



CRESCENT

SOCIAL GUARDIAN

Contextual **R**isk **E**valuation **S**ystem for **C**hild **E**xploitation in **N**etworked **T**ext

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

Faculty Advisor and Client: Dr. King, michaelking@fit.edu (meetings on 8/27/26 and 8/28/26)



Project Goal & 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.



Approach: 3 key features



Detection Model

- Users can access specific case information from the model via our website.
- Users will be able to view specific highlighted predatory messages with attached risk scores as probabilities.
- This feature converts the transformer model's raw output into clear, human-readable explanations.

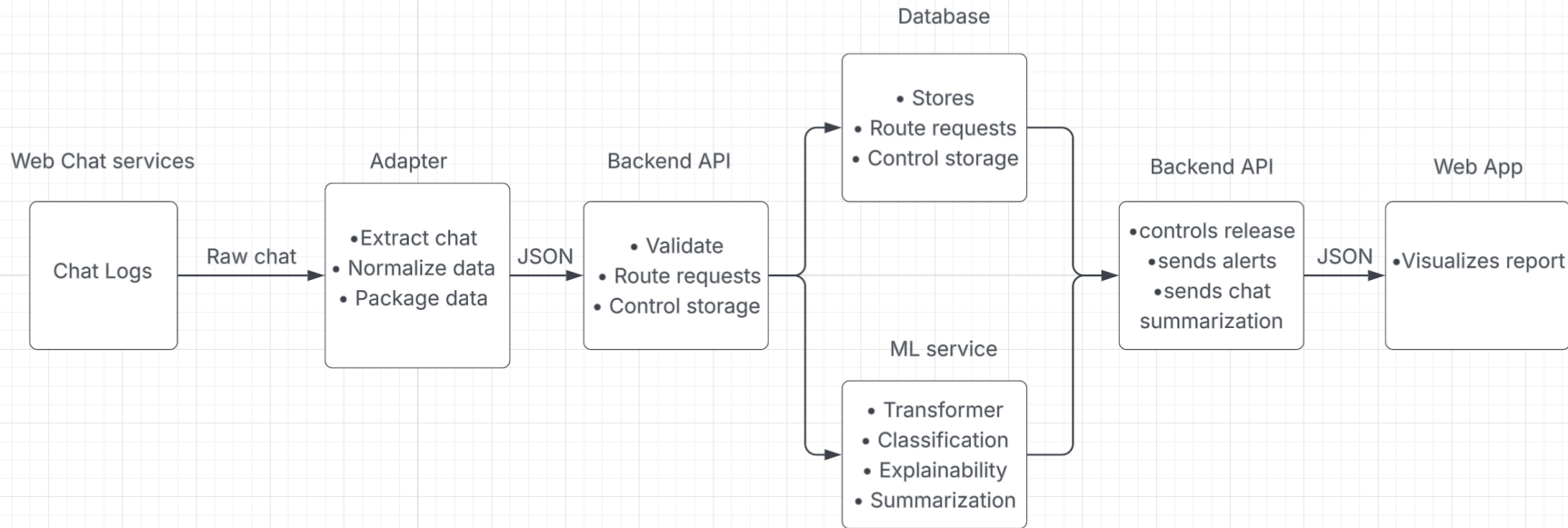
Triage Dashboard

- Users can log in to view a triage dashboard that displays all predatory-chat alerts in a card based system (ranked by urgency and highest-risk indicators).
- Within the alert, the user can view which channel the alert is from, a summarization of flagged chat logs, and recommended actions to take based on analysis.

Notification System

- Users will receive a notification whenever a chatlog meets/exceeds the predetermined predatory-behavior threshold defined by the sysadmin.
- Users have the option to choose between emails or text messages, or to receive both!

System Architecture

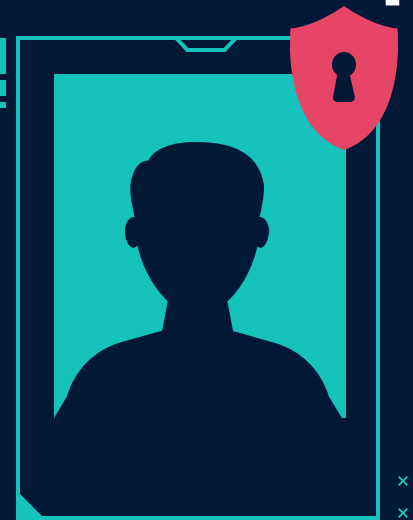




Algorithms & Tools

- BERT
- Transformers Interpret
- Pytorch
- Interpreto
- Captum
- BART

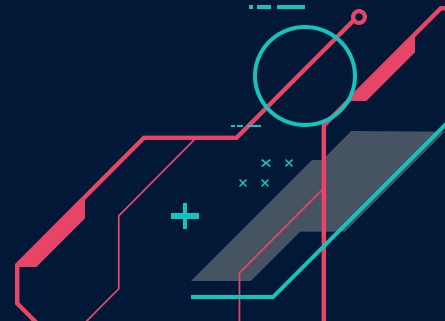
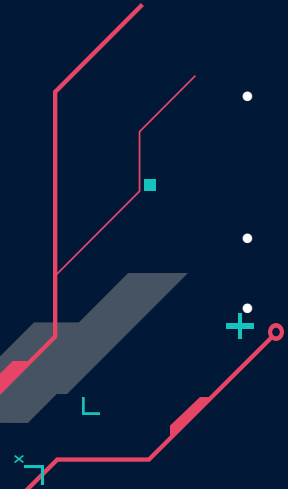
Technical Challenges



- 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.

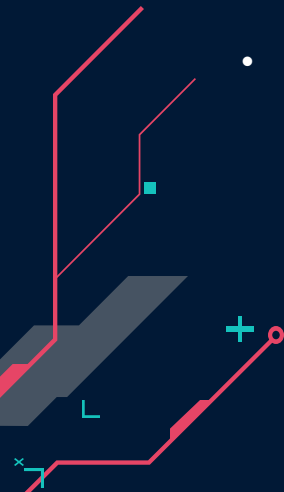
Milestone 1 (Sept 28)

- 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
- Systems engineering
- Requirements Document/Plan Document/Design Document



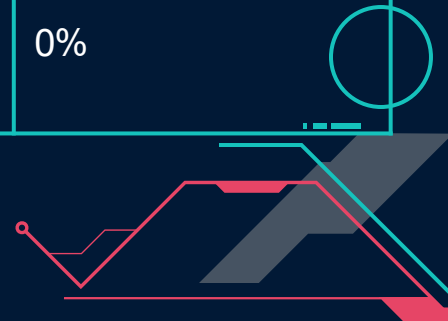
Milestones 2 (Oct 26) & 3 (Nov 23)

- Milestone2
 - Implement, test, and demo transformer
 - Implement, test, and demo adapter
- Milestone 3
 - Implement, test, and demo classifier head
 - Implement, test, and web application



Task matrix for Milestone 1

Task	Adrianna	Bhavik	Michael	Nick
Resolve technical challenges	25%	25%	25%	25%
System Engineering	25%	25%	25%	25%
Data Preprocessing	50%	50%	0%	0%



THANKS!



CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

