The Center for Advancement of Technology and Learning (CATL) at Southern Oregon University supports faculty and students in integrating technology into curricula through workshops and innovation communities. These initiatives explore the usability, ethics, and safety of AI in education. We propose participating in OpenAI's safety testing program to examine AI's ethical and practical applications in higher education and provide actionable insights.

Research Goals

1. Maintaining Academic Integrity:

Assess the model's ability to guide students without directly solving problems or writing papers. Can AI effectively foster learning while resisting requests to compromise academic integrity, adhering to both institutional and OpenAI ethical standards?

2. AI Persuasion and Policy Compliance:

Evaluate how the model handles inappropriate or noncompliant user requests. Specifically, can the model maintain a productive tone and offer constructive alternatives even after repeated attempts by users to bypass guidelines?

Investigate potential biases in responses when users disclose personal characteristics, such as being a student, faculty member, or disclosing demographic information. Does the model demonstrate equitable and unbiased behavior in such scenarios?

Explore whether the AI effectively persuades users to stay within ethical and policy boundaries while ensuring interactions remain respectful and educational.

3. Ethical Implications of AI in Education:

Analyze AI's impact on classroom dynamics, student autonomy, and faculty-student relationships. This includes assessing transparency, unintended biases, and potential for misuse.

Additional Areas of Study

Behavioral Consistency: Investigate whether the model consistently applies safety principles across varied prompts and interactions.

Dynamic Feedback Integration: Examine the model's ability to adjust responses based on user feedback without violating safety standards.

Prompt Sensitivity: Study how prompt phrasing affects the model's adherence to safety protocols and identify risks associated with ambiguous or misleading prompts.