Protecting Patients against Phishing Attacks using AI-enabled Agents

Yizhu Wang[1], Haoyu Zhai[1], Chenkai Wang[1], Qingying Hao[1], Nick Cohen[2], Roopa Foulger[2], Jon A. Handler[2], **Gang Wang[1]***

University of Illinois Urbana-Champaign[1], OSF HealthCare[2]

MassAITC Aging Focus Pilot Core

Motivation

- **SMS phishing** poses a significant threat to **older adults**
- We use large language models (LLMs) to detect SMS phishing
- It also generates **evidence-based explanations** and actionable advice for users
- We develop a prototype called **SmishX** *using* chain-of-thought (CoT) reasoning of LLMs

Challenges

- Short message length
- URL shorteners
- Vague decision boundary (e.g., phishing is different from spam)

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

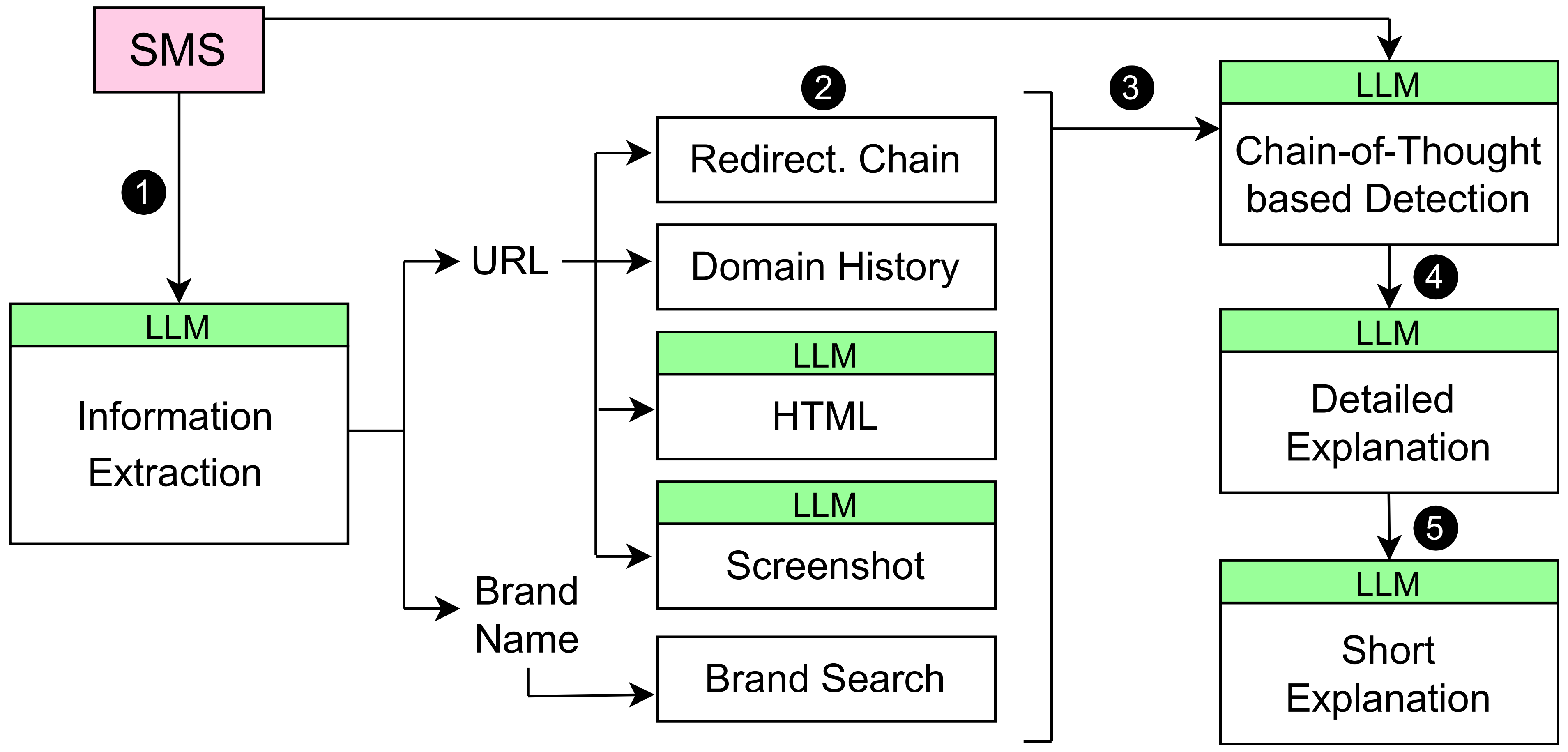
(a) Phishing SMS

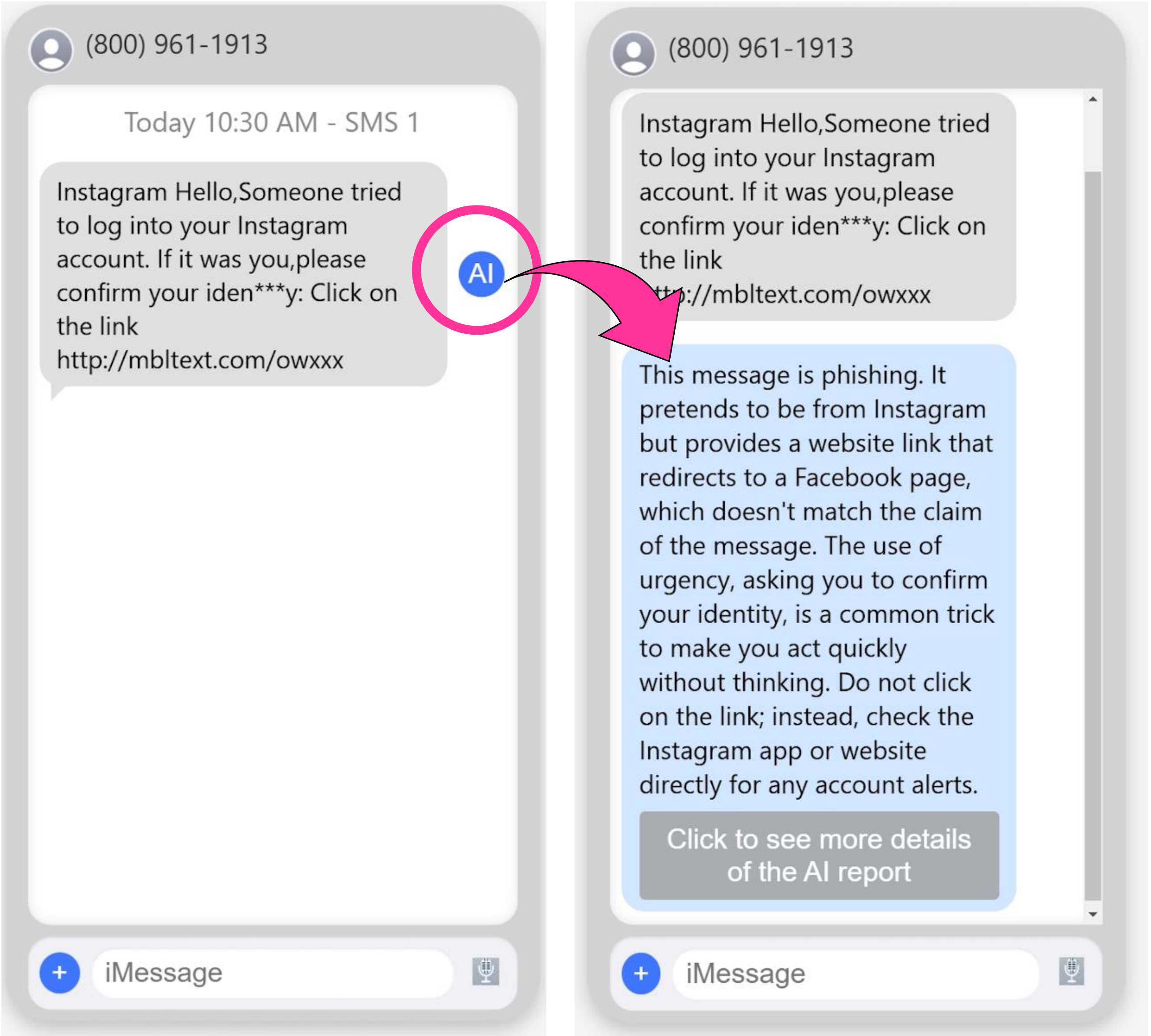
Decision This SMS message is likely a phishing attempt. **Explain.** It pretends to be from Netflix and warns about a payment failure to create urgency, which is a common trick to make people act quickly without thinking. The website link doesn't match Netflix's official website, and the strange spacing and grammar errors add to the suspicion. **Advice** 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.

(b) Short Explanation from SmishX

SmishX: System Design

- **Information Extraction LLM** (❶❷)
  Extract URL and brand, collect context info
- **Phishing Detection LLM** (❸❹)
  Chain-of-Thought reasoning to make a detection decision (*phishing, spam* or *legitimate*) and a detailed report
- **Explanation LLM** (❺)
  Convert the detailed report into a short explanation





Instagram Hello,Someone tried to log into your Instagram account. If it was you,please confirm your iden***y: Click on the link http://mbltext.com/owxxx

This message is phishing. It pretends to be from Instagram but provides a website link that redirects to a Facebook page, which doesn't match the claim of the message. The use of urgency, asking you to confirm your identity, is a common trick to make you act quickly without thinking. Do not click on the link; instead, check the Instagram app or website directly for any account alerts.

System Evaluation

SmishX achieves **98.8% accuracy** on real-world SMS datasets and can affectively suppress LLM hallucinations

User Study Results and Takeaways

User Study

- N=175 participants (40% are older adults of age 65+)

SMS	Accuracy Before AI	Accuracy After AI
Phishing	0.830	0.971
Legitimate	0.594	0.885
All	0.712	0.928
Older Adults	0.710	0.942

Study Findings

- **Effective:** participants' phishing detection accuracy improved from 0.712 to 0.928 after reading AI's explanations.
- **Usable:** SUS score 82.6, indicating "Excellent" usability
- What participants like about **SmishX**: explanation (n=60), ease of use (n=51), link verification ability (n=9), educational (n=7)
- **Human-AI disagreements**: 7.2% cases of disagreements, more common on legitimate SMS
- **Over-reliance on AIs** for decision-making: can be a potential issue if AI made errors

Future Directions

- Internal tests with OSF Healthcare employees/patients
- Integrating OSF-specific contexts

Acknowledgements

National Institute on Aging grant P30AG073107.