EthicsBot: Fine-Tuning Open-Source LLMs to Assist Scientific Investigators in Analyzing Ethical Issues in Research

Spencer Phillips Hey
Founder
Hey Research & Innovation
Boston, MA, USA
heyspencer@gmail.com

Charles Weijer
Department of Medicine
Western University
London, ON, Canada
cweijer@uwo.ca

Julie Walsh
Department of Philosophy
Wellesley College
Wellesley, MA, USA
julie.walsh@wellesley.edu

Eni Mustafaraj
Department of Computer Science
Wellesley College
Wellesley, MA, USA
eni.mustafaraj@wellesley.edu

***Abstract*—Ethical considerations are fundamental to responsible research, yet many investigators struggle to identify and analyze ethical concerns in their study designs. Institutional review boards (IRBs) and other regulatory bodies, while essential, are often perceived as bureaucratic obstacles rather than collaborative partners in ethical inquiry. Recent advances in large language models (LLMs) offer an opportunity to enhance the way researchers engage with ethical analysis. This paper introduces EthicsBot, an innovative project that proposes to leverage open-source LLMs to provide real-time, context-aware ethical guidance for researchers.**

***Index Terms*—large language models, artificial intelligence, fine-tuning, ethics, human-subjects research.**

I. Introduction

Ethical considerations are essential to conducting responsible and impactful research. They help safeguard participant welfare, uphold privacy, balance risks and benefits, and align studies with societal values [1]. However, researchers often struggle to systematically identify and address ethical concerns, particularly when preparing grant proposals, institutional review board (IRB) submissions, or regulatory compliance documents. These challenges are especially pronounced in institutions with limited access to ethics experts or structured guidance [2].

Recent advances in large language models (LLMs) offer an opportunity to enhance the way researchers engage with ethical analysis [4], [5]. Although proprietary AI tools, like ChatGPT, already have the ability to surface ethical issues, reliance on proprietary LLMs presents significant challenges, including privacy vulnerabilities, bias, and intellectual property risks. However, with thoughtful development, we believe open-source LLMs could democratize access to ethical expertise, equipping a broader range of researchers with the resources needed to strengthen the ethical rigor of their work.

In this paper, we introduce EthicsBot, a project aimed at developing a generalizable framework and toolkit for institutions to fine-tune, test, and implement an ethically informed LLM research-support tool.

The structure of this paper is as follows: Section II discusses the motivation and practical considerations behind EthicsBot. Section III presents preliminary empirical findings from our experiments with these ideas using ChatGPT, illustrating the potential of LLMs in ethical analysis. Section IV outlines a proposed methodology and scope of EthicsBot. Sections V and VI explore some of the major challenges and the transformative potential of the project, respectively.

II. Rationale

The current system of ethical review, primarily conducted through Institutional Review Boards (IRBs) in the United States (or "Research Ethics Boards" in Canada and the United Kingdom), is often slow and perceived as a bureaucratic hurdle rather than an integral, value-adding component of the research process [6]. Investigators typically engage with IRBs only in the final stages of protocol development, viewing it as a compliance checkpoint rather than a resource for meaningful ethical reflection and insight that creates better science. Feedback from IRBs is also frequently regarded as opaque, inconsistent, and disconnected from the realities of scientific inquiry [2].

This status quo stands in stark contrast to what scientists say that want from the IRB [3] and what we believe practical research ethics should be. Ethical considerations should not be an afterthought or a regulatory burden—they should be embedded in the foundations of a scientific investigation. Indeed, ethics in research means asking the right questions at the right time and employing the right methods to answer those questions in ways that respect participants, ensure a favorable risk-benefit balance, promote social value, and minimize bias. In short, ethical research is not just about compliance; it is about rigorous, responsible, and socially valuable science [7].

While both scientists and society share an interest in ensuring ethical research, a challenge remains: most researchers receive little to no formal training in rigorous ethical analysis [8]. This lack of training can leave investigators feeling

979-8-3315-3228-4/25/$31.00 ©2025 IEEE

alienated by the ethical review process, particularly when their primary engagement with ethics comes in the form of IRB requirements that emphasize procedural compliance over substantive ethical reasoning.

This is where EthicsBot presents a transformative opportunity. Our goal is to develop a dynamic tool that researchers can use prospectively to identify and address ethical considerations early in the research design process. EthicsBot would function as an accessible, context-sensitive assistant, helping investigators integrate ethical analysis into their proposals and protocols before they reach the IRB stage.

III. Pilot Data

To explore the potential of LLMs in conducting ethical analyses, we conducted a preliminary experiment using ChatGPT [9]. Our objective was to assess whether a state-of-the-art, commercial LLM could identify meaningful ethical concerns in clinical trial protocols and provide relevant ethical guidance.

We selected a cohort of 12 clinical trial protocols from ClinicalTrials.gov, the United States' mandatory clinical trial registry [10], that varied across the following dimensions of study design:

- **Study populations**: Adults, pediatric populations, elderly participants, and indigenous communities.
- **Randomization**: Single-arm trials, individually randomized trials, and cluster-randomized designs.
- **Intervention types**: Drug trials, medical device evaluations, and behavioral interventions.

Each full protocol was downloaded from ClinicalTrials.gov uploaded through ChatGPT's web interface, where we provided a simple prompt: "What are the ethical issues in this protocol?"

The point of this initial experiment was to generate a baseline understanding of a commercial LLM's performance (given a simple prompt), as well as producing pilot data to serve as a point of comparison with open-source models in the context of a more substantial project like EthicsBot.

A. Observations and Key Findings

Two authors (SPH, CW), both with experience in research ethics and protocol review, analyzed ChatGPT's responses. Although we did not conduct a formal quantitative evaluation, we were impressed by the model's ability to surface relevant and significant ethical considerations.

ChatGPT listed out 9 ethical issues on average for each protocol. It frequently surfaced what we would consider to be basic ethical issues concerning informed consent and risk disclosure. However, it also demonstrated the capacity to surface more subtle and context-specific issues that would have been easy to miss. For example:

- **HIV Counseling Trial**: In a study evaluating a counseling intervention for HIV-positive individuals and their families, ChatGPT correctly identified the risk of HIV transmission among serodiscordant couples and advised investigators to ensure participants were adequately educated and supported in minimizing transmission risks.
- **Alzheimer's Prevention in Down Syndrome Trial**: In a trial promoting physical activity to help prevent Alzheimer's Disease in individuals with Down syndrome, the model recognized the potential burden on caregivers, highlighting the ethical need for adequate support and feasibility considerations.

These examples illustrate the kind of ethical reflection that a rigorous review process should facilitate. Yet, such issues can often escape the IRBs notice. But even in cases where the IRB identifies these kinds of challenges, if the investigator is unaware of them, it can add further steps to the IRB review process, requiring lengthy back-and-forth communications between the committee and investigators.

This brings us to one of the chief advantages of an AI-driven ethics assistant, such as EthicsBot: The potential for tireless consistency. Human reviewers may miss certain issues in the course of a ethics review due to distraction, fatigue, lack of expertise, or any number of other reasons. An LLM fine-tuned for ethics review, by contrast, suffers no such limitations. It can, in principle, surface a comprehensive set of ethical issues with greater consistency.

Of course, this small experiment represents only an initial exploration, and more rigorous, systematic evaluation is required. ChatGPT is also owned and controlled by a commercial entity, and its use often requires a data sharing agreement that is in tension with the privacy or intellectual property concerns of users or their research institutions. This need to protect the privacy interests of scientists, research subjects, and their supporting institutions is critical to the motivation for the EthicsBot project.

IV. Proposed Methods and Scope for EthicsBot

The development of EthicsBot requires a structured, multiphase approach, encompassing model selection, dataset creation, fine-tuning, prototype development, evaluation, and ethical safeguards. Below, we outline what we see as the key methodological steps to guide such a project.

A. Selection of Open-Source LLMs

To ensure accessibility and adaptability, EthicsBot should be built upon an open-source large language model (LLM). Several candidate models—including LLaMA [11], DeepSeek [12], and BLOOM [13]—should be evaluated based on their suitability for fine-tuning in ethical analysis.

Model selection should include the following criteria:

- **Architectural Flexibility**: The model must support fine-tuning with domain-specific ethical datasets.
- **Scalability**: It should be capable of handling complex ethical reasoning across multiple research disciplines.
- **Computational Efficiency**: Resource requirements must be balanced to facilitate both large and small research institutions.
- **Developer Community Support**: Active developer and researcher engagement should be prioritized to ensure sustainability and ongoing improvements.

Once an initial set of models is identified, pilot testing will be needed to establish baseline performance on ethical reasoning tasks. This can be achieved by presenting generic ethical prompts to each model and having experts trained in research ethics assess the response quality, relevance, and comprehensiveness [14].

B. Dataset Creation and Curation

A robust, high-quality dataset is critical for fine-tuning LLMs to generate reliable ethical guidance. The dataset constructed should be based on multiple sources, including annotated research protocols, ethics sections from grant proposals, submitted research protocols and the accompanying IRB reviews, and case studies illustrating ethical dilemmas across disciplines, including biomedical research, social sciences, and AI development.

To enhance the dataset's rigor and relevance, we would propose a structured annotation process:

- Ethics experts (e.g., experienced IRB members with appropriate philosophical training), annotating key ethical considerations within protocols, ensuring comprehensive coverage of issues such as informed consent, privacy, risk-benefit trade-offs, conflicts of interest, and societal implications.
- Diversity and representation should be prioritized by including ethical cases from different geographical, cultural, and disciplinary perspectives to account for variations in ethical norms. For example, it will be important to include research from low- and middle-income countries, research conducted in the Global South, research conducted with indigenous populations, and research conducted across different ethical jurisdictions (e.g., research that falls under U.S. regulations vs. Canadian or U.K. regulations).
- Quality assurance measures should be applied throughout, particularly with respect to minimizing bias. We would propose to systematically review the dataset to assess representation across disciplines, study populations, geographical regions, and ethical concerns. Undersampled areas (e.g., research ethics in low-resource settings) will be identified and supplemented with additional data. We would also propose to use statistical tools such as word embeddings analysis [15] and distributional analysis to detect potential biases (e.g., gender, racial, or socioeconomic biases in ethical assessments) [16]. If biases are detected, targeted data augmentation techniques will be used to balance perspectives [17].

C. Model Fine-Tuning

Once a suitable dataset is compiled, EthicsBot should undergo a multi-stage fine-tuning process to enhance its domain-specific expertise:

- **Supervised Learning**: Using expert-annotated data, the model's outputs can be aligned with high-quality ethical reasoning.
- **Human-in-the-Loop Evaluation**: Ethics specialists should periodically review the model's responses, providing feedback to refine its ethical judgment.
- **Reinforcement Learning with Human Feedback**: This technique can be applied to optimize the model's contextual understanding and improve the relevance and accuracy of its recommendations [18].
- **Performance Metrics**: Model assessments should be based on response accuracy, ethical depth, coherence, and user satisfaction to ensure continuous improvement.

D. Prototype Development

To maximize usability, we propose to develop EthicsBot within the context of a chat-based interactive tool with functionalities tailored to research ethics consultation:

- **Automated Protocol Analysis**: Users should be able to upload research proposals or protocols, and EthicsBot will analyze them for key ethical concerns, flagging potential issues for further consideration and review.
- **Ethics Section Drafting**: The system should generate structured drafts of ethics sections tailored to specific funder or institutional guidelines.
- **Resource Recommendations**: EthicsBot should be able to provide links (or summaries) of relevant ethics literature, regulations, and best practices.
- **Customization Features**: Users should be able to specify discipline-specific or region-specific ethical considerations, enabling more tailored and context-sensitive outputs.

Moreover, the user interface should be designed with simplicity and transparency in mind, ensuring that researchers can engage with the tool intuitively while maintaining trust in its recommendations.

E. Evaluation and Testing

A rigorous evaluation process will be needed to assess EthicsBot's effectiveness, reliability, and usability. This should involve:

- **User Testing**: Researchers from diverse fields should interact with the tool, providing qualitative and quantitative feedback on its functionality and ethical reasoning.
- **Scenario-Based Evaluation**: EthicsBot should be tested on real-world research protocols to determine its ability to identify ethical concerns, provide meaningful feedback, and generate coherent ethics sections.
- **Iterative Refinement**: Insights from user testing should be used to inform improvements in both the model's ethical reasoning and the interface's usability.
- **Benchmarking Against Human Experts**: EthicsBot's outputs should be compared to ethics analyses conducted by expert reviewers to assess reliability and validity.

F. Ethical Considerations for EthicsBot Development

Finally, given that EthicsBot is an AI-driven ethics consultation tool, its own development must adhere to rigorous ethical principles. This includes:

- **Bias Mitigation**: As we noted above in the section Dataset Creation and Curation, steps must be taken to identify and reduce biases in training data, ensuring that EthicsBot does not perpetuate harmful or exclusionary ethical assumptions.
- **Transparency, Explainability, and User-Centricity**: At a minimum, users should be able to understand why the model reaches particular ethical conclusions. Requiring "think-out-loud" outputs from the LLM, or other explanation features, will be critical to clarifying the rationale behind its recommendations, and thereby, promoting greater trust and understanding among users. But we would also endorse a more comprehensive approach to LLM tool development that emphasizes user-centricity and heuristic skill development [19].
- **Safeguards Against Over-Reliance**: EthicsBot is designed to complement, not replace, expert ethical review. Clear disclaimers should emphasize that users must treat its recommendations as a support tool rather than a definitive ethical authority.

V. Challenges

The development and deployment of a project like EthicsBot present several key challenges that must be addressed to ensure the tool's effectiveness, reliability, and ethical integrity.

A. Privacy and Intellectual Property

Privacy is a core concern for EthicsBot, as researchers and institutions seek to protect confidential study details and intellectual property (IP) while benefiting from AI-assisted ethical analysis [20]. Research protocols often contain sensitive information about study populations, methodologies, and unpublished findings, which could be at risk if uploaded to third-party AI services. To help mitigate this, EthicsBot should prioritize on-premise and private-cloud deployment options, ensuring users and institutions retain full control over their data.

Beyond confidentiality, protecting IP is essential, as institutions have a vested interest in securing novel research ideas and preventing external entities from leveraging their data for unintended purposes. Many commercial AI models involve data-sharing agreements, creating potential risks for research and IP security. EthicsBot should therefore be designed as an institutionally governed, open-source tool, ensuring that research data remains local, non-public, and free from third-party model training pipelines.

B. Resource and Cost Constraints

Even though EthicsBot is built upon open-source LLMs, significant technical resources are still required for fine-tuning, hosting, and maintaining such a system. Training and optimizing an LLM for ethical reasoning demands substantial computational power, expert oversight, and continuous updates to remain aligned with evolving ethical standards and best practices. Institutions, particularly those with limited funding, may struggle to implement and sustain such a tool without external support. Strategies such as cloud-based shared resources, federated learning, and institutional partnerships will need to be explored to mitigate these costs while ensuring broad accessibility.

C. The Risk of Over-Reliance and the Erosion of Human Expertise

While EthicsBot is designed as a supplementary tool rather than a replacement for human ethical review, there remains a risk that researchers and institutions may over-rely on automated ethical guidance, reducing engagement with deeper ethical reflection [21], [22]. We believe the dynamic, real-time nature of a chatbot interface has the potential to enhance investigators' ethical awareness—through a form of Socratic interaction—but without appropriate safeguards, it is easy to see how it could lead to a passive approach where ethical decision-making is simply offloaded to AI.

To address this concern, EthicsBot must be positioned as a tool for structured ethical inquiry rather than a definitive arbiter of ethical correctness. Clear disclaimers, educational interventions, and integration with human-led ethics consultations will be necessary to reinforce the idea that ethical reasoning remains a fundamentally human responsibility. The tool should provide reasonable guidance, or draft text, as output, but it remains the user's or institution's responsibility to thoughtfully evaluate that output and ultimately decide how to act.

D. Preventing Bias and Ensuring Ethical Rigor

Current LLMs exhibit various biases, including a positivity or "helpfulness" bias—a tendency to generate responses that favor user expectations rather than challenging users' assumptions. In an ethical context, this bias can be particularly problematic [23]. EthicsBot must not simply affirm researchers' perspectives or offer perfunctory justifications. Instead, it should prompt investigators to critically engage with ethical issues.

Additionally, ethical considerations are inherently context-sensitive. EthicsBot must be designed to recognize disciplinary, cultural, and regulatory differences, ensuring that its recommendations are not one-size-fits-all but instead contextually appropriate and adaptable. This requires careful dataset curation, human-in-the-loop reinforcement, and continuous auditing to prevent ethical distortions and ensure robust, principled reasoning in its outputs.

VI. Transformative Potential

While the challenges outlined above are significant, if they are managed appropriately, then we believe they are far outweighed by the potential benefits of EthicsBot as a scalable, democratizing force in research ethics.

A. A Systematic and Scalable Approach to Ethical Analysis

By fine-tuning open-source LLMs, EthicsBot can offer a powerful, structured method for identifying and helping investigators to address ethical concerns across diverse research domains. The ability to automatically generate tailored ethics

sections, highlight potential risks, and provide real-time ethical insights has the potential to streamline the ethics review process, allowing investigators to engage with ethical issues early and effectively rather than treating them as bureaucratic afterthoughts.

B. Democratizing Access to Ethical Expertise

Many institutions, particularly those in resource-limited settings, lack access to experienced research ethics consultants or well-developed research oversight mechanisms. EthicsBot presents an opportunity to bridge this gap. By making expert-level ethical analysis available at scale, the project has the potential to enhance the ethical integrity of research globally, ensuring that investigators—regardless of their institutional resources—have access to rigorous ethical review tools.

C. Enhancing Ethical Awareness and Education

Beyond its immediate utility in research protocol development, EthicsBot can serve as a pedagogical tool that actively enhances ethical literacy among researchers. By providing explanatory reasoning, relevant case studies, and interactive feedback, EthicsBot can help investigators develop a stronger, more intuitive understanding of ethical principles over time. This could foster a cultural shift in research ethics, moving away from compliance-driven approaches toward ethically engaged scientific inquiry.

D. Laying the Groundwork for Future AI-Assisted Ethical Frameworks

The methodologies developed for fine-tuning, bias mitigation, and ethical validation within EthicsBot will serve as a blueprint for future applications of LLMs in specialized domains. The project's findings on how to align AI-generated ethical reasoning with human expertise could inform a broad range of fields, from bioethics and AI governance to legal and policy analysis.

VII. Conclusion

If successful, we believe EthicsBot has the potential to make a profound and far-reaching impact on the global research community. By equipping investigators with an accessible, user-friendly, and ethically rigorous tool, the project can promote researcher understanding of ethics and higher standards of ethical responsibility in research.

Beyond reducing the administrative burden associated with ethical review, EthicsBot can help integrate ethical reflection directly into the scientific process, ensuring that investigators engage with ethical principles at an earlier stage of research design. Indeed, the vision of EthicsBot is more than just a chatbot—it is a step toward a future in which AI actively enhances the integrity, inclusivity, and ethical depth of scientific inquiry. By addressing the challenges head-on and harnessing the opportunities inherent in LLM technology, the EthicsBot project has the potential to reshape the landscape of research ethics for the better.

References

[1] United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Department of Health, Education, and Welfare, 1978.

[2] R. Klitzman, *The Ethics Police?: The Struggle to Make Human Research Safe*. Oxford: Oxford University Press, 2015.

[3] P. Keith-Spiegel, G. P. Koocher, and B. Tabachnick, "What scientists want from their research ethics committee," *J. Empir. Res. Hum. Res. Ethics*, vol. 1, no. 1, pp. 67–81, Mar. 2006.

[4] D. Yigci, M. Eryilmaz, A. K. Yetisen, S. Tasoglu, and A. Ozcan, "Large Language Model-Based Chatbots in Higher Education," *Adv. Intell. Syst.*, vol. 2024, p. 2400429, 2024.

[5] J. Zhou, M. Hu, J. Li, X. Zhang, X. Wu, I. King, and H. Meng, "Rethinking Machine Ethics—Can LLMs Perform Moral Reasoning through the Lens of Moral Theories?," *arXiv preprint arXiv:2308.15399*, Aug. 2023. [Online]. Available: https://arxiv.org/abs/2308.15399

[6] K. D. Haggerty, "Ethics creep: Governing social science research in the name of ethics," *Qualitative Sociology*, vol. 27, no. 4, pp. 391–414, Dec. 2004.

[7] A. Binik and S. P. Hey, "A framework for assessing scientific merit in ethical review of clinical research," *Ethics & Human Research*, vol. 41, no. 2, pp. 2–13, Mar. 2019.

[8] A. Eisen and R. M. Berry, "The absent professor: Why we don't teach research ethics and what to do about it," *The American Journal of Bioethics*, vol. 2, no. 4, pp. 38–49, Sep. 2002.

[9] OpenAI, "ChatGPT: Large Language Model for Conversational AI," 2023. [Online]. Available: https://openai.com/chatgpt. [Accessed: Feb. 6, 2025].

[10] National Library of Medicine, "ClinicalTrials.gov," 2025. [Online]. Available: https://clinicaltrials.gov. [Accessed: Feb. 6, 2025].

[11] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[12] DeepSeek AI, "DeepSeek: Open Large Language Models," 2024. [Online]. Available: https://github.com/deepseek-ai/DeepSeek-V3. [Accessed: Feb. 6, 2025].

[13] T. Le Scao *et al.*, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," *arXiv preprint arXiv:2211.05100*, 2022. [Online]. Available: https://arxiv.org/abs/2211.05100

[14] H. F. Lynch, M. Abdirisak, M. Bogia, and J. Clapp, "Evaluating the quality of research ethics review and oversight: A systematic analysis of quality assessment instruments," *AJOB Empirical Bioethics*, vol. 11, no. 4, pp. 208–222, Aug. 2020.

[15] M. E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, "Understanding the origins of bias in word embeddings," in *Proc. Int. Conf. Machine Learning*, May 2019, pp. 803–811.

[16] A. Slowik and L. Bottou, "Algorithmic bias and data bias: Understanding the relation between distributionally robust optimization and data curation," *arXiv preprint arXiv:2106.09467*, Jun. 2021. [Online]. Available: https://arxiv.org/abs/2106.09467

[17] B. Ding *et al.*, "Data augmentation using LLMs: Data perspectives, learning paradigms and challenges," in *Findings of the Association for Computational Linguistics ACL 2024*, Aug. 2024, pp. 1679–1705.

[18] R. Kirk *et al.*, "Understanding the effects of RLHF on LLM generalisation and diversity," *arXiv preprint arXiv:2310.06452*, Oct. 2023. [Online]. Available: https://arxiv.org/abs/2310.06452

[19] K. G. Barman, N. Wood, and P. Pawlowski, "Beyond transparency and explainability: On the need for adequate and contextualized user guidelines for LLM use," *Ethics and Information Technology*, vol. 26, no. 3, p. 47, Sep. 2024.

[20] A. Cappello, M. Dada, V. Grigoreva, R. Khan, C. Stinson, and H. Stuart, "Large Language Models and the Disappearing Private Sphere," Ethics and Technology Lab, Queen's University, Mar. 2024. [Online]. Available: https://llmprivacy.ca/report.pdf. [Accessed: Feb. 6, 2025].

[21] C. Royer, "Outsourcing Humanity? ChatGPT, Critical Thinking, and the Crisis in Higher Education," *Stud. Philos. Educ.*, vol. 2024, pp. 1–9, May 31, 2024.

[22] L. Stark, "The unintended ethics of Henry K. Beecher," *The Lancet*, vol. 387, no. 10036, pp. 2374–2375, Jun. 11, 2016.

[23] L. Ranaldi and G. Pucci, "When Large Language Models contradict humans? Large Language Models' Sycophantic Behaviour," *arXiv preprint arXiv:2311.09410*, Nov. 15, 2023.