

# Can You Walk Me Through It?

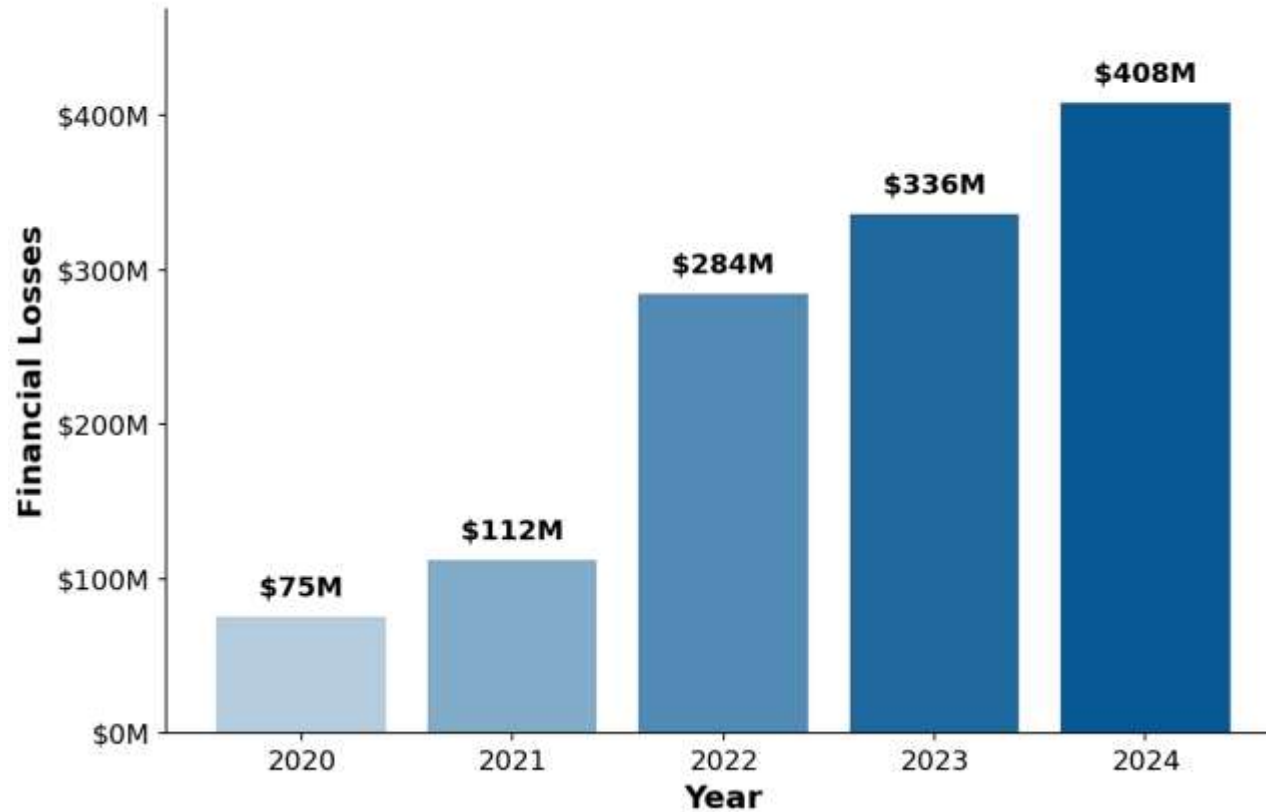
## Explainable SMS Phishing Detection using LLM-based Agents

Yizhu Wang, Haoyu Zhai, Chenkai Wang, Qingying Hao,  
Nick A. Cohen, Roopa Foulger, Jonathan A. Handler, Gang Wang

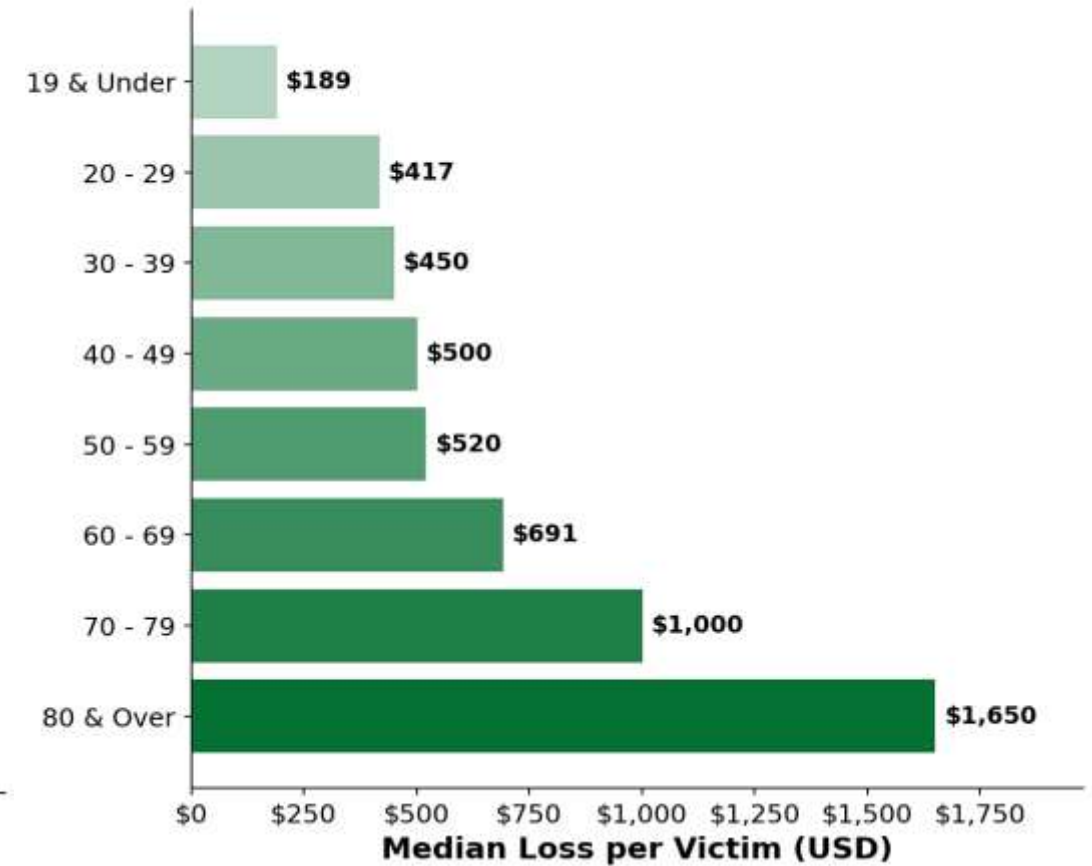


# SMS Phishing is a Serious Threat, Particularly to Older Adults

### SMS Phishing Financial Losses in the United States (2020-2024)



### Fraud Losses by Age Group

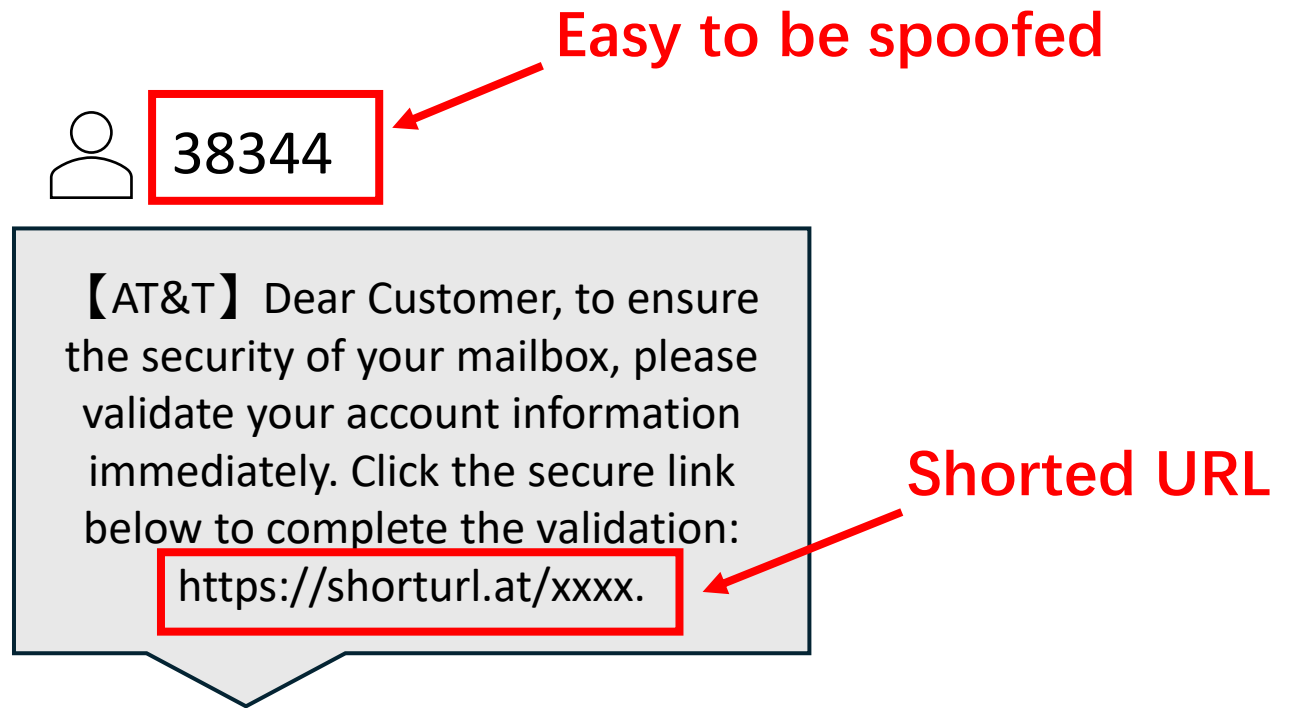


[1] FTC Report on Year: <https://public.tableau.com/app/profile/federal.trade.commission/viz/shared/MNHYNJWNT>

[2] FTC Report on Age: <https://public.tableau.com/app/profile/federal.trade.commission/viz/AgeandFraud/Infographic>

# Challenges of Detection

- **Short Message Length:**
  - Limited information
  - URL shortener abuse
  - Unreliable Caller ID

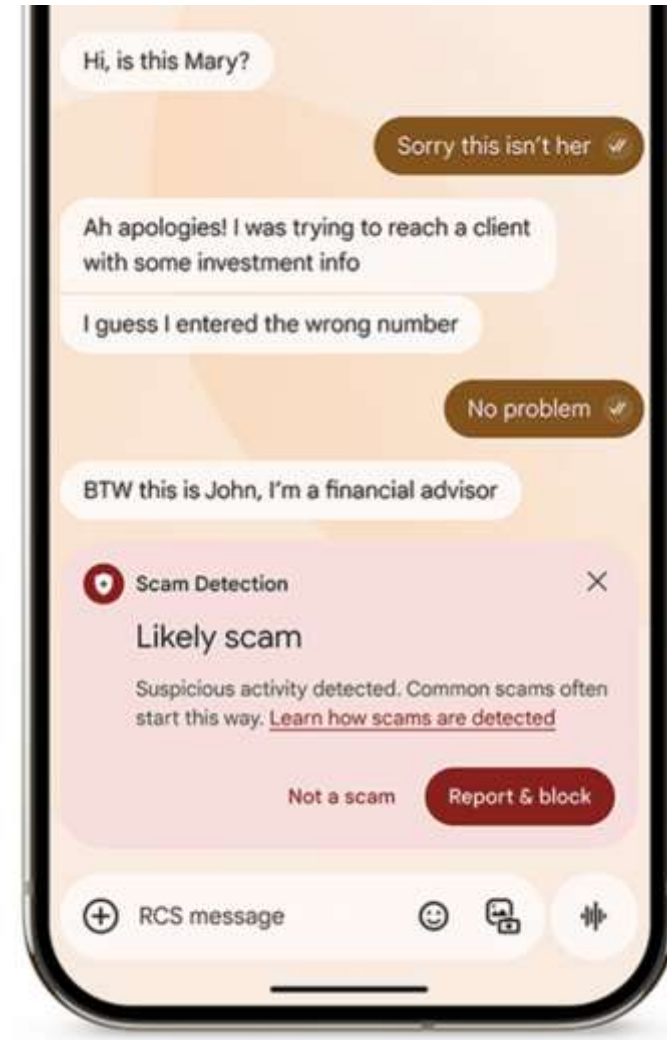
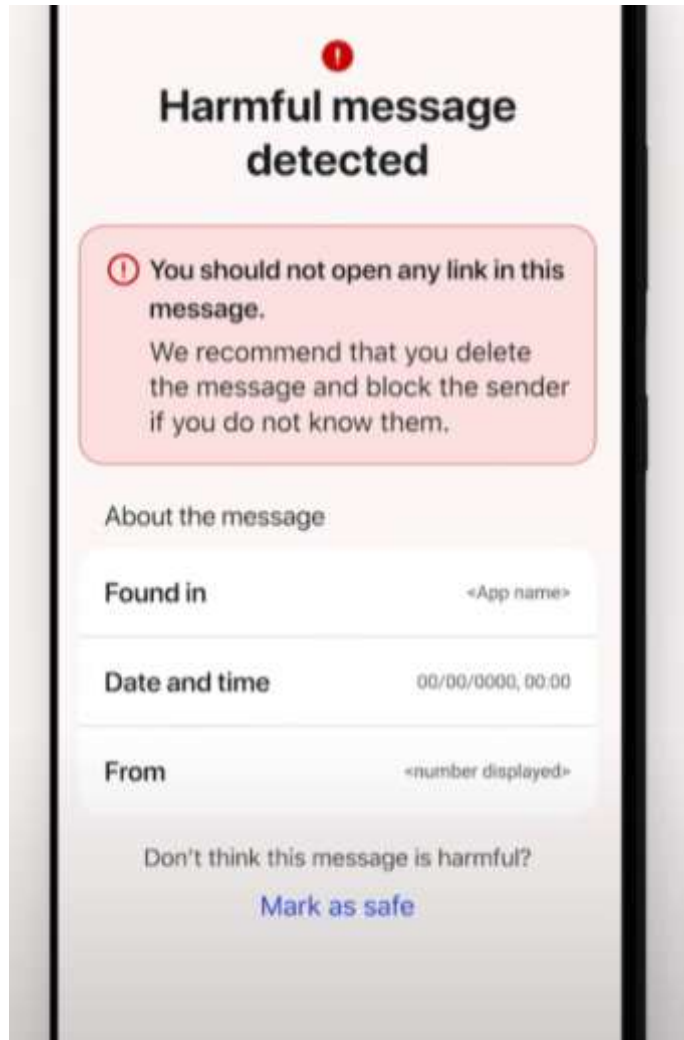


➔ Not enough to only analyze textual features: need more context!

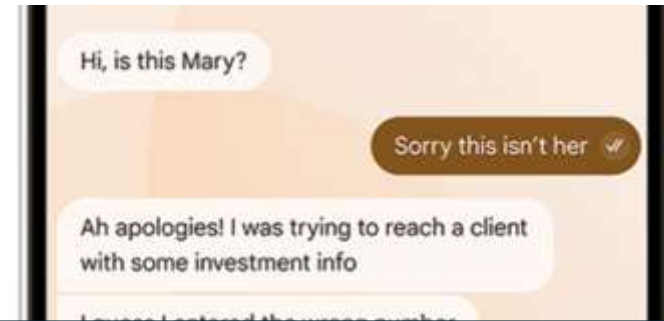
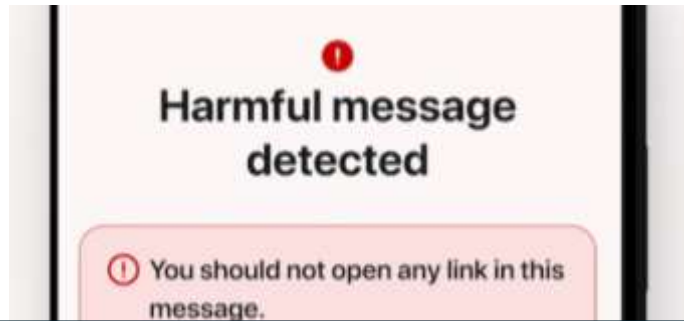
- **Vague Decision Boundary:**



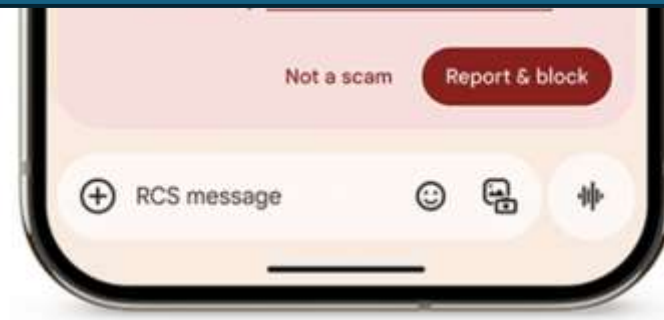
# From Detection to Explanation: Limitations



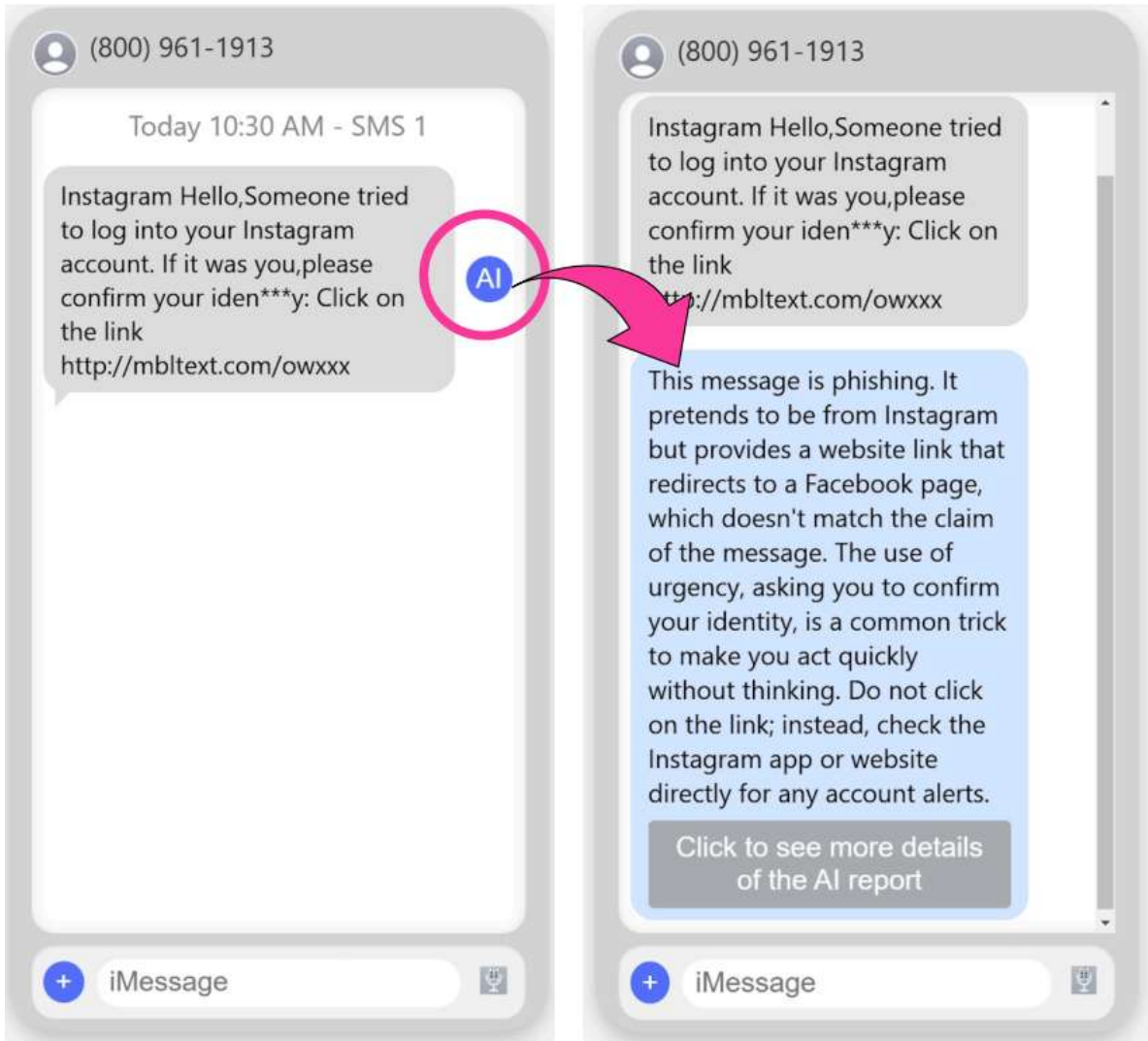
# From Detection to Explanation: Limitations



No message-specific explanations for detection decisions. Users can't understand "why" this particular message is flagged.

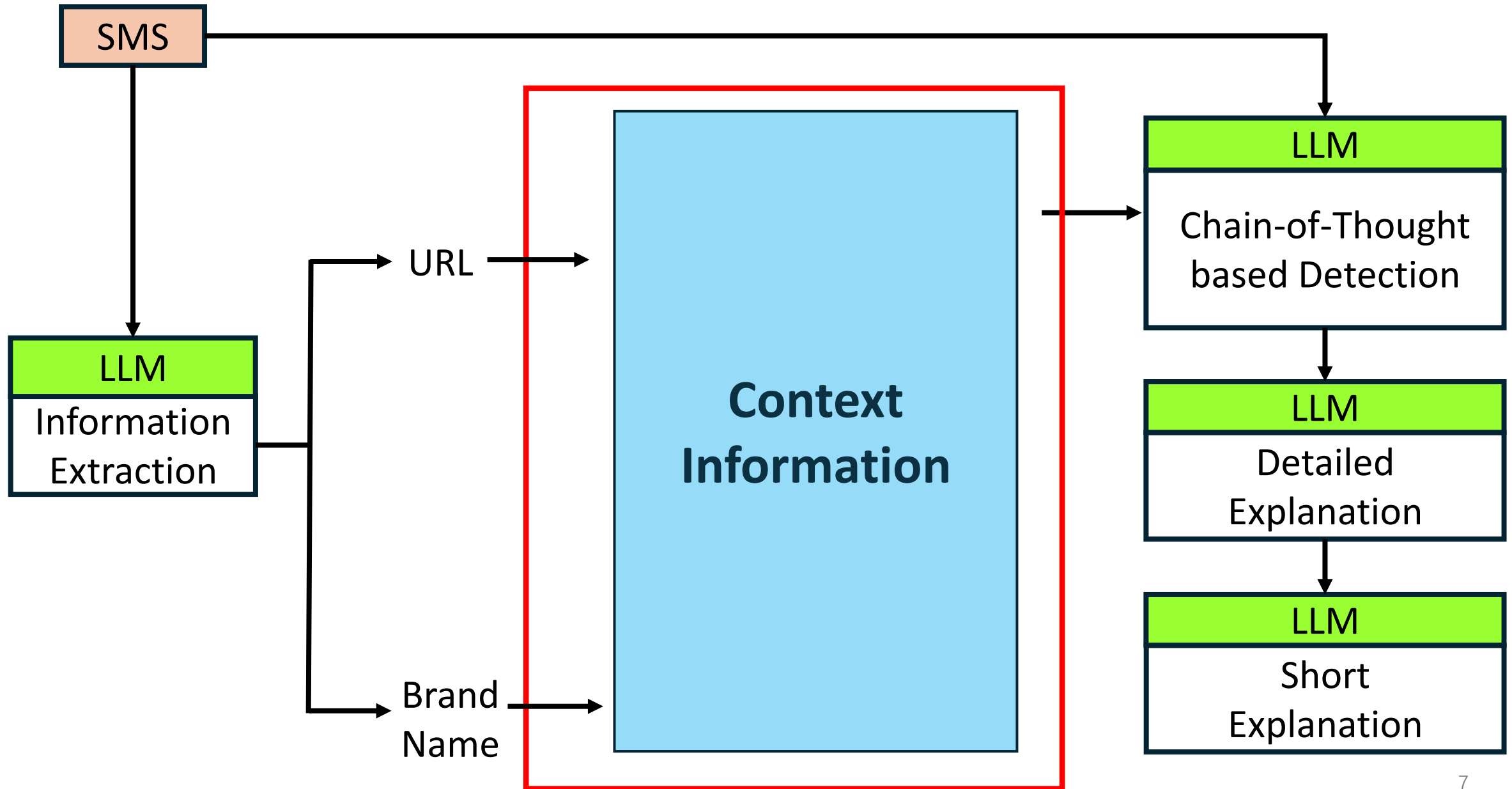


# Our Goal:

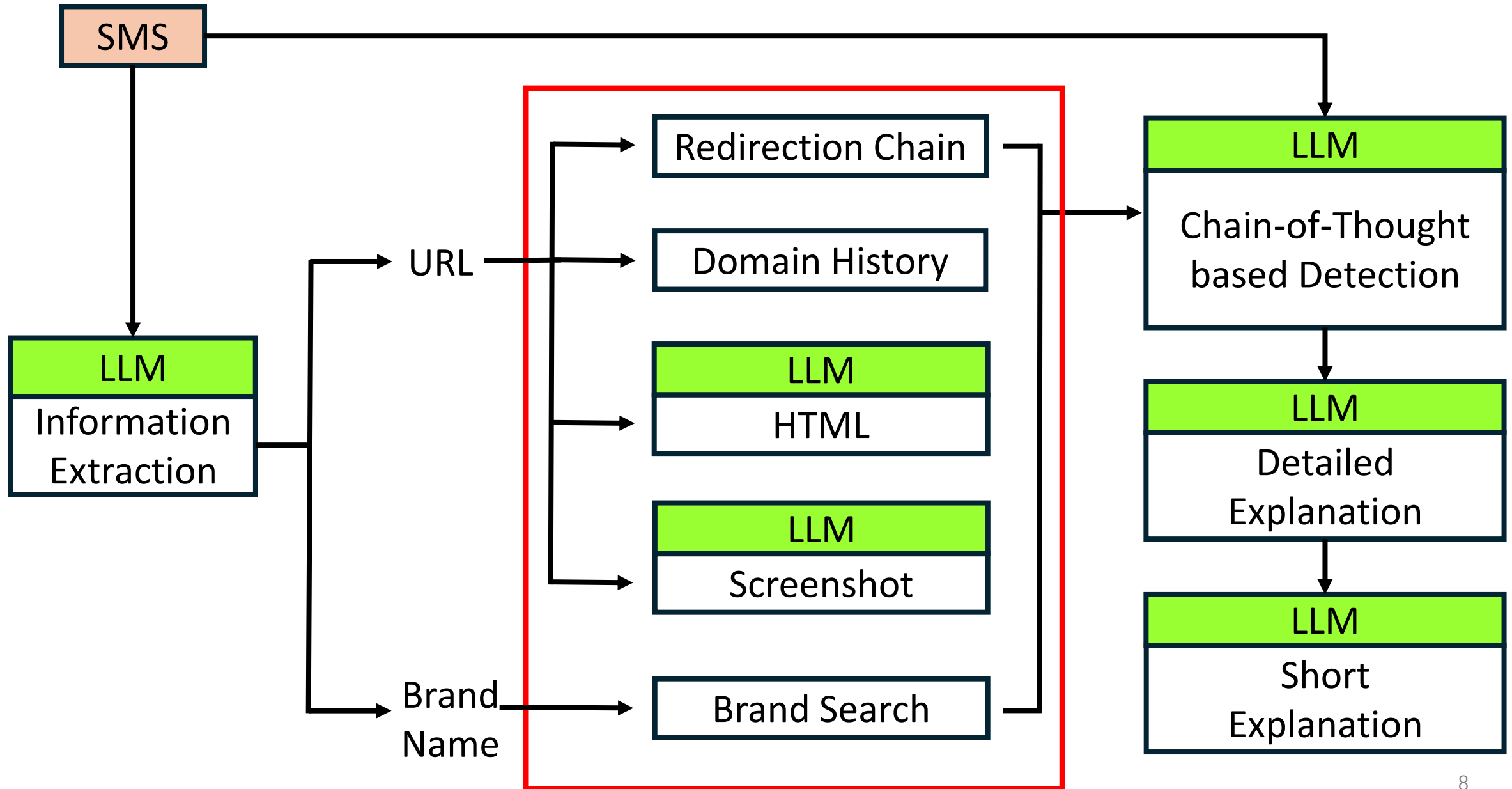


- Using **LLM** to perform both **detection** and **evidence-based explanation**.
- Gather **external context** (URLs, domain info, screenshots)
- **Chain-of-Thought** reasoning

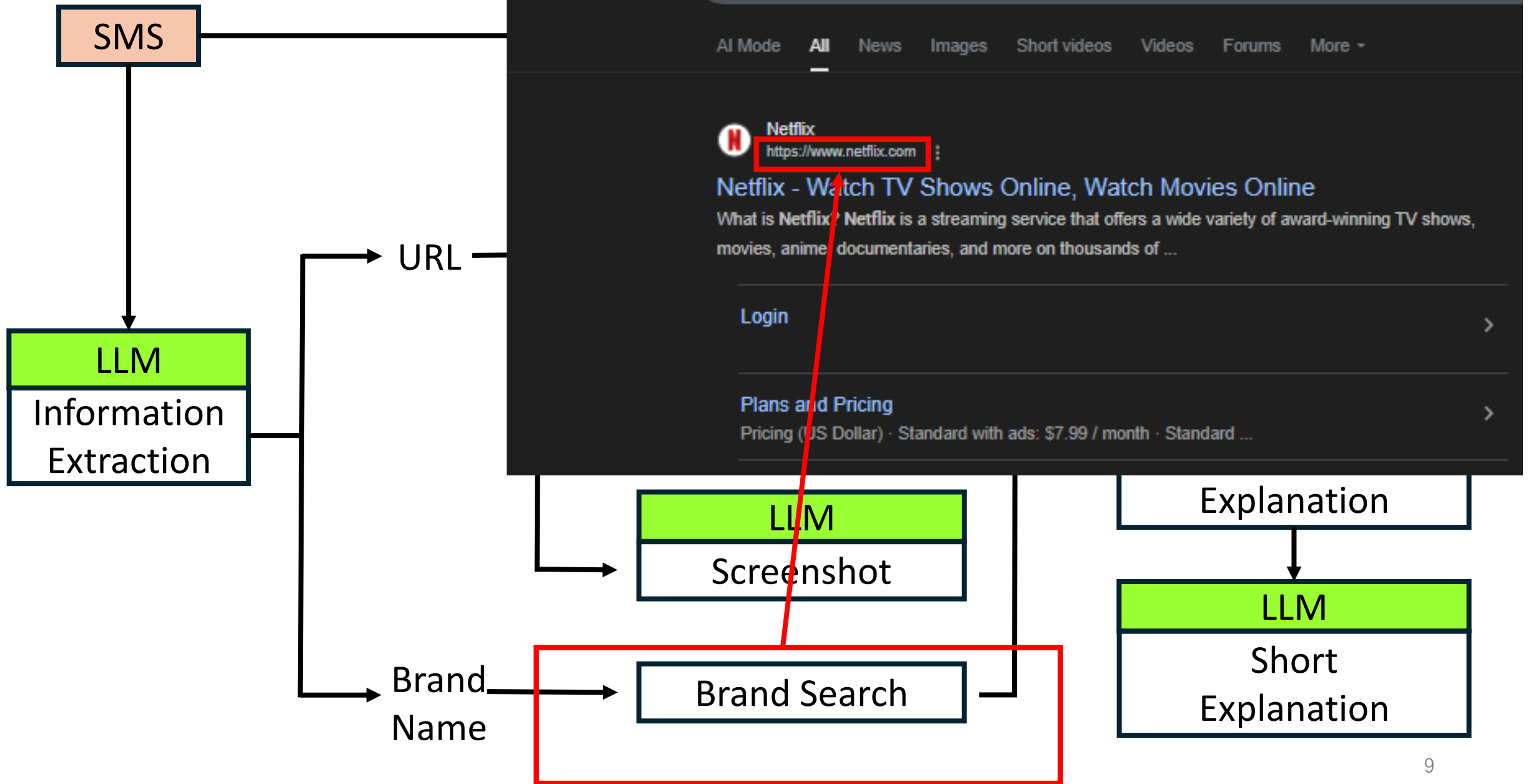
# SmishX System Design



# SmishX System Design



# SmishX System Design



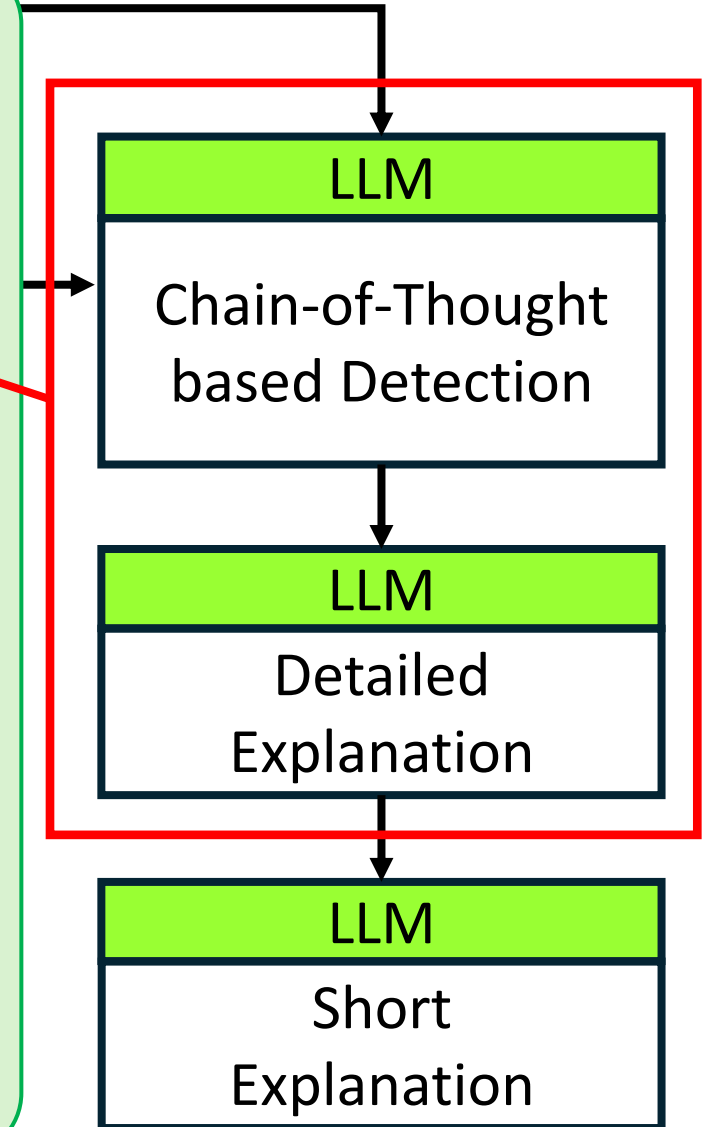
# SmishX System Design

## CoT Detection

I want you to **act as a spam detector** to determine whether a given SMS is phishing, spam, or legitimate. **Your analysis should be thorough and evidence-based.**

### Classification Steps:

1. **Legitimate:** conversations between friends and family members, notifications from known organizations
2. **Spam:** promoting high-risk services such as gambling
3. **Phishing SMS Features:** urgent/alarming languages, embedding suspicious links, requesting personal information, inconsistencies in URLs and domain names, and grammar/spelling errors.
4. **Analyzing the extra context information**
5. **Context information**



# SmishX System Design

Alert - Netflix: Payment failed we will close your account due to missing on payments watch time log-in:  
<https://xxxx>

(a) Phishing SMS

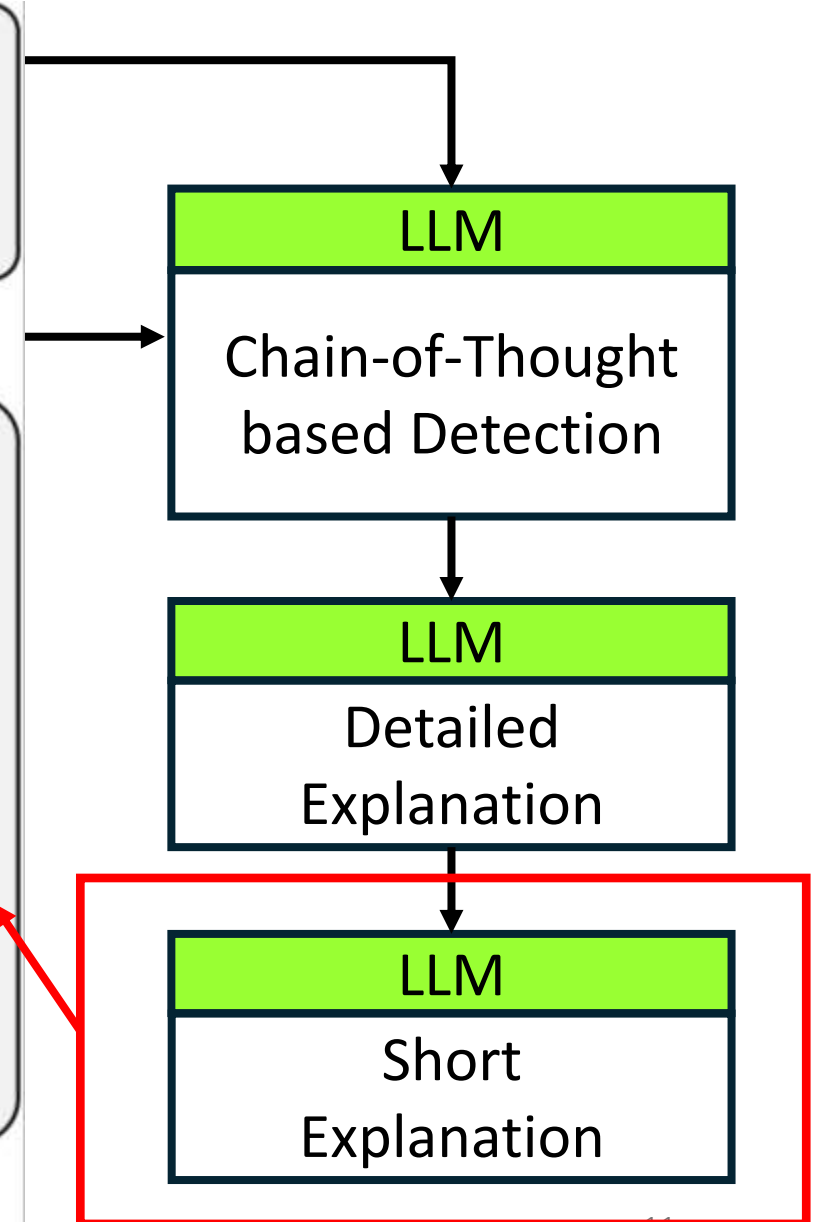
Decision

This SMS message is likely a phishing attempt. It pretends to be from Netflix and warns about a payment failure to create urgency, which is a common trick to make people act quickly without thinking. The website link doesn't match Netflix's official website, and the strange spacing and grammar errors add to the suspicion. Do not click on the link or respond—if you're concerned about your Netflix account, go directly to the official Netflix website or app to check your account status.

Explain.

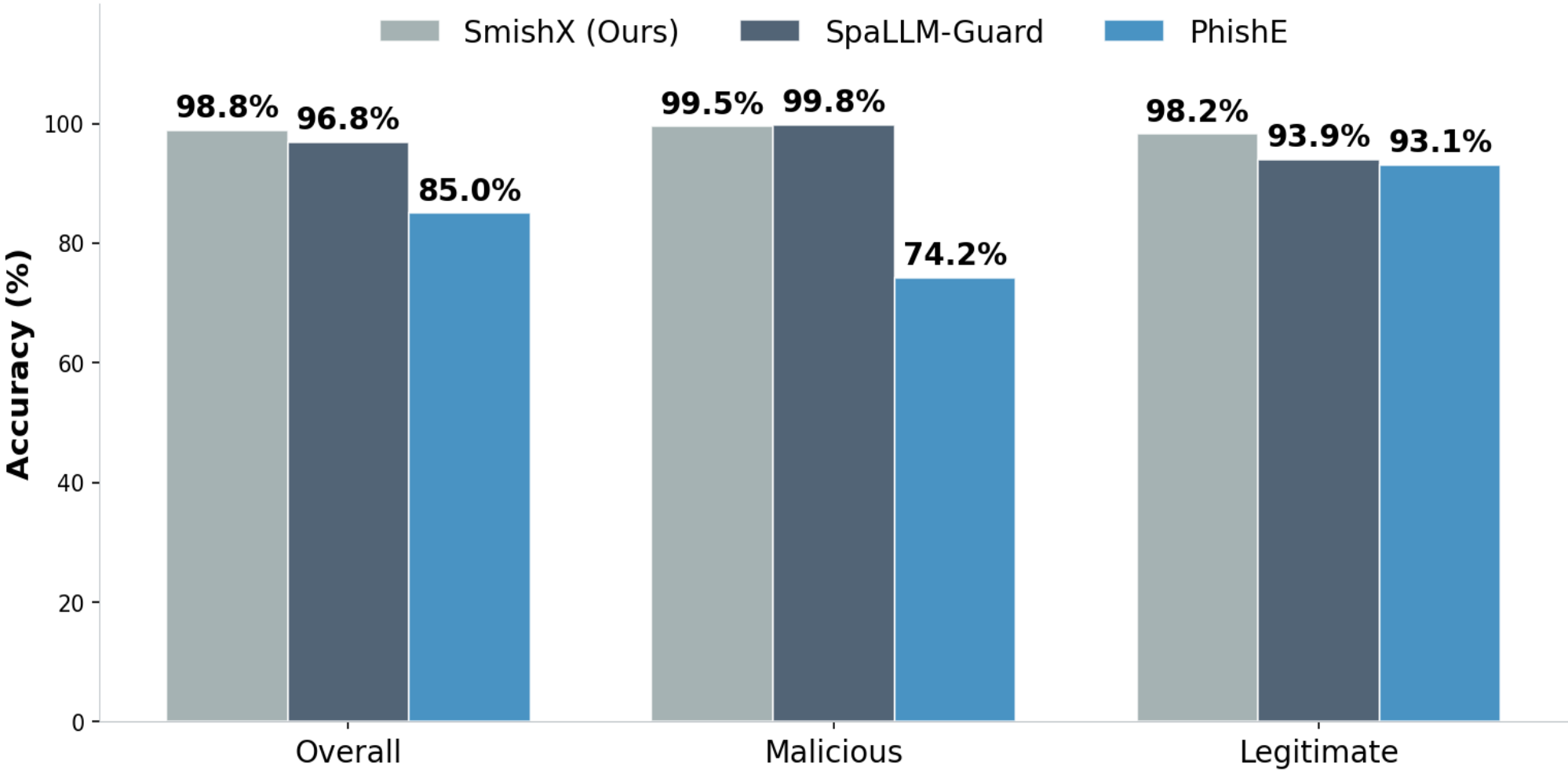
Advice

(b) Short Explanation from SmishX



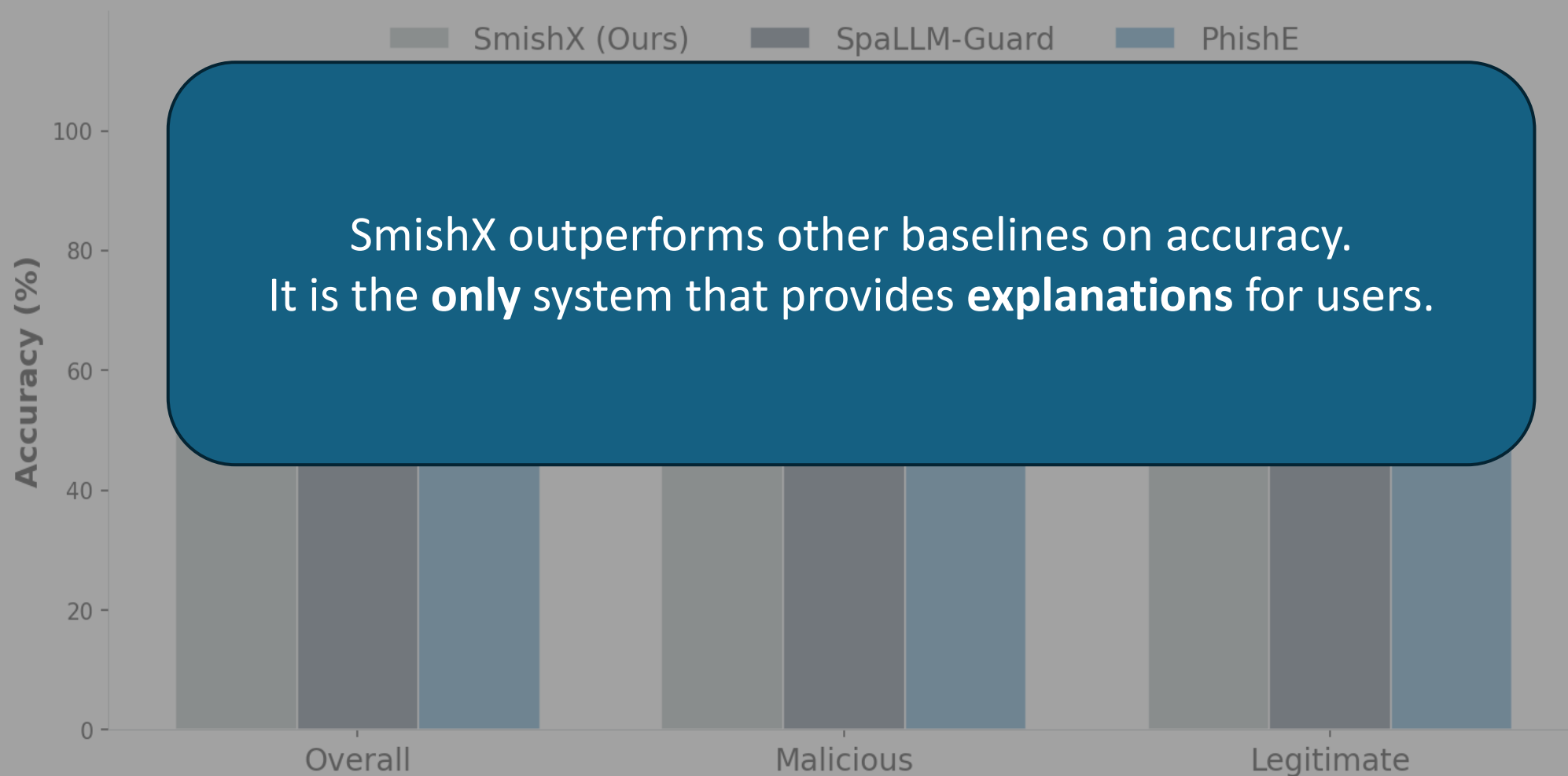
# SmishX Evaluation

- 578 Malicious (Phishing + Spam)
- 622 Legitimate



# SmishX Evaluation

- 578 Malicious (Phishing + Spam)
- 622 Legitimate



# Research Questions

- **RQ1: Effectiveness.** How effective is the AI agent in helping users recognize phishing SMS.
  - **RQ2: Usability.** How do users perceive the usability of the AI agent?
  - **RQ3: Trust and Disagreement.** How does the AI agent influence users' trust and confidence, and under what situation would users disagree with the AI agent?
  - **RQ4: AI Errors.** How do users respond to the AI agent's mistakes?
- Study A (N=125)
- Study B (N=50)



Recruit 175 participants through Prolific (40% are 65+).



Participants take the user study at home with a laptop.

# User Study Design

Pre-Task Survey

Message Evaluation Task

Post-Task Survey

- Age
- Gender
- Ethnicity
- Highest Education Level
- Technology Background in CS
- Cell Phone Usage
- SMS Usage
- SMS Phishing Encounter Experience
- Self-confidence in Identifying Phishing SMS
- Trust in General AI Systems

# User Study Design

Pre-Task Survey

Message Evaluation Task

Post-Task Survey

**Study A:** 10 SMS (N=125, 40% over 65+)

5 Phishing SMS + 5 Legitimate SMS

**Study B:** 11 SMS (N=50, 40% over 65+)

5 Phishing SMS + 5 Legitimate SMS + 1 AI mistake case (False Negative)

# User Study Design

Pre-Task Survey

Message Evaluation Task

Post-Task Survey



Do you think this message is a legitimate message or a fraudulent phishing message?

**Fraudulent**

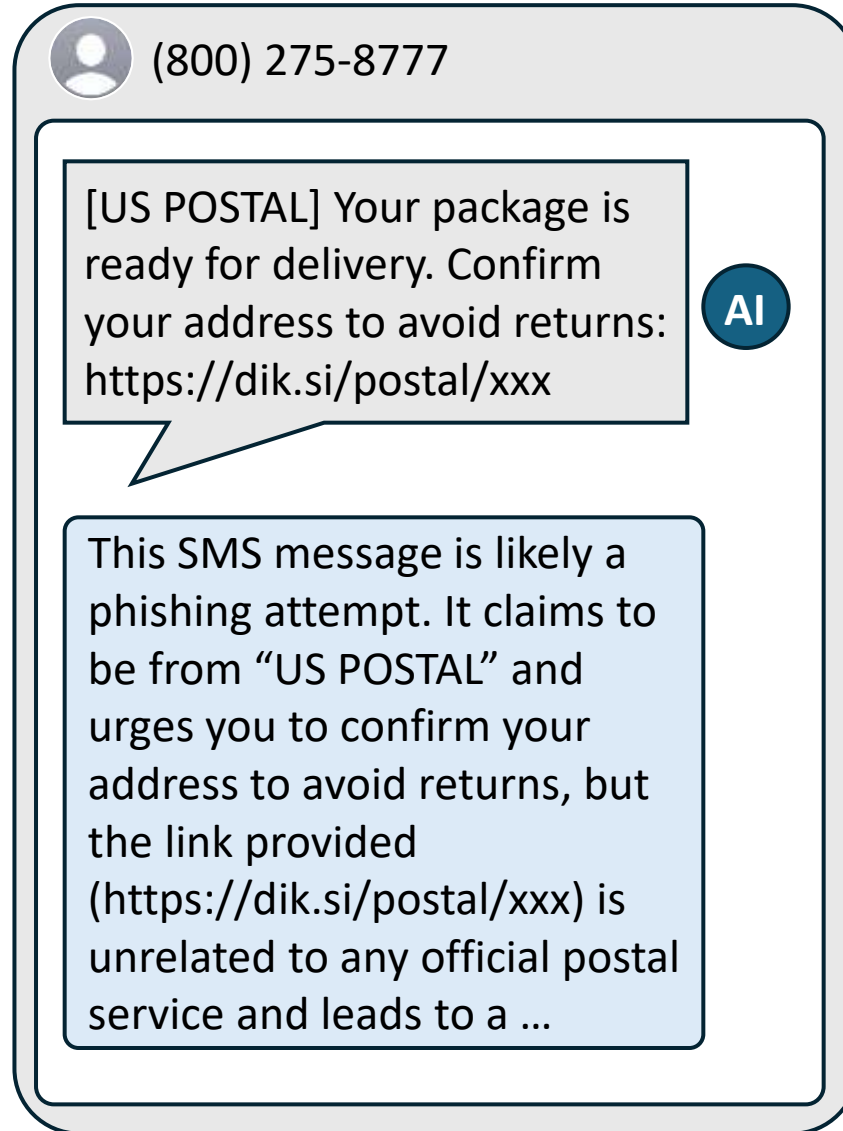
**Legitimate**

# User Study Design

Pre-Task Survey

Message Evaluation Task

Post-Task Survey



(800) 275-8777

[US POSTAL] Your package is ready for delivery. Confirm your address to avoid returns: <https://dik.si/postal/xxx>

AI

This SMS message is likely a phishing attempt. It claims to be from "US POSTAL" and urges you to confirm your address to avoid returns, but the link provided (<https://dik.si/postal/xxx>) is unrelated to any official postal service and leads to a ...



# User Study Design

Pre-Task Survey

Message Evaluation Task

Post-Task Survey

(800) 275-8777

[US POSTAL] Your package is ready for delivery. Confirm your address to avoid returns: <https://dik.si/postal/xxx>

AI

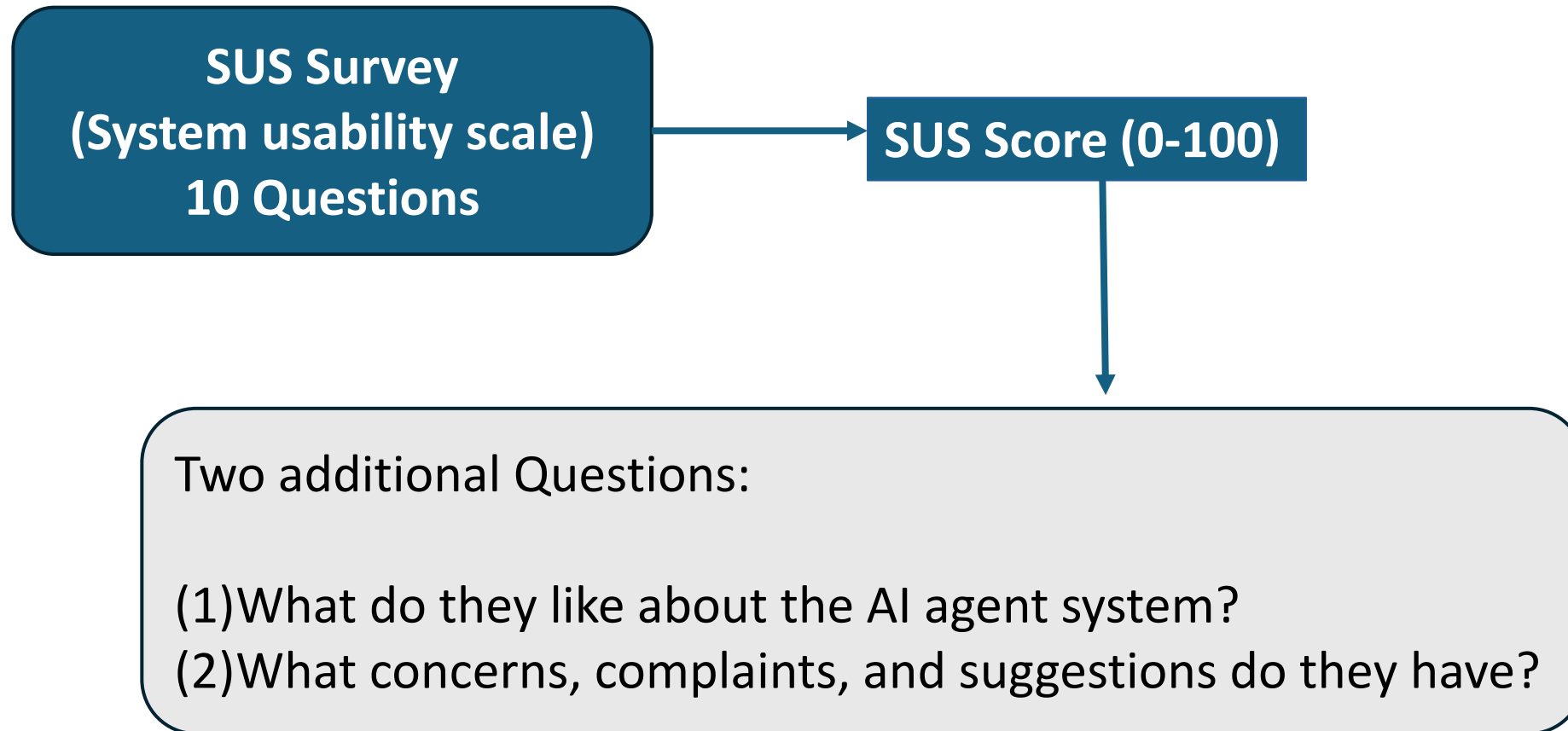
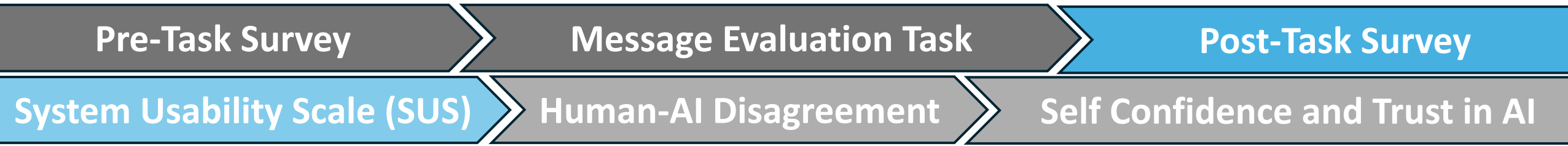
This SMS message is likely a phishing attempt. It claims to be from "US POSTAL" and urges you to confirm your address to avoid returns, but the link provided (<https://dik.si/postal/xxx>) is unrelated to any official postal service and leads to a ...

**After reading the AI report, do you now believe this message is legitimate or fraudulent?**

**Fraudulent**

**Legitimate**

# User Study Design



# User Study Design

Pre-Task Survey

Message Evaluation Task

Post-Task Survey

System Usability Scale (SUS)

Human-AI Disagreement

Self Confidence and Trust in AI



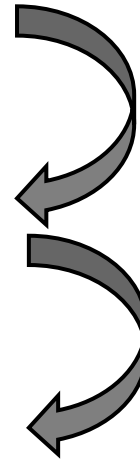
Legitimate



Phishing



Legitimate



**Disagreement!**

Can you explain why you disagree with the AI report?

# User Study Design

Pre-Task Survey

Message Evaluation Task

Post-Task Survey

System Usability Scale (SUS)

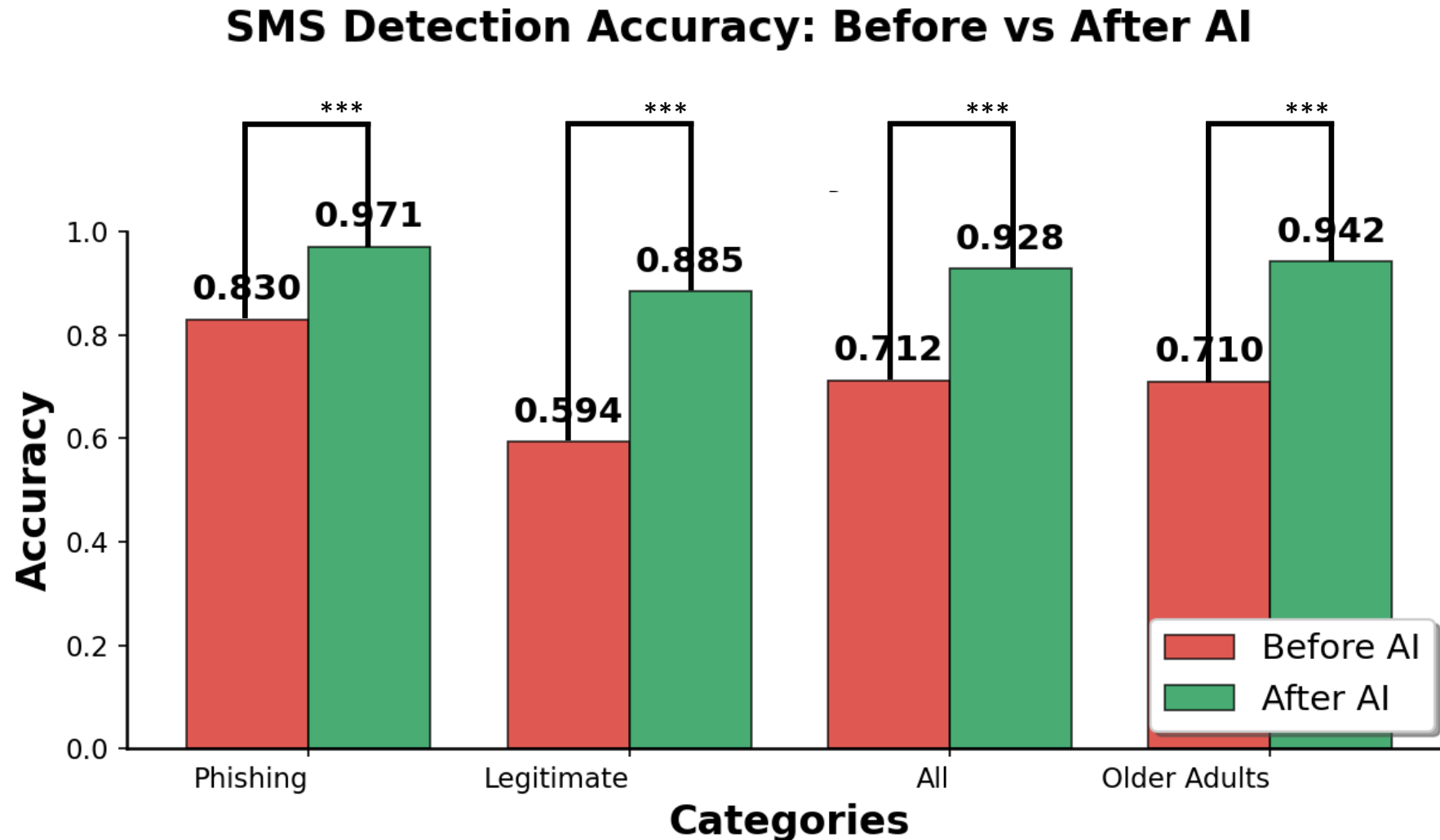
Human-AI Disagreement

Self Confidence and Trust in AI

- How would you rate your ability to recognize phishing (fraudulent) messages?
- How much do you trust the AI agent's judgement on SMS messages?

# User Study Results

- RQ1: Effectiveness.

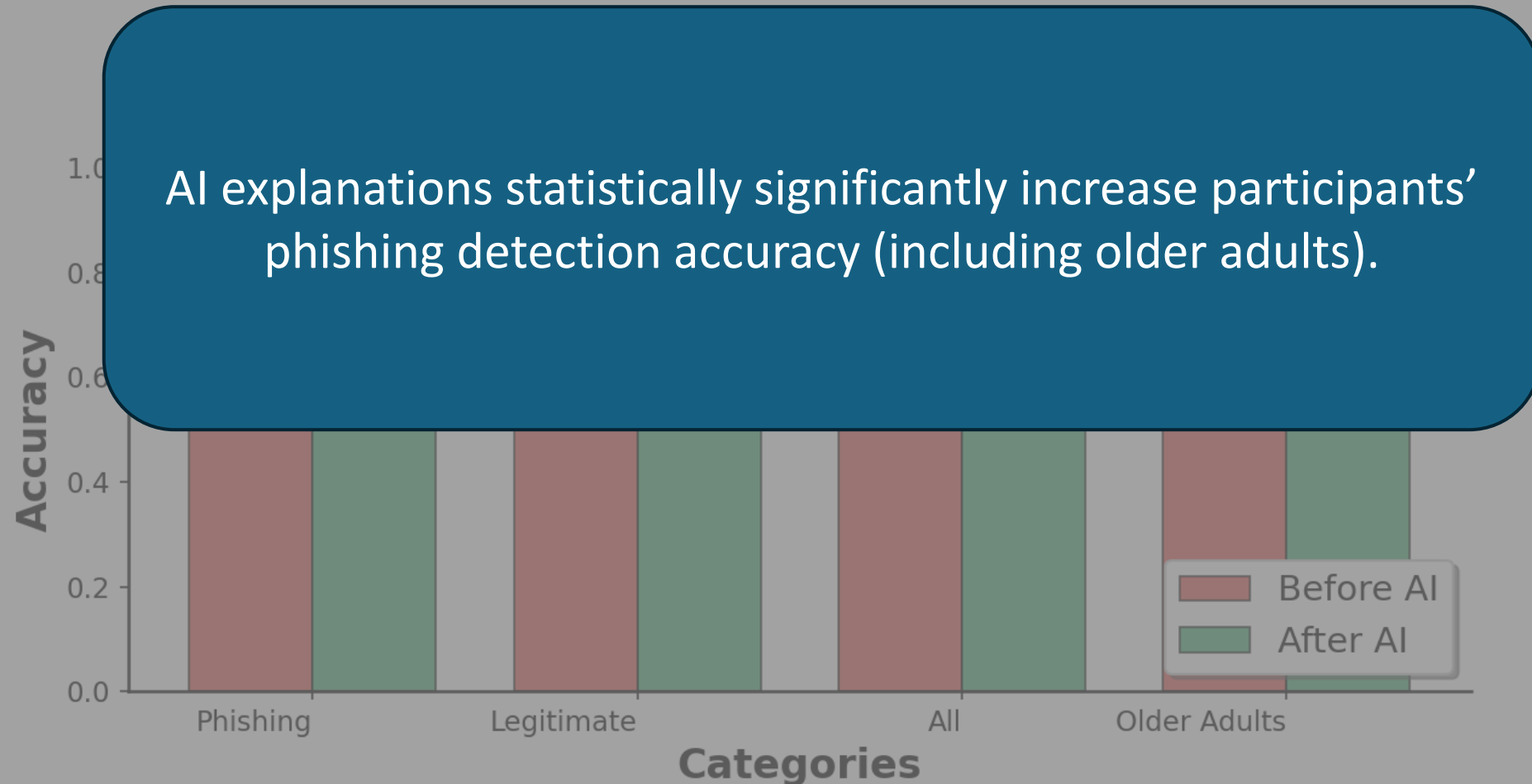


\*,\*\*,\*\*\* = statistically significant under T-test

# User Study Results

- RQ1: Effectiveness.

## SMS Detection Accuracy: Before vs After AI

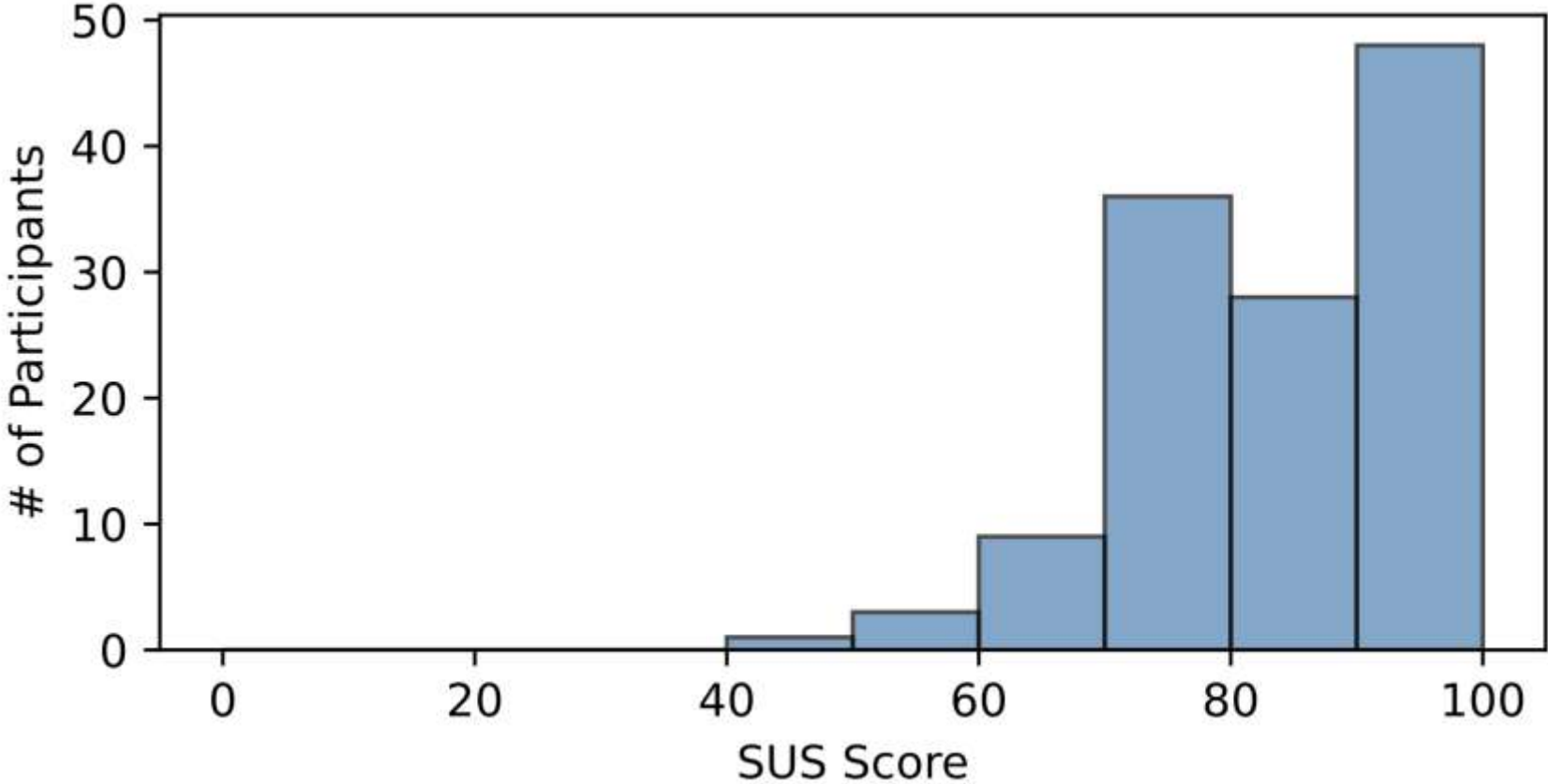


\*,\*\*,\*\*\* = statistically significant under binary mixed effect regression

# User Study Results

➤ RQ2: Usability.

Mean SUS: 82.6 (“Excellent”)

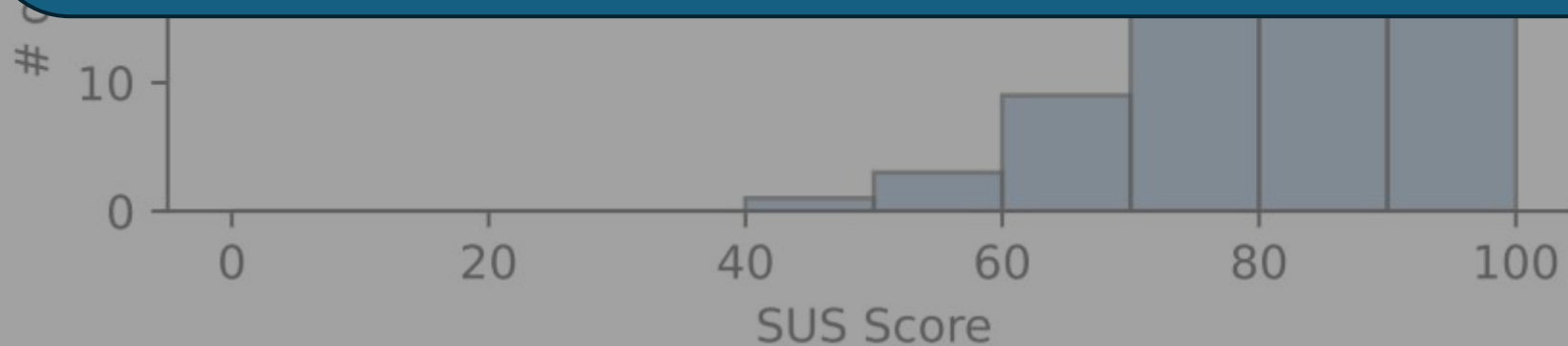


# User Study Results

- RQ2: Usability.

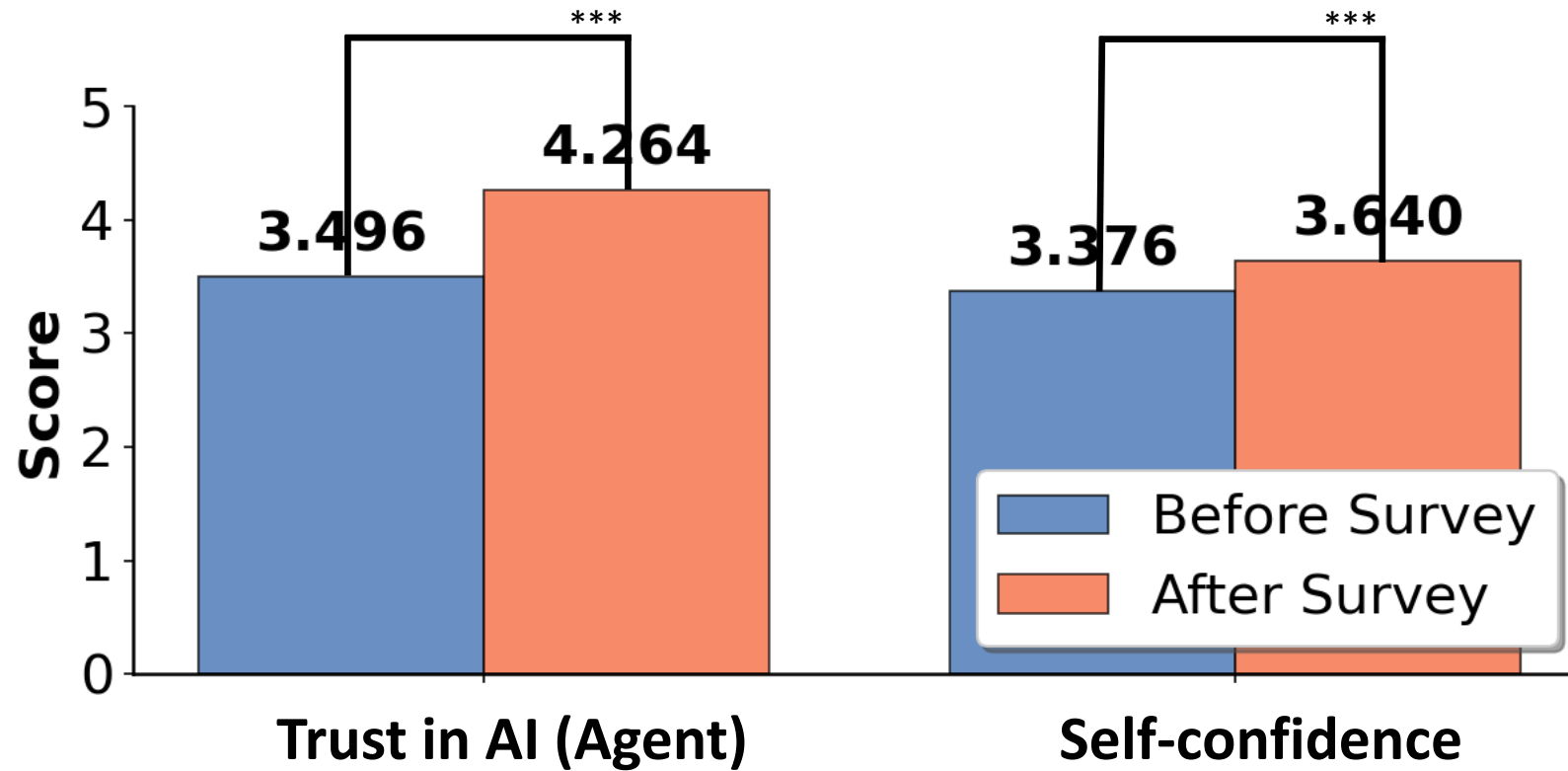
Mean SUS: 82.6 (“Excellent”)

Users give an “Excellent” (82.6) usability rating for our system.



# User Study Results

➤ RQ3: Trust and Disagreement.

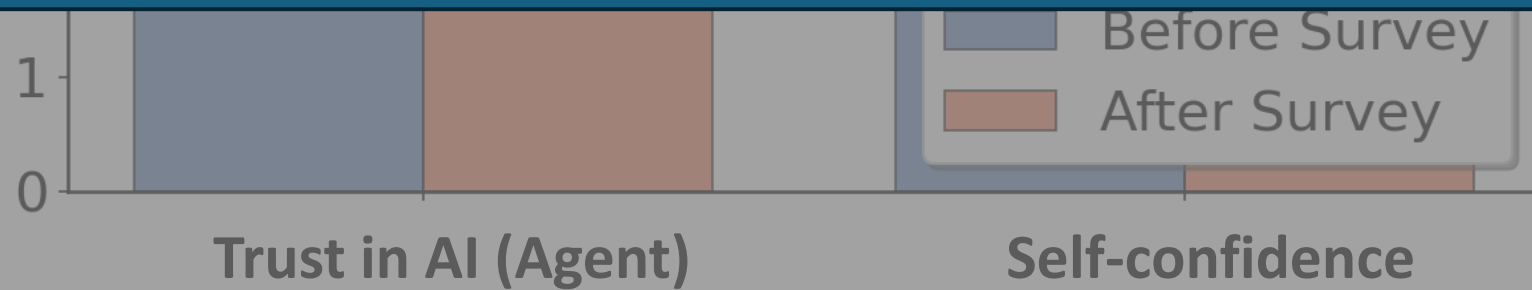


\*,\*\*,\*\*\* = statistically significant under T-test

# User Study Results

- RQ3: Trust and Disagreement.

Users' trust in AI agents and their self-confidence in identifying phishing messages both increase significantly after the survey.



# User Study Results

## ➤ RQ3: Trust and Disagreement.

More disagreement when AI determines **legitimate!**  
Users show **healthy skepticism.**

### 11.5% Disagreement on Legitimate SMS

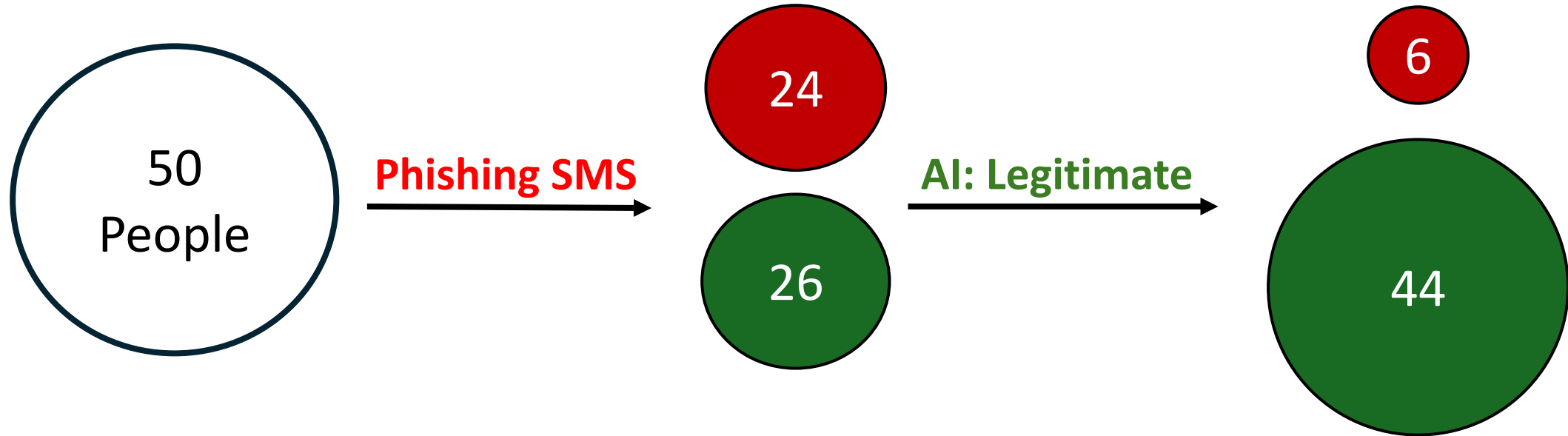
1. **(Shortened) URLs (n = 14)**
2. **Security Habits (n = 15)**
3. **Context Mismatch (n = 11)**

### 2.9% Disagreement on Phishing SMS

1. **Contextual Alignment (n = 3)**  
Expect legitimate notifications  
(package deliveries, etc.)
2. **Focus on Content (n=2)**  
Focus on content rather than  
technical indicators

# User Study Results

➤ RQ4: AI Errors. How do users respond to the AI agent's mistakes?



Users show over-reliance on AI decisions!!!

# User Study Results

Users show **over-reliance** on AI decisions!!!

➔ **Solution:** Adaptive explanations based on message type



Phishing Message



Legitimate Message



The SMS is likely a **phishing** message because...  
Do not click the link or reply the message!



No strong evidence of phishing was found, but please verify through official channels before responding.

# Summary

- **Develop SmishX:**  
**Detection + evidence-based explanations.**
- **Major Performance (N=175, 40% older adults)**  
User detection accuracy: 71.2% → 92.8%  
"Excellent" usability rating: SUS score of 82.6
- **Key Behaviors**  
Users have more disagreement with Agent on legitimate messages.  
Risk: Over-reliance on AI when errors occur

Our code, dataset, materials, codebooks are all available at  
<https://github.com/yizhu-joy/SmishX>

# Thank you!

Check out our paper!

<https://yizhu-joy.github.io/>  
[yizhu-wa21@mails.tsinghua.edu.cn](mailto:yizhu-wa21@mails.tsinghua.edu.cn)

