Trust-Enhanced AI for Education: A Proposal for Interpretability in Educational LLMs

Applicant: Megha Patel

Introduction

This proposal seeks to integrate advanced model interpretability algorithms into the AI for Education (AI4ED) project at Northeastern University, enhancing trust in super-capable Large Language Models (LLMs) for educational purposes. Under the guidance of Professor Nicholas Brown, the technical lead, and in collaboration with the Provost's Office developers, it aims to make AI tools more understandable and transparent. By doing so, we intend to bolster trust among faculty and students in AI-generated results, further democratizing access to cutting-edge AI technologies in education and increasing their adoption.

Background

The AI4ED initiative focuses on leveraging AI to create adaptive, interactive learning experiences tailored to individual student needs. Central to this effort is the use of Generative AI and chatbots for personalized teaching and learning analytics. This proposal seeks to extend the AI4ED project by incorporating model interpretability, ensuring that educational tools not only deliver personalized content but also operate in a manner that is transparent and trusted by users.

Objectives

- Work under the mentorship of Professor Nicholas Brown to navigate the complexities of AI in education.
- Coordinate with AI4ED developers at the Provost's Office to integrate interpretability features into AI4ED tools seamlessly.
- Develop interpretability algorithms to enhance model transparency, making the AI decision-making process clear and understandable.
- Foster a culture of trust and transparency in educational AI applications, ensuring educators and students can rely on AI-generated content.

- Promote AI literacy, increasing the educational community's understanding of AI mechanisms.
- Extend the AI4ED SmartyBots software - An open-source toolkit for Educational AI, incorporating interpretability and trust mechanisms to enhance its utility and adoption in educational settings.

Methodology

- Interpretability Research: Investigate existing interpretability frameworks suitable for educational contexts and adapt them for the specific needs of AI4ED.
- Algorithm Implementation: Develop algorithms that can elucidate the reasoning behind the AI's responses, decisions, and content generation.
- Community Engagement: Work with educators and students to refine interpretability features based on feedback, ensuring they meet actual needs and enhance user experience.

Budget Allocation

- Stipend: $75K for personal support and research activities.
- Compute and Research Funding: $75K dedicated to computational resources, development of interpretability algorithms, and user testing.

Impact and Evaluation

- Educational Transformation: This project aims to revolutionize educational practices by making AI interactions more transparent, enhancing the effectiveness of personalized learning.
- User Trust: By ensuring the interpretability of AI tools, we anticipate increased adoption and trust in AI4ED technologies among faculty and students.
- Scholarly Contribution: The project will contribute to academic discourse on AI in education, particularly on the integration of interpretability to enhance trust.
- Diverse User Base: Northeastern University operates a network of 13 campuses across the U.S., U.K., and Canada, offering a range of unique experiences, opportunities, and connections to enrich and inform learning and research. Additionally, the university's academic offerings are extensive, encompassing more than 290 majors, demonstrating the university's comprehensive academic structure.

Conclusion

Integrating interpretability algorithms into the AI for Education project represents a significant step towards creating more transparent, trustworthy AI tools in educational settings. This initiative will not only advance the AI4ED project's goals but also serve as a model for incorporating interpretability and trust in educational AI systems worldwide. Megha Patel's proposal aligns with the mission of the Superalignment Fast Grants by addressing a crucial aspect of AI safety and alignment in the context of education.