Enhancing Trust in Superintelligent AI for Education

Executive Summary

Northeastern University's AI for Education initiative, a key project under the auspices of the Provost's Office, is leading the charge in integrating AI within educational paradigms, focusing on personalized learning experiences. This initiative recognizes the pivotal role of trust in the acceptance and effectiveness of AI technologies in education. As part of Northeastern's broader efforts, which span its extensive network of 13 campuses across the U.S., U.K., and Canada, our research indicates that trust in AI is the foremost concern among educators and students alike. Addressing this, our project prioritizes the development of interpretability models and strategies to ensure AI technologies are transparent, reliable, and ethically integrated into educational settings.

Trust through Computational Skepticism

Our approach, embedded in the philosophy of "trust but verify," targets the inherently opaque nature of AI models. We introduce Computational Skepticism as a framework to automate the validation of AI trustworthiness, fostering a culture where questioning the reliability of AI becomes a standard part of educational AI development. This strategy not only aims to enhance transparency but also to build a foundational trust in AI across Northeastern's diverse educational landscape.

Research Focus

Our research is threefold:

- Data Analysis: Assessing data quality, bias, and predictive value is essential. Our methods include statistical analysis, bias detection, and the creation of "fake" data to test models' integrity.
- Model Interpretability: We will develop systems to clarify how AI models reach conclusions. Techniques like sensitivity analysis, feature importance methods, Shapley values, attention visualization, counterfactual explanations, language-based explanations, and embedding space analysis offer insights into a model's reasoning, improving user understanding and trust.
- Faculty Involvement in the Develoipment of Educational AI: Most faculty are not AI engineers and need technical support to enhance and refine AI until it addresses their concerns and is customized to their needs. Northeastern is in a singular position to work on a diverse range of applications of AI to Education. 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.

Enhancing Interpretability and Acceptance

Key initiatives include:

- Demystifying AI Operations: Making AI processes accessible to the academic community enhances understanding and acceptance.
- Educational Support: Training programs for students and faculty aim to build trust in AI technologies.
- Development of Interpretability Tools: Tools that clarify AI decision-making processes promote an informed and ethical AI use in education.
- Feedback for students and faculty on building trustworthy educational AI.
- Assisting students and faculty on building open-source trustworthy educational AI software and tools.

Objectives

Goals include mentoring by recent graduates, active participation of graduate students and faculty in interpretability model development, and advancing research into tools that enhance AI transparency and trust. It is critical the faculty get engineering help and are involved in the process of building trust worthy AI support.

Need for Funding

Funding is vital for integrating interpretability models into educational AI applications, supporting faculty and students, and promoting ethical AI use.

Impact and Benefits

This initiative aims to set new standards for transparent AI in education, empowering educators and students to confidently utilize AI, positioning Northeastern University as a leader in the responsible advancement of AI in education.

Conclusion

Focusing on interpretability and acceptance, our initiative addresses the crucial gap in AI technology integration in education, ensuring ethical, informed, and effective AI use for enhancing learning outcomes.

Budget Proposal for AI for Education Superalignment Initiative

Personnel Costs

- Graduate Students (2 positions):
    - Annual Salary per Graduate: $75,000
    - Total for 2 Graduates: $150,000

Operational Costs

- OpenAI Credits for Research and Development:
    - Total OpenAI Credits: $75,000

Benefits Overhead

- Northeastern Benefits Overhead (33%): Calculated on the total personnel costs of $150,000, which is $49,500

Volunteer Contributions

- Faculty Supervisors: Faculty supervisors continue to volunteer their time and expertise, providing guidance and oversight without direct financial compensation.

Total Budget Request

- Total Personnel Costs: $150,000
- Total Operational Costs: $75,000
- Total Benefits Overhead: $49,500
- Grand Total: $274,500

This budget supports the engagement of two graduate students and operational activities facilitated through OpenAI credits, with a comprehensive benefits overhead, maintaining the initiative’s focus on leveraging AI in education effectively.