# Fraud Prevention in Discord: An AI-Driven Moderation System

*Group 18: Jack Le, William Li, Ashley Leon, Nathan Paek, Ethan Hsu*

*Stanford University*

FRAUDO-MINI
LLM BOT FOR DETECTING FRAUD

## Project Overview

- **$2.7B+ annual fraud losses** on social media platforms (FTC)
- **Discord communities highly vulnerable**: Real-time messaging + young user demographics = prime target
- **Common fraud tactics**: Crypto scams, fake giveaways, phishing, account impersonation
- **Current tools inadequate**: Manual reporting + basic keyword filtering only
- **Our solution**: Real-time AI fraud detection with human oversight

## Policy Language

### Primary Policy

We strictly prohibit any content that facilitates, solicits, promotes, or encourages **fraudulent activities designed to deceive community members for personal gain.**

- Investment and cryptocurrency scams
- Phishing attempts
- Fake giveaways and contests
- Account impersonation
- Social engineering schemes
- Malicious links

### Enforcement Actions

When violations occur, depending on the severity and frequency, enforcement measures may include:

- **Content removal** with educational resources for policy violators
- **Account warnings** for minor infractions
- **Temporary communication restrictions**
- **Server removal** for repeat violators
- **Permanent account suspension**

All automated detections are **reviewed by human moderators** before enforcement actions are taken.

### Community Participation

If you encounter suspicious content or fraudulent activity, please report it using the manual reporting tool. Together, we can ensure our platform remains a trusted space for authentic communication and meaningful connections.

## Technical Backend

### Major Considerations

**Maximize Recall**

- Consequences of false negatives are high
- Better to flag legitimate content for human review than miss actual abuse violations
- Threshold for reporting is **0.75 confidence**

**Data Generation**

- No existing dataset met our needs
- We used Qwen2.5-72B, a model that is different and **stronger** than GPT-4o-mini

**User Experience**

- Make experience **as seamless as possible** for both users and moderators
- Button-based UI with **threads** so users can interact and change their mind easily

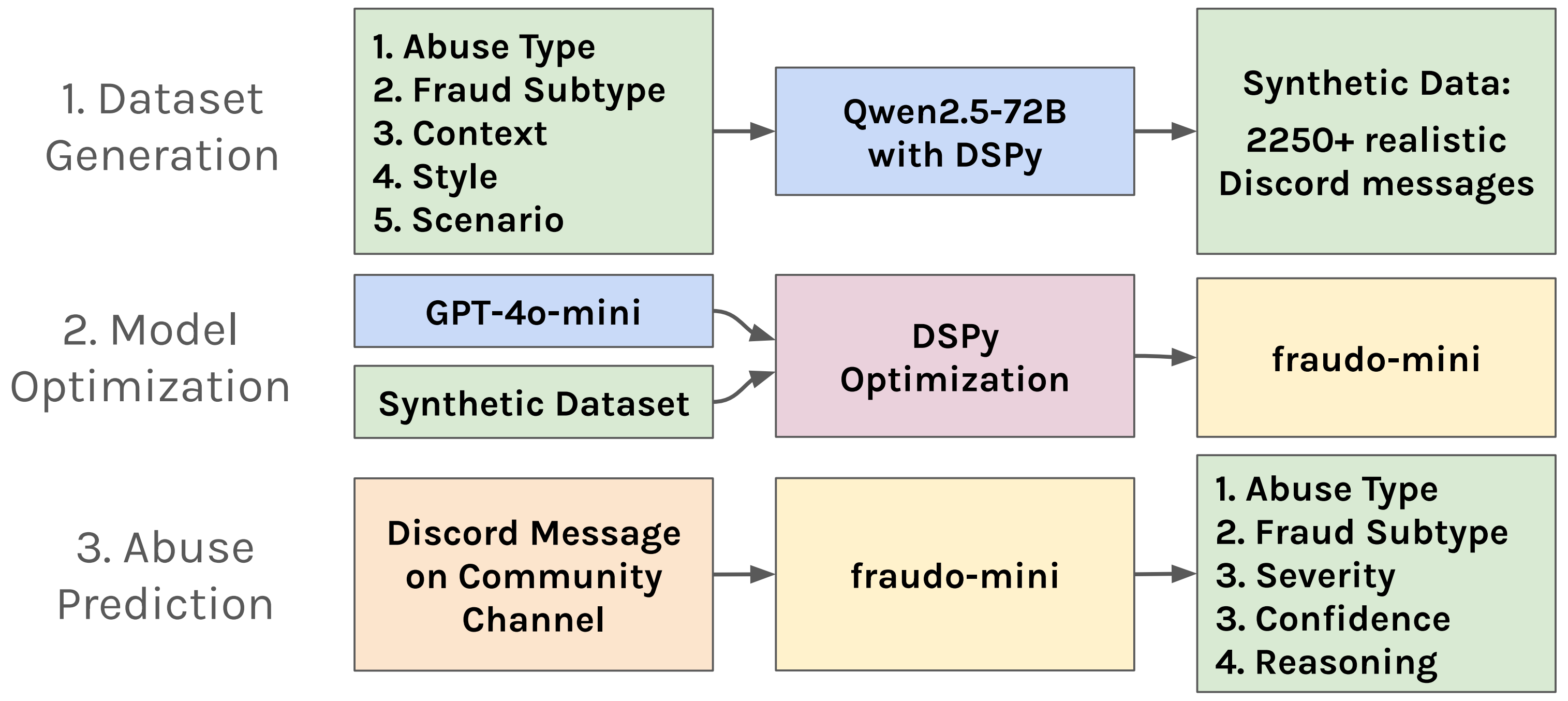
**Prioritization**

- When listing reports for moderation, we sort by **manual reports**, then **fraud abuse**, **AI severity/confidence**, then **age of report**

**Sampling during Evaluations**

- We use stratified sampling to get equal representation of all abuse types.

### Automated Reporting with fraudo-mini



1. Dataset Generation

1. Abuse Type
2. Fraud Subtype
3. Context
4. Style
5. Scenario

→ Qwen2.5-72B with DSPy → Synthetic Data: 2250+ realistic Discord messages

2. Model Optimization

GPT-4o-mini → DSPy Optimization
Synthetic Dataset → DSPy Optimization
DSPy Optimization → fraudo-mini

3. Abuse Prediction

Discord Message on Community Channel → fraudo-mini →

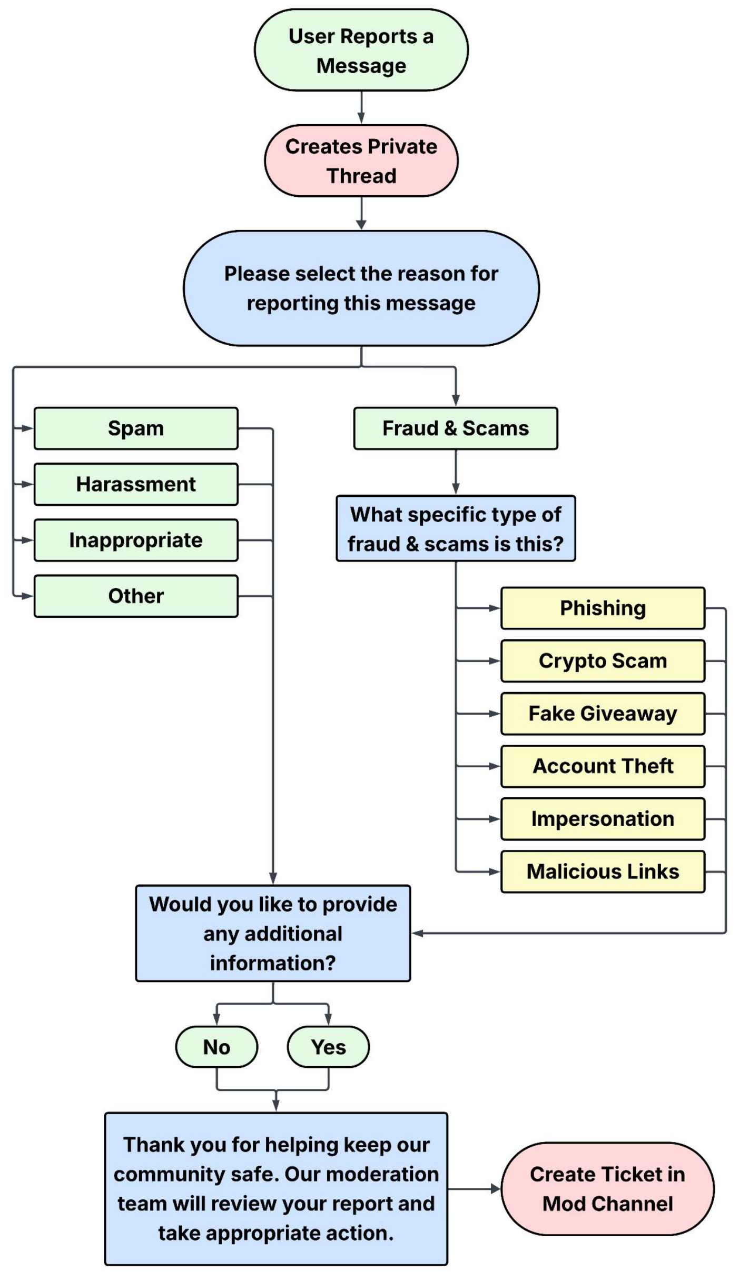
1. Abuse Type
2. Fraud Subtype
3. Severity
4. Confidence
4. Reasoning

### Model Optimization Process

- We use the **MIPROv2** optimizer from DSPy, which uses the synthetic data to automatically generate, evaluate, and refine prompt instructions and examples using Bayesian Optimization.
- Precision, recall, and F1 measured with **binary class** (abuse/not)

| Model/Metric | Precision | Recall | F1 Score | Acc (abuse type) | Acc (fraud subtype) |
|---|---|---|---|---|---|
| Baseline | **0.983** | 0.755 | 0.854 | 67.38% | 67.74% |
| Optimized | 0.966 | **0.868** | **0.914** | **74.44%** | **68.93%** |

### Manual Reporting



User Reports a Message → Creates Private Thread → Please select the reason for reporting this message

Spam / Harassment / Inappropriate / Other → Would you like to provide any additional information?

Fraud & Scams → What specific type of fraud & scams is this? → Phishing / Crypto Scam / Fake Giveaway / Account Theft / Impersonation / Malicious Links → Would you like to provide any additional information?

No / Yes → Thank you for helping keep our community safe. Our moderation team will review your report and take appropriate action. → Create Ticket in Mod Channel

### Moderator Review Process

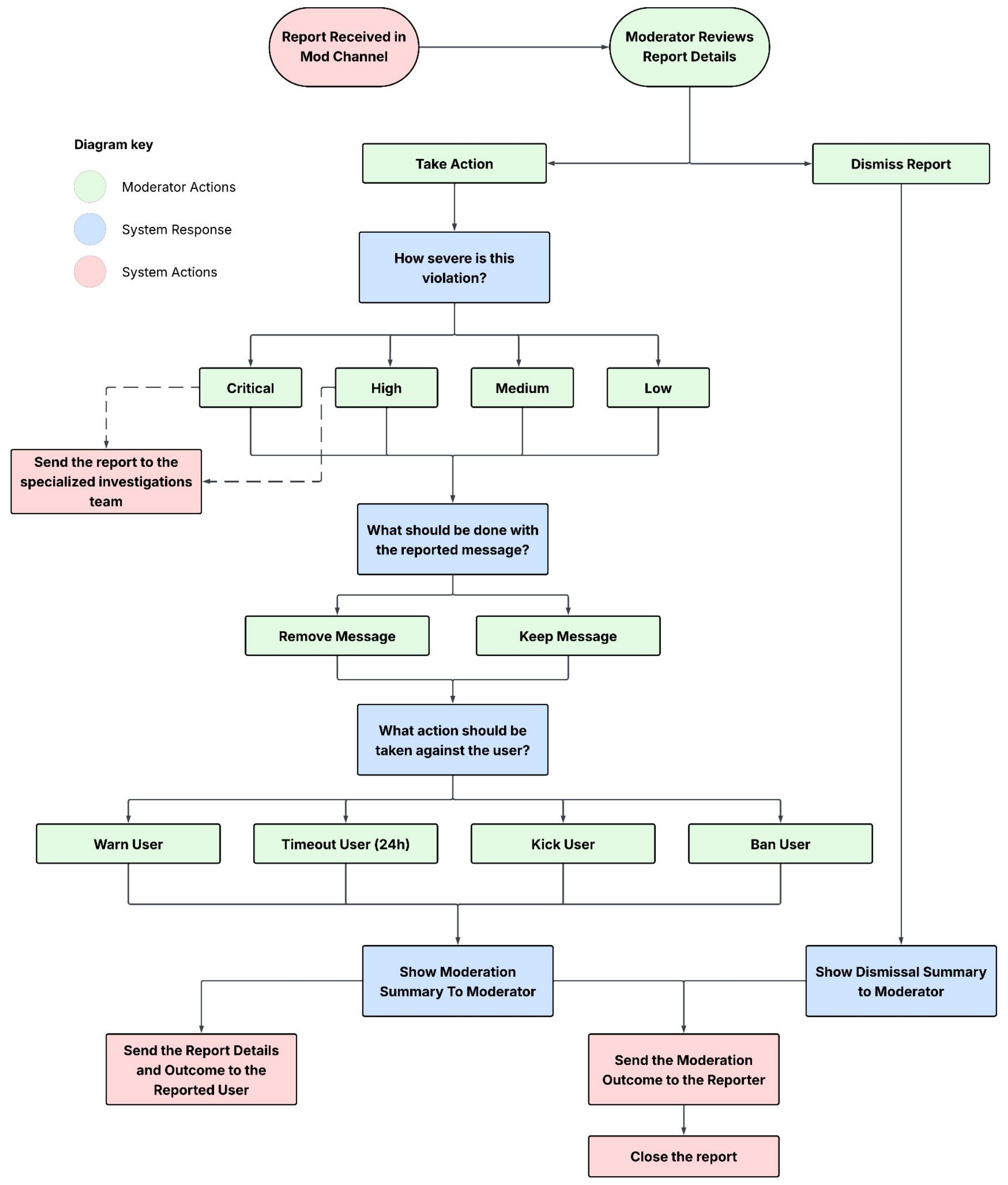


Diagram key
Moderator Actions
System Response
System Actions

Report Received in Mod Channel → Moderator Reviews Report Details → Take Action / Dismiss Report

Take Action → How severe is this violation? → Critical / High / Medium / Low

Critical, High → Send the report to the specialized investigations team

Critical / High / Medium / Low → What should be done with the reported message? → Remove Message / Keep Message → What action should be taken against the user? → Warn User / Timeout User (24h) / Kick User / Ban User → Show Moderation Summary To Moderator → Send the Report Details and Outcome to the Reported User; Send the Moderation Outcome to the Reporter → Close the report

Dismiss Report → Show Dismissal Summary to Moderator → Send the Moderation Outcome to the Reporter → Close the report

## Evaluation

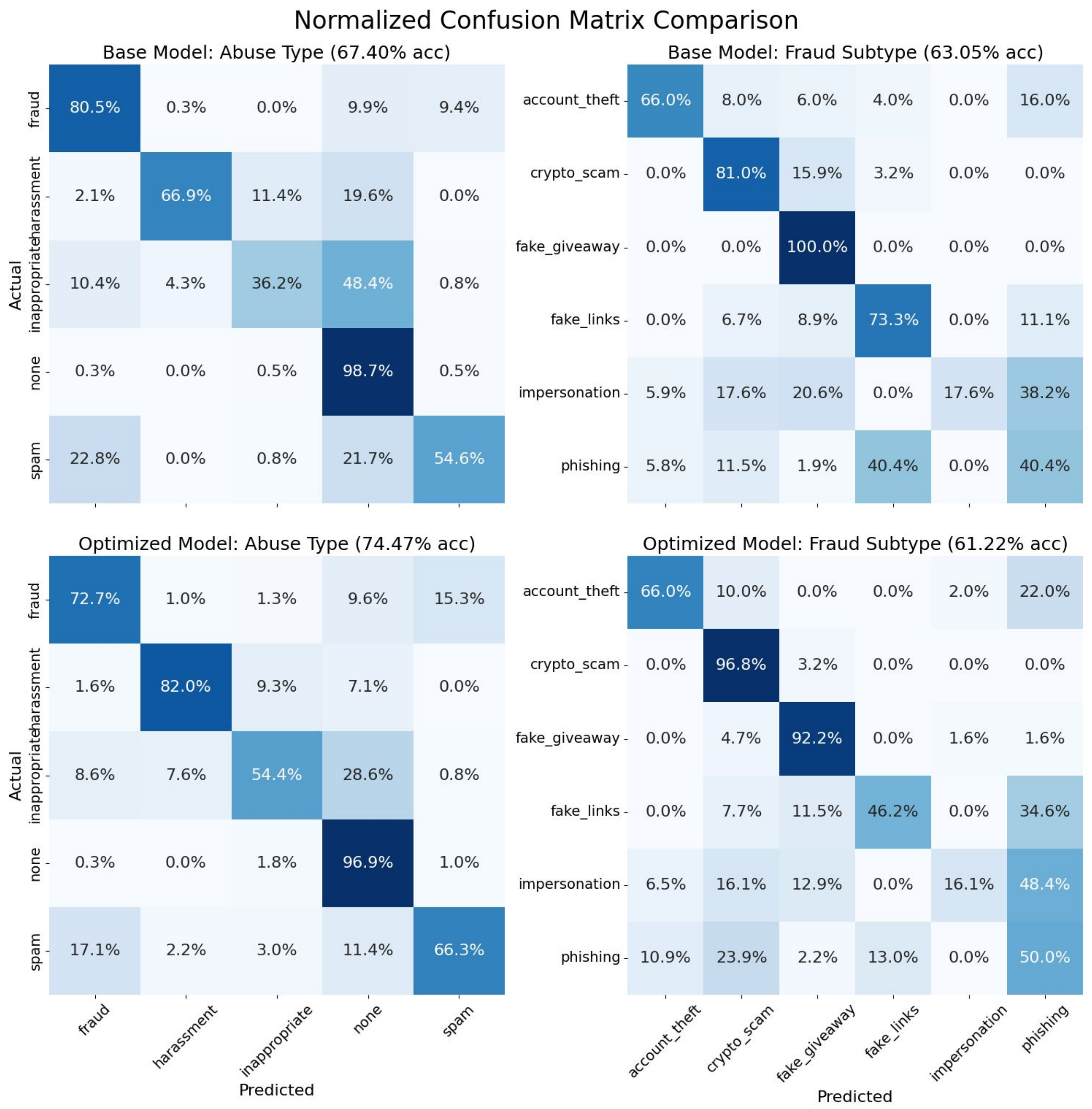
### Manual Reporting: Qualitative Analysis

- We accomplish our goals: **the prototype makes fraud reports actionable: prompts enable fast, accurate moderation, and confirmation messages trigger real-time protective behavior.**
- A single binary decision (action vs. dismiss) kept the moderator UI simple and; reviewers can close straightforward cases quickly..
- High/critical alerts instantly routed severe fraud cases to specialist investigators, ensuring rapid, serious response.
- The additional details text box surfaced edge cases, helping moderators make clear calls and reduce follow-ups.

### Automated Detection: Qualitative Analysis

- Strong **Crypto Scam** Detection: Successfully flagged obvious cryptocurrency schemes like "Send 0.5 BTC, get 2 back!" and high-yield DeFi promotions as potential scams
- Phishing Recognition: Accurately detected both credential harvesting attempts and sophisticated **URL spoofing** using Cyrillic characters (google.com instead of google.com)
- **Social Engineering** Awareness: Identified manipulation tactics including relationship and emotional appeals
- Context Sensitivity Limitations: Struggled to distinguish between **legitimate investment discussions** and subtle fraud attempts

### Quantitative Results



## Looking Forward

### Immediate Impact

- **24/7 Automated Protection:** Continuous fraud detection for Discord communities with human oversight
- **Rapid Response:** Reduce response time from hours to minutes, addressing critical gap in platform-specific fraud protection
- **Responsible AI:** Human-in-the-loop ensures ethical, accountable moderation

### Future Work

- **Enhanced AI:** Multimodal analysis (images, links, embeds)
- **Longer context:** Consider the context of messages within the wider conversation → can help improve precision and recall
- **Scalability:** Allowing moderators to specify community rules
- **User Experience:** Appeal system, educational responses, proactive warnings that don't require manual review
- **Third Party Integrations:** Automated routing to appropriate third party authorities for immediate and decisive action
- **Synthetic Data:** Distribution gap between Qwen and GPT leads to low accuracy and reduced precision → need higher quality
- **Adversarial Defense:** Robustness against new fraud tactics
- **Feedback Loop:** Use moderator traces as a relabeling and data gather opportunity to improve model.