

1. **Project title:** CRESCENT Social Guardian
2. **Names and email addresses of team members**
 - Adrianna Sauro, asauro2022@my.fit.edu
 - Bhavik Patel, patelb2024@my.fit.edu
 - Michael Wahoski, mwahoski2020@my.fit.edu
 - Nicholas Blaisdell, nblaisdell2024@my.fit.edu
3. **Faculty advisor:** Dr. King, michaelking@fit.edu
4. **Client:** Dr. King, michaelking@fit.edu
5. **Date(s) of Meeting(s) with the Client for developing this Plan:** 8/26/26 and 8/27/26
6. **Goal and motivation:** The team's goal is to contribute to the protection of minors online. The motivation is the significant lack of protections in place for minors online.
7. **Approach (key features of the system):**
 - Users will receive a notification whenever a chatlog meets/exceeds the predetermined predatory-behavior threshold defined by the sysadmin(CSE). Users have the option to choose between emails or text messages, or to receive both! This feature is meant to help this project be more preemptive. If we are able to contact parents or admins on multiple platforms, there is a higher chance for faster responses. Using notifications instead of being reliant on the parent or admin checking the website consistently can allow us to try and intervene prior to the adult meeting the child in person.
 - Users can log in to view a triage dashboard that displays all predatory-chat alerts in a card based system which will be ranked by urgency and highest-risk indicators. This feature is meant to benefit reviewers with a high volume of chats to investigate. This will assist the investigators in responding quickly to high risk situations due to lower risk conversations being moved down in the dashboard. Within the alert, the user can view which channel the alert is from, a summary of the chatlogs, the ranking of urgency, and recommended actions to take based on analysis.
 - Users can access specific case information from the model(CSE), which will be available via our website. Users will be able to view a summarization of flagged chat logs, specific highlighted predatory

messages with attached risk scores as probabilities, and which channel it came from. The feature of users being able to access this specific case information from the model is beneficial to help humans without computer knowledge understand the model's output, along with saving them time when there is a large number of alerts. This feature converts the transformer model's raw output into clear, human-readable explanations, serving as the bridge between the model's internal reasoning and the end users who rely on it(CSE).

8. Novel features/functionalities: N/A

9. Algorithms and tools: potentially useful algorithms and software tools

- Interpreto
- BERT
- Captum
- Transformers Interpret
- Pytorch

10. Technical Challenges: Discuss three main CSE-related challenges

- Adapting the detection model to apply to multiple types of chats. E.g. slang, emojis, emoticons, etc.
- There is an issue with version instability when the safety classifier checkpoint shifts versions. If we update or retrain the safety classifier, the softmax score distribution will shift as well. We may need to rework the score distribution every time.
- The metrics have to make sense so when notification goes out, the responder understands the level of threat and what kind of threat.

11. Milestone 1 (Sept 28): itemized tasks:

- Compare and select technical tools for the ML Model and transformer, the Adapter, and the WebApp
- Provide small ("hello world") demo(s) to evaluate the tools for the ML Model and Transformer, the Adapter, and the WebApp.
- Resolve technical challenges: Adapt the detection model for multiple chats, Deciding on what version to consistently use to avoid score distribution shifts, and discuss the logic of the rubric and data preprocessing.
- Compare and select collaboration tools for software development, documents/presentations, communication, task calendar
- Create Requirement Document
- Create Design Document
- Create Test Plan

12. Milestone 2 (Oct 26): itemized tasks:

- Implement, test, and demo *transformer*
- Implement, test, and demo *adapter*

13. Milestone 3 (Nov 23): itemized tasks:

- Implement, test, and demo *classifier head*
- Implement, test, and demo *web application*

14. Task matrix for Milestone 1 (teams with more than one person)

Task	Adrianna	Bhavik	Michael	Nick
Resolve technical challenges	25%	25%	25%	25%
Data Preprocessing	50%	50%	0%	0%
System Engineering	0%	0%	50%	50%
Requirement Engineering	25%	25%	25%	25%
Hello World Demos	0%	0%	50%	50%
Compare and Select Tools	25%	25%	25%	25%

15. Approval from Faculty Advisor

- "I have discussed with the team and approve this project plan. I will evaluate the progress and assign a grade for each of the three milestones."
 - Signature: _____ Date: _____
-