  $P<.001$  for LLaMA). The degree of improvement is listed in [Table 1](#).

**Figure 2.** Pairwise agreement heatmap for illustration purposes only. Note the substantial agreement among large language models and among physicians on the panel (P1-P5) during the evaluation phase (enclosed by red boxes). The improvement phase (discussed later) is however designed to improve the agreement of large language models with physician consensus, not with individual physician responses.



**Table 1.** Proportion of agreement between each large language model and physician consensus before and after improvement. Note that all large language models underwent improvement that was statistically significant.

Model	Before improvement, n	After improvement, n	Degree of improvement (N=44), n (%)	P value
ChatGPT	27	39	+12 (27)	.006
Gemini	22	38	+16 (36)	<.001
LLaMA	26	40	+14 (32)	<.001

## Discussion

### Principal Findings

Medical ethics provides a moral framework for the practice of clinical medicine, particularly as it pertains to physician responsibilities and obligations in the context of patient-physician interactions [1]. Beauchamp and Childress [2] are often credited for laying the foundation for this framework with 4 principles: beneficence, nonmaleficence, patient autonomy, and justice. Although these principles have been extensively criticized [19], debated [20], and defended [21], they nevertheless remain at the center of modern teaching and practice of medical ethics. The principle of patient autonomy holds a particularly important place among the four, often taking precedence over the other 3 principles when an ethical dilemma

arises due to conflicts among the 4 principles [3]. It is a constant reminder to the medical practitioner that “the patient must remain at the heart of his or her own individual battle” [22] and that “the needs of the patient come first” [23].

Despite their relative infancy, LLMs have revolutionized various aspects of health care, ranging from medical education to practice and research [6,7]. LLMs have demonstrated remarkable accuracy on tasks that were once believed to be germane to the rigorously trained human, such as outperforming medical students, and neurosurgery residents on neurosurgery written board-like questions [24]. LLMs are nevertheless prone to limitations, a few of which are particularly concerning in medicine, such as the fabrication of responses not based on evidence—a behavior referred to as hallucination [25]. Moreover, reliance on technology such as LLMs poses

fundamental questions in medical ethics, where human-like reasoning, emotional intelligence, and an understanding of local context and values are of utmost importance [26,27]. A recurrent theme in the proposed approaches to address these concerns is the importance of human expert participation in evaluation and governance of LLMs. This importance of the human factor, which is also emphasized in our study, has prompted the development of guidelines from various governing bodies across continents [9-11] and paved the way for effective interaction such as through prompt engineering [13]. The field of prompt engineering is rapidly evolving with several successful approaches already being adopted, including chain-of-thought, directional stimulus, forcefulness, self-correction and consistency, role playing, reflection, guiding output with rails, and long context prompting [13,15-18].

Although literature on the employment of artificial intelligence in the domain of medical ethics is sparse, some work warrants mention. Before the conception of LLMs, Shalowitz et al [28] proposed a population-based treatment indicator and claimed its superiority to traditional surrogate decision-making in treatment decisions made for incapacitated patients [28]. The population-based treatment indicator was critiqued for disrespecting patient autonomy by relying on statistical evidence rather than individual patient preferences [26]. Leveraging the natural language processing power of LLMs, Earp et al [29] proposed P4, a personalized patient preference predictor, to address this specific concern [29]. P4's claim of superiority is based on its incorporation of material from prior patient treatment decisions, thereby creating a "digital twin" to act on behalf of the incapacitated patient, when such a need arises.

Our study has perhaps the closest resemblance to the expert panel evaluation of GPT-4 by Balas et al [27] who created a set of 8 ethical case vignettes and present to the LLM via a priori

prompt template. LLM responses are then evaluated for the depth of reasoning, ability to consider alternate viewpoints, and sensitivity to nuances of ethical dilemmas. Although our case repertoire is considerably larger, it is also arguably simpler since we do not probe the depth of reasoning. However, the greatest distinction from their study is in the improvement phase. In the strictest sense of the study design, their study can be considered as lacking this phase. By not only evaluating the foundational LLMs but also exploring the ways to improve their performance as it pertains to a subset of medical ethics, we venture to establish some confidence in this promising technology while reminding the adopters of the importance of the human factor.

### Limitations

This study was subject to a few limitations. Although larger than other similar studies, it still involved a limited set of questions covering a relatively small section of medical ethics. It is possible that LLMs could perform differently with different clinical scenarios or ethical topics. Although we employed several types of prompt engineering, other improvement techniques such as RAG and fine-tuning were not used. Over time, work will undoubtedly continue to improve the performance of foundational LLMs in medical ethics; how quickly and whether the need for human supervision will ever be entirely eliminated remains to be seen.

### Conclusion

The use of foundational LLMs in domains such as medical ethics warrants an abundance of caution and intricate involvement of a human expert. With adequate testing and by utilizing established techniques such as prompt engineering, LLM performance can be improved, even in the domain of medical ethics, where human-like reasoning, emotional intelligence, and context awareness are crucial.

### Conflicts of Interest

None declared.

### References

1. Medical ethics. American Medical Association. URL: <https://www.ama-assn.org/topics/medical-ethics> [accessed 2024-09-15]
2. Beauchamp TL, Childress JF. Principles of Biomedical Ethics (Eighth Edition). Oxford, UK. Oxford University Press; 2019.
3. Badger J, Ladd RE, Adler P. Respecting patient autonomy versus protecting the patient's health: a dilemma for healthcare providers. *JONA's Healthcare Law, Ethics, and Regulation*. 2009;11(4):120-124. [FREE Full text] [doi: [10.1097/nhl.0b013e3181c1b542](https://doi.org/10.1097/nhl.0b013e3181c1b542)]
4. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. *ACM Trans Intell Syst Technol*. Aug 18, 2025;16(5):1-72. [FREE Full text] [doi: [10.1145/3744746](https://doi.org/10.1145/3744746)]
5. Scott IA, Zuccon G. The new paradigm in machine learning - foundation models, large language models and beyond: a primer for physicians. *Intern Med J*. May 2024;54(5):705-715. [doi: [10.1111/imj.16393](https://doi.org/10.1111/imj.16393)] [Medline: [38715436](https://pubmed.ncbi.nlm.nih.gov/38715436/)]
6. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. Mar 04, 2023;47(1):33. [FREE Full text] [doi: [10.1007/s10916-023-01925-4](https://doi.org/10.1007/s10916-023-01925-4)] [Medline: [36869927](https://pubmed.ncbi.nlm.nih.gov/36869927/)]
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. Mar 19, 2023;11(6):887. [FREE Full text] [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
8. Kwong J, Wang SCY, Nickel GC, Cacciamani GE, Kvedar JC. The long but necessary road to responsible use of large language models in healthcare research. *NPJ Digit Med*. Jul 04, 2024;7(1):177. [FREE Full text] [doi: [10.1038/s41746-024-01180-y](https://doi.org/10.1038/s41746-024-01180-y)] [Medline: [38965411](https://pubmed.ncbi.nlm.nih.gov/38965411/)]

9. Biden J. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. URL: <https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/> [accessed 2024-09-15]
10. Déclaration de Montréal IA Responsable. URL: <https://montrealdeclaration-responsibleai.com/the-declaration/> [accessed 2024-08-22]
11. Ethics guidelines for trustworthy AI. European Commission. URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> [accessed 2024-08-22]
12. Skryd A, Lawrence K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. *JMIR Form Res.* May 08, 2024;8:e51346. [FREE Full text] [doi: [10.2196/51346](https://doi.org/10.2196/51346)] [Medline: [38717811](https://pubmed.ncbi.nlm.nih.gov/38717811/)]
13. Amatriain X. Prompt design and engineering: introduction and advanced methods. ArXiv. Preprint posted online on May 5, 2024. 2024. [doi: [10.48550/arXiv.2401.14423](https://doi.org/10.48550/arXiv.2401.14423)]
14. Johnston C, Bradbury P. 100 Cases in Clinical Ethics and Law (Second Edition). Boca Raton. CRC Press; 2016.
15. Li Z, Peng B, He P, et al. Guiding large language models via directional stimulus prompting. ArXiv. Preprint posted on October 9, 2023. 2023. [doi: [10.48550/arXiv.2302.11520](https://doi.org/10.48550/arXiv.2302.11520)]
16. Chang K, Xu S, Wang C, et al. Efficient prompting methods for large language models: a survey. ArXiv. Preprint posted on December 2, 2024. 2024. [FREE Full text] [doi: [10.48550/arXiv.2404.01077](https://doi.org/10.48550/arXiv.2404.01077)]
17. Parthasarathy V, Zafar A, Khan A, Shahid A. The ultimate guide to fine-tuning llms from basics to breakthroughs: an exhaustive review of technologies, research, best practices, applied research challenges and opportunities. ArXiv. Preprint posted on August 23, 2024. 2024. [doi: [10.48550/arXiv.2408.13296](https://doi.org/10.48550/arXiv.2408.13296)]
18. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. ArXiv. Preprint posted on May 28, 2020. 2020. [doi: [10.48550/arXiv.2005.14165](https://doi.org/10.48550/arXiv.2005.14165)]
19. Clouser K, Gert B. A critique of principlism. *J Med Philos.* Apr 1990;15(2):219-236. [doi: [10.1093/jmp/15.2.219](https://doi.org/10.1093/jmp/15.2.219)] [Medline: [2351895](https://pubmed.ncbi.nlm.nih.gov/2351895/)]
20. Davis R. The principlism debate: a critical overview. *J Med Philos.* Feb 1995;20(1):85-105. [doi: [10.1093/jmp/20.1.85](https://doi.org/10.1093/jmp/20.1.85)] [Medline: [7738461](https://pubmed.ncbi.nlm.nih.gov/7738461/)]
21. Gillon R. Defending 'the four principles' approach to biomedical ethics. *J Med Ethics.* Dec 1995;21(6):323-324. [FREE Full text] [doi: [10.1136/jme.21.6.323](https://doi.org/10.1136/jme.21.6.323)] [Medline: [8778453](https://pubmed.ncbi.nlm.nih.gov/8778453/)]
22. Mattei JF. [Respecting the patient's autonomy]. *Soins.* Apr 2018;63(824):25-27. [FREE Full text] [doi: [10.1016/j.soins.2018.02.004](https://doi.org/10.1016/j.soins.2018.02.004)] [Medline: [29680133](https://pubmed.ncbi.nlm.nih.gov/29680133/)]
23. Beck CS. The needs of the patient come first. *Mayo Clinic Proceedings.* Mar 2000;75(3):224. [doi: [10.4065/75.3.224](https://doi.org/10.4065/75.3.224)]
24. Guerra G, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. *World Neurosurg.* Nov 2023;179:e160-e165. [doi: [10.1016/j.wneu.2023.08.042](https://doi.org/10.1016/j.wneu.2023.08.042)] [Medline: [37597659](https://pubmed.ncbi.nlm.nih.gov/37597659/)]
25. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res.* May 31, 2023;25:e46924. [FREE Full text] [doi: [10.2196/46924](https://doi.org/10.2196/46924)] [Medline: [37256685](https://pubmed.ncbi.nlm.nih.gov/37256685/)]
26. Sharadin N. Patient preference predictors and the problem of naked statistical evidence. *J Med Ethics.* Dec 2018;44(12):857-862. [doi: [10.1136/medethics-2017-104509](https://doi.org/10.1136/medethics-2017-104509)] [Medline: [29895554](https://pubmed.ncbi.nlm.nih.gov/29895554/)]
27. Balas M, Wadden JJ, Hébert PC, Mathison E, Warren MD, Seavilleklein V, et al. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. *J Med Ethics.* Jan 23, 2024;50(2):90-96. [doi: [10.1136/jme-2023-109549](https://doi.org/10.1136/jme-2023-109549)] [Medline: [37945336](https://pubmed.ncbi.nlm.nih.gov/37945336/)]
28. Shalowitz D, Garrett-Mayer E, Wendler D. How should treatment decisions be made for incapacitated patients, and why? *PLoS Med.* Mar 2007;4(3):e35. [FREE Full text] [doi: [10.1371/journal.pmed.0040035](https://doi.org/10.1371/journal.pmed.0040035)] [Medline: [17388655](https://pubmed.ncbi.nlm.nih.gov/17388655/)]
29. Earp BD, Porsdam Mann S, Allen J, Salloch S, Suren V, Jongsma K, et al. A personalized patient preference predictor for substituted judgments in healthcare: technically feasible and ethically desirable. *Am J Bioeth.* Jul 16, 2024;24(7):13-26. [FREE Full text] [doi: [10.1080/15265161.2023.2296402](https://doi.org/10.1080/15265161.2023.2296402)] [Medline: [38226965](https://pubmed.ncbi.nlm.nih.gov/38226965/)]

---

## Abbreviations

**HIPAA:** Health Insurance Portability and Accountability Act

**LLM:** large language model

---

*Edited by A Benis; submitted 06.May.2025; peer-reviewed by B Bhasuran, C Ma; comments to author 11.Aug.2025; revised version received 21.Aug.2025; accepted 06.Oct.2025; published 23.Oct.2025*

*Please cite as:*

*Mugu V, Carr B, Khandelwal A, Olson M, Schupbach J, Zietlow J, Vu TND, Chan A, Collura C, Schmitz J  
Human-Machine Agreement in Medical Ethics: Patient Autonomy Case-Based Evaluation of Large Language Models  
JMIR Med Inform 2025;13:e77061*

*URL: <https://medinform.jmir.org/2025/1/e77061>*

*doi: [10.2196/77061](https://doi.org/10.2196/77061)*

*PMID: [41056099](https://pubmed.ncbi.nlm.nih.gov/41056099/)*

©Vamshi Mugu, Brendan Carr, Ashish Khandelwal, Mike Olson, John Schupbach, John Zietlow, T N Diem Vu, Alex Chan, Christopher Collura, John Schmitz. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 23.Oct.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.