Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Social Good

An LLM-based Chain-of-Response Counter-Scam System

Heedou Kim[1,2], **Mogan Gim**[3], **Donghee Choi**[4], **Soonil Bae**[5], **Hoonick Lee**[1], **Mi-Young Kim**[6*] and **Jaewoo Kang**[1*]
[1] Korea University
[2] Korean National Police Agency
[3] Hankuk University of Foreign Studies
[4] Pusan National University
[5] Korean National Police University
[6] University of Alberta
{heedou123, hoonick, kangj}@korea.ac.kr, gimmogan@hufs.ac.kr, dchoi@pusan.ac.kr, soonil.bae@gmail.com, miyoung2@ualberta.ca

Abstract

The rapid evolution of online scams, driven by transnational networks and mass-produced social engineering scenarios, has exposed the speed limitations of conventional detection, necessitating tighter inter-agency coordination. While LLMs show promise in scam identification, their role in accelerating integrated response frameworks remains underexplored. We propose Counter-Scam, a unified LLM-based multi-agent framework that orchestrates end-to-end response from initial detection to crime investigation. The framework first proposes safe data guidelines, emphasizing non-public scam data and secure dataset construction via scam-specific NER. Developed with insights from 37 stakeholders to reduce delays and improve analytical efficiency, the system integrates CSRA (multi-agent mitigation), CSRT (nine role-aligned NLP tasks), and CSRD (a corpus of 185,300 scam cases and 38,587 knowledge entries). Experiments show that fine-tuned sLLMs surpass commercial models by over 10% in all CSRT tasks and a 0.24 F1 improvement in scam-specific NER. This proves the framework's capability for enabling rapid, collaborative mitigation of online scam.

1 Introduction

Online scams have become a serious global threat driven by transnational criminal networks, causing substantial financial losses of innocent people and increasing the burden on law enforcement [Interpol, 2024]. Scammers employ sophisticated social engineering scenarios across calls, messaging apps, and social media by impersonating family members, romantic partners, financial institutions, or authorities, inflicting severe psychological harm on victims [FBI, 2025; Yosiandra and Sakariah, 2024]. In this context, LLM-based agents specialized in social engineering content show strong potential as next-generation scam response technologies by enabling real-time detection and delivering warnings to victims [Shen *et al.*, 2025; Kumarage *et al.*, 2025].

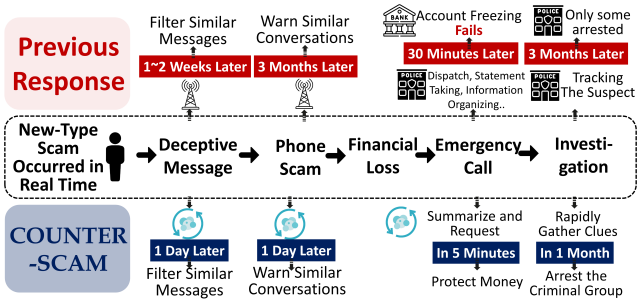

Figure 1: Previous scam response systems often fail to keep pace with the rapid fund extraction of criminals due to fragmented inter-agency coordination and manual analysis. COUNTER-SCAM enables efficient end-to-end scam response.

However, the practical effectiveness of these technologies is often limited by critical delays in institutional response. As shown in Figure 1, when a citizen receives a new type of smishing involving a fake payment, delays in blacklisting allow scammers to complete their deception. Even with emergency reports, account freezing often lags behind the speed of illicit withdrawals. These systemic delays, stemming from inter-agency coordination and manual analysis, span detection, reporting, and intervention, enabling transnational criminal networks to outpace defenses. An integrated framework is urgently needed to enable rapid, coordinated multi-agency responses and close these gaps [Tundis and Mühlhäuser, 2019].

Despite the agentic potential of LLMs, most existing studies remain detection-centric and fail to address coordination delays across institutions and jurisdictions. Even with accurate detection, fragmented downstream responses can lead to failure, forcing citizens to navigate disconnected systems for actions such as account freezing and investigation and often missing the critical window for financial recovery [Westmore *et al.*, 2022; Luong and Ngo, 2024]. This limitation is particularly severe as international criminal networks exploit

*Co-corresponding author

weakly regulated jurisdictions and real-time payment systems to accelerate cross-border fund exfiltration [Mitchell, 2025].

We propose **COUNTER-SCAM** , an LLM-based multi-agent framework for end-to-end online scam response. It proactively blocks or detects scam-related content in real time across channels such as phone calls and text messages and, when prevention fails, rapidly reports relevant information to law enforcement agencies and security teams, integrating prevention, emergency response, and investigation. By enabling multiple stakeholders to jointly leverage the agentic capabilities of LLMs, the framework supports rapid inter-agency sharing and analysis to prevent scams and minimize harm.

COUNTER-SCAM is a unified framework designed to integrate the previously fragmented stages of scam response organically, including scam detection, blocking, emergency intervention, and post-incident handling. To achieve this, it is composed of **C**hain-of-**S**cam-**R**esponse **A**gents (**CSRA**), which explicitly model the full lifecycle of scam response through specialized agents; **C**hain-of-**S**cam-**R**esponse **T**asks (**CSRT**), a suite of NLP tasks that guide LLMs toward role-aligned decision-making for each agent; and **C**hain-of-**S**cam-**R**esponse **D**ataset (**CSRD**), a carefully constructed dataset grounded in real-world scam response knowledge to ensure safe, reliable, and effective task execution. To the best of our knowledge, this work is the first to move beyond *standalone* scam detection and present an *integrated* agent–task–data framework that covers the *entire* online scam response.

CSRA was developed through interviews with experts in cybersecurity and criminal investigation, resulting in an agent-based framework for comprehensive phone and text based scam response. It organizes mitigation into three stages including pre scam, immediate post victimization, and repeated victimization, guided by a situational analysis agent that determines appropriate actions. Three specialized agents then support execution: a scam prevention agent for proactive blocking and real time detection, an emergency response agent for immediate interventions such as reporting and account freezing, and an investigation agent for large scale analysis of organized scam activities. By covering the full scam cycle and linking information across stages, **CSRA** reduces response gaps and enables coordinated mitigation. Through this, it aims to complement law enforcement capacities and strengthen collaboration against international scams.

For practical deployment, we adopted strict safety principles. To prevent misuse, we release only non-scam data, apply scam-specific NER–based de-identification, and actively involve domain experts. Under this framework, the construction of **CSRT** and **CSRD** proceeds by (1) defining additional eight tasks that capture the core capabilities of each agent based on prior work, and (2) collecting original scam cases and scam response knowledge, including legal and operational materials, from the Korean National Police to build expert-annotated training and evaluation datasets. As a result, we construct a large-scale corpus comprising 185,300 scam cases and 38,587 response knowledge entries, providing a strong foundation for comprehensive scam response agents.

Under real-world constraints where commercial models are restricted, we fine-tuned sLLMs on the **CSRD**. Our results show that scam-specific NER consistently outperformed commercial models by 0.24 F1, and the fine-tuned sLLMs exceeded GPT-4o and Gemini's average performance by 10%, showing improvements of up to 0.55 in deceptive message analysis compared to zero-shot baselines. Nonetheless, complex reasoning tasks such as legal analysis and long-context emergency report summarization remain challenging, highlighting the need for further training. These results demonstrate the potential and effectiveness of **COUNTER-SCAM** as an integrated scam response system and point toward directions for further enhancement. The main contributions are:

1. We propose **COUNTER-SCAM** , the first integrated agent–task–data framework covering the entire cycle of scams for an end-to-end response.
2. Involving 37 stakeholders, we developed **CSRA** (agents), **CSRT** (tasks), and **CSRD** (misuse-resistant data) for practical LLM-based scam response.
3. We demonstrate that fine-tuned sLLMs significantly outperform commercial large models on scam-response-specific 9 tasks.
4. By releasing our framework and large datasets,[1], we provide a foundation for scalable deployment and cross-border collaboration in scam mitigation.

2 Related Work

2.1 Evolution of Online Scams and Challenges

Online scams exploit human psychological vulnerabilities through social engineering, using SMS, calls, and social media to impersonate family, acquaintances, recruiters, or authorities and defraud victims [Kumarage *et al.*, 2025; Ai *et al.*, 2024]. Transnational criminal networks based in Southeast Asia, such as Cambodia and Vietnam, orchestrate these schemes, exploiting legal gaps and money-laundering networks [Amnesty International, 2025]. Effective mitigation requires coordinated cross-border and cross-institution responses, yet fragmented law enforcement and weak collaboration allow scams to persist [Shen *et al.*, 2025; Loggen and Leukfeldt, 2022].

2.2 LLM-based Scam Response

Language models have emerged as crucial tools in scam response, capable of accurately identifying scams and providing interpretable explanations [Kim *et al.*, 2026; Koide *et al.*, 2024; Lee and Han, 2024]. Recent advances in multimodal processing and real-time detection have extended their utility beyond mere plausibility checks, enabling more realistic and practical applications in real-world settings [Cao *et al.*, 2025; Shen *et al.*, 2025]. Building on this trend, LLMs are increasingly envisioned as agent-based automated tools that can assist human responders at critical intervention points, supporting faster and more effective scam mitigation [Xue *et al.*, 2025; Afane *et al.*, 2024].

3 Method

Our key insight is that, just as large-scale scam organizations operate through internal coordination and role specializa-

[1]Github repository

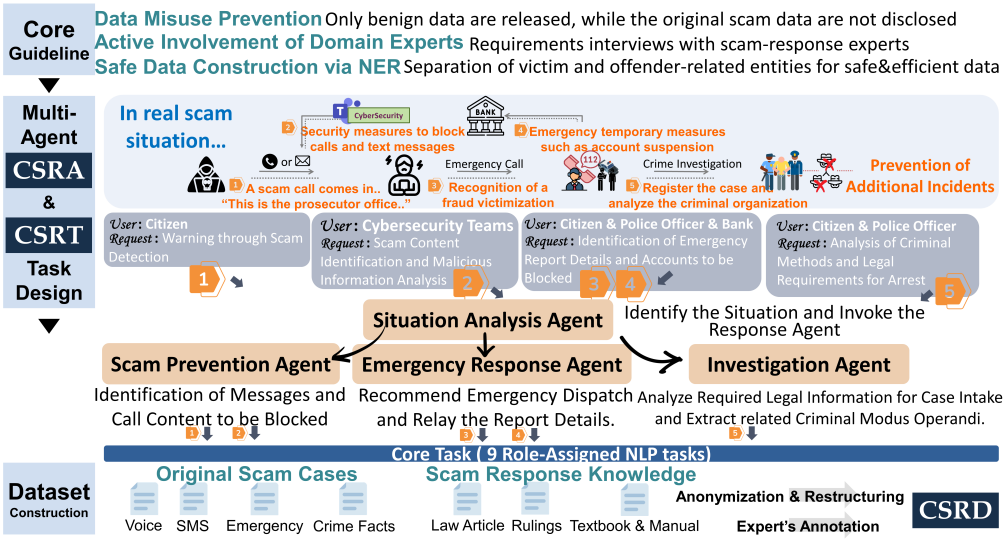

Figure 2: **COUNTER-SCAM** is an agent-based system that, unlike prior LLM studies focusing solely on scam content detection, supports AI-driven coordinated scam response across citizens, security teams, law enforcement, and banks within the real-world scam response workflow.

Agent	Task	Purpose	Cases	Knowledge	Instance Type
All Agents	Case Analysis NER	Crime entity extraction (malicious links, suspect numbers)	D_{Phone}: 14k D_{SMS}: 25k	–	$\{X_{Tok}, Y_{NER}\}$
Situation Analysis	Offense Detection	Crime type identification	D_{Crime}: 144k	–	$\{X_{Cri}, Y_{Cri}\}$
	Operational QA	Response planning	QA	K_{PM}: 5.6k	$\{X_Q, Y_A\}$
Scam Prevention	Scam Scenario Detection	Scam conversation detection	D_{Phone}: 14k	K_{Benign}: 14k	$\{X_{Scam}, Y_{Scam}\}$
	Message Classification	Fraud modus classification	D_{SMS}: 25k	–	$\{X_{Sms}, Y_{M.O.}\}$
Emergency Response	Call Summarization	Rapid victim report summary	D_{ER}: 2.3k	–	$\{X_{Em}, Y_{Em}\}$
Investigation	Hypothesis Eval	Determination of criminality with its rationales	–	K_{CI}: 687 K_{CL}: 3.3k K_{CR}: 15k	$\{X_{Hyp}, Y_{Hyp}\}$
	Statute Mapping	Applicable law prediction	D_{Crime}: 144k	–	$\{X_{Cri}, Y_{Law}\}$
	Element Analysis	Legal element identification	D_{Crime}: 144k	–	$\{X_{Cri}, Y_{Com}\}$

Table 1: The purpose, type and dataset formats of all tasks (**CSRT**) assigned to the **COUNTER-SCAM** Agent (**CSRA**).

tion, effective scam response requires seamless collaboration among all stakeholders including government, victims, and industry. As illustrated in Figure 2, this integration can be realized through a multi-agent framework. Based on interviews with field experts, we derived design principles for such agents. The resulting **CSRA** defines clear roles and responsibilities for each agent, **CSRT** structures LLM-based NLP tasks to operationalize these roles, and **CSRD** provides a safe and reliable dataset for task execution and evaluation. We further validate the framework's practical feasibility through evaluations.

3.1 Establishing Core Guidelines

A key requirement for an effective scam response ecosystem involving multiple stakeholders is the seamless integration of incident data with real-time information. In practice, however, the inherent risks and potential for misuse of scam data have led to restricted access, which in turn has hindered the development of integrated response systems. To address this gap, we propose core guidelines for designing datasets that ensure safety while enabling efficient collaboration among stakeholders, and empirically demonstrate their implementation.

① **Data Misuse Prevention.** We adopt a principle of not releasing real scam conversation data that could be exploited by fraudsters for scenario generation or victim information theft [Permana and Jamaludin, 2023]. Instead, we provide carefully curated benign data, enabling the development of detection models while minimizing false positives. Additionally, officially produced texts such as emergency report dialogues, legal summaries of criminal incidents, and communications between law enforcement and citizens are made available to support the generation of appropriate response actions.

② **Active Involvement of Domain Experts.** We actively engaged domain experts to develop the practical components necessary for implementing **COUNTER-SCAM** , spanning agent design, task definition, and dataset creation. A total of 37 experts from four fields including general police officers, criminal investigators, cyber fraud investigators and policy analysts participated in deriving agent designs, evaluating the

operational suitability of proposed tasks through preliminary interviews, and performing data labeling and quality assessments. The reliability of their annotations and evaluations was ensured through consistent adherence to classification guidelines, supplemented by inter-annotator agreement measurements and correlation analyses to validate the results.

③ **Safe Data Construction via NER.** For efficient and safe scam response, we designed the CASE ANALYSIS NER. This pipeline clearly distinguishes victim information, perpetrator details, and impersonated identities, minimizing exposure of sensitive data while preserving analytically meaningful signals for crime analysis. In particular, it is designed to extract impersonated identities (e.g., fake names) and criminal tools (e.g., malicious links and account numbers), enabling their use for crime prevention and tracking [Kim and Lim, 2022]. We constructed a dataset from 28,806 texts including phone scam conversations, smishing messages, and crime fact data, labeled with 10 crime-domain and 13 general-domain entity types using BIO tagging.

> **CASE ANALYSIS NER**
> Let $D = \{X_{\text{Tok}}, Y_{\text{NER}}\}$. For a scam-related text, each token sequence $X_{\text{Tok}} \in X$, where $X \subseteq \{D_{\text{Phone}}, D_{\text{Sms}}, D_{\text{Crime}}\}$, maps to a BIO-labeled entity sequence Y_{NER} identifying crime-specific (e.g., fake identities, tools) and general entities (e.g., dates).

3.2 LLM-Based Multi-Agent(CSRA)

Goal 1 (End-to-End Scam Response)**.** *The objective of* **COUNTER-SCAM** *is to identify the phase of an online scam and deliver phase-appropriate response to the relevant stakeholders in a timely manner. Since online scams unfold as lifecycle-oriented crimes rather than single events, effective mitigation requires structured, phase-aware responses aligned with specific analytical goals and responsible actors. We define an agent $A = \langle R, G, I, O, S \rangle$ by its role R, objective G, input space I, output space O, and stakeholders S.*

Agent 1 (Situation Analysis)**.** *The Situation Analysis Agent A_{SA} is designed to assess scam situations across heterogeneous users, where R_{SA} denotes situation assessment, G_{SA} is to identify the user's current situation and analytical needs, and I_{SA} consists of scam messages, emergency reports, victim summaries, and contextual metadata such as user type and message timing. Given an input $i \in I_{\text{SA}}$, the agent produces an output $o \in O_{\text{SA}}$ through joint reasoning over input content and metadata as $o = A_{\text{SA}}(i)$, and invokes the downstream agent. Stakeholders are categorized into four groups:*

- ***Citizens**: targets of scam attacks*
- ***Cybersecurity teams**: analyze blacklists and develop detection models to proactively block scams*
- ***Police officers**: respond to scams through emergency actions and investigations*
- ***Banks**: handle financial losses resulting from scam-related fund transfers*

Agent 2 (Scam Prevention)**.** *The stakeholders S_{SP} include citizens and cybersecurity teams. The role R_{SP} aims to prevent scam-related harm via early detection. Given an input message $i \in I_{\text{SP}}$, the agent generates an output $o \in O_{\text{SP}}$ by identifying scam, issuing warnings to citizens, or, for security teams, classifying social engineering techniques and extracting malicious indicators.*

Agent 3 (Emergency Response)**.** *The stakeholders S_{ER} are citizens, police, and banks. The role R_{ER} supports urgent scam response and emergency intervention. Given an input $i \in I_{\text{ER}}$ consisting of urgent victim reports and contextual metadata, the agent outputs $o \in O_{\text{ER}}$ by generating an emergency-focused summary for immediate protective action, forwarding it to relevant law enforcement agencies, and, for police officers, extracting action-critical entities and coordinating with financial institutions and telecom providers.*

Agent 4 (Investigation)**.** *The stakeholders S_{I} are citizens, police. The role R_{I} supports post-incident investigation and legal reasoning. Given an input $i \in I_{\text{I}}$ consisting of detailed victim accounts and contextual information, the agent outputs $o \in O_{\text{I}}$ by assessing criminal liability, generating a statute-grounded legal analysis, and extracting impersonated identities and instrumentalities to trace recurring crimes linked to the same criminal organization.*

3.3 Agent-Specific Core Tasks(CSRT)

We define NLP tasks that underpin each agent's decision-making capabilities. Guided by a review of prior work on scam detection and crime-related NLP, we identified tasks critical for these capabilities and categorized the types of textual inputs that stakeholders may provide during the scam response process. The task inputs include criminal facts X_{Cri}, scam response-related queries X_{Que}, conversations suspected of being scam-related X_{Scam}, scam messages X_{Sms}, general texts containing crime-related information X_{Tok}, victims' urgent statements X_{Em}, and criminal hypotheses X_{Hyp}.

① **Situation Analysis Agent.** Just as police officers rely on foundational procedural knowledge to make situational judgments [Holgersson and Gottschalk, 2008], the Situation Analysis Agent must accurately assess the state of scam victimization and select the appropriate response pathway among prevention, emergency, and investigation.

> **OFFENSE DETECTION** $D = \{X_{\text{Cri}}, Y_{\text{Cri}}\}$. For a description of suspected criminal activity, each input $X_{\text{Cri}} \in X$, where $X \subseteq \{D_{\text{Crime}}\}$, maps to a label Y_{Cri} corresponding to one of 111 offense categories.
>
> **OPERATIONAL QA** $D = \{X_{\text{Que}}, Y_{\text{Ans}}\}$. For procedural questions related to scam response, each input $X_{\text{Que}} \in X$, where $X \subseteq \{K_{\text{PM}}\}$, maps to an answer Y_{Ans} derived from police procedural manuals.

The agent can first determine whether a conduct constitutes a criminal offense and identify its basic offense category through OFFENSE DETECTION, and then use OPERATIONAL QA to decide the appropriate next steps.

② **Scam Prevention Agent.** To prevent imminent scams, the Scam Prevention Agent must detect content in real time, warn citizens, and filter malicious links and phone numbers extracted from previously identified scam cases of the same scenario to block their redistribution.

SCAM SCENARIO DETECTION $D = \{X_{\text{Scam}}, Y_{\text{Scam}}\}$. For a given conversation, $X_{\text{Scam}} \in X$, where $X \subseteq \{D_{\text{Phone}}, K_{\text{Benign}}\}$, maps to a label Y_{Scam} indicating whether the input is fraudulent or benign.

SCAM MESSAGE CLASSIFICATION $D = \{X_{\text{Sms}}, Y_{\text{Sms}}\}$. For SMS messages, $X_{\text{Sms}} \in D_{\text{Sms}}$ represents bait messages by scammers to deceive victims, and maps to a label Y_{Sms} indicating one of seven scam tactics.

The agent can identify conversations in real time and issue warnings to consenting citizens, while classifying scam tactics in reported or large-scale analyzed texts and, through CASE ANALYSIS NER, extracting malicious indicators to be blocked and providing them to cybersecurity teams.

③ **Emergency Response Agent.** When a victim has transferred funds due to a scam, immediate action such as prompt reporting and emergency account freezing by financial institutions is essential to mitigate losses.

EMERGENCY CALL SUMMARIZATION $D = \{X_{\text{Em}}, Y_{\text{Em}}\}$. For an emergency report, input $X_{\text{Em}} \in D_{\text{E.R.}}$ consists of transcripts of emergency calls, and maps to a summary Y_{Em} that appropriately reflects the situation.

The agent supports urgent responses by police and financial institutions through rapid summarization of reports following the occurrence of scam victimization. The summarized report facilitates dispatch, and, together with account extracted via CASE ANALYSIS NER, enables prompt account freezing.

④ **Investigation Agent.** Just as a crime investigator collects evidence and pursues offenders [Roberts, 2012; Stainton *et al.*, 2025], the Investigator Agent extracts criminal methods from texts, including impersonation details and instruments like accounts and phone numbers, to trace repeat offenders and support prosecution under relevant legal statutes.

The agent supports investigations of large-scale scam cases to prevent further harm. It ensures lawful investigation by assessing criminality (CRIMINAL HYPOTHESIS EVALUATION), mapping applicable laws (STATUTE MAPPING), and identifying statutory elements (ELEMENT ANALYSIS), thereby supporting clear investigative reports. Information extracted by CASE ANALYSIS NER further enables case clustering and follow-up investigation.

CRIMINAL HYPOTHESIS EVALUATION $D = \{X_{\text{Hyp}}, Y_{\text{Hyp}}\}$. Each input $X_{\text{Hyp}} \in X$, derived from $K_{\text{CI}}, K_{\text{CR}}$ and reformulated as a hypothesis requiring legal assessment, maps to an output Y_{Hyp} consisting of a binary judgment (true or false) and a corresponding rationale.

STATUTE MAPPING $D = \{X_{\text{Cri}}, Y_{\text{Law}}\}$. Given a crime description X_{Cri}, , each input $X_{\text{Cri}} \in X$, where $X \subseteq \{D_{\text{Crime}}\}$, the task maps it to one or more applicable criminal law articles, where Y_{Law} is selected from 160 statutorily defined provisions.

ELEMENT ANALYSIS $D = \{X_{\text{Cri}}, Y_{\text{Com}}\}$. Given a crime description X_{Cri}, each input $X_{\text{Cri}} \in X$, where $X \subseteq D_{\text{Crime}}$, is mapped to one or more legal elements constituting the offense, where Y_{Com} is selected from 129 statutorily defined components extracted from law articles.

3.4 Dataset Construction

We built benchmark datasets for all tasks through data processing and expert annotation. Data were collected from two sources, Original Scam Cases for detection and Response Knowledge for situational and legal reasoning.

Original Scam Cases. As summarized in Table 1, the Original Scam Case data comprise four subsets: $D_{\text{Phone}}, D_{\text{Sms}}, D_{\text{E.R.}}$, and D_{Crime}. D_{Phone} includes 14,030 real voice phishing conversations collected from the Korean National Police Agency, substantially larger and more realistic than prior datasets [Boussougou *et al.*, 2024; Ma *et al.*, 2025]. D_{Sms} consists of 24,947 victim-reported scam SMS messages, representing a significantly larger real-world smishing corpus than existing benchmarks [Timko and Rahman, 2024; Tanbhir *et al.*, 2024]. $D_{\text{E.R.}}$ is built from 2,361 anonymized call transcripts between victims and police officers absent from public dispatch datasets [Clinic, ; Data,]. Finally, D_{Crime} contains descriptions of criminal victimization derived from legal benchmarks and supplemented with synthetic cases [Hwang *et al.*, 2022; Law&Company, 2022].

Scam Response Knowledge. Effective scam response requires not only domain-specific knowledge of scams but also alignment with general law-enforcement protocols and legal frameworks. Accordingly, our knowledge base is designed to cover comprehensive crime-response knowledge and is organized into five components: $K_{\text{PM}}, K_{\text{Benign}}, K_{\text{CI}}, K_{\text{CL}}$, and K_{CR}. K_{PM} captures 5,660 police procedural knowledge, constructed from internal manuals spanning 380 crime types and 57 response domains. K_{Benign} is introduced to reduce false positives in scam detection by augmenting dialogue data that reflect legitimate police-initiated calls to citizens, based on real investigator examples. $K_{\text{CI}}, K_{\text{CL}}$, and K_{CR} provide criminal investigation, criminal law, and judicial precedent knowledge, respectively, leveraging the publicly available LAPIS knowledge base [Kim *et al.*, 2024].

Expert-Annotated Task Dataset. For OPERATIONAL QA, questions and answers were generated from K_{PM} and reviewed by police experts, resulting in 2,624 high-quality QA pairs. Unlike prior legal QA benchmarks [Zhong *et al.*, 2020; Goebel *et al.*, 2024], our benchmark captures real-world, manual-grounded police procedures. For OFFENSE DETECTION and STATUTE MAPPING, offense labels assigned to all 143,893 D_{Crime} instances were consolidated using the standard charge classification table [Agency, 2023], mapping them to 111 unified offense categories and corresponding legal labels. For SCAM SCENARIO DETECTION, we constructed

balanced benign counterparts to D_{Phone} by collecting non-fraud dialogues from financial consultations [Agency, 2025] and the Korean Dialogue Corpus [of Korean Language, 2024], and incorporating simulated official calls from K_{Benign}. For SCAM MESSAGE CLASSIFICATION, three crime analysts annotated all messages with one of seven fraud tactic categories. For EMERGENCY CALL SUMMARIZATION, experts summarized dual-perspective summaries (caller and officer) from anonymized $D_{\text{E.R.}}$ transcripts. For ELEMENT ANALYSIS, we analyzed 160 legal provisions and extracted 129 common crime elements, covering both conduct and result.

3.5 Evaluation under Real-World Constraints

We evaluated whether **COUNTER-SCAM** can operate under real-world constraints such as closed-network environments, limited computation, and the unavailability of commercial LLM APIs. We compared commercial LLMs, including chatgpt-4o-latest and gemini-2.0-flash, as performance upper bounds against deployable open-source sLLMs across our tasks and datasets, including a 1B-scale model to test extreme resource limits. Multilingual sLLMs such as Llama-3.1-8B-Instruct, SOLAR-10.7B-Instruct, and Llama-3.2-1B-Instruct were evaluated for deployment on police intranet servers or on-device police phones, alongside a Korean-tuned model, EEVE-Korean-Instruct-10.8B, for Korean crime-response settings. Commercial models were evaluated in zero-shot settings, while sLLMs were tested under zero-shot inference, domain-specific fine-tuning, and RAG, with fine-tuning performed for five epochs using QLoRA and the Paged AdamW optimizer on two A100 80GB GPUs.

4 Experiment Results

Metrics. For OPERATIONAL QA, to reflect deployment constraints where commercial models cannot access police manuals, we applied RAG-based QA prompts only to open-source models. For classification tasks, we used rule-based post-processing to correct format errors and align outputs with predefined labels [Xia *et al.*, 2024], and evaluated performance using Accuracy and F1 score. For open-ended tasks such as summarization and QA, we adopted an LLM-as-a-Judge framework [Chiang and Lee, 2023], using expert-annotated gold references and instructing the judge model to evaluate outputs strictly against those references [Zheng *et al.*, 2023]. Prior studies comparing expert and automatic evaluations on police manuals have demonstrated the validity of this approach [Lee *et al.*, 2026], and in our study, correlations with human evaluations exceeded 0.8, confirming strong agreement with experts.

4.1 Crime-domain NER Is Effective, Yet Human Validation Is Still Essential For Safety

As shown in Table 2, fine-tuned sLLMs for CASE ANALYSIS NER consistently outperformed commercial models, achieving substantially higher F1 scores when both B- and I-tokens were correctly identified. While commercial models averaged an F1 of 0.35 across all entity tokens, fine-tuned sLLMs reached 0.59, with the best-performing models showing a large gap (GPT-4o 0.42 vs. Llama8B 0.79). In the crime domain, newly defined entities such as fake position, accused name, account, and crime tools were recognized far more accurately by fine-tuned sLLMs, indicating that even small models can effectively support safe, partially automated crime data processing. Commercial models also struggled with precise span detection for general entities, reflecting limited adaptability to crime-specific texts. However, fine-tuned sLLMs still showed weaknesses in extracting highly variable malicious elements such as bad links (maximum F1 0.54), suggesting the continued need for human review.

Entity	Gemini	GPT4	EEVE	Llama8B	Llama1B	SOLAR
Crime Domain						
fake position	.39	.63	**.70**	**.88**	.52	**.72**
accused name	.22	.28	**.82**	**.92**	.70	**.81**
case number	.22	.76	**.83**	**.92**	.76	.79
account	.14	.35	.09	**.94**	**.59**	.03
product	.08	**.44**	.21	**.72**	.32	.12
bad link	.11	.25	**.43**	**.54**	.34	**.48**
fakename	.38	.38	**.54**	**.72**	.49	.38
crime tool	.38	.61	**.74**	**.90**	.62	.55
law	.22	.24	**.69**	**.83**	.49	**.71**
fake org	.23	.18	**.48**	**.64**	**.52**	.51
General Domain						
relation	.15	.27	**.75**	**.86**	.60	.66
birth	.49	**.65**	**.59**	**.75**	.45	.29
identity	.00	.24	**.33**	**.33**	.30	**.36**
age	.32	.40	**.76**	**.97**	.54	.59
location	.32	.28	**.51**	**.67**	.48	.37
datetime	.31	.64	**.82**	**.94**	**.81**	.79
occupation	.22	.28	**.47**	**.58**	.33	.41
price	.35	.35	.52	**.94**	**.76**	.38
organization	.38	**.43**	**.48**	**.64**	.42	.42
plate	**.63**	**.78**	.50	**.84**	.29	.50
transport	**.66**	.52	.52	**.95**	.58	.13
name	.40	.61	**.72**	**.92**	**.76**	.80
position	.04	.10	.29	**.80**	.19	**.29**

Table 2: Experiment Results of CASE ANALYSIS NER.

4.2 A Scam-specific Labeled Dataset Is Essential For Reliable Detection

Accurate detection and fine-grained classification of scams are essential, as they determine downstream actions such as warnings and blocking. Although commercial LLMs achieve high performance in SCAM SCENARIO DETECTION (0.97 and 0.87), Table 4 indicates that non-negligible false negatives persist. Notably, Gemini misclassified 351 non-scam messages as scams, which can undermine system reliability. In contrast, the fine-tuned sLLM misclassified only 4 out of 1,426 instances, demonstrating substantially greater stability. The SCAM MESSAGE CLASSIFICATION task further reveals consistent misclassification patterns across commercial models. Both GPT-4o and Gemini frequently confused specific scam types, with errors concentrated in *Overseas payment fraud* and *Voice spam*. Such errors limit the system's ability to provide clear grounds for scam intervention by security teams.

4.3 While Fine-tuning Is Highly Effective In The Scam Response Task, More Fine-grained Training Is Still Required

Effect of Fine-tuning sLLMs. We compared the zero-shot performance of sLLMs with their fine-tuned counterparts on 150 small samples for each task. On average across all sLLMs, fine-tuning improved performance by 0.26, and the ability

Task	Metric	GPT-4o	Gemini 2.0	EEVE SFT	SOLAR SFT	Llama8B SFT	Llama1B SFT
Situation Analysis							
Operational QA	LLM Judge	0.69	0.66	**0.87**	0.85	**0.88**	0.64
Offense Detection	ACC	0.86	0.86	**0.87**	**0.98**	0.50	0.21
	F1	0.9	0.93	**0.95**	**0.99**	0.77	0.61
Scam Prevention							
Scam Scenario Detection	ACC	0.97	0.87	**0.99**	**0.99**	0.86	0.63
	F1	0.97	0.88	**0.99**	**0.99**	0.85	0.58
Scam Message Classification	ACC	0.88	0.93	**0.97**	**0.99**	**0.97**	0.88
	F1	0.7	0.76	0.91	**0.98**	**0.95**	0.73
Emergency Response							
Emergency Call Summarization	LLM Judge	**0.89**	**0.75**	0.62	0.56	0.51	0.2
Investigation							
Criminal Hypothesis Eval	ACC	**0.73**	0.62	**0.74**	0.62	0.62	0.62
	F1	**0.79**	0.77	**0.79**	0.77	0.77	0.77
Statute Mapping	ACC	0.43	0.4	**0.86**	**0.88**	0.19	0.07
	F1	0.65	0.69	**0.92**	**0.95**	0.35	0.12
Element Analysis	ACC	0.67	**0.81**	0.66	**0.71**	0.64	0.12
	F1	0.84	**0.93**	0.81	**0.88**	0.82	0.24

Table 3: Comparative Performance Evaluation of Fine-tuned sLLMs and Commercial LLMs

True Label	Inst.	GPT4	Gemini	EEVE	SOLAR
Scam Scenario Detection					
Non Scam	1426	38	**351**	4	4
Scam Message Classification					
Impersonation of acquaintances	1845	**3**	2	1	1
Job recruitment fraud	197	**7**	4	4	6
Overseas delivery scam	100	15	**16**	7	7
Impersonation of institutions	415	3	3	5	**6**
Voice spam	50	**50**	**50**	21	0
Overseas payment fraud	501	**327**	162	5	10
Illegal loan offers	638	36	40	**66**	10

Table 4: False Negatives by Task and Class.

to follow task-specific instructions also increased by 0.28, demonstrating a clear benefit of fine-tuning. In particular, the SCAM MESSAGE CLASSIFICATION showed an average improvement of 0.55 compared to zero-shot, confirming that fine-tuning is an effective approach.

Fine-tuned sLLMs vs Commercial Models. Table 3 shows performance across all tasks. Fine-tuned sLLMs outperformed commercial LLMs in 6 out of 9 tasks, highlighting their strong adaptability to the scam response. In particular, fine-tuned EEVE 10.8B and SOLAR 10.7B achieved an average score of 0.85 and 0.87, exceeding commercial LLMs by 10%.

Risk Without Scam Response Knowledge. Without useful manual knowledge enhanced by K_{PM}, scam response systems may hallucinate legal facts or provide operationally misleading answers, even when factually correct. We present concrete examples in the supplementary material. For instance, when asked how to delete harmful content, a commercial model merely provides a single deletion link, whereas the manual-based sLLM explains the actual procedure used in practice, including that reporting speed does not differ between individual and institutional submissions. This shows that commercial models can sometimes hinder efficient response.

Limitations from a Practical Usability Perspective. Despite improvements, some tasks still fall short of practical performance levels. OPERATIONAL QA achieved only scores in the 0.8 range even with manual-based RAG, indicating the need for additional evaluation metrics such as logical completeness. EMERGENCY CALL SUMMARIZATION also underperformed compared to commercial models despite fine-tuning, suggesting that smaller models lack the ability to effectively process long-context text. Finally, although tasks like ELEMENT ANALYSIS and CRIMINAL HYPOTHESIS EVALUATION showed significant gains, they still do not reach mature levels of accuracy, highlighting the need for more refined training methods for complex reasoning tasks.

5 Conclusion

We propose **COUNTER-SCAM** , the first multi-agent framework to integrate fragmented online scam response processes. By establishing a comprehensive agent-task-data model, we extend the scope of LLMs beyond simple detection to include blocking, emergency intervention, and post-incident investigation. Leveraging a large-scale dataset of over 185,300 scam cases and 38,587 legal knowledge entries, we demonstrated that fine-tuned sLLMs outperform commercial models in scam-specific NER and all tasks, proving their practical utility. While challenges remain in complex reasoning tasks such as legal analysis, this work confirms the potential of AI agents as practical partners in scam mitigation. Through open-source deployment, we aim to provide a foundation for cross-border collaboration against global scams.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00653, Development of a Voice Phishing Information Collection, Processing, and Big Data–Based Investigation Support System).

References

[Afane *et al.*, 2024] Khalifa Afane, Wenqi Wei, Ying Mao, Junaid Farooq, and Juntao Chen. Next-generation phishing: How llm agents empower cyber attackers. In *2024 IEEE International Conference on Big Data (BigData)*, pages 2558–2567. IEEE, 2024.

[Agency, 2023] Korean National Police Agency. Criminal charge classification table. https://www.police.go.kr/component/file/ND_fileDownload.do?q_fileSn=156181&q_fileId=0e28f484-9ad5-4ebf-b3d2-0ce03c4eeb3c, 2023. Accessed: 2025-09-25.

[Agency, 2025] National Information Society Agency. Aihub. https://www.aihub.or.kr/, 2025. Accessed: 2025-09-25.

[Ai *et al.*, 2024] Lin Ai, Tharindu Sandaruwan Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael S Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, et al. Defending against social engineering attacks in the age of llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12880–12902, 2024.

[Amnesty International, 2025] Amnesty International. Cambodia: 'i was someone else's property': slavery, human trafficking and torture in cambodia's scamming compounds. https://www.amnesty.org/en/documents/asa23/9447/2025/en/, June 26 2025. Accessed: 2026-01-16.

[Boussougou *et al.*, 2024] Milandu Keith Moussavou Boussougou, Prince Hamandawana, and Dong-Joo Park. Korean voice phishing detection dataset with multilingual back-translation and smote augmentations, 2024. IEEE Dataport.

[Cao *et al.*, 2025] Tri Cao, Chengyu Huang, Yuexin Li, Wang Huilin, 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*, volume 39, pages 27869–27877, 2025.

[Chiang and Lee, 2023] Cheng-Han Chiang and Hung-Yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, 2023.

[Clinic,] Data Clinic. Vera 911 consolidated dataset. https://github.com/tsdataclinic/Vera. Accessed: 2025-09-25.

[Data,] San Francisco Open Data. San francisco law enforcement dispatched calls for service. https://data.sfgov.org/Public-Safety/Law-Enforcement-Dispatched-Calls-for-Service-Real-/gnap-fj3t. Accessed: 2025-09-25.

[FBI, 2025] FBI. Fbi internet crime report, 2025. Accessed on 2025-10-04.

[Goebel *et al.*, 2024] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024. In *JSAI International Symposium on Artificial Intelligence*, pages 109–124. Springer, 2024.

[Holgersson and Gottschalk, 2008] Stefan Holgersson and Petter Gottschalk. Police officers' professional knowledge. *Police Practice and Research: An International Journal*, 9(5):365–377, 2008.

[Hwang *et al.*, 2022] Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551, 2022.

[Interpol, 2024] Interpol. Interpol annual report, 2024. Accessed on 2025-10-04.

[Kim and Lim, 2022] Hee-Dou Kim and Heuiseok Lim. A named entity recognition model in criminal investigation domain using pretrained language model. *Journal of the Korea Convergence Society*, 13(2):13–20, 2022.

[Kim *et al.*, 2024] Heedou Kim, Dain Kim, Jiwoo Lee, Chanwoong Yoon, Donghee Choi, Mogan Gim, and Jaewoo Kang. Lapis: language model-augmented police investigation system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4637–4644, 2024.

[Kim *et al.*, 2026] Heedou Kim, Changsik Kim, Sanghwa Shin, and Jaewoo Kang. SCRIPTMIND: Crime script inference and cognitive evaluation for LLM-based social engineering scam detection system. In Yevgen Matusevych, Gülşen Eryiğit, and Nikolaos Aletras, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 11–38, Rabat, Morocco, March 2026. Association for Computational Linguistics.

[Koide *et al.*, 2024] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Chatspamdetector: Leveraging large language models for effective phishing email detection. *arXiv preprint arXiv:2402.18093*, 2024.

[Kumarage *et al.*, 2025] Tharindu Kumarage, Cameron Johnson, Jadie Adams, Lin Ai, Matthias Kirchner, Anthony Hoogs, Joshua Garland, Julia Hirschberg, Arslan Basharat, and Huan Liu. Personalized attacks of social engineering in multi-turn conversations–llm agents for simulation and detection. *arXiv preprint arXiv:2503.15552*, 2025.

[Law&Company, 2022] Law&Company. Klaid. https://huggingface.co/datasets/lawcompany/KLAID, 2022. Accessed: 2025-09-25.

[Lee and Han, 2024] 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*, pages 642–656, 2024.

[Lee *et al.*, 2026] S. Lee, H. Kim, and H. Kim. Evaluating llms for police decision-making: A framework based on police action scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38817–38825, 2026.

[Loggen and Leukfeldt, 2022] Joeri Loggen and Rutger Leukfeldt. Unraveling the crime scripts of phishing networks: an analysis of 45 court cases in the netherlands. *Trends in Organized Crime*, 25(2):205–225, 2022.

[Luong and Ngo, 2024] Hai Thanh Luong and Hieu Minh Ngo. Understanding the nature of the transnational scam-related fraud: Challenges and solutions from vietnam's perspective. *Laws*, 13(6):70, 2024.

[Ma *et al.*, 2025] Zhiming Ma, Peidong Wang, Minhua Huang, Jinpeng Wang, Kai Wu, Xiangzhao Lv, Yachun Pang, Yin Yang, Wenjie Tang, and Yuchen Kang. Teleantifraud-28k: An audio-text slow-thinking dataset for telecom fraud detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5853–5862, 2025.

[Mitchell, 2025] Ottilie Mitchell. Nearly 60 south koreans repatriated from cambodia over alleged scams, October 2025.

[of Korean Language, 2024] National Institute of Korean Language. Nikl korean dialogue corpus(transcription) 2023(v.1.0). https://kli.korean.go.kr/corpus, 2024. Accessed: 2025-09-25.

[Permana and Jamaludin, 2023] Flea Akbar Permana and Ahmad Jamaludin. Personal data vulnerability in the digital era: Study of modus operandi and mechanisms to prevent phishing crimes. *Jurnal Al-Hakim: Jurnal Ilmiah Mahasiswa, Studi Syariah, Hukum dan Filantropi*, pages 201–216, 2023.

[Roberts, 2012] Paul Roberts. Law and criminal investigation. In *Handbook of criminal investigation*, pages 92–145. Willan, 2012.

[Shen *et al.*, 2025] Zitong Shen, Sineng Yan, Youqian Zhang, Xiapu Luo, Grace Ngai, and Eugene Yujun Fu. " it warned me just at the right moment": Exploring llm-based real-time detection of phone scams. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

[Stainton *et al.*, 2025] Iain Stainton, Robert Ewin, and Tony Blockley. *Criminal investigation*. Routledge, 2025.

[Tanbhir *et al.*, 2024] Gazi Tanbhir, Md Farhan Shahriyar, Khandker Shahed, Abdullah Md Raihan Chy, and Md Al Adnan. Hybrid machine learning model for detecting bangla smishing text using bert and character-level cnn. In *2024 13th International Conference on Electrical and Computer Engineering (ICECE)*, pages 57–62. IEEE, 2024.

[Timko and Rahman, 2024] Daniel Timko and Muhammad Lutfor Rahman. Smishing dataset i: Phishing sms dataset from smishtank. com. In *Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy*, pages 289–294, 2024.

[Tundis and Mühlhäuser, 2019] Andrea Tundis and Max Mühlhäuser. The role of information and communication technology (ict) in modern criminal organizations. In *Organized Crime and Terrorist Networks*, pages 60–77. Routledge, 2019.

[Westmore *et al.*, 2022] Kathryn Westmore, Simon Miller, Jonathan Frost, and Diana Foltean. Enabling cross-sector data-sharing to better prevent and detect scams, September 2022.

[Xia *et al.*, 2024] Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. Fofo: A benchmark to evaluate llms' format-following capability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–699, 2024.

[Xue *et al.*, 2025] Yinuo Xue, Eric Spero, Yun Sing Koh, and Giovanni Russello. Multiphishguard: An llm-based multi-agent system for phishing email detection. *arXiv preprint arXiv:2505.23803*, 2025.

[Yosiandra and Sakariah, 2024] Syakira Yuniar Yosiandra and Dewi Saraswati Sakariah. Unveiling the romance scam scheme: Psychological manipulation and its impact on victims. *Humanika*, 31(2):185–199, 2024.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[Zhong *et al.*, 2020] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. How does nlp benefit legal system: A summary of research on legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, 2020.