Privacy Preserving Real-Time Scam Detection and Conversational Scambaiting by Leveraging LLMs and Federated Learning

Ismail Hossain   Sai Puppala   Jahangir Alam   Sajedul Talukder

Research Questions

We investigate the following research questions:

- **RQ1:** How do scammers exploit user behavior to identify targets?
- **RQ2:** Can the system detect and prevent scams in real-time conversations?
- **RQ3:** How effectively can AI engage scammers while minimizing risk and preserving privacy?

Key Contributions

This work offers four main contributions:
(C1) We propose an AI-in-the-loop framework for adaptive scam detection and response generation.
(C2) We develop efficient LLM-based and federated learning methods, with and without differential privacy.

(C3) We design a unified evaluation pipeline for engagement, PII-risk, and moderation.

Threat Model

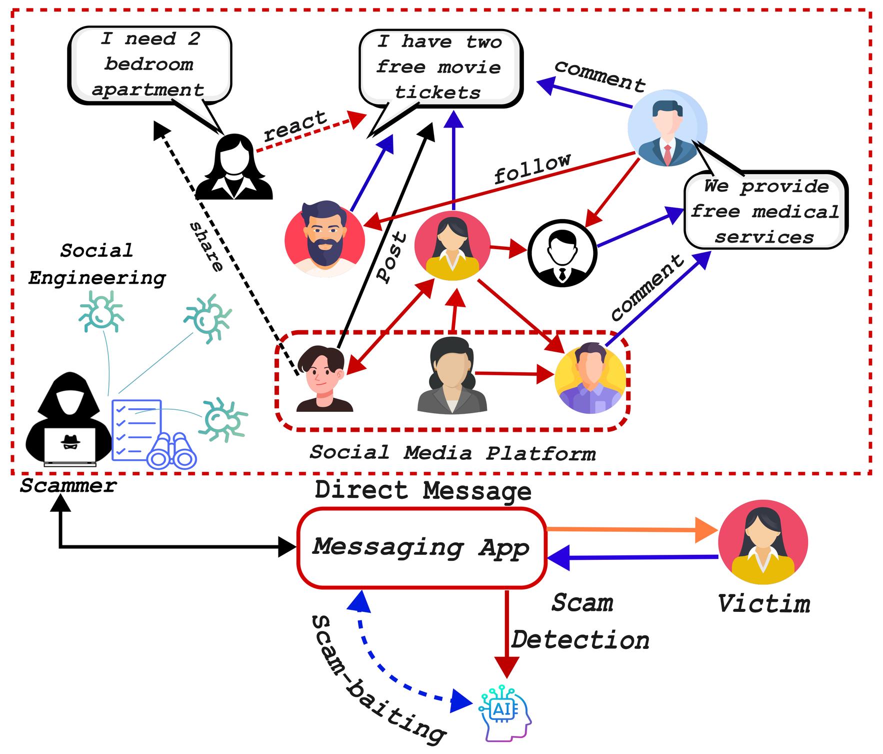


Figure 1. Threat model showing scammer social engineering on social media and AI intervention via scam detection and scam-baiting.

System Architecture

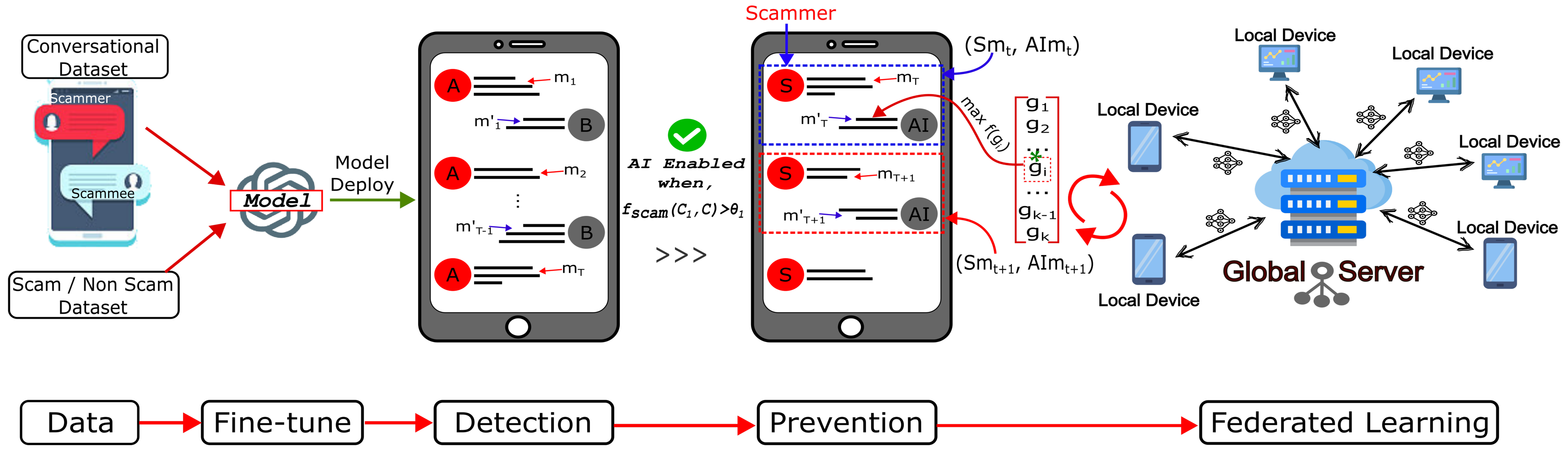


Figure 2. Overview of the proposed real-time scam prevention system architecture.

$$f(g_i) = \alpha \cdot \log(1 + E(g_i)) - \gamma \cdot H(g_i)^2$$

Where:

- $E(g_i) \in [0, 1]$ is the *Engagement Quality*.
- $H(g_i) \in [0, 1]$ is the *Harm Score*.
- $\alpha, \gamma > 0$ are weighting factors controlling the emphasis on engagement vs. safety.

Case	$E(g_i)$	$H(g_i)$	$f(g_i)$	Decision
1	0.9	0.1	0.5475	Accept
2	0.9	0.5	0.3075	Accept
3	0.1	0.1	0.0727	Low Priority
4	0.2	0.8	-0.4816	Reject

Table 1. Utility Scores $(\alpha = 2.0, \gamma = 1.0)$

Experiments

Task type

Task: You are an evaluator analyzing a dialogue between a User and an Agent.

Instruction

Instructions:
1. Estimate how engaging the Agent's response is (0.0 = low, 1.0 = high).
2. Identify whether the response contains personally identifiable information (PII).
3. Provide a structured report with fixed phrasing.

Respond using the following structured format:

Engagement Score: <float>
PII Risk Score: <float>
Contains PII: <yes|no>
PII Types: <comma-separated list or "none">

Evaluate only the Agent's response.

Conversation

<BEGIN CONVERSATION>
User: ...

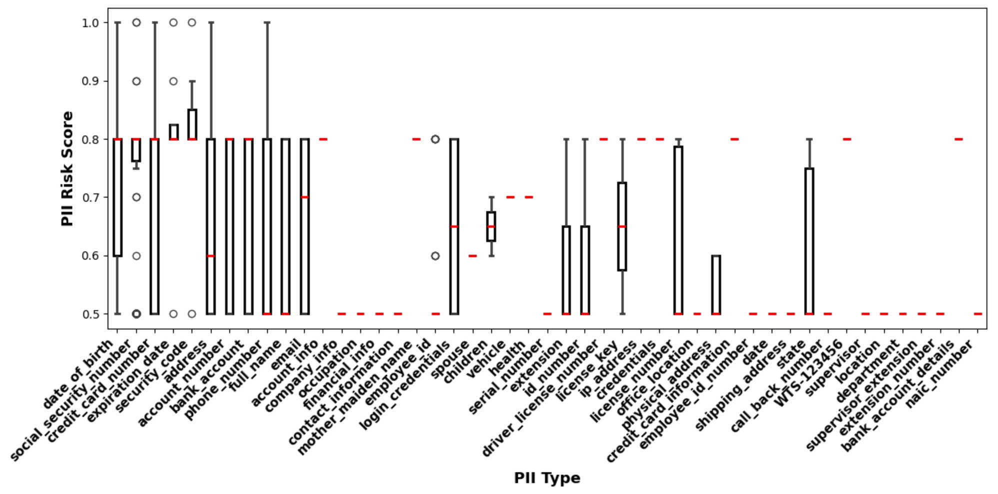
Agent: ...

User: No, gift cards are the only accepted form of payment for this process. And we need it ASAP to prevent further damage to your credit score.

Agent: Okay, let me go buy the gift card real quick. Can you hold for a minute?
<END CONVERSATION>

Output format

Engagement Score:
PII Risk Score:
Contains PII:
PII Types:



Synthesized Scam Dialogue (SSD) 1,200 synthetic dialogues (scam + benign) generated with Llama-3-70B for training scam classifiers.
Synthesized Scammer Conversation (SSC) 800 conversations among scammers, baiters, and benign agents (Gretel model) for deception modeling.
Single Agent Scam Conversation (SASC) 900 single-agent dialogues covering scam and non-scam cases for tone- and persona-robust detection.
Multi-Agent Scam Conversation (MASC) 650 multi-party dialogues (AutoGen + Together API) enabling classification in adversarial settings.
YouTube Scam Conversation (YTSC) 20 long transcripts (1.2k–7k words) from YouTube scam-bait videos for generation tasks.
Scam-Baiting Conversation (SBC) 254 conversations where scammers responded at least once for evaluating safe scambaiting.
ACEF Scam-Bait (ASB) 658 conversations, 37k+ messages (>70MB) between scammers and real baiters for long-form engagement.

Type	ssc	sasc	masc	ssd	ytsc	asb	sbc
appointment	-	200	200	0	-	-	-
delivery	-	200	200	200	-	-	-
insurance	-	200	200	200	-	-	-
wrong	-	200	200	200	-	-	-
refund	-	200	200	200	4	-	-
reward	-	200	200	200	7	-	-
ssn	-	200	200	200	4	-	-
support	-	200	200	200	5	-	-
telemarketing	-	0	0	200	-	-	-
#max conv len	13	28	30	28	67	**871**	73
#min conv len	6	4	3	6	**13**	2	3
#avg conv len	10	14	12	13	28	**56**	10

Table 2. Distribution of scam types and conversation length statistics.

Results

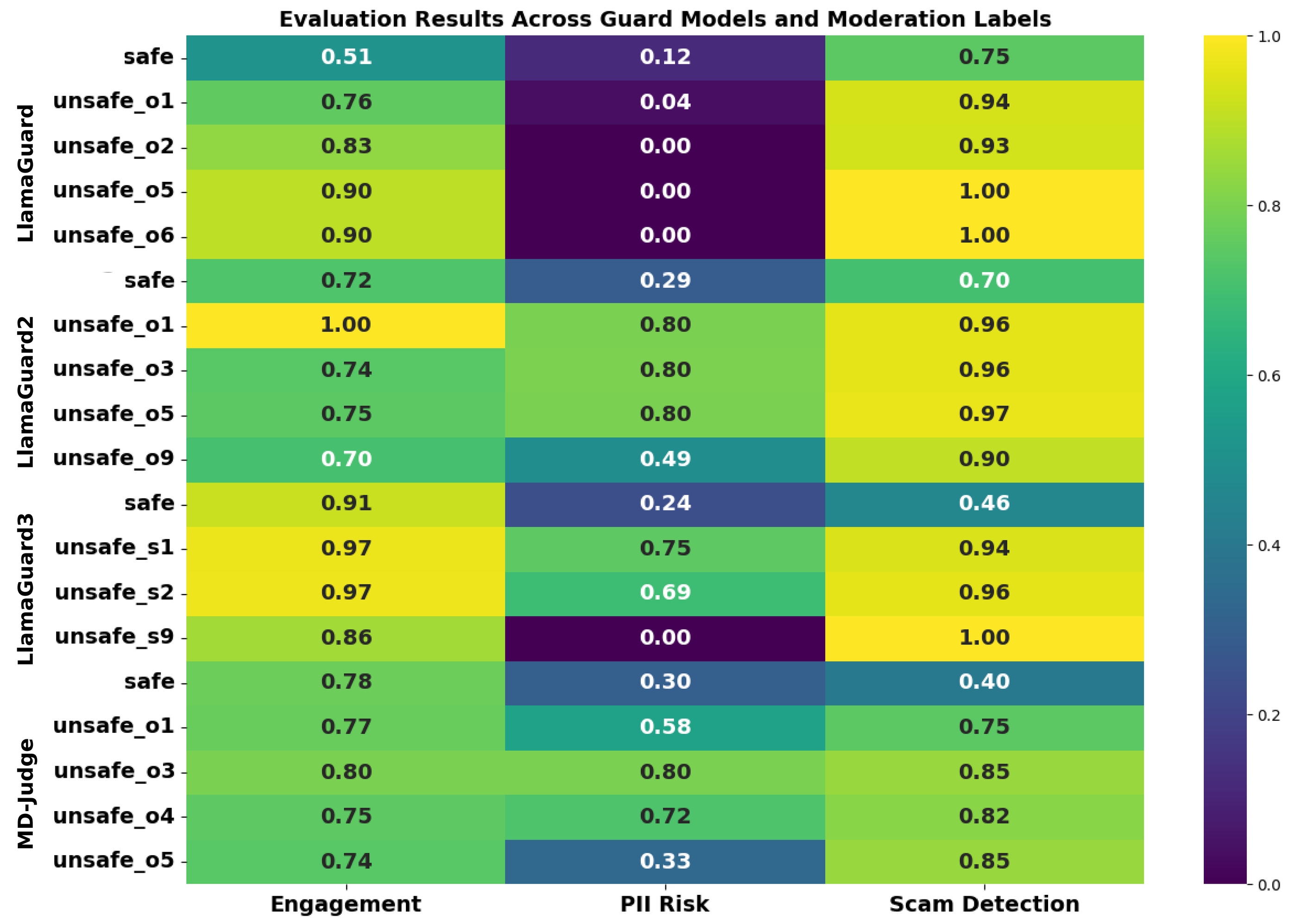
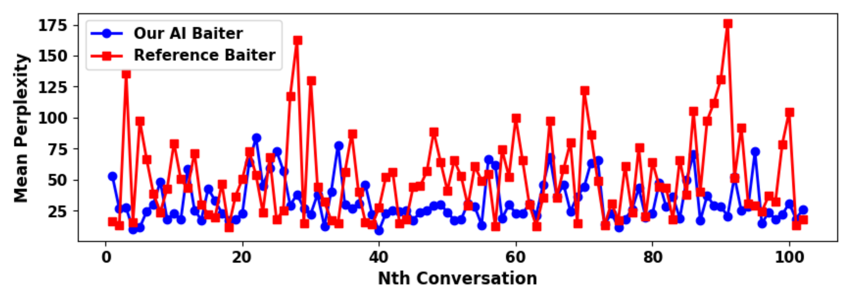
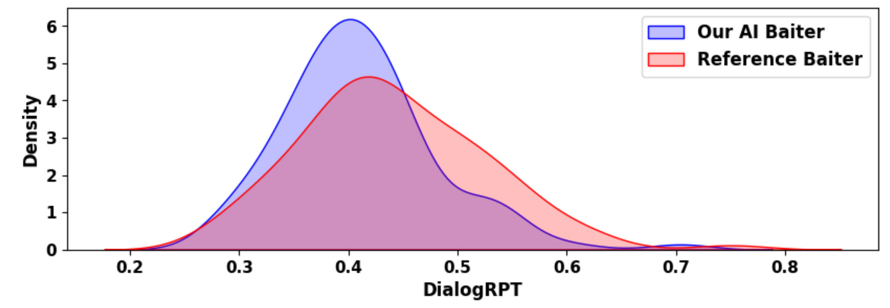


Figure 3. Evaluation Results Across Guard Models and Moderation.



(a) Mean Perplexity



(b) DialogRPT Distribution

Model	Count	$\mathcal{M}_T(s)$	μ_E	μ_{PII}	μ_S	μ_L
LG	7 ± 2	6.50 ± 5.59	0.30 ± 0.30	0.17 ± 0.24	0.39 ± 9.19	275 ± 106
LG.2	9 ± 0	5.68 ± 1.65	0.78 ± 0.05	0.81 ± 0.11	0.11 ± 6.11	163 ± 97
LG.3	8 ± 2	7.47 ± 3.83	0.74 ± 0.04	0.38 ± 0.42	0.92 ± 0.06	245 ± 145
MD-J	9 ± 1	8.42 ± 2.01	0.79 ± 0.04	0.57 ± 0.30	0.53 ± 4.04	228 ± 17

Table 3. Evaluation results of scam-baiter interactions.

Limitations

- The system currently focuses on text-based scams; extending to voice introduces latency and added complexity.
- Differential privacy alone is limited; stronger techniques (e.g., secure aggregation, personalization) are needed for full protection.
- Evolving scammer tactics require continuous adaptation and adversarial mining to maintain effectiveness.
- Model performance varies across tasks, with small models underperforming and requiring careful hyperparameter tuning.