UTEP COMPUTER SCIENCE

PhD Proposal Presentation

Evolutionary Dynamics of Online Behavior: Modeling Emotional Drift, Vulnerability, and Robust AI Defenses

SUPREME LAB | FALL 2025

Advisor: Dr. Sajedul Talukder

Presenter: Ismail Hossain

Committee Members:

Dr. Aritran Piplai

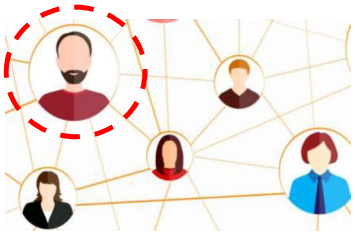
Dr. Anantaa Kotal

Dr. Md Fashiar Rahman

Outlines

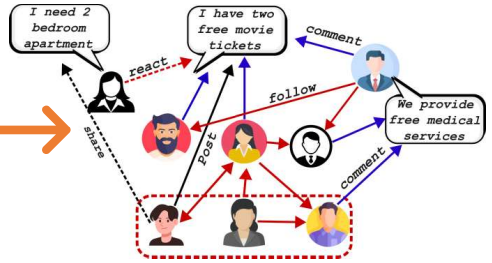
- Background
- Threat Model
- Problem Statement
- Research Questions
- Research Gap
- Research Objective
- Overview of Proposed Framework
- Preliminary Works
- Future Works
- Conclusion

Background



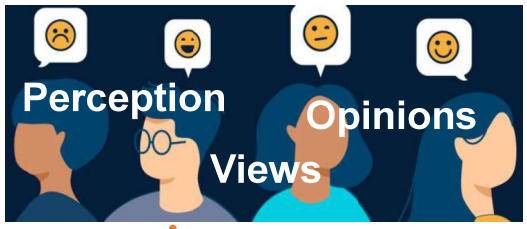
Initial Phase

Behavioral Shift



User Engagement

Emotional Shift

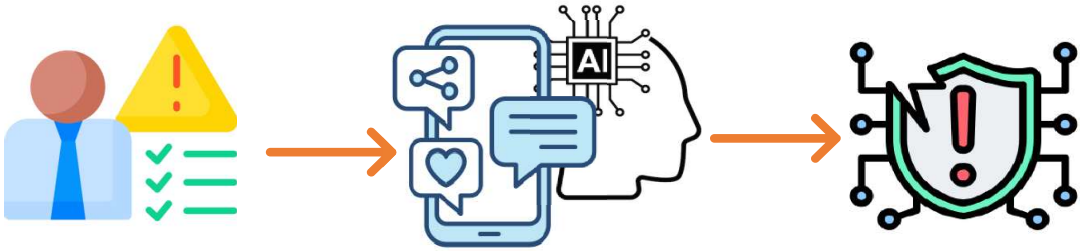


Sentiment Shift

User Activities

Trap by Adversaries

Background (Contd.)



Background (Contd.)

- User behavior in social media is **dynamic**, not static. So, users evolve due to:

Shifting Interests

Changing Emotional States

Demographic and life-stage Transitions

Influence from Connected Peers and Highly Influential Users

- These behavioral shifts appear in:

Posting History

Engagement Patterns

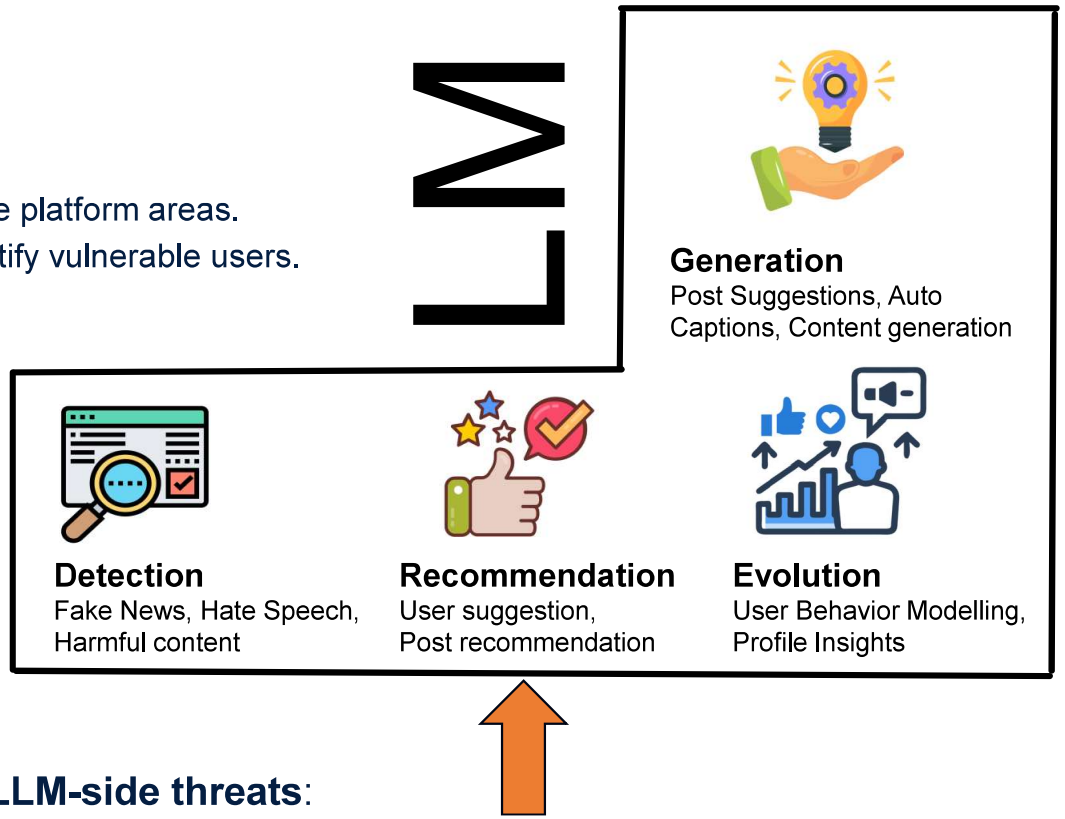
Comment Tone

Network Expansions

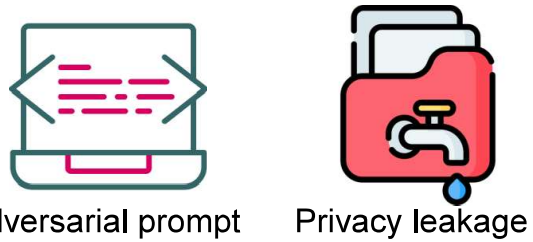
Threat Model

- Behavior-Driven Vulnerabilities on Social Media
  - Users' evolving activities leave traces across multiple platform areas.
  - Attackers monitor these behavioral footprints to identify vulnerable users.

- Examples of risk exposure:
  - fake giveaways / ticket scams
  - phishing through conversation
  - engagement-based traps targeting emotionally reactive users



- Integration of LLMs on social media introduces **LLM-side threats**:



Problem Statement

Given user-network dynamics:

$$s_i(t+1) = f\big(s_i(t), G\big), \qquad e_i(t+1) = g\big(s_i(t), a_i(t)\big)$$

adversarial prompts: $p_a(t) = \phi\big(s_a(t)\big)$

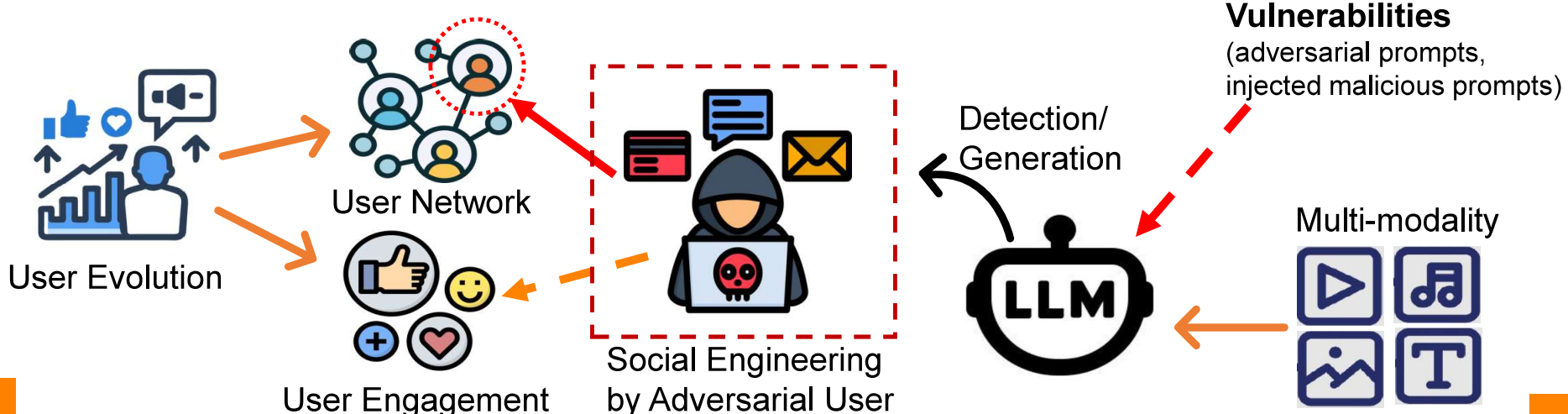
and,

multimodal content: $Z_j(t) = h\big(X_j(t)\big)$

the goal is to **detect or mitigate adversarial perturbations:**

$$\delta(t): \quad Z'_j(t) = Z_j(t) + \delta(t)$$

$$F_{\text{LLM}}\big(Z'_j(t)\big) \notin Y_{\text{unsafe}}$$



Research Questions

RQ1: Can we effectively model how users evolve over time?

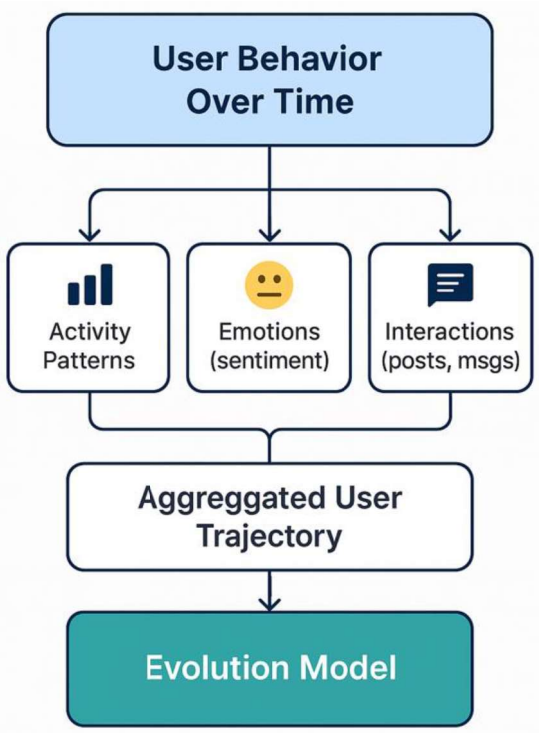
RQ2: Does user evolution leads their vulnerability to adversarial threats?

RQ3: If users become vulnerable, what robust methods can detect and protect them effectively?

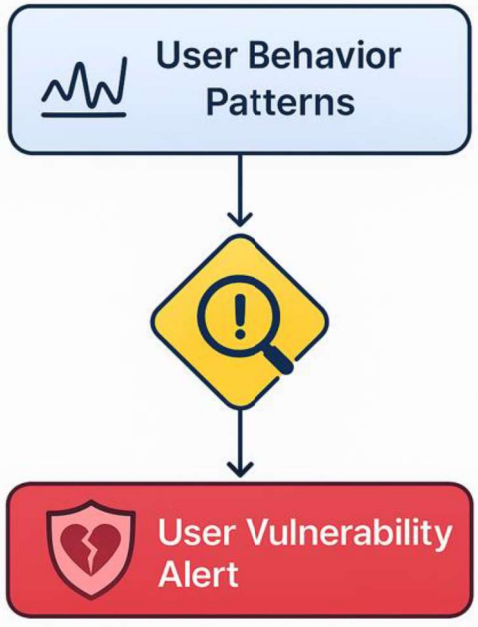
Research Gap

We currently lack a unified framework that simultaneously:

(1) Models user evolution and behavioral shifts



(2) Detects user vulnerability based on these patterns



(3) Protects both users and deployed LLMs from adversarial exploitation

Research Objective

Model User Evolution

Understand Behavioral Shift

Identify Vulnerability Windows

Design LLM-based Threat Detection

Improve the LLM capability on detecting Adversarial activities

Develop Robust LLM Defenses

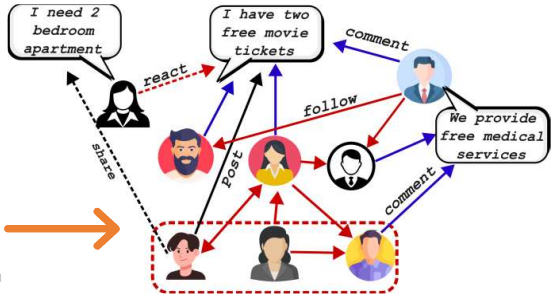
Overview of Proposed Framework

User Evolution

1. What causes the evolution?
2. The impact of evolution?

Behavior or Emotion Shift

1. What causes this shift ?
2. How important to analyse this shift?
3. The impact of this shift?



User activities on social Media

1. Why user activities are different?
2. Does it have any negative consequences?
3. How can user get rid of if they fall in trap because of social media activities?

LLM robustness and Defense

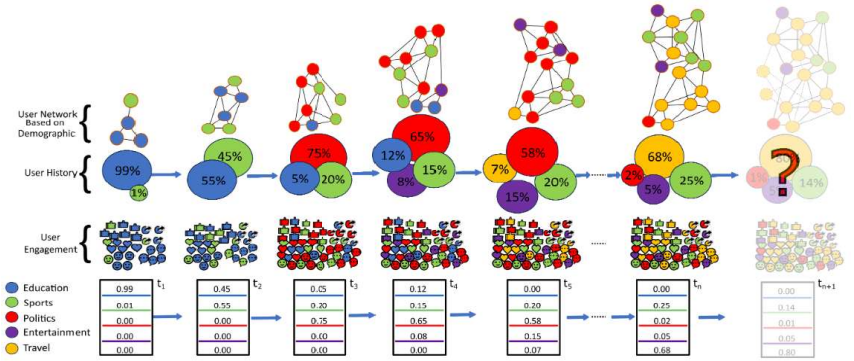
1. Why do we need LLMs in this context?
2. How can we enhance LLM robustness?
3. What security risks arise from LLM usage?
4. How LLM can be more defensive?

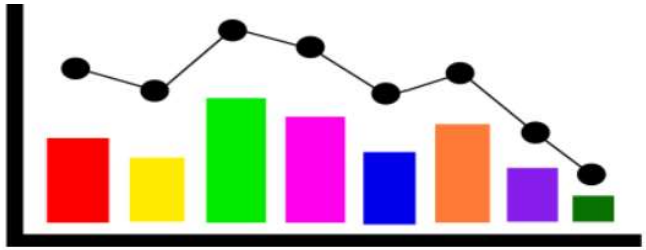
Past studies

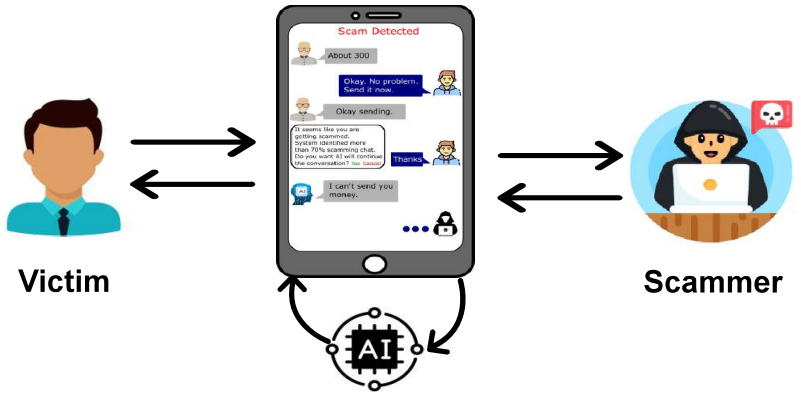
Future studies

Preliminary Works

- **EVOLVE — Predicting User Evolution**
  - GPT-based model predicting next behavioral stages.
  - Captures network growth, posting shifts, engagement patterns.
  - Shows how user attributes drive long-term behavioral trajectories.
- **EMOVIS — Emotional Sentiment Tracking**
  - Visualizes 28 emotions + 8 conversation sentiment.
  - Shows how user emotions and viewpoints shift over time.
- **AI-in-the-Loop — Scam Detection & Prevention**
  - LLM-based framework for detecting and disrupting scams.
  - Demonstrates how adversaries exploit user behaviour trends.
  - Reveals the emergence of LLM vulnerabilities.



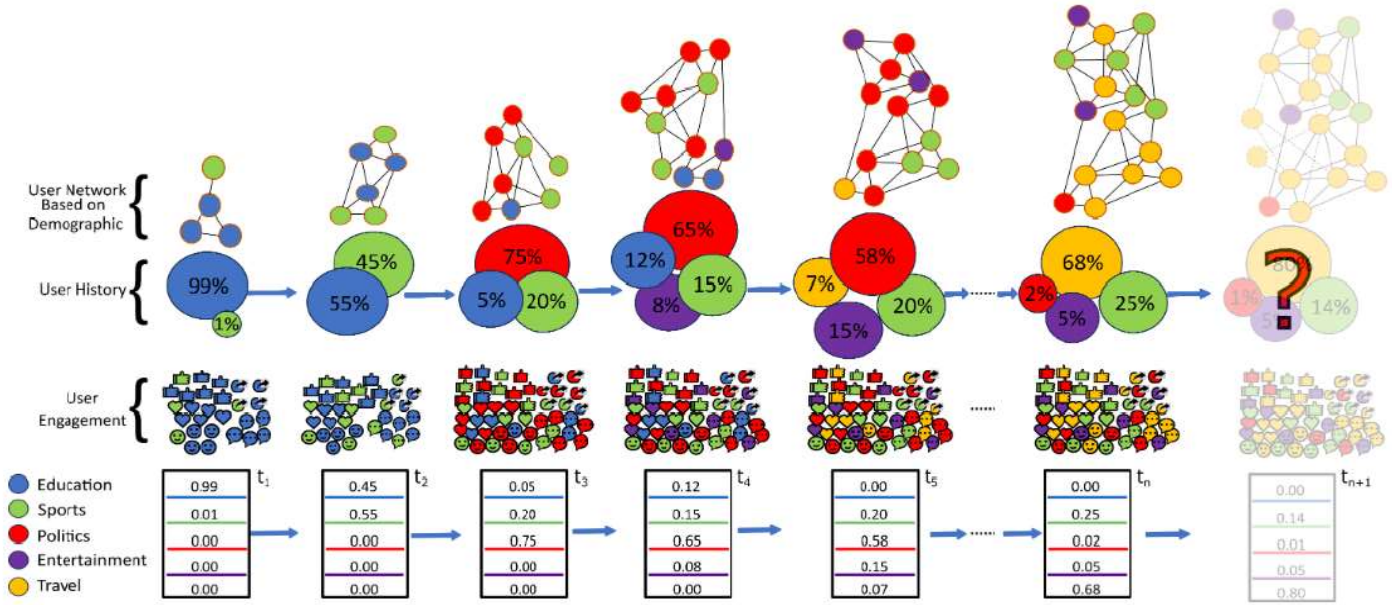




Project 1

"Understanding How Users Change Over Time"

- **Paper:** *Predicting User Evolution and Network Dynamics Using a GPT-like Model*
- **Goal:** Forecast the *next stage* of a user's evolution in social media.
- **Key Features Modelled:**
  - Demographics
  - Post history
  - Engagement patterns
  - Network graph changes
  - Behavioural roles



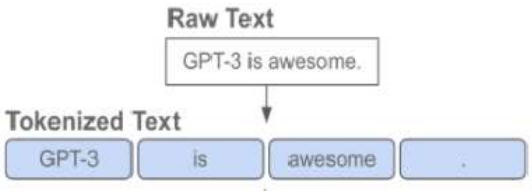
EVOLVE: Predicting User Evolution and Network Dynamics in Social Media Using Fine-Tuned GPT-like Model. [**ASONAM 2024**]

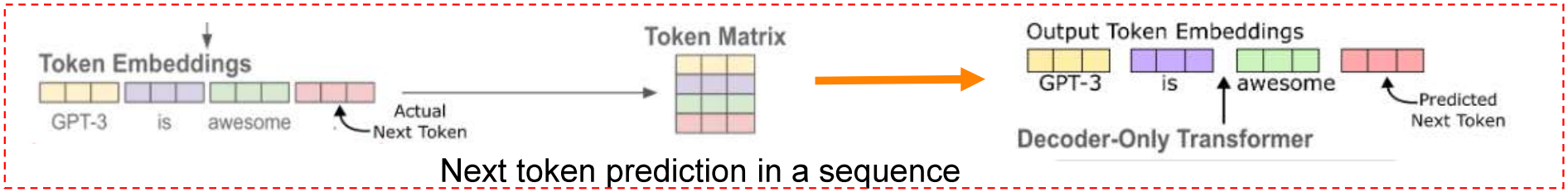
EVOLVE: Model Architecture

- Combined feature matrix = Demographics + History + Engagement + Network
- Sequential modelling using GPT-style decoder

Insight:

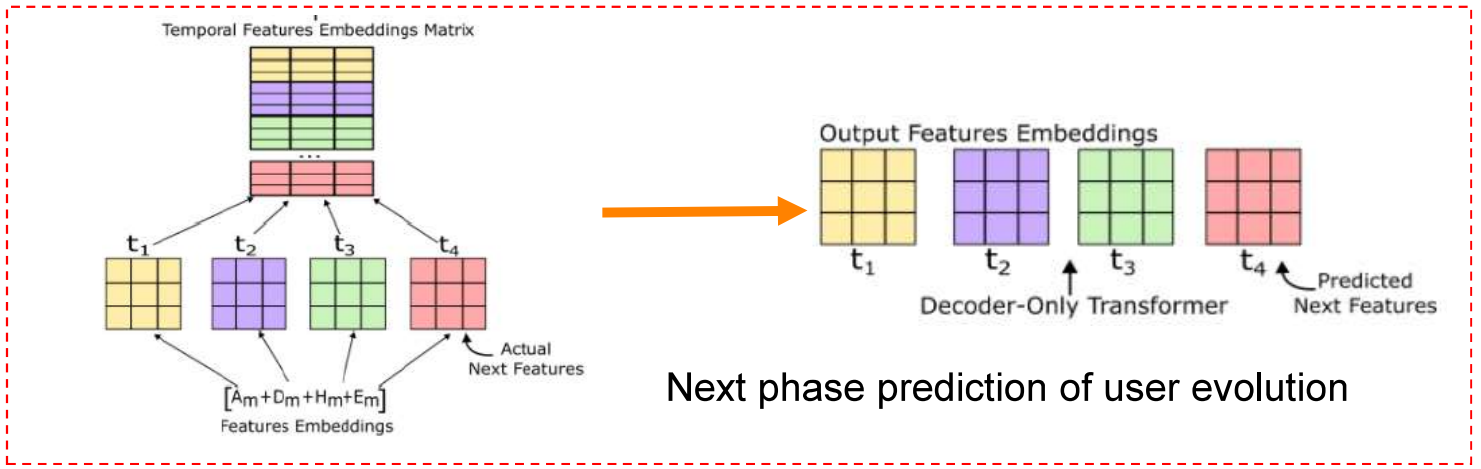
User behavior evolves in predictable sequences—just like language.





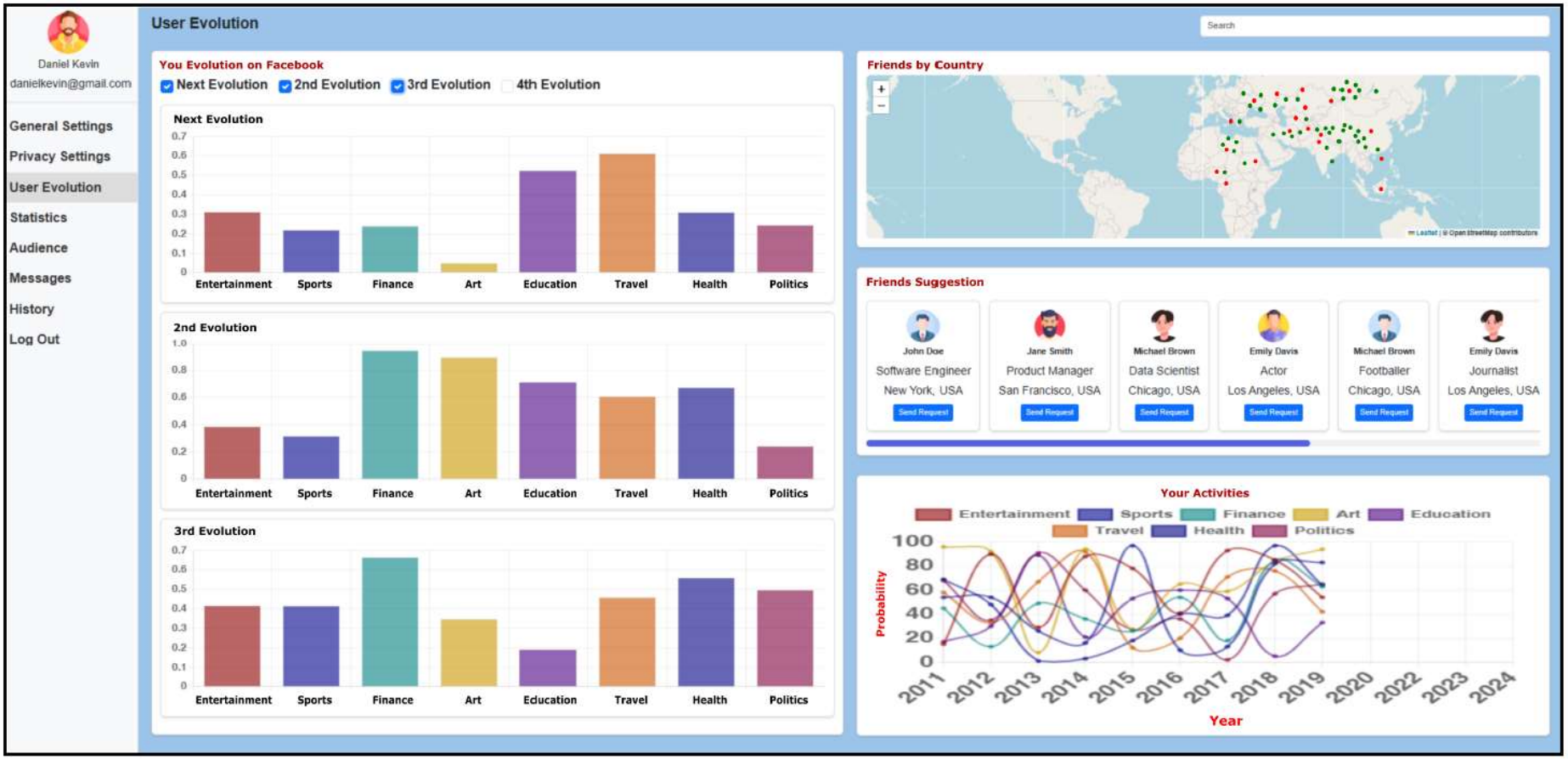
Next token prediction in a sequence

Inspired by GPT architecture



Next phase prediction of user evolution

EVOLVE: Application

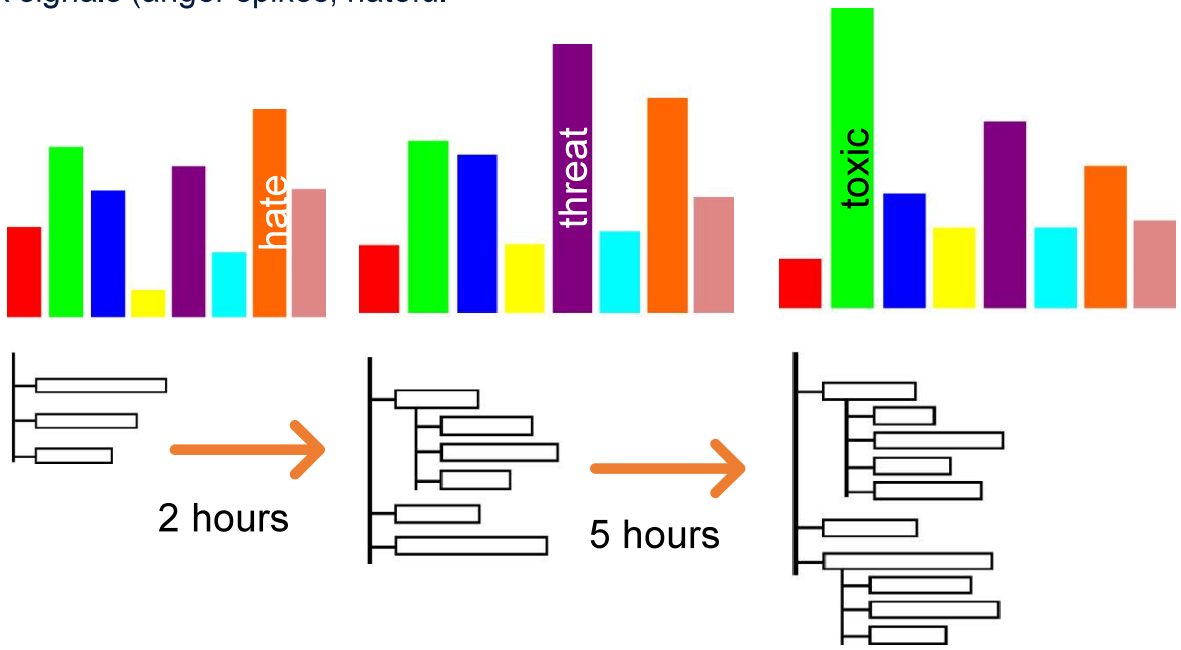


Project 2

User Emotion Understanding (EMOVIS)

- Track emotional drifts in user comments
- Extract sentiment polarity, hostility, supportiveness
- Identify early risk signals (anger spikes, hateful conversations)

1. approval 2. toxic 3. obscene 4. threat
5. insult 6. hate 7. offensive



A Visual Approach to Tracking Emotional Sentiment Dynamics in Social Network Commentaries. [**ICWSM 2024**]

EMOVIS: Model Inference

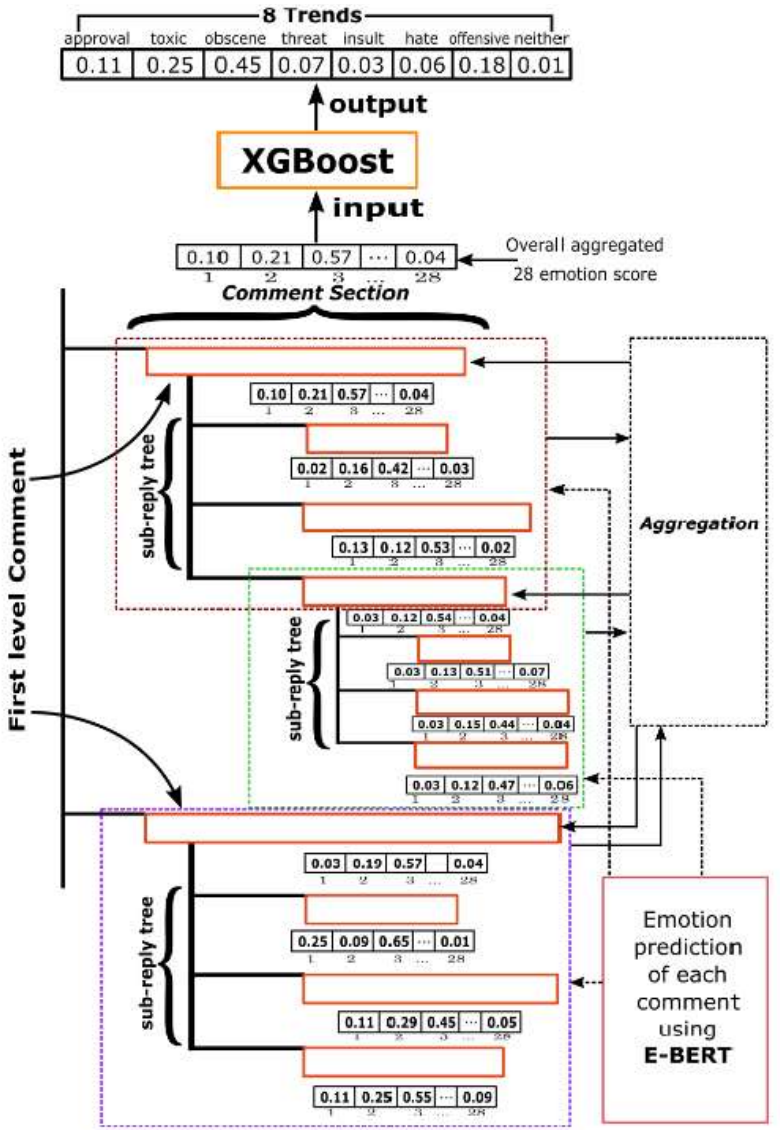
"Understanding What Users Express"

Goal: Track and visualize emotional and opinion trends in comment sections.

Key Components:

- RoBERTa-based multi-emotion classifier (28 emotions)
- Trend category detection
- The aggregation strategies:

  (i) Z-score, (ii) Weighted



EMOVIS: Results

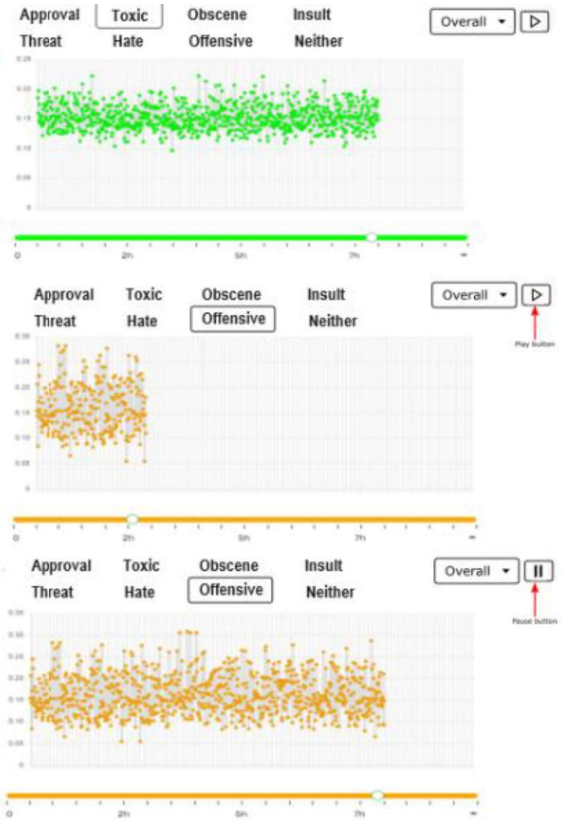
Model	F1	Recall	Precision	AUC
$BERT_{base}$	0.74	0.73	0.77	0.90
$BERT_{large}$	0.74	0.72	0.78	0.91
$RoBERTa_{large}$	**0.75**	**0.72**	**0.79**	**0.92**

Table 1: Performance of Transformer Models

	Micro-Avg			Macro-Avg		
Models	Precision	Recall	F1	Precision	Recall	F1
Bagging-ORC	0.86	0.70	0.77	0.58	0.33	0.38
Boosting-ORC	0.85	0.69	0.76	0.55	0.32	0.35
DecisionTree	0.73	**0.73**	0.73	0.43	**0.42**	0.42
LR-CC	0.79	0.65	0.72	0.22	0.24	0.23
LR-LP	0.80	0.67	0.73	0.38	0.26	0.26
MultiNB-BR	0.82	0.57	0.68	0.23	0.19	0.20
MultiNB-CC	0.78	0.62	0.69	0.40	0.26	0.29
MultiNB-LP	0.72	0.59	0.65	0.22	0.19	0.19
RandomForest	**0.88**	0.70	0.78	**0.68**	0.32	0.36
SVC-MOC	0.86	0.63	0.72	0.46	0.24	0.25
XGB-ORC	0.86	**0.73**	**0.79**	0.65	0.38	**0.44**

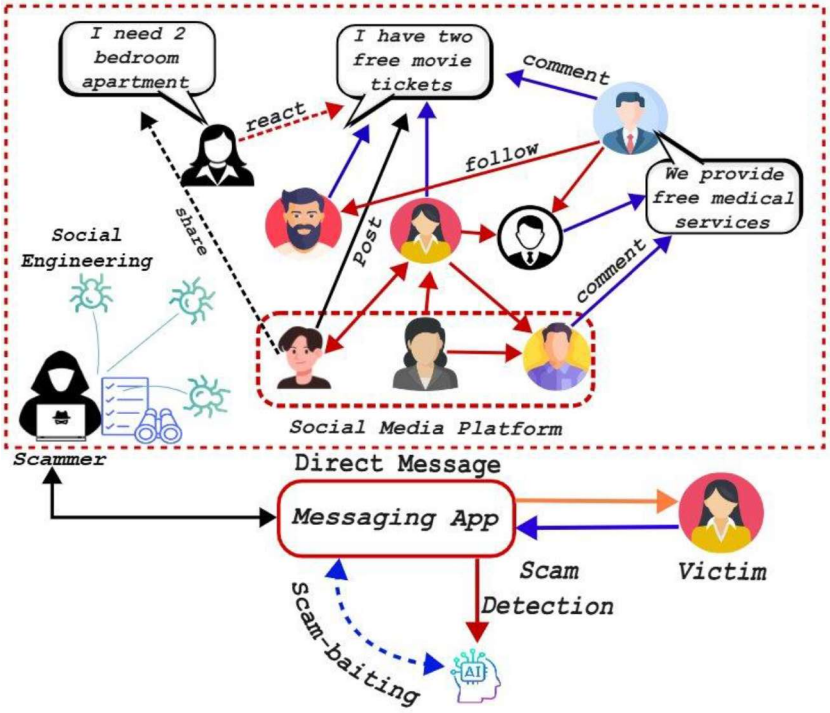
Table 2: Performance of ML Models

EMOVIS: Application



Project 3

Vulnerability & Scam Detection (AI-in-the-Loop)- Threat Model

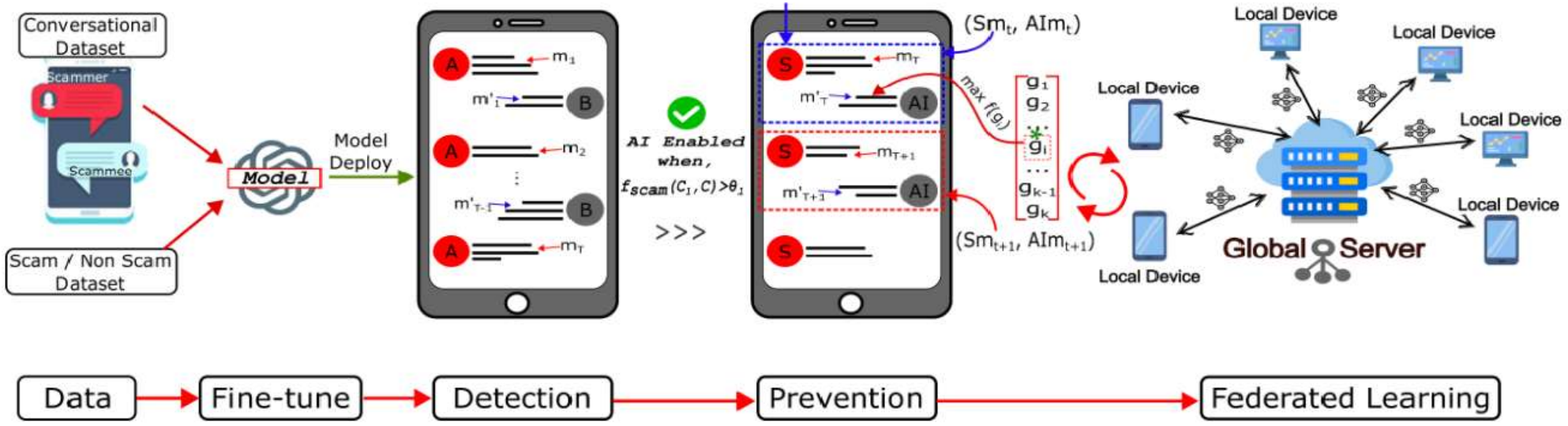


AI-in-the-Loop: Privacy Preserving Real-Time Scam Detection and Conversational Scambaiting by Leveraging LLMs and Federated Learning. [Accepted in cycle 1 of **PoPETS 2026**]

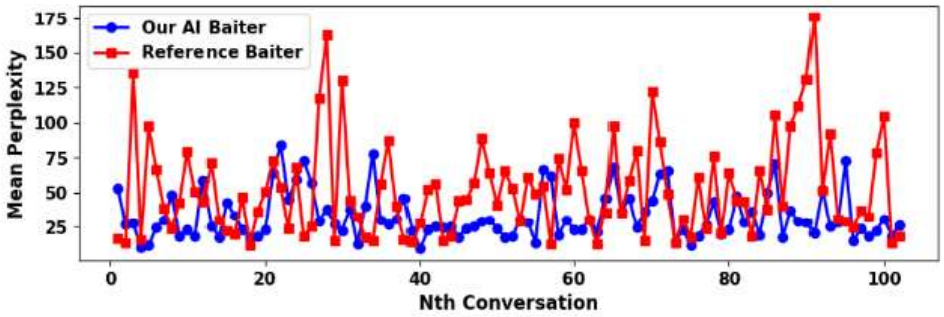
Project 3: AI-in-the-Loop

- **Goal: Combine LLMs + privacy-preserving techniques to detect and disrupt scam attempts.**
- **Core Components:** (1) Scam detection using fine-tuned LLMs

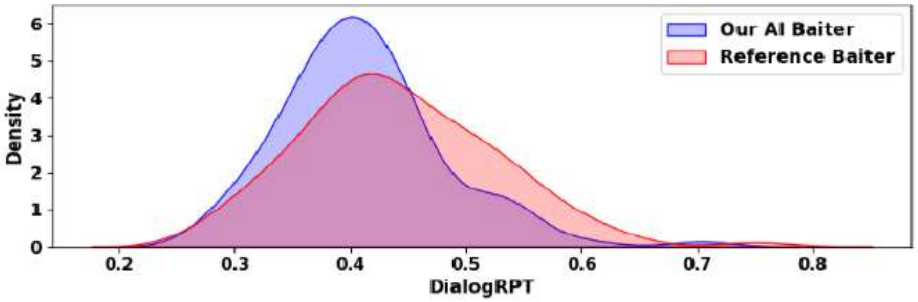
(2) Real-time conversational scambaiting agents (3) Federated Learning to preserve user privacy



AI-in-the-Loop: Results



(a) Mean Perplexity

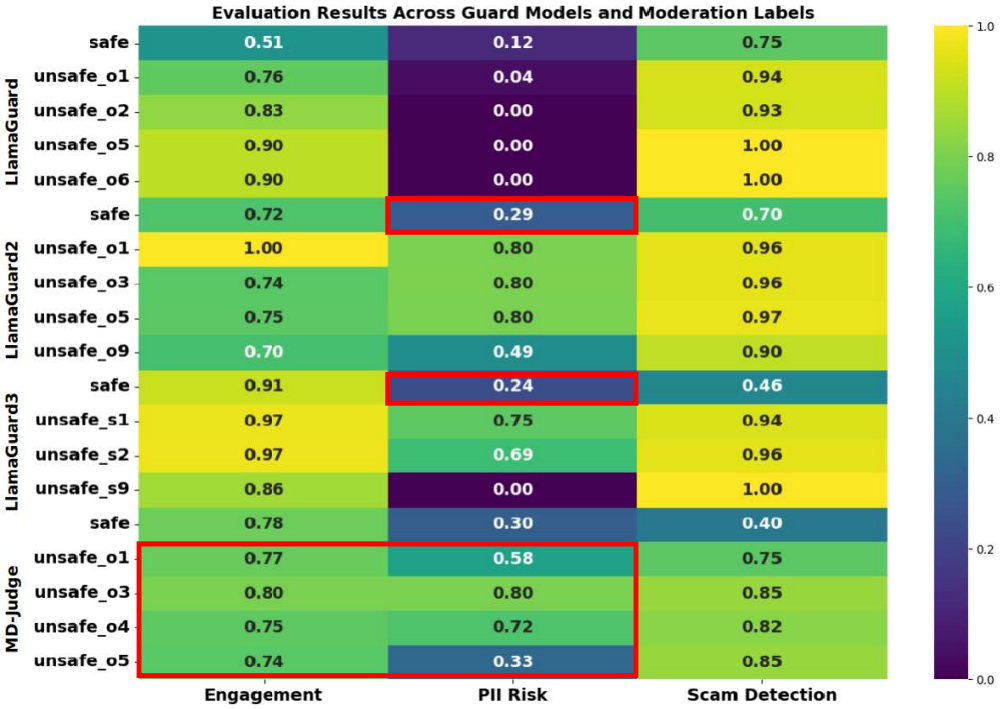


(b) DialogRPT Distribution

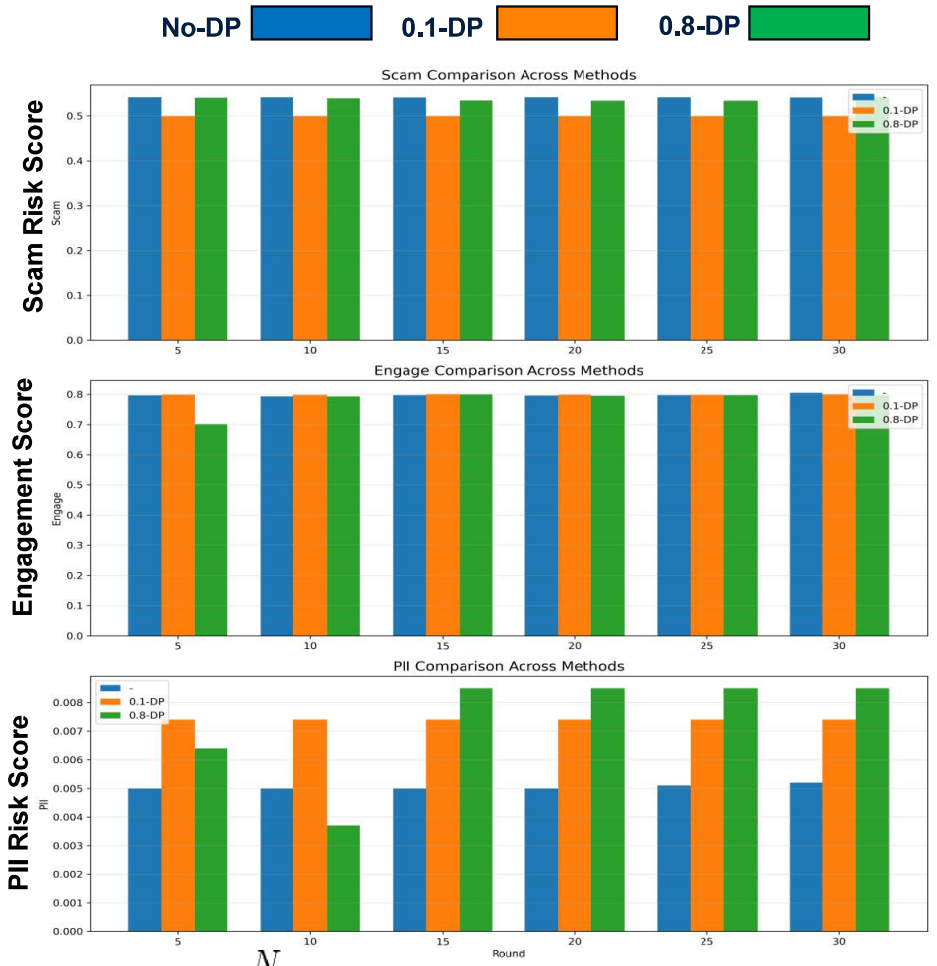
Model	Count	$\mathcal{M}_T$ (s)	μ_E	μ_{PII}	μ_S
LG	7 ± 2	6.50 ± 5.59	0.30 ± 0.30	0.17 ± 0.24	0.39 ± 9.19
LG.2	9 ± 0	5.68 ± 1.65	0.78 ± 0.05	0.81 ± 0.11	0.11 ± 6.11
LG.3	8 ± 2	7.47 ± 3.83	0.74 ± 0.04	0.38 ± 0.42	0.92 ± 0.06
MD-J	9 ± 1	8.42 ± 2.01	0.79 ± 0.04	0.57 ± 0.30	0.53 ± 4.04

Table 1: Evaluation results of scam-baiter interactions.

AI-in-the-Loop: Results (Contd.)



Safeness and Risk Awareness Evaluation



$$\tilde{g} = \frac{1}{N}\sum_{i=1}^{N} \text{clip}(g_i, C) + \mathcal{N}(0, \sigma^2 C^2 I)$$

How AI is vulnerable?

1. Jailbreak attacks bypass LLM safety rules using adversarial prompts.

Refs:

- Wei et al., *"Jailbroken: How Does LLM Safety Training Fail?"*, 2023
- Zou et al., *"Universal and Transferable Adversarial Attacks on Aligned LLMs"*, 2023

2. Prompt injection manipulates the model to ignore system instructions.

Refs:

- Greshake et al., *"More Than You've Asked For: Novel Prompt Injection Threats"*, IEEE S&P 2024
- Hubinger et al., *"Prompt Injection: Attacks and Defenses"*, ARC 2023

3. Training data poisoning inserts malicious patterns.

Refs:

- Carlini et al., *"Poisoning the Training Data of Large Language Models"*, USENIX 2024
- Wang et al., *"Backdoor Attacks on NLP Models"*, ICML 2021

How AI is vulnerable? (Contd.)

4. Privacy leakage occurs when LLMs memorize and reveal sensitive training data.

Refs:

- Carlini et al., *"Extracting Training Data from Large Language Models"*, USENIX 2021
- Nasr et al., *"Comprehensive Privacy Analysis of Training Data in LLMs"*, IEEE S&P 2023

5. LLM-generated content can enable social engineering, phishing, or impersonation.

Refs:

- Kumar et al., *"Automating Social Engineering Using LLMs"*, 2023
- Shrestha et al., *"LLM-Generated Phishing Emails: Dangerously Effective"*, 2023

Prior Arts on Adversarial Prompt generation

Category	Key Idea	Reference Work
Multi-turn Escalation	Gradually escalates benign prompts into harmful ones	Crescendo (USENIX 2025)
Task Decomposition	Decomposes harmful intent into subtasks to increase success rate	Exploiting Task-Level vulnerabilities (USENIX 2025)
DSL-Based Composition	Compositional generation using fixed prompt primitives	h4rm3l (ICLR 2025)
Open-Ended Red Teaming	Broad tactic exploration without intent-wise ranking	WildTeaming (NurIPS 2024)
Gradient-Based Optimization	Token-level adversarial optimization	AutoDAN (NurIPS 2025)
Self-Prompting Attacks	LLM recursively optimizes its own jailbreak prompts	An LLM Can Fool Itself (ICLR 2024)

Prior Arts on Jailbreak Evaluation Metric

Metric Type	Key Idea / Limitation	Reference Work
Binary ASR	Measures attack success only; overestimates effectiveness	Shen et al. [ACM CCS 2024]
Human Annotation	Labels such as “Detailed” or “Denial”; costly and subjective	Yu et al. [USENIX 2024]
Multi-Agent Judges	LLM judges score outputs on ordinal scales	JailJudge (Liu et al.) [ICLR 2025]
Information-Based Metrics	Evaluates semantic content rather than refusal	StrongReject (Souly et al.) [NeurIPS 2024]
Transferability Metrics	Measures cross-model success; ignores harm magnitude	One Model Transfer to All [ICLR 2025]
Standard Benchmarks	Fixed prompts with inconsistent judge behavior	JailbreakBench (Chao et al.) [NeurIPS 2024]

Ongoing and Future work