USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Can You Walk Me Through It? Explainable SMS Phishing Detection using LLM-based Agents

Yizhu Wang, Haoyu Zhai, Chenkai Wang, and Qingying Hao, *University of Illinois Urbana–Champaign;* Nick A. Cohen, Roopa Foulger, and Jonathan A. Handler, *OSF Healthcare;* Gang Wang, *University of Illinois Urbana–Champaign*

https://www.usenix.org/conference/soups2025/presentation/wang

This paper is included in the Proceedings of the Twenty-First Symposium on Usable Privacy and Security.

August 11–12, 2025 • Seattle, WA, USA

ISBN 978-1-939133-51-9

Can You Walk Me Through It? Explainable SMS Phishing Detection using LLM-based Agents

Yizhu Wang[1], Haoyu Zhai[1], Chenkai Wang[1], Qingying Hao[1], Nick A. Cohen[2], Roopa Foulger[2], Jonathan A. Handler[2], Gang Wang[1]
[1]*University of Illinois Urbana-Champaign*
[2]*OSF Healthcare*

Abstract

SMS phishing poses a significant threat to users, especially older adults. Existing defenses mainly focus on phishing detection, but often cannot explain *why* the SMS is malicious to lay users. In this paper, we use large language models (LLMs) to detect SMS phishing while generating evidence-based explanations. The key challenge is that SMS is short, lacking the necessary context for security reasoning. We develop a prototype called `SmishX` which gathers external contexts (e.g., domain and brand information, URL redirection, and web screenshots) to augment the chain-of-thought (CoT) reasoning of LLMs. Then, the reasoning process is converted into a short explanation message to help users with their decision-making. Evaluation using real-world SMS datasets shows `SmishX` can achieve an overall accuracy of 98.8%, outperforming existing methods. Through user studies ($N = 175$), we show that `SmishX`'s explanation can significantly improve users' phishing detection efficacy across age groups. Its usability is rated "excellent" by participants (SUS score 82.6). We conclude by discussing open challenges in resolving human-AI disagreements and safely handling AI errors.

1 Introduction

Phishing attacks through Short Message Services (SMS) [1] pose a significant threat to Internet users, especially older adults. The Federal Communications Commission (FCC) reports that the annual financial losses from SMS phishing have surpassed $231 million in recent years [24]. In particular, *older adults* are among the highly targeted user populations. The recent FBI IC3 reports [13,14] show that individuals aged 60+ filed over 100,000 fraud complaints a year and incurred the highest financial losses among all age groups.

Researchers have investigated defense methods against SMS phishing, with a focus on the *detection problem*. They aim to detect phishing SMS by analyzing textual/linguistic features in the message [7, 33, 34, 39, 75] or verifying the sender identity [84]. They often formulate the problem as a classification problem between legitimate and phishing SMS, and apply machine learning (ML) [7,34,39] (or large language models [72, 82]) as a solution. There are two critical issues with these approaches. First, real-world SMS messages are usually very short, lacking sufficient context information for ML models to make reliable detection decisions. Second, it is difficult for users to understand *why* a message is flagged as phishing due to a lack of explanation on the detection result.

Our Motivation. The recent development of Large Language Models (LLMs), especially in their reasoning and tool-use capabilities, makes us wonder whether it is feasible to develop a system that can *explain* phishing detection results to *lay users*. Our vision is that the explanation should be *evidence-based*, i.e., derived from a series of analyses of the specific input message (instead of making generalized claims). As such, the explanation process must be closely coupled with the phishing detection steps. In addition, given the short length of SMS messages (and their use of shortened URLs), the analysis may need to go beyond the linguistic features of the message content and require external contexts to reason about the message's intent. Finally, the generated explanations should be tested with lay users to make sure they are understandable and useful.

System Design. With the above considerations, we design a prototype system called `SmishX`. We take an Agentic AI approach to perform phishing detection and explanation, instructing LLM agents to analyze SMS messages in a similar way as *a human security analyst*. The analysis contains three steps: (1) information extraction (e.g., to extract URLs and brand names), (2) collecting context information by calling

USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

external tools (e.g., collecting URL redirection chains, domain history, webpage screenshots, and brand information), and (3) using the original SMS and the extra context data to perform chain-of-thought (CoT) reasoning to determine if the message is a phishing SMS.

To explain the result to lay users, we first prompt LLM to generate a detailed report based on the step-by-step analysis, and then summarize the report into a short explanation message. Figure 2 shows an example. The explanation message is semi-structured, starting with a statement sentence on the phishing detection *decision*, followed by an *explanation* of the analysis result, and also the *advice* to help users take safe actions. Compared with prior works that use LLMs for phishing detection [20, 40, 47], the main difference is that `SmishX` goes beyond just relying on linguistic features. It also leverages LLMs' tool-use capabilities to collect external contexts to aid the reasoning and explanation generation processes. Our evaluation (Section 4) and user studies (Section 6) show that the context information is critical to ensure a high detection accuracy and the quality of the explanation.

Evaluation. To evaluate the proposed system, we have compiled and manually labeled a real-world SMS dataset of 1,200 messages. We show `SmishX` can detect phishing SMS accurately with an overall accuracy of 98.8%, outperforming existing methods. We find that context information is particularly important in reducing false positives, and there is a benefit of separately handling spam (SMS that promotes products and services) and phishing messages (SMS that deceives users to take unsafe actions), given their different threat models. Finally, we show that the evidence-based reasoning process effectively suppresses LLM hallucinations [35] in the generated explanations.

User Studies. We have conducted two user studies. The first study (N=125) is to assess the effectiveness (RQ1) and usability (RQ2) of AI-generated explanations and explore how users resolve potential disagreements with AI (RQ3). Then we run a second study (N=50) where we intentionally insert AI errors to examine user response (RQ4). For both studies, we have oversampled *older adults of age 65+* (40% of our participants) to ensure the system works well for them.

Our studies have led to several key findings. *First*, we find that `SmishX`'s explanation can significantly improve users' phishing detection efficacy. Participants' detection accuracy improves from 0.712 (before reading AI explanations) to 0.928 (after reading AI explanations). In particular, older adults have reached an accuracy of 0.942 after AI explanation. *Second*, the usability of our system is rated "excellent" based on the System Usability Scale (SUS) [8, 11]. The mean SUS score is 82.6 (85.3 for older adults). We observe that "users' trust in AI" is significantly correlated with the perceived usability of the proposed system. *Third*, we find users do not always agree with AI decisions (on 7.2% of the messages), which happen more on *legitimate SMS* (11.5%) than on phishing SMS (2.9%). A common reason is that users tend to remain skeptical of legitimate SMS containing *shortened URLs* despite that the AI has verified the URLs. *Fourth*, we also observe an over-reliance on AI for decision-making, and only a small portion of participants can rectify the intentionally inserted AI errors. We further discuss possible ways to mitigate the impact of AI errors and the open challenges in this direction.

The prototype of `SmishX` has received interest from industrial partners. We are working with our partners at a healthcare system to perform internal tests on their employees and patients to improve existing phishing defense. We also release our code and datasets [85].

Contributions. Our paper has three main contributions.

- First, we proposed a new Agentic AI system to detect and explain SMS phishing messages for lay users.
- Second, using a real-world SMS dataset, we demonstrated that the system was highly accurate and could suppress LLM hallucinations in the generated explanations.
- Third, we conducted user studies to verify the effectiveness and usability of the explanation. We additionally explored how users responded to human-AI disagreements and AI errors under the context of phishing SMS detection.

2 Related Work

Phishing is a common attack in which adversaries *impersonate* a trusted entity to lure the victim into revealing personal information or taking insecure actions [32, 83]. Most existing works are focused on phishing emails [25, 31, 43, 46, 51], phishing websites [48, 64, 89, 90], and phishing URLs [79]. Below, we mainly discuss *SMS* phishing detection methods.

2.1 Phishing SMS Detection

Sender-based Defenses. To detect phishing SMS, one direction is to check the authenticity of the sender's phone number (caller ID) using caller ID authentication mechanisms [55, 84], e.g., those in STIR/SHAKEN [65]. However, this cannot fundamentally prevent phishing—adversaries can still claim that the message is from a trusted party (e.g., a well-known bank) *without spoofing* the caller ID.

Content-based Defenses. Another direction is to analyze the SMS content to identify indicators of phishing, such as the use of shortened URLs [36, 44] and emotional tactics (fear, curiosity, a sense of urgency, and greed) [28, 58]. Existing phishing detection methods typically craft *features* based on such adversaries' behaviors/tactics, and then use rule-based methods [33] or machine learning models [7, 34, 39, 75] to classify phishing messages from legitimate ones. These methods can be further enhanced by URL analyses (to look for indicators such as file downloading [36] and domain name mismatches [27]).

Our method belongs to "content-based" defense. Compared with the existing effort, a key difference is that we do not need to *manually* craft the detection features. We explore the possibility of using an LLM agent to perform these tasks automatically. More importantly, we focus on explaining the reasons behind the phishing detection results.

2.2 LLM for Phishing Detection

Large language models (LLMs) have been used for both assisting phishing attacks [29, 49, 71, 73] and augmenting phishing defenses [12, 15, 20, 30, 40, 41, 45, 50, 57] (most of which are focused on phishing emails and websites). Below, we focus on existing *defense* solutions against *phishing SMS*, and discuss our key differences with them.

Using LLMs to Detect Phishing SMS. LLM has been used to *enhance* conventional machine learning models via data augmentation [74]. Others have used an LLM as a static classifier to perform *spam SMS detecting* [72, 82]. Spam messages (e.g., promoting a product or service) have a different threat model compared with phishing attacks (e.g., deceiving users to perform unsafe actions) and should be handled differently (see Section 3.1). More importantly, we advance the state-of-the-art by directly using LLMs' tool-use and reasoning capabilities to collect external contexts, perform phishing analyses, and generate explanations for users.

Using LLMs to Explain Phishing SMS. A few recent works have used LLMs to generate *explanations* for the phishing detection results. Uddin et al [82] use post-hoc explanation tools such as LIME to highlight keywords that most contribute to ML classification results. The most related work to ours is a Korean-centric LLM system [47] customized for the Korean language. The main technical difference is that their detection and explanation are solely based on *linguistic features* within the SMS content. We find this approach is insufficient for real-world SMS messages and *extra contexts* (on URLs, brands, webpages) are needed for evidence-based phishing detection and explanation (see Section 4.2).

2.3 Human Factors

Human factors are important to understand how and why phishing attacks work [21]. Existing research on *email* and *web*-based phishing has looked into related topics such as user awareness training [69], behavioral interventions [16], phishing warnings [23, 88], and user susceptibility [59, 91] and resilience [86].

Regarding human factors in SMS phishing, a recent study measures the impact of SMS phishing by sending phishing messages to their participants [68]. They find that 16% of the participants are vulnerable. In another study [77], researchers investigate how users evaluate SMS legitimacy, finding that message content, formatting, and embedded URLs receive more attention than the sender information. They also find that participants are often confused by spam messages (that promote products/services). A more recent study [80] has reported that users are better at recognizing phishing SMS than legitimate SMS, and rely on different heuristics to detect them. Our work is complementary to existing efforts since we study how AI-generated explanations can help users better recognize phishing attempts.

Phishing against Older Adults. Older adults are a highly targeted user population by SMS phishing attacks [13, 14]. A recent survey has reviewed 82 phishing studies [6]: some studies have concluded that older adults have higher accuracy in identifying phishing threats [22, 63] while others have observed the opposite [78]. In this paper, we over-sampled older adults to test our system.

3 System Design

In this section, we describe the `SmishX` designs for phishing detection and explanation. Figure 1 presents an overview.

3.1 Challenges and System Overview

Challenges. SMS phishing detection has several unique challenges. First, *short message length*: SMS messages are often extremely short, with limited information. Also, URL shorteners are commonly used in both legitimate and phishing messages, making it difficult to determine the message's legitimacy based on the content alone. Metadata such as caller IDs (phone numbers) are also unreliable since they can be easily spoofed [55]. Second, *vague decision boundary*: Between legitimate and phishing messages, there is a wide spectrum of messages in the gray area. A typical example is *spam* messages that promote both legitimate services/products and potentially harmful ones.

Threat Model and Design Goals. We define *phishing SMS* as messages that impersonate a trusted brand/organization (e.g., a bank or a government agency) to lure users to take harmful actions (e.g., clicking on a malicious URL, calling the scammer's phone number, or giving away sensitive data such as pin codes, passwords, and credit card information).

We define *spam SMS* as messages sent *in bulk* to promote potentially harmful services/products. A key difference from phishing is that spam SMS *does not deceive users regarding who the sender is* (i.e., no impersonation of another brand). During data labeling (Section 4.1), we classify promotional messages for services such as gambling, digital currency, and sex services as "spam." However, messages sent in bulk from legitimate businesses/organizations (e.g., delivery notifications) are still classified as "legitimate."

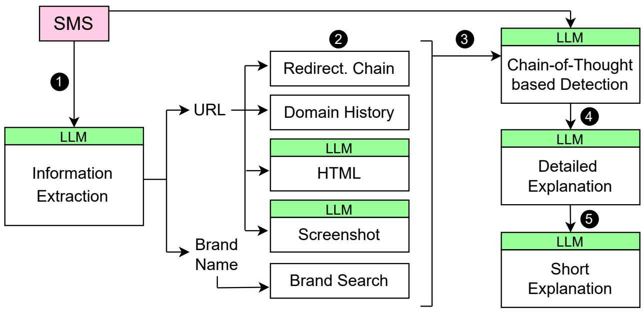


Figure 1: Workflow of `SmishX`: an LLM-based phishing SMS detection and explanation system.

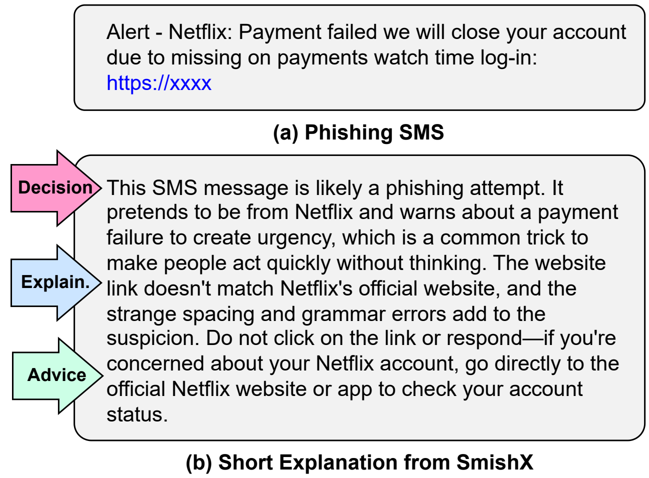


Figure 2: Example SMS and our system's short explanation.

The primary goals of `SmishX` are to detect and explain *phishing SMS*. However, considering the potential harm of spam, `SmishX` will be tuned to flag harmful *spam SMS* too.

System Overview. The workflow of `SmishX` is shown in Figure 1. Given an input SMS, the *Information Extraction LLM* first extracts information such as brand names and URLs (❶) and then collects further context information (❷). Then the *Phishing Detection LLM* uses the context information and the original SMS to perform analysis with Chain-of-Thought (CoT) prompts. It decides whether this message is legitimate or not (❸) and generates a detailed report about the analysis results (❹). However, considering the details report is difficult for users to read, the *Explanation LLM* summarizes the report into a *short* explanation message (❺). Figure 2 shows an example SMS and its explanation.

We use ChatGPT-4o [61] to implement `SmishX`. We have tested other commercial LLMs including ChatGPT-4 [60], Meta AI [52], Gemini [2], Bing Chat Copilot [53] and found that ChatGPT-4o (gpt-4o-2024-08-06) is better for our purposes. We also tested *open-source* LLMs including Qwen2.5 [67] and DeepSeek-v3 [19] to show that `SmishX` can be implemented with a local LLM as well (see Appendix A).

3.2 Information Extraction LLM

Considering an SMS message is usually short, we first use an LLM to extract URLs and brand names from the messages, and then call external tools to collect additional context information. For example, from the message in Figure 2 (a), the agent will extract the URL `https://xxxx` (masked for readers' safety) and the brand name "Netflix." Then we collect further information that can be potentially helpful to phishing detection, including URL redirection chain, domain history, web HTML and screenshots, and brand information. Our innovation is *not* these information collection steps themselves—the novelty is about integrating the holistic context to generate *well-grounded* explanations and reduce LLM hallucinations.

Redirection Chain. Based on the *URL*, the agent traces the complete redirection chain using the `Requests` library [70]. Phishing SMS often contains "shortened" URLs where the true destination and domain names are hidden [56]. Therefore, expanding the URL to reveal the final destination is needed. Also, attackers often apply cloaking techniques to hide their phishing websites from security companies through a chain of intermediate sites (even including legitimate sites) [56]. Analyzing the redirection chain can provide useful context.

Domain History from WHOIS. Based on the *URL* (and the final destination), the agent collects the domain name history from WHOIS [18]. The intuition is that attackers often register new domain names for phishing attacks [9, 38, 56]. As a result, recently registered domains (under less reputable registrars) that claim to represent well-established brands are strong indicators of phishing.

HTML Content. Given the final destination, the agent analyzes whether the website matches with what is claimed in the SMS. This is done by LLM's content summarization capability. It calls the JINA Reader API [3] to extract HTML content, remove HTML tags, and only retain the plain text. Then the agent summarizes the text to infer its identity (e.g., whether this is Netflix's website).

Web Screenshot. The web screenshot analysis has a similar purpose as the above HTML analysis. The reason to run both is to improve the robustness. The agent uses Puppeteer [17] scripts to capture the full webpage screenshots and analyze them using a vision-enabled LLM. The LLM is instructed to identify logos, brands, and key visual elements, and describe the website regarding its content and purpose.

Brand Information. Finally, given the *brand* name (e.g., "Netflix"), the agent seeks to obtain the true domain name associated with this claimed brand. This is useful context because the Phishing Detection LLM (in Section 3.3) can check whether the URL in the SMS will lead users to a website under the brand's *true domain name*. More specifically, the agent first queries the search engine (e.g., Google) with the claimed brand name ("Netflix"). Then it obtains the domain name of the top-ranked website on the search engine (`netflix.com`) as the true domain name for this brand. The Phishing Detection LLM can later match the domain name "`netflix.com`" with that of the destination of URL "`https://xxxx`". A mismatch would be an indicator of phishing. The intuition is that the true website of a well-established brand should rank higher than the phishing website under the target brand name keyword [90].

3.3 Phishing Detection LLM

The phishing detection LLM takes both the extracted context information (from ❷) and the original SMS message as the input. To detect phishing SMS and more importantly explain the reasoning process, we employ a Chain-of-Thought (CoT) approach, guiding the LLM through a structured, step-by-step reasoning process to ensure *evidence-based* analysis. The goal is to reduce hallucination and improve the explanation quality. This corresponds to steps ❸ and ❹ in Figure 1. We present the full prompts in Appendix D.

Assigning Roles and Tasks. Assigning specific roles tailored to the task to the LLM is known to enhance the model's response [42]. Here, we assign LLM the role of a phishing/spam detector to analyze SMS messages and classify them as either *malicious* or *legitimate*.

Prompting to Handle "Spam." The initial version of `SmishX` was only prompted to detect "phishing" SMS, which had a subpar performance on "spam". Considering spam messages are *unsolicited* and can be potentially harmful, we put them into the "malicious" category and prompt the LLM to detect them. This is done by feeding LLM with the definition and description of spam messages (see Section 3.1).

Chain-of-Thought (CoT) Reasoning. We implement Chain-of-Thought (CoT) prompting [87] by dividing the task into seven subtasks. Subtasks 1–4 involve identifying characteristics of *legitimate* messages (e.g., conversations between friends and family members, notifications from known organizations) and *spam* (e.g., messages promoting high-risk services such as gambling). Subtask 5 is focused on identifying *phishing* cues such as using urgent/alarming languages, embedding suspicious links, requesting personal information, inconsistencies in URLs and domain names, and grammar/spelling errors. Subtask 6 is about analyzing shortened URLs (and their legitimate use cases). Subtask 7 is about analyzing the extra context information provided by the *Information Extraction LLM* (from ❷). All the context information is attached at the end of the prompt template to support the LLM's analysis.

Classification and Reporting. We prompt the LLM to classify the SMS based on its analysis and generate a *detailed report* (in a JSON format). The detailed report includes the extracted brand name (if applicable), URLs (if applicable), detailed reasons for the classification result, and the final decision on whether the message is malicious or legitimate. We intentionally put the *reasoning* steps before the *decision* step to encourage evidence-based decision-making.

3.4 Explanation LLM

The Explanation LLM transforms the detailed analysis report into a concise, user-friendly explanation message with *no more than four sentences*. As shown in the example in Figure 2 (b), we instruct the LLM to include three key components. First, *decision*: it starts with a clear statement on the detection decision (i.e., whether the message is phishing, spam, or legitimate). Second, *explanation*: it summarizes key reasons behind the detection decision. We prompt the LLM to use non-technical languages and provide evidence to support their reasoning. Third, *advice*: the explanation LLM provides users actionable advice based on the analysis results. For example, if the SMS is flagged as phishing/spam, the LLM will advise users not to click on any links in the message, and use the official channels to verify the information. Extra examples are presented in Appendix E.

4 Evaluation

In this section, we evaluate `SmishX`'s detection accuracy and explanation quality.

4.1 Dataset

We have compiled and re-labeled a real-world SMS dataset of 1,200 SMS messages. The malicious set is sampled from publicly available datasets [4, 54, 56, 81]. The legitimate set is primarily sampled from existing datasets [37,54] and researchers' personal archives (22 extra messages). Considering `SmishX`'s analysis relies on URLs, we sample and include messages (both benign and malicious) whose URLs are still active. The rationale is that a phishing SMS with an expired URL is no longer a threat. We do not exclude messages without URLs. The final dataset contains 1,200 messages, including 622 legitimate messages and 578 malicious messages (259 phishing and 319 spam messages). 636 messages (53%) contain URLs.

When manually inspecting existing datasets, we find many incorrectly labeled messages. For example, notification messages from benign businesses/organizations are often incorrectly labeled as "phishing." Also, not all the existing datasets explicitly differentiate spam from phishing messages in their labeling. To ensure the reliability of labeling, one researcher coded the entire dataset of 1,200 messages into three categories ("phishing", "spam", and "legitimate") and developed a codebook. A second researcher then independently coded the full dataset using this codebook. After independent labeling, they meet to discuss their results and resolve disagreements. The codebook and dataset is shared [85].

SMS	SmishX	Baseline 1 (w/o Context)	Baseline 2 (w/o Spam P.)
Phishing	**100.0%**	100.0%	98.5%
Spam	**99.1%**	99.4%	64.6%
Legitimate	**98.2%**	92.6%	98.9%
Overall	**98.8%**	96.0%	89.7%

Table 1: SMS classification accuracy. We compare SmishX (results in bold) with two ablation baselines.

System	Overall	Malicious	Legitimate
SpaLLM-Guard [72]	96.8%	99.8%	93.9%
PhishE [39]	85.0%	74.2%	93.1%
SmishX **(Ours)**	**98.8%**	99.5%	98.2%

Table 2: Accuracy comparison with existing methods. SmishX outperforms existing methods, with the additional advantage of generating explanations.

4.2 Accuracy Evaluation

Baselines. First, we include two ablation baselines by *omitting* certain modules from SmishX (Table 1). *Baseline 1* directly applies the Phishing Detection LLM on the input SMS *without collecting the extra context information* (i.e., omitting ❶ and ❷). This mimics an existing LLM-based method to detect *phishing emails* [40]. *Baseline 2* removes the spam prompt in ❸. Additionally, we compare SmishX against existing solutions whose code/models are available (Table 2), including PhishE [39] (a ML-based detector) and SpaLLM-Guard [72] (an LLM-based detector).

Classification Accuracy. As shown in Table 1, SmishX has an overall accuracy of 98.8% (classifying legitimate and malicious messages), outperforming baseline 1 (96.0%) and baseline 2 (89.7%). In particular, SmishX has an accuracy of 100% in classifying phishing SMS, and an accuracy of 99.1% on spam. Its accuracy on legitimate messages is 98.2%. The few errors are caused by messages between family members requesting sensitive information, e.g., "*Sent me ur email ID soon? I will send that*." Such errors can be potentially addressed by incorporating an allow-list of trusted contacts.

Baseline 1 confirms the impact of the *context information*. Without the extra context, the accuracy on legitimate SMS is dropped from 98.2% to 92.6%. Within this category, there are 70 messages with URLs. We find that the baseline 1 accuracy on these messages is even lower (only 71.4%). These results confirm that directly applying LLMs [40] for phishing SMS analyses is challenging without the external contexts.

Baseline 2 confirms the importance of spam-specific prompts. Without spam prompts, the accuracy on spam messages drops from 99.1% to only 64.6%. It is necessary to teach the LLM the knowledge (or definition) of harmful spam to facilitate accurate detection.

SMS	Completeness	Factual Consistency
Phishing	3.00	1.00
Spam	2.96	1.00
Legitimate	2.99	0.98
Overall	2.98	0.99

Table 3: Evaluation of explanation quality. The completeness score has a range of 0–3. The factual consistency score has a range of 0–1. Higher scores are better for both.

Table 2 shows that SmishX achieves the best performance in comparison with PhishE [39] and SpaLLM-Guard [72]. Note that these existing solutions do not provide explanations (which is our main contribution).

4.3 Quality of Explanation

We next evaluate the quality of the *short explanation*. Out of the 1,200 SMS messages, we have sampled 180 messages to manually inspect their explanation quality (including 40 phishing, 50 spam, and 90 legitimate messages). Two coders have *independently* rated these explanation messages from two aspects: *explanation completeness* and *factual consistency*. The agreement rate between the two researchers is 98.3% across their ratings, and Table 3 reports the average ratings between the two coders.

Completeness. As discussed in Section 3.4, we instruct the Explanation LLM to construct the outputs with three components: decision, explanation, and advice. The completeness rating, with a scale from 0 to 3, measures how many components are included in the explanation message. As shown in Table 3, SmishX achieves a high completeness score of 2.98 (out of 3), confirming its ability to follow the instructions to generate the desired structure for the explanation message.

Factual Consistency. Factual consistency assesses whether the explanations contain fabricated or misleading information. Such information could be introduced by LLM hallucinations [35]. For factual consistency, we annotate the explanation with a binary score. The score is 1 if all the provided explanations are factually correct. The score is 0 if the explanation contains at least one hallucinated argument. Table 3 shows SmishX achieves a factual consistency score of 0.99 (out of 1). Both phishing and spam categories have a full score of 1.00. There were two error cases on *legitimate messages* due to failed URL redirections.

5 User Study Methodology

We conduct user studies to answer four research questions:

- **RQ1: Effectiveness.** How effective is the AI agent in helping users recognize phishing SMS, and what factors influence its effectiveness?
- **RQ2: Usability.** How do users perceive the usability of the AI agent, and what factors affect their experience?
- **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?

We designed two studies: Study A is focused on RQ1–RQ3, where participants read SMS messages and the explanations from `SmishX` to perform phishing detection tasks. Study B is a follow-up study to answer RQ4 where we intentionally insert errors in the AI explanation to study user reaction. The two studies have a near-identical workflow.

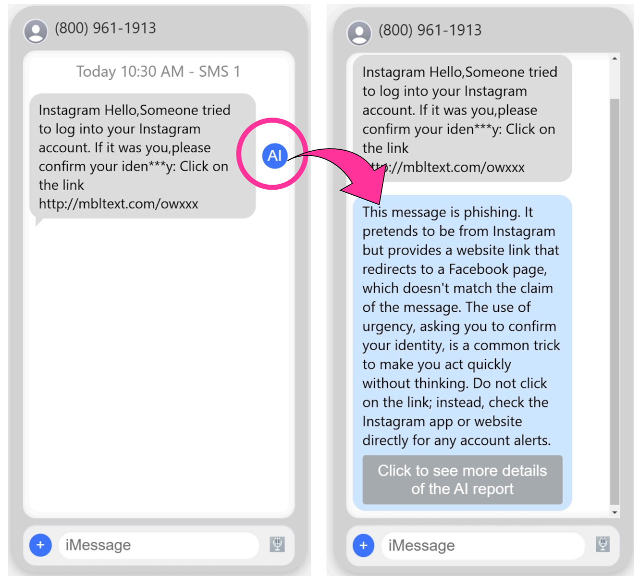
5.1 Study Workflow

We design the user study in the form of online surveys. Figure 5 (Appendix) describes the workflow from participants' perspectives. The full question list used in our survey is provided in the supplementary materials [85].

Participant Onboarding. Participants join the study by visiting the survey website, reading the consent form, and providing their consent. They will then read a brief introduction about the study, familiarizing themselves with the definitions/concepts of "SMS", "phishing", and "AI". Then they answer 9 questions related to their demographics (i.e., age, gender, race/ethnicity, and education level), their technical background in computer science and engineering, SMS usage and prior encounters of SMS phishing, self-confidence in their phishing detection ability, and their general trust in AI. After that, participants will go through a *tutorial* on how to interact with `SmishX` prototype to evaluate SMS messages.

SMS Message Evaluation. After the tutorial, participants will perform SMS evaluation tasks. They will read a sequence of SMS messages and determine whether each message is legitimate or phishing. The order of the messages is *randomized* for each participant to minimize the potential biases from the order effects [76]. Under each message, participants will answer two questions.

- *Before AI Assistance:* At first, participants will read the SMS message (Figure 3 (a)) and answer question Q1: "Do you think this message is a legitimate message or a fraudulent phishing message?"
- *After AI Assistance:* Only after the above question is answered, an "AI" button will pop up, which allows the participant to click and view the AI-generated explanation (Figure 3 (b)). Then the participant will answer question Q2: "After reading the AI report, do you now believe this message is legitimate or fraudulent?" Here, we present



(a) SMS **(b) SMS + AI Explanation**

Figure 3: Screenshots of the UI used in the user study.

the *short explanation* from `SmishX`. At the bottom of the explanation, there is a button that users can click to read the *detailed* AI report which contains extra analysis details and a screenshot of the webpage behind the URL in the SMS (see an example in Appendix E). To make sure participants pay attention to the explanation, we use a *typewriter animation* to type out the text in the explanation message (20 ms per character).

Usability Evaluation. After the message evaluation, the participants will rate the usability of the system using the System Usability Scale (SUS) [11]. SUS contains 10 standard questions to assess the usability of a system and each question has selection opinions on a 5-point Likert scale from "Strongly Agree" to "Strongly Disagree." Participants read questions such as "I think that I would need the support of a technical person to be able to use this system" and "I found the system unnecessarily complex" and then provide their ratings. The ratings will be aggregated to produce an SUS score ranging from 0 to 100. To ensure the participants are attentive, we insert one attention-check question (which asks the participants to select "Strongly Disagree" directly).

After the participants provide their SUS ratings, we further ask them two open-ended questions to describe (1) what they like about the AI agent system; and (2) their concerns, complaints, and suggestions for improvements.

Exit Questions. We briefly discuss the exit questions of our survey—the full questions and options are available here [85].

- **Disagreement with AI.** To understand *why* users may disagree with the AI *after reading the AI explanation*, we ask a follow-up question here. We randomly pick two SMS messages where the participant's determination is different from that of AI after reading the AI explanation. The disagreements fall into two categories: (1) the participant initially disagreed with the AI and stuck to their original decision after reading the explanation; (2) the participant initially agreed with the AI decision but changed their opinion after reading the explanation. Here, we present the SMS, the AI explanation, and the participant's original answer to this participant again, and ask them to reflect on the disagreement and explain their reasons (using a text box). If a participant has more disagreement messages, we only present two (one for each type). If a participant only has one type of (or no) such disagreement message, we will present the one type they have (or skip this question).

- **Self-Confidence.** To understand whether interacting with the AI agent has changed the participants' self-confidence in their phishing detection ability, we insert a question here. Recall that, before the survey (during "onboarding"), we have asked participants a question to self-assess their phishing detection ability. Here we ask the question again to measure potential changes.
- **Trust Towards AI.** We also want to understand whether the participant's trust in AI has changed after the interaction with the AI agent. Before the survey (during "onboarding"), we have asked about their general trust towards AI as the baseline. Here, after the survey, we ask about their trust level towards the AI agent in the study, for a comparison.
- **Detailed vs. Short Explanations.** Finally, we ask the participants regarding their preferences for the *detailed AI report*. This is to determine whether the detailed AI report is necessary (for future iterations of the system design).

5.2 Study A vs. Study B

We run two versions of the studies under the same workflow to answer different research questions.

Study A. This study (for RQ1-RQ3) directly uses SMS messages sampled from our evaluation dataset and the explanations produced by `SmishX`. In total, we use 10 SMS messages including 5 phishing and 5 legitimate messages. We do not consider spam messages since they are not the main focus of the study. These SMS messages are selected to cover diverse topics including online shopping, delivery services, banking, mobile services, app notifications, and government notices (see the supplementary materials [85]).

Study B. This study (for RQ4) is to evaluate how users react to potential AI errors. We use the same 10-message set from Study A and then intentionally insert a new message that can trigger AI errors. Here, we only insert a false negative (FN) error (i.e., a phishing SMS is determined to be legitimate by the AI) because this type of error leads to harm to users. Recall that `SmishX` did not make any FN error on phishing messages (Section 4.2), and thus we need to hand-craft an error SMS ourselves. After multiple rounds of experimentation, we successfully crafted an FN message (see Appendix C).

Other Considerations. We use realistic (phishing) SMS messages in the study and take active steps to ensure the safety of the participants. In particular, URLs are not directly clickable in the messages. We also convert parts of the URLs into "`xxxx`" so that users cannot accidentally visit a phishing site even if they type the URL into the browser's address bar. When possible, "`xxxx`" is only applied to the *path* part of the URL so that participants can still see the domain names (e.g., `tinyurl.com/3p7exxxx`). Finally, we have proactively informed the participants that parts of the URLs are intentionally masked for safety considerations and asked them not to guess or visit the URLs.

5.3 Participant Recruitment and Ethics

In total, we recruited 175 participants from Prolific [66]. The demographic information of participants is shown in Table 10 in the Appendix. We recruited 125 participants for Study A, and a different group of 50 participants for Study B. All participants were based in the United States and were compensated at a rate of $3 per survey (equivalent to 12 USD/hour). The study received approval from our Institutional Review Board (IRB). Informed consent was obtained from all participants prior to their involvement. Participant data was anonymized to uphold privacy and confidentiality standards.

Since we wanted to test our system with *older adults*, we "oversampled" this user population using Prolific's targeted recruiting function. Out of the 125 participants in Study A, 50 participants are at the age of 65 or older (40%). Similarly, 20 out of 50 participants in Study B (40%) are older adults.

6 User Study Result Analysis

In this section, we analyze the data collected from the user studies to answer our research questions.

6.1 RQ1: Effectiveness

We use the data from Study A to examine how effective the AI agent is in helping users identify phishing SMS (RQ1).

Phishing Detection Accuracy. First, we compare the user's phishing detection accuracy *before* and *after* reading the AI-generated explanations. The result is presented in Table 4. Overall, before reading the AI explanation, participants' mean detection accuracy is 0.712. After AI's explanation, their mean detection accuracy is improved to 0.928. We run a T-test to confirm the difference is statistically significant (t = -15.01, p <0.001). The accuracy improvement applied to both phishing and legitimate SMS. The accuracy of phishing SMS improves from 0.830 to 0.971 (t = -7.69, p <0.001),

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

Table 4: Participants' phishing detection accuracy before and after using the AI Agent.

Variable	Estimate (β)	*P*-value
Intercept	-0.951	0.040*
Age (Ref=18-64)		
65+	0.155	0.692
Gender (Ref="Male")		
Female	0.406	0.363
Tech. Experience (Ref=No)		
Yes	1.012	0.043*
Trust in AI (Ref="Neutral" or lower)		
"Trusting" or higher	0.111	0.775

Table 5: Factors that influence the accuracy improvements after using the AI agent. Significance is denoted by *** ($p<0.001$), ** ($p<0.01$), and * ($p<0.05$).

and the accuracy of legitimate SMS improves from 0.594 to 0.885 (t = -15.49, p <0.001). The relatively low accuracy on *legitimate SMS* before AI explanation (0.594) also indicates the participants' cautious and skeptical tendency when evaluating messages. Note that the accuracy improvement of older adults is also statistically significant, from 0.710 to 0.942 (t = -11.92, p <0.001). The results confirm the positive impact of AI explanations on users' phishing detection efficacy.

Another observation is that participants are better at identifying phishing SMS messages than legitimate ones, which is consistent with the observation in a previous study [80].

Influencing Factors. Moreover, we investigate factors that potentially influence the effectiveness of AI explanations. As shown in Table 4, participants' average accuracy improvement is around 0.2 (from 0.712 to 0.928). Here, we divide participants into two groups: those with an accuracy improvement above 0.2 are placed in the *high-improvement* group, and the rest are placed in the *low-improvement* group. We seek to understand how different factors, including participants' age, gender, technical experience, and trust in AI (before the survey), correlate with their accuracy improvement. We select age, gender, and technical experience because these are potentially influential factors according to prior works [77, 80]. We also include "trust in AI" since the AI explanation is the focus of our study. We run a logistic regression model, a common method to perform statistical significance tests.

The result is shown in Table 5. The Estimates (β) are the regression coefficients with a positive estimate indicating a positive correlation. We do not find statistically significant evidence that age, gender, or trust in AI has influenced the accuracy improvement of participants. However, technical experience is a significant factor (β=1.012, p=0.043). This means participants with technical backgrounds have benefited more from the AI explanation. A closer examination shows that participants with technical backgrounds had more errors on legitimate messages before reading AI explanations (e.g., probably because they were more cautious). They had a bigger improvement in the overall accuracy after the AI explanation.

Sanity Check of the Order Effect. Our study has a "within-subject" design where participants are exposed to both conditions (with and without AI explanation) for a series of messages. A potential concern is whether the AI explanations of earlier messages would help participants perform better on later messages under the "without AI" condition. Therefore, we run a quick analysis: for the "without AI" condition, we divide messages of each user into "first-half" (5 messages) and "second-half" (5 messages) groups. Then we aggregate and compute the overall accuracy for these two groups. For the "without AI" condition, the average accuracy is 0.71 for the "first-half", and the average accuracy is 0.72 for the "second-half." A T-test returns p=0.81 (i.e., not a statistically significant difference), which alleviates this concern.

6.2 RQ2: Usability

Next, we assess the perceived usability of the system (RQ2) based on participants' SUS ratings and open-ended answers.

6.2.1 System Usability Scale (SUS) Analsyis

SUS Score. The SUS score has a range from 0 to 100. Typically, a score above 68 is considered "above average to good" and a score over 80 indicates "excellent" usability [8]. Our system receives a mean SUS score of 82.6 and a median score of 82.5, both are at the "excellent" usability level. Note that older adults give a mean SUS score of 85.3 which is at the same level. Figure 4 shows the distribution of participants' SUS scores. We further analyze the scores of the 10 questions in the SUS survey. We find that the question that receives the lowest score (7.557/10) is Question 1: "*I will use this system frequently.*" This is understandable since users may primarily use the system when they encounter suspicious SMS messages, which may not occur frequently for most users.

Influencing Factors. We further examine which factors may be correlated with perceived usability. We use a similar methodology used in Section 6.1 to perform this analysis. We first divide the participants into two groups: the "high-SUS" group contains participants whose SUS score is above 80, and the "low-SUS" group contains the rest of the participants.

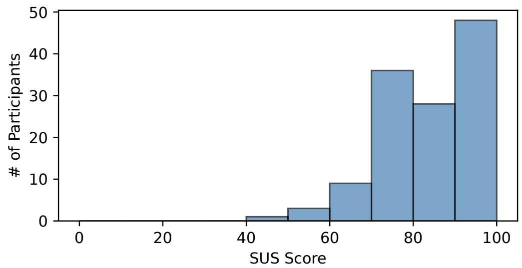


Figure 4: SUS score distribution.

Variable	Estimate (β)	*P*-value
Intercept	-1.395	0.004**
Age (Ref=18-64)		
65+	0.604	0.143
Gender (Ref=Male)		
Female	0.887	0.057
Tech. Experience (Ref=No)		
Yes	-0.065	0.901
Trust in AI (Ref="Neutral" or lower)		
"Trusting" or higher	1.448	<0.001***

Table 6: Factors that influence the usability rating (SUS score) from the participants. Significance is denoted by *** (p<0.001), ** (p<0.01), and * (p<0.05).

We use 80 as the cutoff threshold because it is a common threshold for distinguishing between "good" and "excellent" usability. We consider the same set of variables as before including age, gender, technical experience, and trust in AI.

The regression analysis result is presented in Table 6. We do not observe statistically significant differences in the SUS scores between older and younger adults, between male and female participants, or between people with and without technical experience. We consider this as a positive result because it suggests that the system is generally applicable to different demographic groups in our user study. However, the "trust in AI" (before the survey) is a statistically significant predictor (β=1.448,p <0.001). This means people who are more trusting towards AI give the system a higher usability score.

6.2.2 User Feedback Analysis

Next, we analyze their open-ended answers regarding what they like and dislike about the system.

Coding Method. We analyze the data using thematic analysis [10]. The first researcher codes all the data and develops a codebook. Then the second researcher uses this codebook and independently re-codes all the data again. The inter-rater reliability (IRR) measured by Cohen's Kappa is 0.896. The codebook contains 7 codes and 19 subcodes. The detailed codebook is included in the supplementary materials [5], and a brief version is presented in Table 13 in the Appendix.

What Participants Like About the AI Agent. The *explanation* provided by the AI was the most commonly mentioned reason (n=60 mentions). One participant noted: "*the AI agent explains or give(s) enough reason as to why you may be at threat. It gives me good reason to trust it by explaining into detail for me.*" The second most commonly mentioned reason was the ease of use of the system (n=51). Some participants (n=9) appreciated the link verification capabilities of the system which analyzed the URLs without requiring users to click on the potentially harmful links themselves. One participant mentioned: "*I love that the AI agent was able to view the links and provide information on where the page leads without the user clicking onto the link.*" Some participants (n=7) also found the system educational, suggesting that it could be used to teach others: "*I found myself making more confident choices as to the legitimacy of the SMS simply by reading the explanations as a tutorial.*" A few participants (n=3) explicitly mentioned that the system can be helpful to older adults.

Aspects to Improve. Participants also provided constructive feedback and explained their concerns. Some participants ($n = 14$ mentions) still felt that the explanation was too long, despite that we already presented the "short version". Making the explanation even shorter without losing important details is a potential future direction for improvement. Some participants also expressed their concerns regarding the systems' accuracy (n=11). Moreover, some participants mentioned the *speed* of explanation generation. Recall that we used a typewriter animation to type out the text in the explanation message (20 ms per character). This was a programmed delay to ensure participants paid attention. Participants usually waited for 10-15 seconds to read the explanation. While some participants (n =12) mentioned their appreciation of the "quick" response of AI, a few participants ($n = 3$) still felt it was slow. This suggests that minimizing the delay caused by the AI analysis will be important for real-world deployment. Finally, a few participants ($n = 6$) suggested using color-coded messages to indicate the phishing detection results before showing the explanation: "*...perhaps highlighting potentially dangerous messages (in) red with an exclamation, or yellow for unsure messages would accomplish the goal of safety better. Then, for more info, the AI explanation button could still be available.*"

Detailed AI Report. In the study, we only showed the short explanation to users while having a button at the end for the *detailed AI report* (see Figure 3 (b)). During the study, we logged user clicks on this button. As shown in Table 8 (Appendix), 97 out of 125 participants (77.6%) had never clicked the button to view the detailed report. Participants who did read the report praised the extra details: "*I liked that it could provide evidence if I wasn't confident about the*

Score	Before Survey	After Survey
Trust in AI (Agent)	3.496	4.264
Self Confidence	3.376	3.640

Table 7: Participants' trust in AI and self-assessed phishing detection ability. Both scores range from 1 to 5.

results that it was yielding. The screenshot(s) of the pages were helpful for me to see what the website looked like and made it easier to trust AI." To understand users' preferences for the detailed AI report, we ask this question explicitly in the exit questions. We find that the majority of the participants still prefer to keep the detailed report available (85 out of 125, 68%) despite they do not (often) read them.

6.3 RQ3: Trust and Disagreement

We now analyze our data to understand the trust and potential disagreement between human users and the AI agent (RQ3).

Impact on Trust and Confidence. As mentioned in Section 5.1, participants were asked to report their trust in general AI *before* the survey, and then their trust in our AI agent *after* they interacted with the prototype. Both ratings are using a 5-Likert scale from "Very distrusting" to "Very trusting." As shown in Table 7, the mean trust score before the survey is 3.496 (out of 5) and the score is 4.264 after the survey. A T-test shows the difference is statistically significant (t=-9.364, p <0.001). This suggests that after interacting with the AI agent, their trust level becomes higher.

We run a similar analysis to compare their self-assessed phishing detection ability (i.e., self-confidence) before and after the survey. Table 7 shows the rating has also significantly increased, from 3.376 to 3.640 (t=-3.813, p <0.001).

Disagreement with the AI Agent. During the study, we observed that participants occasionally disagreed with the AI agent's decision even *after* reading the AI explanations. Among 125 participants, 44 (35.2%) disagreed with the AI at least once. Out of a total of 1,250 message assessments, we observed 90 (7.2%) cases of disagreements.

At the end of the survey, we asked a follow-up question where participants explained their reasons for the disagreement. We analyze their open-ended answers using the same method as in Section 6.2.2. The inter-rater reliability (IRR) measured by Cohen's Kappa is 0.915 (for legitimate SMS) and 1.00 (for phishing SMS). The detailed codebook is in [85]. For all the disagreement cases, AI's decisions are always aligned with the ground truth (i.e., AI is correct).

Disagreement on Legitimate SMS. We find it more common for participants to classify a message as phishing even though AI determines it is legitimate (72 out of 625, 11.5%). The most commonly mentioned reason (n=14) was the URLs in the message, especially shortened URLs. Even though the shortened URLs were used by *legitimate* services (and verified by the AI agent), some participants still expressed their distrust. This suggests that legitimate services/organizations should avoid using shortened URLs in their messages to users. Some participants (n=15) mentioned their security habits, stating that they took an (overly) cautious approach to assessing suspicious SMS messages. Some participants ($n = 11$) mentioned that the context of the SMS was misaligned or they were unfamiliar with the service. Overall, we find that users tend to be cautious and such disagreements would not put users at risk.

Disagreement on Phishing SMS. It is less common for participants to insist that a message is legitimate after the AI has classified it as phishing (18 out of 625, 2.9%). A common reason ($n = 3$) was that participants believed the SMS context (e.g., package delivery) was aligned with reality and determined it should be legitimate. Another participant mentioned that the message (that offered discounts on energy bills) did not ask to provide any personal information and thus determined it as low-risk. Finally, one participant expressed distrust towards the AI's analysis of the redirection chain between Instagram and Facebook: "*I think the AI incorrectly labeled the URL redirection as fraudulent.*" The result suggests that while the AI agent can reduce the risk, it may not be able to convince all users under certain situations.

6.4 RQ4: AI Errors

Finally, to understand how users respond to AI errors (RQ4), we analyze the data from Study B. As described in Section 5.2, Study B includes the same 10 messages used in Study A and one additional "AI error" message (false negative). We first check participants' accuracy on the 10 SMS messages (where AI was correct). This returns an accuracy of 0.706 before reading the AI explanation, and an accuracy of 0.940 after the explanation. The result is consistent with that in Study A, confirming its reliability.

On the "AI-error" message, we observed that 24 participants (48%) initially determined the message was phishing. After reviewing the AI explanation, 18 of them changed their decisions to mark the message as legitimate. Only 6 participants maintained their original decision (i.e., the correct determination). We further coded the open-ended responses from these participants regarding why they disagreed with AI on this message. Two participants explained that they remained skeptical due to abnormal word choices in the message. Two other participants distrusted the shortened URL in the message. One participant did not trust the message as it was about an unfamiliar service. The last participant did not provide a specific reason. We discuss the implications next in Section 7.

7 Discussion and Conclusion

AI Explanation. We develop `SmishX` for phishing SMS detection, and more importantly, explanation. Through data-driven evaluation and user studies, we show that the AI agent is highly accurate, and the generated explanations are effective and usable in helping users improve their phishing detection efficacy. Importantly, we confirm the system works well with *older adults* (one of our target demographic groups). Finally, we find interacting with the AI agent helps to improve participants' trust in AI as well as their self-confidence in phishing detection ability.

During our user studies, we have observed that users do not always agree with the AI's determination (on 7.2% of the messages). This happens more often on legitimate messages, especially when the legitimate messages contain shortened URLs. Note that our participants have expressed suspicion about both generic URL shorteners (bit.ly) and brand-specific ones (e.g., w-mt.co for Walmart). The implication is legitimate businesses/services should *avoid* using shortened URLs in their messages to improve user trust.

AI Accuracy. Our evaluation (Section 4.2) shows `SmishX` achieves a high detection accuracy of 98.8%. One may ask, is "users-in-the-loop" still needed if the AI can block all phishing messages accurately? We believe that users need a sense of control over their communication channel to build trust in the system. More importantly, a perfect 100% accuracy is not a practical expectation (we discuss potential overfitting issues in Apepndix B). A future version of the system can benefit from generating "confidence scores." The idea is that high-confidence phishing messages can be blocked/quarantined directly while low-confidence ones may need users to read the message and AI explanations.

AI Latency. The latency of the AI agent is considered as an important factor by our participants. As a prototype, the system currently takes about 60 seconds to analyze and explain an SMS. For latency improvements, we are actively discussing ideas with our industry collaborators. Possible directions include locally hosting open-source LLMs (instead of using APIs), incorporating allow-lists for message filtering, pre-fetching and caching information for domain-brand matching and domain history checks, and using vLLM acceleration.

AI Errors. The result in Section 6.4 indicates that the AI explanation has a strong influence on users' decisions. When the AI makes a mistake, it's difficult for users to correct it. This echoes a prior work [26] that has highlighted the challenge of "algorithm-in-the-loop" decision making when users are unable to evaluate the accuracy of their own or model's predictions. The implications are two-fold. First, it is important to maintain the high accuracy of the detection model. Our current model has a high accuracy, especially in detecting phishing SMS (Section 4.2), which helps to alleviate this issue. Second, we need to further study how to avoid users' over-reliance on AIs for decision-making and help users effectively resolve disagreements with the AI. For example, one possible direction is to phrase the explanation messages for phishing and legitimate SMS differently. For phishing SMS, the AI can directly state its decision and evidence (e.g., "*the SMS is likely a phishing message because...*"). For legitimate SMS, AI can make more conservative decisions, inform users that AI can make mistakes, and suggest safe options (e.g., "*Our AI agent did not find sufficient evidence to determine whether this is a phishing message or not. Please avoid directly clicking on the link. When possible, you should access your account through the official website/app.*"

Privacy Considerations for Deployments. We recommend deployment of `SmishX` through SMS service providers (or messaging app providers) that already have permission to access users' SMS. They can use *local* open-source LLMs to further minimize privacy risks (performance validated in Appendix A). If they prefer commercial/closed-source models, e.g., ChatGPT-4o from OpenAI, they may use their enterprise/business plan to ensure their private data is not stored/used by OpenAI.

Limitations and Future Directions. This paper presents a preliminary effort in this direction, and we want to acknowledge our limitations and point out future directions. First, we did not experiment with adversarial attacks against our system. Appendix C presents a manually crafted false negative example, which suggests the possibility of such attacks. Second, our user study has several limitations too. For example, we recruited U.S. participants from Prolific which may have introduced biases (e.g., participants from Prolific may be more familiar/comfortable with computers and technology than the general population). In addition, the participants knew that the survey was about phishing, which may have made them more careful/cautious during the study (than they would be in real life). Finally, during the study, we used a web interface to present the UI of the messaging app, which can be different from the real-life experience of users. Also, we only tested one particular interface design for AI explanations. Future work can explore other variants of the explanation structure and the UI design (e.g., the use of colors and buttons) to further improve user experience.

Acknowledgments

We would like to thank Margie Lachman and anonymous reviewers for their valuable feedback on this work. This work was in part supported by a MassAITC pilot award, and NSF grants 2229876 and 2055233. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or sponsors.

References

[1] Max Adler. Sms marketing statistics 2025 for usa businesses. https://www.smscomparison.com/sms-statistics/, 2025.

[2] Google AI. Gemini. https://gemini.google.com/, 2024.

[3] Jina AI. Reader. https://jina.ai/reader/, 2025.

[4] Tiago A. Almeida, José María G. Hidalgo, and Akebo Yamakami. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering (DocEng)*, 2011.

[5] Anonymous Authors. Supplementary materials for paper submission. https://github.com/SmishX/SmishX-Material, 2025.

[6] Shahryar Baki and Rakesh M. Verma. Sixteen years of phishing user studies: What have we learned? *IEEE Transactions on Dependable and Secure Computing*, 20(2):1200–1212, 2023.

[7] Caner Balim and Efnan Sora Gunal. Automatic detection of smishing attacks by machine learning methods. In *Proceedings of the 1st International Informatics and Software Engineering Conference (UBMYK)*, 2019.

[8] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[9] Marzieh Bitaab, Haehyun Cho, Adam Oest, Zhuoer Lyu, Wei Wang, Jorij Abraham, Ruoyu Wang, Tiffany Bao, Yan Shoshitaishvili, and Adam Doupé. Beyond Phish: Toward Detecting Fraudulent e-Commerce Websites at Scale. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2023.

[10] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[11] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[12] Tri Cao, Chengyu Huang, Yuexin Li, Huilin Wang, Amy He, Nay Oo, and Bryan Hooi. Phishagent: A robust multimodal agent for phishing webpage detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[13] FBI Internet Crime Complaint Center. Federal bureau of investigation elder fraud report 2023. https://www.ic3.gov/AnnualReport/Reports/2023_IC3ElderFraudReport.pdf, 2023.

[14] FBI Internet Crime Complaint Center. Federal bureau of investigation internet crime report. https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf, 2023.

[15] Robin Chataut, Prashnna Kumar Gyawali, and Yusuf Usman. Can ai keep you safe? a study of large language models for phishing detection. In *Proceedings of the 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, 2024.

[16] Xiaowei Chen, Sophie Doublet, Anastasia Sergeeva, Gabriele Lenzini, Vincent Koenig, and Verena Distler. What motivates and discourages employees in phishing interventions: An exploration of Expectancy-Value theory. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS)*, 2024.

[17] Crawlee. Puppeteer crawler. https://crawlee.dev/api/puppeteer-crawler, 2025.

[18] Leslie Daigle. WHOIS Protocol Specification. RFC 3912, September 2004.

[19] DeepSeek. Deepseek-v3. https://api-docs.deepseek.com/news/news1226, 2025.

[20] Giuseppe Desolda, Francesco Greco, and Luca Viganò. Apollo: A gpt-based tool to detect phishing emails and generate explanations that warn users. *arXiv preprint arXiv:2410.07997*, 2024.

[21] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2006.

[22] Alejandra Diaz, Alan T Sherman, and Anupam Joshi. Phishing in an academic community: A study of user susceptibility and behavior. *Cryptologia*, 44(1):53–67, 2020.

[23] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2008.

[24] Federal Communications Commission (FCC). In the matter of targeting and eliminating unlawful text messages. https://docs.fcc.gov/public/attachments/FCC-23-21A1.pdf, 2023.

[25] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the World Wide Web Conference (WWW)*, 2007.

[26] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 2019.

[27] Brij B Gupta and Ankit Kumar Jain. Phishing attack detection using a search engine and heuristics-based technique. *Journal of Information Technology Research (JITR)*, 13(2):94–109, 2020.

[28] Brynne Harrison, Elena Svetieva, and Arun Vishwanath. Individual processing of phishing emails: How attention and elaboration protect against phishing. *Online Information Review*, 40(2):265–281, 2016.

[29] Julian Hazell. Spear phishing with large language models. *arXiv preprint arXiv:2305.06972*, 2023.

[30] Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S. Park. Devising and detecting phishing emails using large language models. *IEEE Access*, 12:42131–42146, 2024.

[31] Grant Ho, Aashish Sharma, Mobin Javed, Vern Paxson, and David Wagner. Detecting credential spearphishing in enterprise settings. In *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2017.

[32] Hang Hu and Gang Wang. End-to-end measurements of email spoofing attacks. In *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2018.

[33] Ankit Kumar Jain and BB Gupta. Rule-based framework for detection of smishing messages in mobile environment. *Procedia Computer Science*, 125:617–623, 2018.

[34] Ankit Kumar Jain, Sumit Kumar Yadav, and Neelam Choudhary. A novel approach to detect spam and smishing sms using machine learning techniques. *International Journal of E-Services and Mobile Applications (IJESMA)*, 12(1):21–38, 2020.

[35] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), 2023.

[36] Jae Woong Joo, Seo Yeon Moon, Saurabh Singh, and Jong Hyuk Park. S-detector: an enhanced security model for detecting smishing attack for mobile computing. *Telecommunication Systems*, 66:29–38, 2017.

[37] Kaggle. Sms-data. https://www.kaggle.com/datasets/dshah1612/sms-data/data, 2022.

[38] Ilker Kara, Murathan Ok, and Ahmet Ozaday. Characteristics of understanding urls and domain names features: the detection of phishing websites with machine learning methods. *IEEE Access*, 10:124420–124428, 2022.

[39] Hadi El Karhani, Riad Al Jamal, Yorgo Bou Samra, Imad H. Elhajj, and Ayman Kayssi. Phishing and smishing detection using machine learning. In *Proceedings of the 2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2023.

[40] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Chatspamdetector: Leveraging large language models for effective phishing email detection. In *Proceedings of the EAI International Conference on Security and Privacy in Communication Networks (SecureComm)*, 2024.

[41] Takashi Koide, Hiroki Nakano, and Daiki Chiba. Chatphishdetector: Detecting phishing sites using large language models. *IEEE Access*, 12:154381–154400, 2024.

[42] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2024.

[43] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting people from phishing: The design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2007.

[44] Sophie Le Page, Guy-Vincent Jourdan, Gregor V Bochmann, Jason Flood, and Iosif-Viorel Onut. Using url shorteners to compare phishing and malware attacks. In *Proceedings of the 2018 APWG Symposium on Electronic Crime Research (eCrime)*, 2018.

[45] Jehyun Lee, Peiyuan Lim, Bryan Hooi, and Dinil Mon Divakaran. Multimodal large language models for phishing webpage detection and identification. In *Proceedings of the Symposium on Electronic Crime Research (eCrime)*, 2024.

[46] Jehyun Lee, Farren Tang, Pingxiao Ye, Fahim Abbasi, Phil Hay, and Dinil Mon Divakaran. D-fence: A flexible, efficient, and comprehensive phishing email detection system. In *Proceedings of the 2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2021.

[47] Yunseung Lee and Daehee Han. KorSmishing explainer: A Korean-centric LLM-based framework for smishing detection and explanation generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track (EMNLP)*, 2024.

[48] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2011.

[49] Zilong Lin, Jian Cui, Xiaojing Liao, and XiaoFeng Wang. Malla: Demystifying real-world large language model integrated malicious services. In *Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24)*, 2024.

[50] Ruofan Liu, Yun Lin, Xiwen Teoh, Gongshen Liu, Zhiyong Huang, and Jin Song Dong. Less defined knowledge and more true alarms: reference-based phishing detection without a pre-defined reference list. In *Proceedings of the 33rd USENIX Conference on Security Symposium (USENIX Security 24)*, 2024.

[51] D. Kevin McGrath and Minaxi Gupta. Behind phishing: An examination of phisher modi operandi. In *Proceedings of USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2008.

[52] Meta. Meta ai. https://www.meta.ai/, 2024.

[53] Microsoft. Bing chat copilot. https://copilot.microsoft.com, 2024.

[54] Sandhya Mishra and Devpriya Soni. Sms phishing dataset for machine learning and pattern recognition. In *Proceedings of the International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, 2022.

[55] Hossen Mustafa, Wenyuan Xu, Ahmad-Reza Sadeghi, and Steffen Schulz. End-to-end detection of caller id spoofing attacks. *IEEE Transactions on Dependable and Secure Computing*, 15(3):423–436, 2018.

[56] Aleksandr Nahapetyan, Sathvik Prasad, Kevin Childs, Adam Oest, Yeganeh Ladwig, Alexandros Kapravelos, and Bradley Reaves. On sms phishing tactics and infrastructure. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2024.

[57] Quan Hong Nguyen, Tingmin Wu, Van Nguyen, Xingliang Yuan, Jason Xue, and Carsten Rudolph. Utilizing large language models with human feedback integration for generating dedicated warning for phishing emails. In *Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems (SecTL)*, 2024.

[58] Mathew Nicho, Hussein Fakhry, and Uche Egbue. When spear phishers craft contextually convincing emails. In *Proceedings of the International Conferences on WWW/Internet and Applied Computing (ICWI)*, 2018.

[59] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam

Soliman, Tian Lin, and Natalie Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI)*, 2017.

[60] OpenAI. Chatgpt-4. https://chatgpt.com/?model=gpt-4, 2024.

[61] OpenAI. Chatgpt-4o. https://chatgpt.com/?model=gpt-4o, 2024.

[62] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

[63] Kathryn Parsons, Marcus Butavicius, Paul Delfabbro, and Meredith Lillie. Predicting susceptibility to social influence in phishing emails. *International Journal of Human-Computer Studies*, 128:17–26, 2019.

[64] Peng Peng, Chao Xu, Luke Quinn, Hang Hu, Bimal Viswanath, and Gang Wang. What happens after you leak your password: Understanding credential sharing on phishing sites. In *Proceedings of ACM ASIA Conference on Computer and Communications Security (Asia CCS)*, 2019.

[65] Jon Peterson and Sean Turner. Secure Telephone Identity Credentials: Certificates. RFC 8226, 2018.

[66] Prolific. Prolific | easily collect high-quality data from real people. https://www.prolific.com/, 2024.

[67] QWen. Qwen2.5-vl-72b-instruct. https://qwenlm.github.io/, 2025.

[68] Md Lutfor Rahman, Daniel Timko, Hamid Wali, and Ajaya Neupane. Users really do respond to smishing. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy (CODASPY)*, 2023.

[69] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS)*, 2020.

[70] Kenneth Reitz. Redirection and history. https://requests.readthedocs.io/en/latest/user/quickstart/#redirection-and-history, 2025.

[71] Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP)*, 2024.

[72] Muhammad Salman, Muhammad Ikram, Nardine Basta, and Mohamed Ali Kaafar. Spallm-guard: Pairing sms spam detection using open-source and commercial llms. *arXiv preprint arXiv:2501.04985*, 2025.

[73] Ashfak Md Shibli, Mir Mehedi A. Pritom, and Maanak Gupta. Abusegpt: Abuse of generative ai chatbots to create smishing campaigns. In *Proceedings of the 12th International Symposium on Digital Forensics and Security (ISDFS)*, 2024.

[74] Ho Sung Shim, Hyoungjun Park, Kyuhan Lee, Jang-Sun Park, and Seonhye Kang. A persuasion-based prompt learning approach to improve smishing detection through data augmentation. *arXiv preprint arXiv:2411.02403*, 2024.

[75] Gunikhan Sonowal and KS Kuppusamy. Smidca: an anti-smishing model with machine learning approach. *The Computer Journal*, 61(8):1143–1157, 2018.

[76] Fritz Strack. *"Order effects" in survey research: Activation and information functions of preceding questions*, pages 23–34. Springer New York, 1992.

[77] Sarah Tabassum, Cori Faklaris, and Heather Richter Lipford. What drives smishing susceptibility? a u.s. interview study of how and why mobile phone users judge text messages to be real or fake. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2024.

[78] Ronnie Taib, Kun Yu, Shlomo Berkovsky, Mark Wiggins, and Piers Bayl-Smith. Social engineering and organisational dependencies in phishing attacks. In *Proceedings of the IFIP Conference on Human-Computer Interaction (INTERACT)*, 2019.

[79] Christopher Thompson, Martin Shelton, Emily Stark, Maximilian Walker, Emily Schechter, and Adrienne Porter Felt. The web's identity crisis: Understanding the effectiveness of website identity indicators. In *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2019.

[80] Daniel Timko, Daniel Hernandez Castillo, and Muhammad Lutfor Rahman. A quantitative study of sms phishing detection. *arXiv preprint arXiv:2311.06911*, 2024.

[81] Daniel Timko and Muhammad Lutfor Rahman. Commercial anti-smishing tools and their comparative effectiveness against modern threats. In *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WISEC)*, 2023.

[82] Mohammad Amaz Uddin, Muhammad Nazrul Islam, Leandros Maglaras, Helge Janicke, and Iqbal H. Sarker. Explainabledetector: Exploring transformer-based language modeling approach for sms spam detection with explainability analysis. *arXiv preprint arXiv:2405.08026*, 2024.

[83] Amber van der Heijden and Luca Allodi. Cognitive triaging of phishing attacks. In *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2019.

[84] Chenkai Wang, Zhuofan Jia, Hadjer Benkraouda, Cody Zevnik, Nicholas Heuermann, Roopa Foulger, Jonathan A Handler, and Gang Wang. Verisms: A message verification system for inclusive patient outreach against phishing attacks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, 2024.

[85] Yizhu Wang, Haoyu Zhai, Chenkai Wang, Qingying Hao, Nick A. Cohen, Roopa Foulger, Jonathan A. Handler, and Gang Wang. VeriSMS website: Data, code, and supplementary materials. https://github.com/yizhu-joy/SmishX, 2025.

[86] Rick Wash, Norbert Nthala, and Emilee Rader. Knowledge and capabilities that Non-Expert users bring to phishing detection. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, 2021.

[87] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language

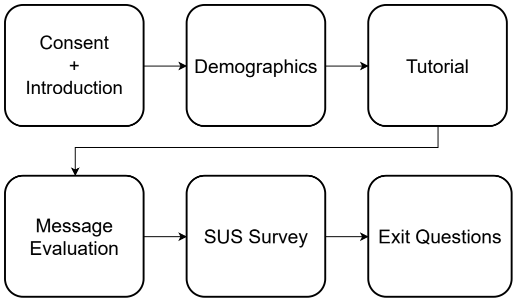


Figure 5: User study workflow.

models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[88] Min Wu, Robert C. Miller, and Simson L. Garfinkel. Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2006.

[89] Yue Zhang, Serge Egelman, Lorrie Cranor, and Jason Hong. Phinding Phish: Evaluating Anti-Phishing Tools. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2007.

[90] Yue Zhang, Jason I. Hong, and Lorrie F. Cranor. Cantina: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, 2007.

[91] Sijie Zhuo, Robert Biddle, Yun Sing Koh, Danielle Lottridge, and Giovanni Russello. Sok: Human-centered phishing susceptibility. *ACM Trans. Priv. Secur.*, 26(3), 2023.

Options	**# Participants**
Clicked to see the detailed report	
0	97
1-5 times	20
6-10 times	8
Keep the detailed report available?	
No, definitely not	2
Probably no	20
Neutral	18
Probably yes	41
Yes, definitely	44

Table 8: User preference for the "detailed AI report."

A Testing Open-source LLMs

SMS Type	GPT-4o	Qwen2.5*	DeepSeek-v3
Phishing	100.0%	94.6%	98.8%
Spam	99.1%	70.8%	93.4%
Legitimate	98.2%	98.7%	96.5%
Overall	98.8%	87.4%	96.2%

Table 9: Using different LLMs to implement `SmishX`. The result confirms that the `SmishX` can be implemented using open-source LLMs such as DeepSeek-v3, which can be hosted locally. *For Qwen2.5, we used Qwen2.5-vl-72b-instruct.

We tested two open-source models to implement `SmishX` and the results are presented in Table 9. DeepSeek-v3 shows promising results in comparison with the GPT-4o version (using commercial APIs). An open-source LLM allows an organization/service to run `SmishX` *locally* without sending any data to a commercial service (e.g., OpenAI), which can mitigate potential privacy complications.

B Risk of Model Overfitting

Our evaluation (Section 4.2) shows that `SmishX` achieves a high detection accuracy of 98.8%. Given that OpenAI models are trained on public data on the Internet, the question is whether the high accuracy is a result of overfitting due to data leakage (i.e., OpenAI trained with our data). We believe this is not a concern because we used GPT-4o-2024-08-06 for our evaluation, and this model had its training cutoff in October 2023 [62]. Meanwhile, about 48% of malicious messages in our dataset were from a recent study [56], which were kept private until May 2024. This means GPT-4o could not have been trained on this data. We did not see accuracy differences between the newer and the older data during evaluation. That being said, our study can benefit from evaluating with an even larger and more diverse dataset to mitigate potential overfitting concerns.

C SMS Message used in Study B

We manually crafted a phishing SMS message to trigger a false negative error in the AI agent for user study B.

> Hi, Thank you for signing up for our open, decentralized bridge! 876544 is your verification code. Hope you enjoy using our product to unlock the power of interoperability & liquidity between blockchain ecosystems. Click to see your demos: `https://[ShortenedURL]`

This message can bypass AI detection for three reasons. First, it does not contain an *explicit brand name*. Thus the agent cannot use the brand name to search for and match the corresponding domain name to determine the trustworthiness of the URL. Second, we insert a sentence at the beginning of the message to indicate a prior interaction with the sender service: "Thank you for signing up for our ..." This sentence helps to avoid being flagged as spam. Finally, we set the shortened URL to point to a phishing website that looks similar to the official legitimate website (`gravitybridge.net`). The website is also hosted under a *similar* hostname as the legitimate one.

D Chain of Thought Prompt

Prompt for the Information Extraction LLM:

> Extract any URLs and brand names from the following SMS message.
> Your output should be in JSON format and should not have any other output:
> - is_URL: true or false
> - URLs: If no URL in SMS, answer non. If there are URLs, the response should be a list. Each element is a URL extracted from the SMS.
> - is_brand: true or false
> - brands: If no brand name in SMS, answer non. If there are brand names, the response should be a list. Each element is a brand name extracted from the SMS. You can extract the brand name from the SMS content and the URL.

Prompt for the Phishing Detection LLM:

> 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. Analyze the SMS by following these steps:
> 1. If the SMS is promoting any of the following categories: Online gambling, bets, spins, adult content, digital currency, lottery, it is either spam or phishing.
> 2. The SMS is legitimate if it is from known organizations, such as appointment reminders, OTP (One-Time Password) verification, delivery notifications, account updates, tracking information, or other expected messages.
> 3. The SMS is considered legitimate if it involves a conversation between friends, family members, or colleagues.
> 4. Promotions and advertisements are considered spam. The SMS is spam if it is promotion from legitimate companies and is not impersonating any brand, but it is advertisements, app download promotions, sales promotions, donation requests, event promotions, online loan services, or other irrelevant information.
> 5. The SMS is phishing if it is fraudulent and attempts to deceive recipients into providing sensitive information or clicking malicious links. Phishing SMS may exhibit the following characteristics:
> Promotions or Rewards: Some phishing SMS offer fake prizes, rewards, or other incentives to lure recipients into clicking links or providing personal information.
> Urgent or Alarming Language: Phishing messages often create a sense of urgency or fear, such as threats of account suspension, missed payments, or urgent security alerts.
> Suspicious Links: Phishing messages may contain links to fake websites designed to steal personal information.
> Requests for Personal Information: Phishing SMS may ask for sensitive information like passwords, credit card numbers, social security numbers, or other personal details.
> Grammatical and Spelling Errors: Many phishing messages contain grammatical mistakes or unusual wording, which can be a red flag for recipients.
> Expired Domain: Phishing websites often use domains that expire quickly or are already listed for sale.
> Inconsistency: The URL may be irrelevant to the message content.
> 6. Please be aware that: It is common to see shortened URLs in SMS. You can get the expanded URL from the provided redirection chain. Both phishing and legitimate URLs can be shortened. And both phishing and legitimate websites may use a robot-human verification page (CAPTCHA-like mechanism) before granting access the content.
> 7. I will provide you with some external information if there is a URL in the SMS. The information includes:
> - Redirect Chain: The URL may redirect through multiple intermediate links before reaching the final destination; if any of them is flagged as phishing, the original URL becomes suspicious.
> - Brand Search Information: The top five results from a Google search of the brand name. You can compare if the URL's domain matches the results from Google.
> - Screenshot Description: A description of the website's screenshot, highlighting any notable visual elements.
> - HTML Content Summary: The title of HTML, and the summary of its content.
> - Domain Information: The domain registration details, including registrar, creation date, and DNS records, which are analyzed to verify the domain's legitimacy.

8. Please give your rationales before making a decision. And your output should be in json format and should not have any other output:
- brand_impersonated: brand name associated with the SMS, if applicable.
- URL: any URL appears in SMS, if no URL, answer "non".
- rationales: detailed rationales for the determination, up to 500 words. Directly give sentences, do not categorize the rationales. Only tell the reasons why the SMS is legitimate or not, do not include the reasons why the SMS is spam or phishing.
- brief_reason: brief reason for the determination.
- category: True or False. If the SMS is legitimate, output False. Else, output True.
- advice: If the SMS is phishing, output potential risk and your advice for the recipients, such as "Do not respond to this message or access the link."

Below is the information of the SMS:
- SMS text: ...
- Redirect Chain: ...
- Brand Search Information: ...
- Screenshot Description: ....
- Html Content Summary: ...
- Domain Information: ...

Prompt for the Explanation LLM:

You are an SMS phishing analysis advisor.
Below is an SMS message along with an explanation of whether it is legitimate, spam, or phishing. Assess the message and provide a clear and easy-to-understand response to the user, specifying whether the message is legitimate. Provide appropriate advice. For example, if it is phishing, remind the user not to click links or reply. If it is spam, highlight risks like online gambling, betting, adult content, digital currency, or lottery schemes. Use simple words and language (e.g., say 'website link' instead of 'URL'). Include clear reasons for your conclusion and keep the response concise for non-technical users. Limit your answer to four sentences.

E Example Messages and Explanations

Table 11 shows extra examples of SMS and their short explanations. Table 12 shows an example of a detailed AI report.

# of Participants	Study A	Study B
Total	125	50
Age		
18 - 24	19	6
25 - 34	25	7
35 - 44	16	8
45 - 54	10	8
55 - 64	5	1
65+	50	20
Sex		
Female	77	31
Male	47	19
Non-binary	1	0
Ethnicity/Race		
White	96	35
Black	21	11
Asian	3	2
Hispanic or Latino	3	2
Mixed or Multiracial	2	0
Education Level		
Some High School	1	2
High School Diploma	33	4
Associate Degree	17	4
Bachelor's Degree	48	23
Master's degree	18	10
Doctoral degree	6	6
Other	2	1
Technical Experience in Computer Science		
Yes	29	22
No	94	27
Prefer not to say	2	1
SMS Usage		
Daily	106	39
Weekly	16	8
Rarely	1	2
Never	2	1
Phishishing SMS Encounter		
Yes	109	44
No	5	3
Unsure	11	3

Table 10: Demographics of participants.

Type	SMS	Short Explanation
Legitimate	Chase: You successfully validated your device and signed in. Call us, visit chase.com/xxx or use the app to manage your devices.	This SMS message appears to be legitimate. It mentions Chase, a well-known bank, and provides a direct link to Chase's official website, which is confirmed to be safe and matches their official domain. The message content is consistent with typical account security notifications and does not contain any signs of phishing, like urgent language or requests for personal information. You can safely follow the suggestions to visit Chase's website or use the app if you need to manage your devices.
Legitimate	Great news! You have 2 Walmart packages out for delivery, and they should arrive today. Track: https://w-mt.co/g/cZxxxx . Reply HELP for info; STOP to opt out.	The message appears legitimate and aligns with typical delivery notifications from Walmart. The link included in the SMS is associated with Walmart's official website, indicating that it likely directs you to a genuine page. However, if you weren't expecting a delivery, it's safest to check your orders by going directly to the Walmart website rather than clicking on the link in the message. This approach helps protect your information and ensures you're interacting only with verified sources.
Phishing	GOV: You are eligible for a discounted energy bill under the Energy Bills Support Scheme. You can apply here at: https://rebate.energy-uk-xxxxx.com	This message is likely a phishing attempt. The SMS impersonates the "Energy Bills Support Scheme" to sound trustworthy, but the link provided doesn't lead to an official or legitimate website related to the scheme. Instead, the website is suspicious, with signs that it was created recently and primarily exists for resale—typical traits of phishing sites. Do not click on the link or respond to the message; delete it immediately and consider reporting it to your mobile provider or relevant authorities.
Phishing	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.
Spam	A warm welcome for you. Get your Bonus & Free Spins. https://tx.vc/r/3xxxx/ T&Cs Apply, 18 Plus. Play Responsibly. Unsubscribe At https://tx.vc/r/3xxxx/	This SMS message is spam. It's promoting online gambling by offering a bonus and free spins, and it includes website links that could lead to gambling sites. Such messages are generally sent to many people and aren't personal and it's best not to click on any links or respond to this message. Stay cautious with unsolicited offers, especially those related to online gambling.

Table 11: Examples SMS and the corresponding short explanations generated by `SmishX`.

Detailed AI Report	The SMS claims to be associated with the "Energy Bills Support Scheme." However, the provided URL does not lead to any official or legitimate domain related to this scheme. The redirect chain ends at a domain that appears to be available for sale, indicating that it is not associated with any legitimate service. The webpage contains minimal content and lacks any detailed information related to energy claims or rebates. This aligns with typical characteristics of phishing attempts, where fraudulent domains are used. Furthermore, the domain was only recently created on January 7, 2024, and it seems to be primarily intended for resale or domain parking, which is a common indicator of a non-legitimate site. Additionally, the domain's registrar, TravelDomains, Incorporated, is not known for energy-related services. The absence of any identifiable connections to the actual Energy Bills Support Scheme reinforces that the SMS is likely a phishing attempt. It attempts to lure users by using a legitimate-sounding scheme name to deceive recipients into thinking the offer is real. The presence of a suspicious link aimed at misleading users aligns with phishing characteristics.

Table 12: An example of a detailed AI report, for the third SMS in Table 11.

Code	SubCode	Freq.	Examples
Explanation Quality and Trust Building	Explanation Builds Understanding and Boosts Confidence	60	"...the AI agent explains or give enough reason as to why you may be at threat. It gives me good reason to trust it by explaining into details for me..."
	Concerns about Explanation Length	14	"... I would improve upon the system by supplying a tad less information, unless requested."
	Additional Detailed Explanation	9	"...I especially liked being able to see the additional information - seeing the screenshots and knowing that the AI agent looked into several different aspects helped me feel better about the validity of the decisions."
	Educational Value and Learning Support	7	"... The explanations as to why a SMS was legitimate or phishing was very clear and educational as well. I found myself making more confident choices as to the legitimacy of the SMS simply by reading the explanations as a tutorial..."
	Issues with Clarity and Context	4	"The formatting of the message could be a bit long and tedious to read. Some of the terms used could be confusing as well."
Ease of Use and Intuitive Design		51	"It was very simple to use as it just required clicking the AI button..."
Accuracy	Accuracy and Alignment with User Judgments	12	"I liked its accuracy a lot it matches a lot with my opinion about the sms samples"
	Concerns about Accuracy and Trust in AI	11	"only concern is how accurate it really is. i am a little weary to trust technology but i liked having the option of whether i wanted to agree with the AI or not."
Speed and Efficiency	Fast and Efficient	12	" I really like the quick analysis of whether a message was legitimate or fraudulent..."
	Slow	3	"Only complaint is the relatively slow typeback from the AI. Over numerous texts I would find that tedious."
Link Verification Ability		9	"...I love that the AI agent was able to view the links and provide information on the where the page leads without the user clicking onto the link."
Help Older Adults		3	"It would be very helpful to people that are easily duped, like old ladies, etc..."
Suggestions for Improvement	Improved Visual Indicators	6	"...Perhaps highlighting potentially dangerous messages red with an exclamation, or yellow for unsure messages would accomplish the goal of safety better. Then, for more info, the AI explanation button could still be available."
	Suggestion for the Survey itself	5	"...I do need to recommend that you tell us not to put ourselves into the message..."
	Clarity in Responses	4	"Clarity in Responses: AI responses can sometimes be overly complex or lack human warmth. Improving the conversational tone to feel more natural, empathetic, and engaging could enhance user satisfaction. Suggestion: Incorporate language that feels more intuitive and tailored to the user's style."
	More Automatic	4	"Requiring users to click an additional button to access detailed evidence may slow the process and discourage thorough review."
	Source Transparency	3	"...An improvement to help with the credibility of this resource could be to include sites and sources that back up the facts and evidence used in the responses."
	Warnings for All Shortened URLs	2	"The system should warn for any and all URL shorteners. No organization with minimally competent leadership should be sending out any bit.ly urls; a company should have their own dedicated URL shortener that clients/customers can immediately recognize as a legitimate URL..."
	Hide Links At First	1	"If it works and is more integrated so it automatically checks messages and hides links unless you click "show link" then it could help prevent people from getting scammed..."
	Access to Report Phishing	1	"...possibly a way to report malacious links."
	Privacy Concern	1	"I still feel uncomfortable at the thought of AI reading my text messages."
	How to Integrate in Phones	1	"I'd want to know more about how it would function and interface with my phone."

Table 13: Codebook for analyzing participants' feedback on the system design and usability.