FacadeGuard: A Multi-Turn Impersonation Scam Detection System

Team 23: Komal Vij, Breanna Burke, Alice S. Kim, George Hasnah, Arvind Salem

Problem Description

Multi-turn impersonation scams are a growing class of fraud in which a malicious human actor creates or hijacks a social media account to impersonate a trusted individual - most commonly a celebrity, public figure, or a personal acquaintance. Over the course of an extended conversation, the scammer builds trust and manipulates the victim into sending money or personal information. These scams are a widespread phenomenon on social media, especially in the form of investment and romance scams through platforms that facilitate direct messaging.

Victims of impersonation scams can broadly include any user of a social media platform, but vulnerable populations such as the elderly (ages 50+) are particularly susceptible. Unlike traditional spam or phishing, impersonation scams evolve gradually across multiple seemingly harmless messages, making it difficult for existing moderation systems to detect. Through this project, we aim to detect high-risk conversational patterns in social media DMs before financial harm occurs.

Impersonation Scam Policy

Overview:
Our platform prohibits impersonation of any individual or organization with the intent to deceive. This includes creating fake accounts using another person's name, likeness, or photos to solicit money or personal information.

With the rise of AI-assisted fraud, impersonation scams have become faster, more personalized, and harder to detect, while causing great financial, personal, and emotional harm to victims. To combat this behavior, our platform uses automated scam detection systems and user reporting tools, alongside human moderation review.

Common classes of impersonation scams and fraudulent behaviors prohibited by our platform include, but are not limited to:

- Celebrity or influencer impersonation scams– posing as a public figure to solicit payments or private information.
- Romance scams– building false emotional trust before fabricating an emergency requiring money
- Investment scams– impersonating a financial mentor and sharing fabricated success stories to solicit payments.

An extended conversation is not required for content to be prohibited under this policy. A single message requesting money with an impersonating profile picture may still be subject to moderation enforcement.

If you violate this policy:
Depending on the severity of the violation and the previous history of violations, we reserve the right to:

1. Add automated warnings to suspicious conversations
2. Flag suspicious conversations for moderator review
3. Restrict accounts that show high-risk scam behavior
4. Escalate suspicious accounts for additional moderator investigation or identity verification review

Moderators can review prior account violation history when evaluating reports, and repeat offenders may be subject to harsher penalties.

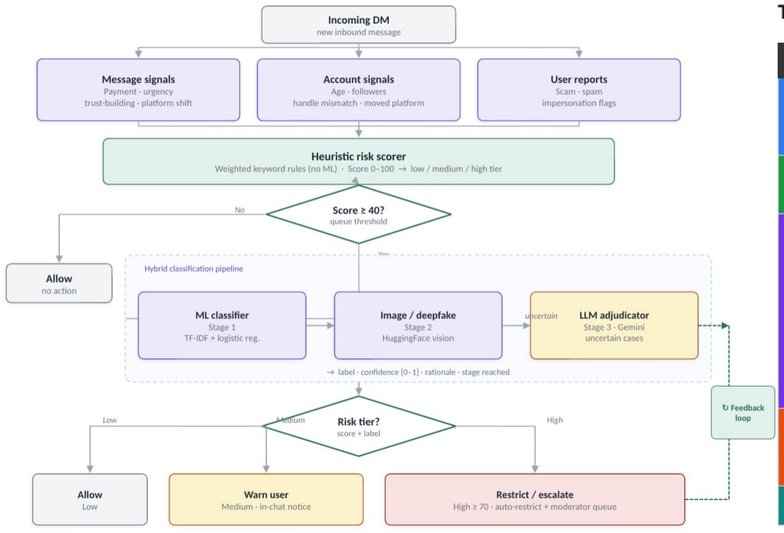
How to help:
If you receive a message that feels suspicious or attempts to impersonate another person, please use our in-app reporting system to flag the conversation for moderator review.

Looking Forward

Given more time and resources, we would pursue the following improvements:

- Expand test sets to cover scams across different languages, diversify impersonation scam scenarios beyond celebrity scenarios for product generalization.
- Conduct user studies with older adults (50+) who are disproportionately targeted to gather feedback on clarity of warning messages and reporting flows.
- **Privacy:** Automatically redact sensitive financial or personal information that a victim types during a scam conversation before it is routed to moderators (ex: Social Security number).
- **Adversarial concerns:** ML and heuristic classification stages can be evaded by scammers who learn signal keywords used for detection. Iterate and improve ML / heuristic classifier logic for adversarial robustness.
- **Scalability:** LLM classification will be costly at scale (median latency = ~ 2.2 s). A potential solution is running LLM classifiers in the background rather than aiming to detect and block each message by LLM in real-time.
- **Techniques not pursued:** Real-time deepfake image / video detection is implemented but untested in evaluation due to text-only test sets. Implement live ML retraining based on moderator action (currently simulated asynchronously).

Technical Backend



Technology Stack & Justification

Category	Technology	Justification
Frontend	Vanilla HTML/CSS/JavaScript	Allowed for quick prototyping without need for configuration considerations; lack of DOM layer allowed us to efficiently facilitate real time updates in UI
Backend	Node.js & Express.js	Allowed to use same language as frontend & seamlessly integrate with Vercel and Supabase (built around Node.js and JavaScript respectively)
ML / AI (Hybrid Classifier)	Python & JavaScript	Python is used for training (scikit-learn for TF-IDF vectorization, NumPy for ML calculations) and JavaScript is used for inference since server is Node.js
	Google Gemini 2.5 Flash	Gemini was most cost-efficient for tokens needed to analyze multi-turn conversations to build the LLM classifier
	HuggingFace (Llama 3.1 8B)	Free API allowed us to simulate demo chats and generate context for full conversation threads (necessary to evaluate multi-turn scams)
Database & Authentication	Supabase (PostgreSQL)	Allowed for easy implementation of row-level security to ensure users can't access moderator dashboard & also enabled Google OAuth for additional user protections
Hosting	Vercel	Allows us to automatically deploy updates that we push to main

- Incoming DMs are scored by a heuristic risk scorer which assigns weighted points for scam-related keyword signals across the entire conversation – for example, payment requests (+30) / platform shift attempts (+8). Anything below 40 points is immediately allowed with no further processing.
- Messages above the 40 point threshold enter a three-stage hybrid pipeline: TF-IDF + logistic regression, HuggingFace vision for image/deepfake detection, and a Gemini LLM classifier call on the remaining uncertain cases.
- Risk-tier triggers action: Low (<40) = allow, Medium (40-69) = in-chat warning, High (≥70) = auto-restrict, as well as moderator escalation.

Evaluation

Test Set – 270 AI-generated Instagram DM conversations. 120 benign, 150 scam examples (generated based on real documented scam conversation examples from online public sources), all manually reviewed.

Which Approach Wins Where?

- **Hybrid** wins overall, and would most likely be the best choice for real-world deployment. Accuracy is matched to the LLM classifier on a 60% lower cost, and achieved a median latency of approximately 1 ms nearly identical to the ML classifier – which is the fastest. (Hybrid P99 latency = ~ 2.8 s)
- **ML** wins on speed and cost – on our narrow test set, it achieved perfect accuracy with $0 cost and 1 ms latency. This is likely best as a first-pass filter for a high number of cases.
- **LLM** wins on edge / ambiguous cases – LLM-as-classifier can handle ambiguous / unusual cases, and provides rationale for its decision. However, it has the highest cost at $0.057/1k and a 2.2s median latency, and would be impractical if used as a standalone classifier in deployment.
- **Heuristic** classifier alone is insufficient – 88.7% recall. 1 in 9 scams are missed. Subtle scams with no explicit keyword signaling are able to pass through.
- **Asymmetric Decision Rule:** Missing real scam risks is more costly than false positives, so we set a lower threshold for scam classification to prioritize recall. (In the hybrid pipeline, ML needs 92% confidence to classify benign, and only 55% confidence to classify as scam.)

Feedback Loop: Moderator decisions made on the moderator dashboard persist to the Supabase data store and feed back into our classifiers.

1. The ML classifier is re-trained on the labels updated by the moderator actions – improved F1 from 96.8% -> 100% and precision from 93.8% -> 100%.
2. The previous 10 moderator decisions are injected as examples before the LLM classifier decides on a new case.

Limitations

- Classifiers trained and tested on a narrow, AI-generated dataset – 100% F1 score on ML / Hybrid is most likely unrealistic for real-world deployment.
- Datasets contain conversations only in English; the scam set skews to celebrity impersonation scenarios as a data stand-in for the broader category of impersonation scams.

Method	Precision	Recall	F1	Median Latency	P99 Latency	Cost / 1k
Heuristic	99.3%	88.7%	93.7%	1 ms	8 ms	$0
Traditional ML	100%	100%	100%	1 ms	5 ms	$0
Hybrid	100%	100%	100%	1 ms	2,816 ms	$0.023
LLM	100%	100%	100%	2,241 ms	6,393 ms	$0.057

Table 1. (Left) Detection Performance and Deployment Cost by Classifier

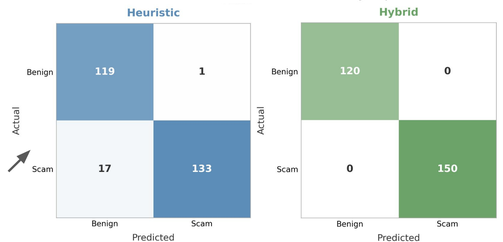


Figure 1. (Right) Confusion Matrices: Heuristic v.s. Hybrid Classifier