The Need for Human-AI Collaborative Methods for Conducting Audits of Machine Learning Models

Yao Rong, Vaibhav Unhelkar

Department of Computer Science, Rice University
6100 Main St, Houston, TX 77005, USA
{yao.rong, vaibhav.unhelkar}@rice.edu

Abstract

Conducting application audits of ML models is essential for ensuring their safe and responsible deployment, particularly in high-stakes applications. However, the auditing of ML models deployed in domain-specific applications remains largely a manual process, relying on domain experts to identify model errors. The manual nature of the process limits scalability of audits and hinders the discovery of problematic model behaviors. We posit that a Human-AI Collaborative paradigm is essential for conducting effective application audits. In this abstract, we propose a research agenda to develop Human-AI collaborative methods for conducting application audits of ML models.

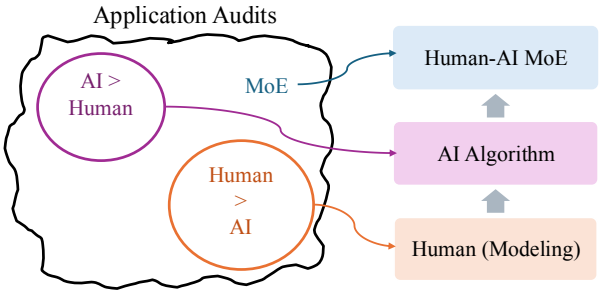

Figure 1: A potential Mixture-of-Experts (MoE) auditing system that leverages complementary strengths of humans and AI to conduct application audits of ML models.

Introduction

Auditing of machine learning (ML) models can be defined as the systematic and independent process of evaluating the model for potential errors, biases, risks and other unintended side effects. For instance, in the context of large language models (LLMs), Mökander et al. (2023) propose a three-layered auditing framework consisting of *governance* audits, *model* audits, and *application* audits. While the first two types of audits are designed for model technology providers – i.e., the organizations developing ML models – the third focuses on downstream applications of ML models and must be conducted within domain-specific contexts.

Application Audits Application audits are defined as the "impact-oriented assessment of the risks posed by products and services built on pre-trained ML models" (Mökander et al. 2023). Conducting application audits of ML models is essential for ensuring their safe and responsible deployment, particularly in high-stakes applications (Zhang et al. 2020). For instance, in autonomous vehicles, perceptual errors can lead to accidents or loss of life (Cummings 2023). Similarly, in healthcare, errors in diagnostic models may endanger patient safety (Sheliemina et al. 2024; Yu et al. 2024). Undetected issues, such as incorrect predictions or feature usage, can have severe consequences.

While recent research has significantly advanced model audits (such as techniques for detecting bias and hallucinations prior to deployment) comparatively less attention has been given to developing solutions and best practices for application audits. As a result, *application auditing today remains largely a manual process*, driven by domain experts of the specific application. Moreover, application audits introduce unique challenges. *They need to be conducted by end-users or application developers* who rely on end-user ML tools, such as Microsoft Azure or Amazon Bedrock, to build and evaluate customized applications by fine-tuning general-purpose ML models. Unlike large organizations that develop the pre-trained ML model, these users may have limited computational resources and may lack dedicated teams for model auditing. Additionally, *application audits must be conducted routinely* to address distribution shifts and evolving requirements within specific application domains.

The Need for Human-AI Collaboration Conducting application audits, thus, poses increasing demands on the already scarce time of domain experts. For instance, to audit an ML model $\mathcal{M} : X \rightarrow Y$, a human auditor – typically a domain expert – iteratively creates test cases, where each test case consists of a model input $x_i \sim X$ and annotations of the corresponding expected output y_i. The creation of test cases is guided by domain expert's intuition to maximize the detection of problematic model behaviors. Identified behaviors are then categorized based on factors such as severity and error type, culminating in an audit report. The manual nature of the process limits scalability and hinders the discovery of problematic model behaviors.

We posit that a *Human-AI Collaborative paradigm is es-*

sential for conducting effective application audits.[1] Our thesis is based on the observation that while certain aspects of the auditing process can be handled more efficiently by AI algorithms, others are better suited for human expertise. For example, in auditing ML models deployed in medical imaging, the expertise of radiologists is crucial for generating accurate ground-truth annotations and identifying relevant features. However, once this information is established, AI methods (as opposed to heuristics) can be more effective at exploring the high-dimensional input space of imaging models and detecting areas where the model is prone to errors.

As illustrated in Figure 1, we propose a research agenda to develop collaborative methods for conducting application audits of ML models. The left portion of the figure represents the space of problematic model behaviors when it is deployed in a specific application; within this space, human auditors are better equipped to identify certain types of errors (shown as human $>$ AI), while AI auditors are more effective at detecting others. We envision a *Mixture-of-Experts (MoE) auditing system* that leverages the complementary strengths of human and AI auditors. To design such a system, we propose three key research thrusts: (1) *Characterizing how human experts conduct audits* to identify components of the auditing process that are better suited for automation; (2) *Developing AI auditing algorithms* to automate the components identified in Thrust 1; and (3) *Designing a Mixture-of-Experts framework* that enhances the AI algorithms developed in Thrust 2 by strategically incorporating human expertise to improve audit effectiveness. Next, we briefly discuss each of these research thrusts.

Characterizing How Humans Conduct Audits

To design a collaborative algorithm that effectively leverages the complementary strengths of humans and AI, we must first understand how humans conduct audits. Studying human auditing processes is crucial for developing AI auditing algorithms that align with and complement human auditing strategies. While prior human-centered research has explored how humans conduct audits (Balayn et al. 2022), a systematic analysis of the complementary strengths of human and AI auditors in this process remains absent. To address this gap, we propose mixed-methods analyses of human auditing strategies. By employing qualitative data collection methods, such as interviewing human auditors, we can gain deeper insights into their decision-making processes that can be used for feature design of a data-driven AI auditor. On the other hand, quantitative data collection can provide the necessary training and validations datasets for training AI auditors. We anticipate two key challenges: documenting and organizing human auditing strategies into structured rules; and representing these strategies in an algorithmic framework to enable AI auditors to effectively complement human expertise.

[1]While Human-AI Collaboration largely caters to assisting humans in established tasks (e.g., healthcare, manufacturing), we explore its use for safe and responsible deployment of ML models.

AI Algorithms for Automated Audits

As noted in the introduction, relying on heuristics to select maximally informative test cases can lead to an inefficient use of domain experts' time. Hence, while AI can assist with various aspects of the auditing process, one particularly promising application is automating test case generation with minimal human feedback. In our recent work (currently under double-blind review), we have developed reinforcement learning (RL)-based methods to automate test case generation. Our experiments validate this approach on both tabular and image datasets, including a real-world medical imaging dataset, demonstrating its effectiveness in automating test case generation. As we scale this method to larger ML models, a key research challenge lies designing the AI auditor's state representation within RL.

In our current method, the auditing policy utilizes a state that captures knowledge from past test cases, making it dataset-specific and limiting its generalizability across applications. A potential solution is to develop a more flexible state representation that captures higher-level abstractions. Meta-learning techniques could help train RL agents on diverse tasks, improving their ability to generalize auditing policies across domains. Another promising approach is leveraging model explanations (or chain-of-thought reasoning, when available) to augment the state representation with insights into the inner workings of the ML model.

Human-AI Collaborative Auditing

The proposed AI algorithms for automated audits, while beneficial, will be limited in their expertise and domain knowledge, especially when compared with domain experts. Building on the human studies on auditing behaviors and development of automated auditing algorithms, we next propose the development of Human-AI Collaborative methods for auditing ML models. A promising approach for realizing such a collaborative system is the Mixture of Experts (MoE) design, with a particular emphasis on the routing algorithm to effectively allocate tasks between AI and human auditors. However, several research challenges must be addressed, including the integration of human expertise into this algorithmic framework, the design of mechanisms to balance AI automation with human oversight, and the development of effective routing strategies for optimal task distribution between human and AI auditors. To tackle these challenges, insights from human-in-the-loop learning and existing MoE frameworks offer a good starting point, enabling dynamic task routing based on expertise and confidence levels. Incorporating these elements has the potential to lead to a robust collaborative auditing system that enhances both the accuracy and efficiency of application audits.

References

Balayn, A.; Rikalo, N.; Lofi, C.; Yang, J.; and Bozzon, A. 2022. How can explainability methods be used to support bug identification in computer vision models? In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–16.

Cummings, M. 2023. What self-driving cars tell us about AI risks. *IEEE Spectrum*.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*, 1–31.

Sheliemina, N.; et al. 2024. The use of artificial intelligence in medical diagnostics: Opportunities, prospects and risks. *Health Economics and Management Review*, 5(2): 104–124.

Yu, F.; Moehring, A.; Banerjee, O.; Salz, T.; Agarwal, N.; and Rajpurkar, P. 2024. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, 30(3): 837–849.

Zhang, J. M.; Harman, M.; Ma, L.; and Liu, Y. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 48(1): 1–36.