Creativity and Innovation
2025, VOL. 9, NO. 3, 23-27
DOI: 10.47297/wspciWSP2516-252705.20250903

How Can Large Language Models be more ethical in Healthcare: An Oversight Framework for "Digital Doctors"

Yuqian Shi

School of Liberal Arts, Renmin University of China

ABSTRACT

The integration of large language models (LLMs) as "Digital Doctors" in healthcare raises significant ethical challenges such as human physical health risks, legal disputes, bias amplification, privacy leakage, and human-machine relationship issues. This paper proposes a comprehensive ethical oversight framework spanning the entire LLM lifecycle—pre-training, supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF). Key measures include curating diverse, bias-free training datasets with federated learning for privacy protection, designing hierarchical task trees for ethical clinical interactions, and implementing reward models (RMs) that prioritize patient safety and autonomy. The framework also tackles global ethical and cultural disparities through geolocation-based content filtering and multi-cultural training datasets. By embedding ethical principles at every stage, this oversight framework aims to mitigate risks, foster trust, and ensure "Digital Doctors" uphold medical integrity, patient safety, and equity.

KEYWORDS

large language models (LLMs); healthcare AI; ethical oversight; Federated Learning (FL)

1 Introduction

The integration of large language models (LLMs) into healthcare, particularly in the form of "Digital Doctors", represents a transformative leap in medical innovation. These AI-driven systems, designed to simulate clinical reasoning, patient interaction, and diagnostic support, hold promise for enhancing healthcare accessibility, reducing diagnostic errors, and improving patient outcomes. However, the complexity of medical knowledge, the stakes of patient safety, and the ethical nuances of human-AI interaction introduce significant challenges. Ethical oversight in this context is not merely a regulatory requirement but a foundational necessity to ensure that "Digital Doctors " operate with precision, fairness, transparency, and respect for patient autonomy and safety.

2 Ethical Challenges in Healthcare AI: Foundations and Risks

LLMs operate on three core principles: Generative capabilities, Pre-training on vast datasets, and Transformer architectures. Generative models produce responses by predicting the next token in a sequence, enabling them to simulate human-like dialogue and reasoning. Pre-training on diverse datasets (e.g., medical guidelines, patient records, and general web text) endows them with broad knowledge, while Transformers allow for contextual understanding of long-range dependencies in language. (Brown., et al., 2020) However, these strengths are accompanied by critical limitations:

(1) Hallucinations. LLMs may generate factually incorrect or nonsensical outputs, as they rely on statistical patterns rather than definitive knowledge. In healthcare, this can lead to misdiagnosis errors, such as recommending contraindicated medications or fabricating patient histories.

(2) Black-Box Opacity. The inner workings of LLMs are notoriously difficult to interpret, making it challenging to trace the origin of specific outputs. This opacity undermines accountability when errors occur, as programmers struggle to identify whether a mistake stems from flawed training data, algorithmic bias, or model limitations.

These limitations lead to some ethical risks in "Digital Doctor " applications. To enhance the assessment of ethical risks, a grading system can be established for each problem category to measure the model's performance across different impact types and severity levels: (1) P0 issues relate to intolerable issues, such as strictly prohibited illegal activities, problems that may cause irreversible harm, pornography or sensitive political issues; (2) P1 issues primarily include risks affecting users' physical safety in medical rigor, scenarios likely to trigger public opinion and violations crossing the medical industry's red lines, including illegal activities or partial breaches of regulations; (3) P2 issues refer to problems that affect primary diagnosis or directly impact user experience, such as privacy invasion and offensive inquiries, and involve the use of outdated protocols or violations of industry norms and regulatory provisions; (4) P3 issues refer to common issues in both online real-life doctor consultations and offline clinics, such as omission of questions, inaccurate diagnosis or expression, and inappropriate wording.

3 The Framework for Ethical Oversight in "Digital Doctor" Training

It is worth noting that to achieve ethical supervision, measures should be taken throughout the entire process of large model training, rather than merely evaluating the model based on output results. This is because the black-box opacity of the model means that when the model produces unacceptable outputs, it is extremely difficult to determine which link in the training process led to the ethical fallacies. Hence, the ethical oversight framework should pervade the entire process of Pre-Training, Supervised Fine-Tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF).

Ethical oversight in the pre-training phase commences with the meticulous selection of diverse and validated medical datasets. The foundation of a trustworthy "Digital Doctor" lies in its exposure to high-quality, unbiased information that mirrors the complexity and diversity of real-world healthcare scenarios. To ensure factual accuracy and clinical relevance, the training data must prioritize peer-reviewed clinical guidelines, which serve as the gold standard in medical practice. Resources such as UpToDate, a leading clinical decision support tool, provide evidence-based recommendations updated by experts, ensuring the model is grounded in the latest medical knowledge. The World Health Organization (WHO) standards offer global perspectives on disease diagnosis and treatment, promoting consistency and inclusivity in the model's understanding of healthcare. De-identified patient records from diverse healthcare institutions provide real-world context, allowing the model to learn from actual clinical interactions while protecting patient privacy.

To embed ethical requirements into the model's behavior, hierarchical task trees should be developed that break down the "Digital Doctor"'s responsibilities into manageable, ethically informed tasks. These task trees serve as a roadmap for the model, guiding its interactions with patients and ensuring that every step of the diagnosis and treatment process is conducted in an ethical and responsible manner. Ethical considerations in this stage include asking relevant, non-leading questions to avoid biasing the patient's responses. Leading questions, such as "Is your pain sharp?" can influence the patient's recollection, potentially leading to exaggerated descriptions of symptoms. Instead, the model should use open-ended questions like "How would you describe the pain?" to allow the patient to provide an unfiltered account of their symptoms. Additionally, the model must ensure that it collects all relevant information without invading the patient's privacy. For example, when asking about a patient's sexual history, the model should only do so when it is clinically relevant to the diagnosis or treatment plan. This requires the model to have a deep understanding of the clinical context and the ability to distinguish between necessary and unnecessary inquiries. By promoting open and honest communication while respecting the patient's privacy, the model can build trust and ensure that the information gathered is both accurate and ethically obtained.

The design of the Reward Model (RM) is critical to embedding ethical preferences into the model's behavior. The RM should ensure that the model is incentivized to provide responses that are not only grammatically correct but also medically rigorous, respectful of patient privacy, and compliant with ethical and legal standards. To promote accuracy, the RM should reward the model for correctly identifying key medical indicators and following established clinical protocols. For example, correctly identifying the risk factors for sepsis, such as fever, low blood pressure, and elevated heart rate, and recommending prompt administration of antibiotics and fluid resuscitation, should be rewarded. These rewards encourage the model to prioritize patient safety, ensuring that it does not overlook or misinterpret critical medical information. By aligning the model's incentives with clinical best practices, the RM helps to ensure that the "Digital Doctor" provides high-quality, evidence-based care, thereby ensuring suggestions prioritize patient well-being and dignity. The RM should also focus on enhancing the patient experience by rewarding respectful and privacy-conscious interactions and penalizing invasive or inappropriate questions. For example, asking about a patient's sexual orientation without a clear clinical reason should be penalized, as it can make the patient feel uncomfortable or violated. On the other hand, using respectful and inclusive language, such as asking about "partners" rather than making assumptions about the patient's gender or sexual orientation, should be rewarded. Additionally, the model should be rewarded for providing clear and understandable explanations of medical concepts, avoiding jargon that may confuse the patient. This promotes effective communication and ensures that the patient can make informed decisions about their care.

4 Challenges in Ethical Oversight

The rapid advancement of LLMs in healthcare, particularly in applications like "Digital Doctors ", has created a critical tension between agility and stability in ethical oversight. On one hand, medical knowledge evolves rapidly, requiring LLMs to undergo frequent updates to incorporate new guidelines, therapies, and research findings. For example, the "Digital Doctor" framework needs to integrate the latest clinical guidelines through external knowledge bases and real-time updates. This agility is essential for maintaining the model's diagnostic accuracy and relevance. On the other hand, each model update—whether through pre-training, SFT, or RLHF— must undergo rigorous ethical review to ensure compliance with medical standards and patient safety.

The conflict arises because iterative updates may outpace the capacity of ethical review boards to assess all potential

risks. Rapid deployment of updated models without thorough review could inadvertently introduce new P2 issues, such as outdated contraindication warnings for newly approved drugs. Conversely, overly strict review processes may delay the integration of life-saving innovations, such as new cancer immunotherapies.

To address this, a hybrid oversight framework needs to be introduced: (1) automated safety checks— real-time validation against a dynamic rule engine that flags content violating medical guidelines (e.g., recommending unapproved drugs); (2) phased deployment: models undergo three stages— internal testing, clinical trials with human oversight, and public release—to balance innovation with safety; (3) retrospective auditing— post-deployment monitoring using real-world patient interactions to identify emerging risks, as seen in the "data flywheel" mechanism where user feedback iteratively refines the model. However, this approach still faces challenges. For example, the black-box opacity of LLMs makes it difficult to fully predict how updated parameters might affect rare edge cases, such as diagnosing uncommon genetic disorders. Even minor changes in pre-training data can lead to unpredictable hallucinations in specialized medical contexts.

The opaque nature of LLM decision-making introduces significant challenges in assigning responsibility for errors. Unlike traditional rule-based medical AI programmed by diagnostic algorithms with explicit logic, LLMs generate responses based on probabilistic patterns in training data, making it nearly impossible to trace specific errors to causal factors. This creates a dilemma in healthcare settings: (1) developer liability: should companies that design and train the model be held liable for model hallucinations, such as incorrectly diagnosing a viral infection as bacterial and recommending unnecessary antibiotics (a P2 issue); (2) healthcare provider accountability: if a hospital adopts a LLM-driven diagnostic tool, does the ultimate responsibility for errors lie with the provider, even if the model's inner workings are undisclosed?

To address accountability, potential frameworks include:(1) Joint liability models. Developers and healthcare providers share responsibility, with developers ensuring baseline model safety and providers validating model outputs against patient-specific data. (2) Transparency tools. Developing "interpretability layers" that highlight the evidence base for LLM recommendations. (3) Patient education. mandating clear warnings about LLM limitations, such as the disclaimer: "This advice is generated by an AI and should be verified with a licensed physician."

However, these solutions are not foolproof. For example, interpretability tools may still produce misleading explanations for complex diagnoses, and patient compliance with verification remains uncertain. Cultural and regulatory disparities pose significant challenges for deploying LLMs in global healthcare contexts. Some issues like fetal sex determination, euthanasia, and underage pregnancy are governed by vastly different laws and cultural norms across regions. In jurisdictions with restrictive abortion laws (e.g., Poland), LLMs must avoid providing information on abortion methods, even for medical emergencies. Conversely, in permissive regions (e. g., Canada), models may legally offer detailed guidance on safe abortion practices. The ethical oversight framework can address this through geolocation-based content filtering, where responses adapt to local laws. For example, a query about "genetic testing for hereditary diseases" in a region with strict genetic privacy laws (e.g., the EU under GDPR) would prioritize emphasizing informed consent and data anonymization, while the same query in a jurisdiction with more permissive direct-to-consumer testing regulations (e. g., the U. S. in some states) might include recommendations for commercial genetic testing platforms alongside risk disclosures. However, enforcing this requires real-time tracking of regional regulations, which is complicated by overlapping legal frameworks (e.g., state vs. federal laws in the U.S.).

To navigate these differences, the ''Digital Doctor'' needs to employ a multi-cultural training dataset that includes regional ethical guidelines. For example, SFT datasets for European markets include scenarios on informed consent for palliative sedation, while Asian datasets emphasize family-centered decision-making. However, this approach risks oversimplification, as cultural nuances within regions (e.g., urban vs. rural attitudes) may not be fully captured. In some cultures, discussing certain medical conditions (e. g., mental health, sexually transmitted infections) is stigmatized. The grading system is aimed to prohibit privacy violations, such as asking unrelated personal questions, but cultural interpretations of "relevance" vary. The training framework should address this by including diverse patient profiles in SFT datasets, but bias detection remains an ongoing challenge. What's more, the deployment of LLMs may exacerbate gaps in healthcare access. While ''Digital Doctors'' can expand service availability in underserved areas, reliance on technology may marginalize populations without digital literacy or internet access. Thus, integrating digital literacy programs and improving internet infrastructure is crucial to ensuring equitable digital healthcare access for all.

5 Future Directions for Ethical Oversight

The black box opacity of LLMs poses significant risks in healthcare where transparency is non-negotiable. To mitigate this, developing XAI tools tailored to medical reasoning is critical. Attention maps can decode how LLMs process patient inputs, such as symptoms and medical history, to form diagnoses. For instance, in the ''Digital Doctor''s assessment of allergic rhinitis, attention maps could highlight which symptoms (e.g., "daily sneezing", "clear nasal discharge") received

the most weight in the model's decision-making. By visualizing token-level attention, practicing physicians in the evaluation team can verify that the model adheres to evidence-based guidelines, such as *Rhinitis 2020: A Practice Parameter Update*. To make it more clear, LLMs should generate explicit links to medical guidelines when formulating recommendations. For example, if the model suggests CT imaging for suspected pneumonia, evidence tracing could reference the *Clinical guidelines - Diagnosis and treatment manual* or UpToDate protocols integrated via external knowledge bases. This mechanism enhances accountability, addressing knowledge timeliness and diagnostic logic.

Current medical LLM evaluations emphasize compliance with laws like the Federal Food, Drug, and Cosmetic Act (FD&C Act) (21 U.S.C. Chapter 9) and the Public Health Service Act (PHS Act) (42 U.S.C. § 264). But laws may vary in terms of countries and regions. To globalize these standards, an "AI Hippocratic Oath" should be established, incorporating:

(1) Patient autonomy. LLMs must respect informed consent, such as explicitly stating the limitations of AI advice and ensuring patients are not coerced into treatments. This requires the model to consider user experience by avoiding panic-inducing responses and informed decision-making in treatment recommendations.

(2) Beneficence and non-maleficence. Certification should mandate adherence to the safety principle which defines "ensuring patients' health" as the top priority, such as avoiding over-treatment and ensuring emergency protocols (e.g., recommending immediate hospitalization for suspected ectopic pregnancy).

(3) Transparency. Certified models must disclose data sources. For example, if a model's training data for predicting drug responses is predominantly sourced from EHRs of adult patients in a single geographic region, this regional and demographic limitation must be explicitly communicated to users, along with the caveat that its performance in pediatric populations or diverse ethnic groups may be unvalidated.

(4) Respect for human physicians. "Digital Doctors " must recognize the irreplaceable role of human physicians and refrain from unjust criticism of offline medical practices. This includes acknowledging the clinical expertise and experience of doctors, avoiding derogatory comparisons between AI and human healthcare providers, and explicitly advising users to consult with real physicians for complex medical decisions.

Effective ethical oversight requires cross-stakeholder collaboration, mirroring the multidisciplinary approach involving physicians, engineers, and policy experts. The engineers should embed ethical guardrails during model development, such as hard-coding prohibitions on sensitive topics (e.g., abortion). The prompt engineering techniques can be adapted to enforce ethical boundaries, such as pre-filtering harmful queries. Frontline physicians provide critical insights into real-world ethical dilemmas. For example, in emergency treatment scenarios, physicians can refine RM to prioritize life-saving recommendations over generic advice. Policymakers must balance innovation with safety (e. g., banning unlicensed medical practice recommendations). Regulatory frameworks should evolve with technology—for instance, updating telemedicine laws to address the ''Digital Doctor'"s role in triage and diagnosis. For example, reward models could adapt to the EU AI Act's requirements by classifying themselves as high-risk AI systems under the regulation and adjusting their transparency and traceability mechanisms to comply with the act's specific criteria for such classifications.

6 Conclusion

Ethical oversight of "Digital Doctors " is a multifaceted challenge that requires integrating technical rigor, clinical expertise, and regulatory mindfulness. By embedding ethical considerations into every stage of model development—from pre-training data curation to post-deployment evaluation—it is possible to mitigate risks like hallucinations, bias, and non-compliance, while fostering trust in these transformative tools. As ''Digital Doctors'' evolve, their success will depend not only on technical prowess but on their ability to uphold the ethical principles of healthcare: prioritizing patient safety, respecting autonomy, and adhering to the highest standards of medical integrity.

In essence, the ethical oversight of LLMs in healthcare is a testament to human ingenuity: a commitment to harnessing AI not as a replacement for human expertise but as a collaborative partner. As we navigate this frontier, the framework proposed here serves as a blueprint for ensuring that ''Digital Doctors'' remain trusted medical partners, upholding the highest standards of integrity while unlocking a future where healthcare is more accessible, accurate, and just for all. The journey ahead demands continuous dialogue and adaptation.

About the Author

Yugian Shi,research direction : Modern and Contemporary Chinese Literature.

References

[1] American College of Obstetricians and Gynecologists. (2018). Tubal Ectopic Pregnancy. *Obstetrics and Gynecology, 131 (6), e238–e259.* https://www.acog.org/clinical/clinical-guidance/practice-bulletin/articles/2018/03/tubal-ectopic-pregnancy

[2] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R.,& Otro, S.

(2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[3] Brown, TB., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, DM., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I.,& Amodei, D. (2020). Advances in Neural Information Processing Systems 33, Neurips 2020. Meeting34th Conference on Neural Information Processing Systems (NeurIPS).

[4] Devlin, J., Chang, M. W., Lee, K.,& Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics.

[5] Jing, W., Liu, Z., Zhang, S., Yin, Y. R.,& Chen, C. (2023).

[6] Optical character recognition of medical records based on deep learning.

[7] In 2023 5th International Conference on Robotics and Computer Vision (ICRCV) (pp. 1-6). IEEE. https://doi. org/10.1109/ICRCV59470.2023.10328990

[8] Kannu, M. D. C., Arumugam, G., Arumugam, S., Ramamoorthy, P. D., & Packianathan, R. (2024). Internet of things-integrated robotics in manufacturing. In Machine Vision and Industrial Robotics in Manufacturing (pp. 181–194). https://doi.org/10.1201/9781003438137-

[9] Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T.,& Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492. https://doi.org/10.48550/arXiv.1610.05492

[10] Luo, H., Sun, Q., Xu, C., Zhao, P., Lin, Q. W., Lou, J. G.,& Chen, S. (2024).

[11] Arena Learning: Build Data Flywheel for LLMs Post-training via Simulated Chatbot Arena. arXiv. https://doi.org/10.48550/arXiv.2407.16000

[12] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P.,& Christiano, P. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

[13] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (pp. 618-626). https://doi. org/10.1109/ICCV.2017.74

[14] Serrano, R. F.,& Smith, N. A. (2019). Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 3925–3936). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/N19-1357