SCRIPTMIND: Crime Script Inference and Cognitive Evaluation for LLM-based Social Engineering Scam Detection System

Heedou Kim[1,2], Changsik Kim[2], Sanghwa Shin[3,*], Jaewoo Kang[1,*]
[1]Korea University, Seoul, Republic of Korea,
[2]Korean National Police Agency, Seoul, Republic of Korea,
[3]Inje University, Gimhae, Republic of Korea
Correspondence: heedou123@korea.ac.kr
[*]Co-corresponding authors.

Abstract

Social engineering scams increasingly employ personalized, multi-turn deception, exposing the limits of traditional detection methods. While Large Language Models (LLMs) show promise in identifying deception, their cognitive assistance potential remains underexplored. We propose **SCRIPTMIND**, an integrated framework for LLM-based scam detection that bridges automated reasoning and human cognition. It comprises three components: the Crime Script Inference Task (**CSIT**) for scam reasoning, the Crime Script–Aware Inference Dataset (**CSID**) for fine-tuning small LLMs, and the Cognitive Simulation-based Evaluation of Social Engineering Defense (**CSED**) for assessing real-time cognitive impact. Using 571 Korean phone scam cases, we built 22,712 structured scammer-sequence training instances. Experimental results show that the 11B small LLM fine-tuned with **SCRIPTMIND** outperformed GPT-4o by 13%, achieving superior performance over commercial models in detection accuracy, false-positive reduction, scammer utterance prediction, and rationale quality. Moreover, in phone scam simulation experiments, it significantly enhanced and sustained users' suspicion levels, improving their cognitive awareness of scams. **SCRIPTMIND** represents a step toward human-centered, cognitively adaptive LLMs for scam defense.
Data & Code: anonymous/ScriptMind

1 Introduction

Preventing social engineering scams is essential for financial security, psychological protection, and societal trust. Online scams have grown sophisticated, demanding more adaptive defenses. In this context, Large Language Models (LLMs) have emerged as interpretable, cognitively assistive tools capable of detecting deception and enhancing user awareness

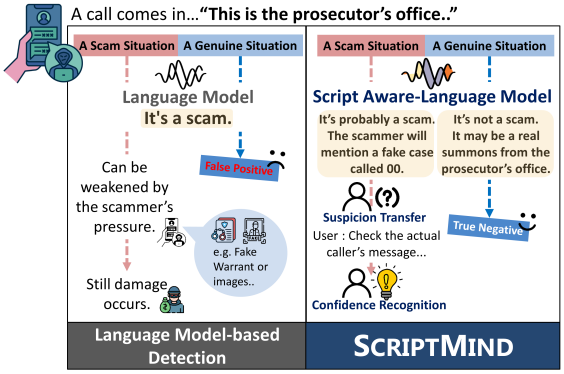

Figure 1: Scam alerts provide accurate detection and explanations but can be neutralized by new tactics. **SCRIPTMIND** overcomes these limits through a crime script inference and simulation-based evaluation, enabling cognitively effective scam defense.

(Lim et al., 2025), proving effective in brand impersonation, fake webpage detection, and phone scam monitoring as real-time defense systems against evolving social engineering scams (Koide et al., 2024; Lee et al., 2024; Shen et al., 2025).

However, social engineering scams have become increasingly sophisticated, using psychological tactics that neutralize suspicion. Scammers exploit user's anxiety and shifting doubt through multiple strategies, leading to psychological submission (Han et al., 2024; Wang et al., 2021). Thus, detection must move beyond scam identification to consider how and when warnings are cognitively delivered. As shown in Figure 1, alerts themselves can be manipulated. For example, scammers may counter a "fake prosecutor" warning by invoking legal pressure by presenting fabricated court documents. False positives in benign interactions can also erode system trust. Therefore, effective defense requires modeling the temporal dynamics of suspicion and designing adaptive cognitive assistants that respond to users' evolving mental states.

Most existing LLM-based social engineering scam detection studies focus primarily on identify-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 5: Industry Track, pages 11–38

ing deceptive content, without adequately reflecting how users' cognitive and behavioral responses evolve in scam situations. In addition, little attention has been paid to how warning alerts influence users' suspicion levels or what types of feedback effectively enhance scam awareness (Kumarage et al., 2025). These limitations have become increasingly critical as social engineering scams grow more personalized, weakening the distinction between LLM-based methods and traditional approaches such as blacklists, phishing campaigns, or conventional automated detection. Therefore, it is essential to validate the cognitive capability of LLMs and develop a framework that dynamically strengthens user suspicion and systematically evaluates its effectiveness in real-world scam scenarios.

We introduce **SCRIPTMIND**, a framework designed to integrate LLM-based inference for effective social engineering scam detection with user-centered evaluation of acceptability. **SCRIPTMIND** supports users cognitively during real-time interactions with scammers by introducing a novel detection task that predicts the scammer's crime script. Through this process, it observes meaningful changes in the user's level of suspicion at each conversational stage and evaluates its effectiveness.

Core three components of **SCRIPTMIND** are: the **C**rime **S**cript **I**nference **T**ask (**CSIT**), which models reasoning over scam scripts; the **C**rime **S**cript–Aware **I**nference **D**ataset (**CSID**), designed to support efficient and secure fine-tuning of LLMs; and the **C**ognitive **S**imulation-based **E**valuation of Social Engineering **D**efense (**CSED**) for evaluating LLM-driven scam detection from a user acceptance perspective. To the best of our knowledge, this is the first approach that unifies scam detection with cognitive effectiveness evaluation.

The **CSIT** and **CSID** was designed to model how an LLM assists users through the cognitive shift from suspicion to conviction during scam interactions. Using crime script analysis, we extracted scammer behavior patterns and formulated a task enabling the model to infer and explain scam intent at each dialogue stage. From publicly available phone scam cases in Korea, we built 22,712 crime script prediction instances for LLM training, including a benign dataset that contains scenarios of legitimate police summons. We then verified the statistical significance of extracted patterns and evaluated the fine-tuning performance gains.

To evaluate user cognitive effects through **CSED**, we conducted a phone scam simulation experiment in which participants were instructed, *"The caller may be a real prosecutor or a scammer; listen carefully and decide."* Participants were sequentially presented with 40 structured utterances representing key criminal intents of a scammer, delivered in both audio and text. The experiment consisted of three conditions: no AI intervention, a single AI warning, and LLM-based stepwise explanatory assistance. Participants' suspicion levels at each major stage of the dialogue were measured using a Likert scale to assess how different forms of intervention influenced suspicion escalation.

Experimental results show that **SCRIPTMIND**'s finetuned sLLM achieved notable gains in scam detection, utterance prediction, and intent explanation over the baseline, outperforming commercial zero-shot models by about 13%. Moreover, leveraging next-utterance prediction to guide users' cognitive shift from suspicion to conviction, the **SCRIPTMIND** significantly raised participants' suspicion levels compared to single-warning and no-intervention conditions. These results demonstrate **SCRIPTMIND**'s effectiveness in enhancing user engagement and advancing LLM-based defenses toward practical crime prevention. Building on this, we present a UI/UX-oriented design for detection system(Appendix A). Contributions are:

1. **SCRIPTMIND**: We propose the first framework that unifies LLMs' cognitive assistance in elevating user suspicion during real social engineering scams, together with a simulation-based evaluation method for user acceptance.

2. We constructed a **Crime Script Inference Task** and 22,712 training instances for LLM.

3. We empirically analyze the effectiveness of fine-tuned smaller LLMs operating in resource-constrained environments.

4. Through a phone scam simulation, **SCRIPTMIND** demonstrated a significant increase in users' suspicion compared to control groups.

2 Related Work

2.1 Evolution of Social Engineering Scams

Online scam exploits human psychological vulnerabilities through social engineering to steal sensitive information (Wang et al., 2021). With advances in LLMs and generative AI, such scams

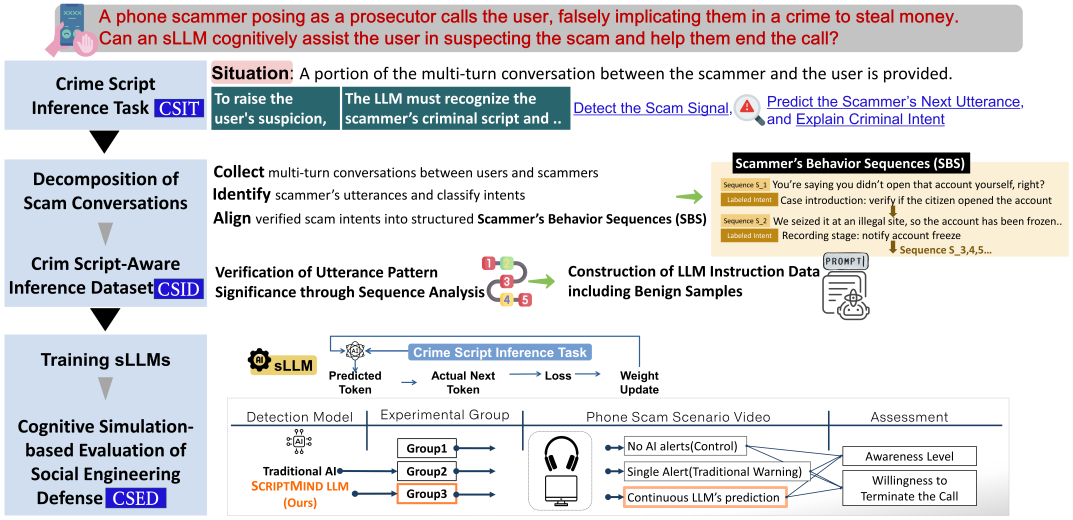

Figure 2: The uniqueness of **SCRIPTMIND** lies in modeling and evaluating tasks that elicit and reinforce users' suspicion throughout scam interactions. Unlike prior studies that evaluate models primarily based on accuracy(Koide et al., 2024; Lee et al., 2024; Shen et al., 2025), we design a framework that trains LLMs to predict scammers' next scripted actions and assess whether such predictions meaningfully enhance user suspicion in realistic scam contexts.

Category	Original Dataset	Initial Dataset	Scammer's Behavior Sequences (ours)	Crime Script-Aware Instruction Dataset (ours)
Purpose	Scam Conversations Collection	Identifying Scammer's Intention in Partial Conversation	Statistical Validation of Scammer's Utterance Pattern	Scam Prediction, Utterance Prediction, Rationale Explanation using LLMs
Instance Type	$D_o = \{C\}$	$D_I = \{(U_S, Y_{intent})\}$	$D_{\text{SBS}} = \{(U^{\text{scm}}, Y_{intent})\}$	$D_{\text{CSID}} = \{(U_S, Y_a, U^{\text{scm}}_{t+1}, Y_{intent})\}$
Scam Cases	571	571	571	571
Benign Conversations	–	–	–	11,356
All Data Instances	571	23,771	571	22,712

Table 1: Summary of Scam Scenario Data Used for the **C**rime **S**cript–Aware **I**nference **D**ataset (**CSID**) Construction.

have evolved into large-scale, organized, and multi-channel attacks (INTERPOL, 2024; AntiPhishing-WorkingGroup, 2025; FBI, 2023; Proofpoint, 2024; Abraham, 2024). Scammers now use SMS, calls, social media, and deepfakes in multi-turn conversations, impersonating acquaintances, recruiters, or officials to build trust and extract data (Tsinganos et al., 2018; Zheng et al., 2019; Reuters, 2023; Kumarage et al., 2025; Ai et al., 2024; Financial Supervisory Service, 2024). Global damages include a $600K deepfake scam in China, $1.3B in U.S. elder fraud, and ₩190B in South Korean voice phishing cases (Reuters, 2023; Financial Supervisory Service, 2024; FBI, 2023).

2.2 LLM-based Scam Detection

Language model–based scam detection serves as the core engine of modern anti-phishing systems (Cao et al., 2025; Koide et al., 2024, 2023; Lee and Han, 2024; Yu et al., 2024). Traditional classifiers lack explainability, whereas recent dialog-based frameworks enhance interpretability through scenario-driven detection grounded in social engineering contexts (Lee and Han, 2024; Koide et al., 2024; Lim et al., 2025). Such detection has also expanded into multimodal domains, including fake website analysis, brand impersonation detection, and real-time conversational alerts, demonstrating strong potential for AI-driven warnings. (Lee et al., 2024; Kulkarni et al., 2025; Shen et al., 2025).

Still, prior studies overlook user cognition and behavioral responses. This approach fails to address overconfidence and alert fatigue often seen in traditional defenses (Redmiles et al., 2016; Wang et al., 2016; Vishwanath et al., 2018; Merete Hagen et al., 2008; Wang and Song, 2021).

3 Method

Our core idea for preventing social engineering scams is to use LLMs to strengthen users' scam awareness and guide their decision to end the conversation. While detection accuracy is important, it is ultimately the user who decides to terminate the interaction, and scammers exploit psychological manipulation to stop them from doing so. We propose that enabling users to anticipate scammers' next actions, much like scammers anticipate users'

vulnerabilities, can turn suspicion into conviction. To realize this, as illustrated in Figure 2, the **CSIT** models scammers' behavioral sequences to train LLMs to predict their next moves. The fine-tuned models, based on the **CSID**, are then evaluated through scam simulations (**CSED**). This section details the construction of these three components.

3.1 Task Formulation

Problem 1 (Enhancing Users' Scam Awareness) *The purpose of the social engineering scam prevention system is to build a cognitive assistance–based detection service that helps users recognize and confirm fraudulent intent during real-time conversations without third-party intervention. Given a conversation context $\mathcal{C}$ within an unknown stage of a scam scenario $\mathcal{S}$, the model generates supportive inference that foster user suspicion and confidence to safely terminate the conversation.*

Task 1 (Crime Script Inference Task, CSIT) *Given an input* ***prompt*** $\mathbf{X}$ *containing a conversation $\mathcal{C}$ under a potential scam scenario $\mathcal{S}$, the model F jointly performs scam detection, next-utterance prediction, and intent inference to simulate cognitive reasoning in real-time scam interactions. The task is defined such that* $F(\mathbf{X}) = \{y_a, U_{t+1}, y_i\}$, *where* $\mathbf{X} = \{\mathcal{S}, \mathcal{C}\}$, $y_a \in [0, 1]$, $U_{t+1} \in \mathcal{U}$, *and* $y_i \in \mathcal{Y}_{intent}$.

3.2 Dataset Construction

3.2.1 Decomposition of Scam Conversations

Scam Conversations The original data were collected from the publicly available Law&Order Benchmark (PSI, 2025). The dataset was constructed based on voice phishing call records released by the Korean National Police Agency, comprising conversations in which scammers impersonate prosecutors to fabricate criminal cases and exert financial pressure on victims. Such impersonation and coercive tactics, using false legal threats, are typical scam strategies observed across multiple countries (FBI, 2023; INTERPOL, 2024). We utilized a total of 571 cases and 48,229 utterances included in the the dataset(see Appendix B).

Scammer's Behavior Sequences To implement the **CSIT** using the given scam conversations, we first separated the scammer's utterances from each dialogue and labeled those that explicitly conveyed fraudulent intent. Through this process, utterances that repeatedly appeared across confirmed scam cases were organized into structured data referred to as Scammer's Behavior Sequences(SBS). This sequence-based analysis is theoretically grounded in Crime Script Analysis, which models social engineering as sequential behavioral scripts (Cornish, 1994; Hutchings and Holt, 2015; Loggen and Leukfeldt, 2022; Choi et al., 2017; Lwin Tun and Birks, 2023), and the MITRE ATT&CK framework, which systematically categorizes phishing techniques (Strom et al., 2018; Shin et al., 2022; Abo El Rob et al., 2024). As shown in Table 1, the initial dataset contained 23,771 scammer utterances, each potentially associated with multiple intents. We segmented the dialogues and mapped verified intents to individual utterances, normalizing the data into a single-utterance–single-intent format (Appendix B.3, E.5).

3.2.2 Crime Script-Aware Inference Dataset

Statistical Validation of Sequences To verify that the classified utterance sequences reflected consistent criminal patterns, we performed statistical validation to distinguish scripted behaviors from improvised statements. Weak sequence associations, even with expert-labeled intents, can undermine the reliability of key behavioral patterns. To address this, we applied the Standardized Residual (SR) method from Behavior Sequence Analysis (Everitt and Skrondal, 2010). The SR score quantifies the normalized difference between the observed and expected frequencies of intent transitions across utterances. A higher SR indicates more consistent and scripted scammer behavior, aligning with criminological perspectives that interpret repetitive behaviors as indicators of intentional or patterned actions (Cornish, 1994). Details of the SR analysis are provided in Appendix C.

Crime Script-Aware Inference Dataset Using the Scammer's Behavior Sequences, we constructed the **C**rime **S**cript-aware **I**nference **D**ataset (**CSID**), designed to enable the LLM to perform tasks based on partial conversational(Appendix B.4). In addition, an equal number of benign instances were added under the scenario of "a legitimate police officer issuing a summons" (Table 7). For scam instances, original dialogues were segmented into input–output pairs, where only preceding utterances were provided as input. Consequently, it supports (1) scam detection, (2) utterance prediction, and (3) intent explanation.

3.3 Training Smaller Large Language Model

We trained an open-source *compact LLM (cLLM)* to support closed-network deployment and lightweight, privacy-preserving operation in restricted environments. We hypothesized that additional training on expert-labeled scammer interaction sequences is necessary to compensate for the model's limited domain-specific prior knowledge and to better align its reasoning with real-world scam scenarios.

We evaluated models at three capacity ranges: **1–2B**, **7–11B**, and **large commercial models**, including those specialized for Korean. The open-source cLLMs were fine-tuned for 5 epochs using the Paged AdamW optimizer (learning rate = $1e{-}4$). Parameter-efficient adaptation was performed using QLoRA (Dettmers et al., 2023), with low-rank adapters applied to the attention and feed-forward layers. Training was conducted on two NVIDIA A100 80 GB GPUs, requiring approximately 30 hours (Appendix D).

3.4 Cognitive Simulation-based Evaluation

Aim and Hypotheses We aim to examine whether **SCRIPTMIND** can serve as a cognitive assistant that supports users' real-time judgment during scam. We hypothesize that *real-time LLM warnings and explanations enhance users' suspicion levels.* To test this, a five-stage conversational script based on phone scam cases was designed, and participants reported their suspicion levels in real time while listening to both audio and text.

Experimental Stimuli and Procedure We used *prosecutor impersonation scam scenario* from the **CSID** as the experimental stimulus, and the structured scam script is provided in Appendix E.5. 98 participants were recruited and evenly assigned across conditions. We used repeated-measures ANOVA and t-tests to assess the impact of AI intervention. All procedures were IRB approved and details regarding the purpose and scope of the experiment (Appendix E.1), the design of experimental stimuli and procedures (Appendix E.2), the questionnaire items (Appendix E.3), the statistical analysis methods used for interpretation (Appendix E.4), and the ethical review and approval for human research (Appendix F.2) are all provided.

> **EXPERIMENTAL CONDITIONS**
> *Participants were told each call might be legitimate or fraudulent, prompting cognitive judgment. Conditions: (1)* ***a control group*** *with no alerts, (2)* ***a single-warning group*** *with one alert during the financial information stage, (3)* ***a SCRIPTMIND's LLM group*** *providing real-time predicted utterances at each scam stage.*

4 Experiment Results

Metrics Scam detection was evaluated using Accuracy, F1, FP, FN, while utterance prediction and intent inference were assessed via the LLM-as-a-Judge (Chiang and Lee, 2023; Lee et al., 2026). Strong correlation with expert evaluations confirmed the reliability of the automatic assessment. To validate the reliability of the LLM-as-a-Judge evaluation, we measured its agreement with human judgment on a randomly sampled subset of the test data. Specifically, two domain experts independently rated 200 instances each for the zeroshot and finetuned settings, following the same evaluation criteria as the LLM. Pearson correlation analysis ($p < 0.05$) showed strong alignment between human ratings and automatic scores, supporting the validity of the proposed evaluation protocol(Appendix G).

Fine-tuned Model Performance SCRIPTMIND consistently outperformed commercial models, demonstrating its effectiveness in enhancing scam awareness. As shown in Table 2, EEVE-Korean-10.8B achieved the best overall performance(Scam Detection 0.98, Next Utterance 0.68, Intent Inference 0.80), exceeding GPT-4o by 13%. On average, fine-tuned small models showed a 51% improvement over zero-shot (Table 16). It also reduced false positives and improved explanatory quality: commercial models averaged FP 0.24, while fine-tuning achieved 0.02.

Cognitive Effect of SCRIPTMIND As shown in Figure 3 and Table 3, participants' suspicion levels across the five-stage scam scenario demonstrated that **SCRIPTMIND** next-utterance prediction warnings achieved the strongest cognitive resilience. Since real scam calls could not be ethically replicated, participants were informed in advance of potential scam, resulting in high initial suspicion. To control for this, statistical analyses were conducted to isolate the true effects of LLM intervention(Appendix H). First, we found that suspi-

Model	Method	Scam Detection			Next Utterance	Intent Inference
		ACC	F1	FP/FN	LLM-as-a-Judge	LLM-as-a-Judge
Llama-3.2-1B-Instruct	ZS	0.55	0.56	0.24/0.21	0.04	0.17
	SCRIPTMIND-FT	0.91	0.92	0.06/0.02	0.39	0.57
EXAONE-3.5-2.4B-Instruct	ZS	0.58	0.65	0.31/0.11	0.30	0.51
	SCRIPTMIND-FT	**0.94**	**0.94**	0.06/**0.01**	**0.53**	0.73
Midm-2.0-Mini-Instruct	ZS	0.48	0.44	0.23/0.29	0.11	0.29
	SCRIPTMIND-FT	0.74	0.67	0.03/0.23	0.33	0.53
Llama-3.1-8B-Instruct	ZS	0.50	0.67	0.50/**0.00**	0.19	0.54
	SCRIPTMIND-FT	0.80	0.76	0.01/0.19	0.41	0.54
SOLAR-10.7B-Instruct	ZS	0.64	0.72	0.33/0.03	0.16	0.44
	SCRIPTMIND-FT	0.85	0.82	**0.00**/0.15	0.44	0.50
EEVE-Korean-Instruct-10.8B	ZS	0.71	0.74	0.21/0.09	0.42	0.52
	SCRIPTMIND-FT	**0.98**	**0.98**	**0.01**/**0.01**	**0.68**	**0.80**
EXAONE-3.0-7.8B-Instruct	ZS	0.64	0.72	0.32/0.04	0.26	0.63
Midm-2.0-Base-Instruct	ZS	0.55	0.6	0.28/0.17	0.22	0.57
	SCRIPTMIND-FT	0.73	0.64	**0.00**/0.26	0.33	0.5
chatgpt-4o-latest	ZS	0.90	0.91	0.1/**0.00**	0.45	**0.78**
gemini-2.0-flash	ZS	0.70	0.77	0.29/**0.00**	0.39	0.77
claude-3-5-haiku	ZS	0.67	0.75	0.33/**0.00**	0.31	0.73

Table 2: Evaluation results of LLMs on our tasks. **ZS** indicates zero-shot and **SCRIPTMIND-FT** refers to finetuning.

Stage	SCRIPTMIND	Single_Warning	Control	F	P
1.Introduction of a fake case	**5.63**±1.69	5.40±1.92	5.07±2.07	0.67	.512
2.Explanation of alleged criminal involvement	**5.60**±1.65	4.77±2.08	4.80±1.90	1.88	.159
3.Setup of a recorded investigation	**4.63**±1.99	3.87±2.21	4.13±1.96	1.07	.346
4.Request for financial information	**6.27**±1.60	6.00±1.49	5.23±1.72	3.36	**.039**
5.Notice of summons for investigation	**5.73**±2.05	5.63±1.69	4.43±2.25	3.88	**.024**

Table 3: Suspicion scores by stage and group. We conducted stage-wise ANOVA. The **SCRIPTMIND** showed significantly higher suspicion at 4~5 ($p = .039, .024$), indicating statistical significance at the $p < .05$ level. This supports our research hypothesis that *the LLM warning increases the level of suspicion more effectively than in other groups*. Detailed analysis of experimental results is presented in Appendix H.4.

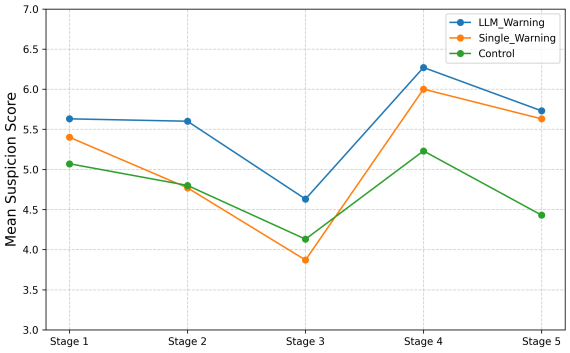

Figure 3: Changes in Suspicion Levels by Script Stage.

cion varied significantly across stages, whereas anxiety and relevance remained stable, confirming that suspicion can serve as a key indicator of cognitive resilience(Table 18). It declined through Stages 1~3 but rose sharply at Stage 4, where monetary information was requested. Second, no significant difference was found between the single-warning and control groups (Table 21), suggesting that one-time alerts yield only temporary awareness. However, a significant stage–group interaction (Table 22), together with the stage-wise one-way ANOVA results, showed that **SCRIPTMIND** maintained higher suspicion compared to all other groups, particularly at Stages 4~5 (Table 3). Overall, **SCRIPTMIND**'s real-time, context-aware warnings promote stronger and more lasting cognitive defense than static alerts.

5 Discussion

Qualitative Analysis Qualitative analyses revealed that fine-tuned models outperformed their zero-shot counterparts by accurately identifying detailed scam patterns that zero-shot models often misinterpreted or overlooked (Table 17). They also achieved lower false positive and false negative rates, with more coherent scammer utterance predictions and rationale explanations, highlighting their strong potential for real-world deployment in localized scam detection systems (Appendix G).

Design Direction for the Scam Detection SCRIPTMIND significantly heightened and sustained suspicion during scams. In particular, be-

havioral prediction of scammers effectively facilitated the transition from suspicion to conviction, demonstrating that LLMs can function as dynamic cognitive companions. Building on this insight, our proposed on-device LLM system continuously monitors scam dialogues, predicts deceptive intent across conversational stages, and provides adaptive notices to sustain user vigilance. A corresponding UI/UX prototype embodies this interactive flow, guiding the development of cognitively adaptive Scam Detection systems in the future(Appendix A).

Ethical Considerations LLM-based scam detection and on-device deployment entail inherent privacy and misuse risks. Scam datasets and models could be exploited by malicious attackers, so we implemented multiple safety measures to mitigate such threats. First, all experiments were conducted on de-identified and anonymized data derived from verified voice phishing cases, and no original audio or personally identifiable information was used. Second, to mitigate the risk of adversarial misuse, model training and inference were performed exclusively within a secure, closed police network, and the model itself was not released; only textual outputs were analyzed. Third, the system was designed as a human-in-the-loop decision-support tool rather than an autonomous surveillance mechanism. Finally, the human-subject study received institutional review board approval, and all procedures complied with applicable privacy regulations and emerging AI governance frameworks for high-impact public-sector AI systems(Appendix F.1).

6 Conclusion

We presented **SCRIPTMIND**, an integrated framework for crime script inference and cognitive evaluation in LLM-based social engineering scam detection. Unlike prior systems, it models the cognitive dynamics of user–AI interaction, connecting automated detection with human-centered defense. Experiments showed that fine-tuning improved accuracy, reduced false positives, and produced more interpretable explanations than baselines. Cognitive simulations revealed that **SCRIPTMIND** interventions strengthened users' suspicion, turning scam detection into an awareness-driven defense.

Despite these results, limitations remain. Emotional states could not be fully measured due to ethical constraints, and the study focused only on phone scams, excluding multimodal attacks. Future work will optimize on-device performance and expand beyond the Korean dataset. Overall, **SCRIPTMIND** advances cognitively adaptive LLMs for scam prevention, showing how crime script inference can enhance model reasoning and awareness against evolving social engineering threats.

Limitations

We validated a real-time, script-aware LLM model, identifying suspicion as a reliable cognitive marker. Despite uniformly high initial suspicion from ethical disclosure, a three-step validation confirmed the marker's validity, the null effect of single warnings, and significant LLM effects at Stages 4–5. However, high baseline suspicion constrained affective measures such as anxiety and trust. Future work should incorporate multimodal sensing to capture subtler emotional responses.

We further validated the model's predictive reasoning through crime script analysis and statistical evaluation, focusing on prosecutor-impersonation scams. Yet the findings remain limited to phone based cases. As social engineering evolves with deepfakes, voice cloning, and multimodal impersonation, future research should develop cross modal script for broader threat understanding.

We focused on enhancing users' cognitive resilience rather than optimizing inference speed or latency for deployment. Nonetheless, fine-tuning 1–2B-parameter models showed practical efficiency and clear trade-offs compared to larger ones, highlighting their potential for lightweight implementation.

The data used in this study was constructed based on Korean prosecutor impersonation phone scam cases. For broader applicability, future research should extend to cases from other countries.

Acknowledgment

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

Mustafa Farouk Abo El Rob, Mohammad Anwar Islam, Sriteja Gondi, and Oula Mansour. 2024. The application of mitre att&ck framework in mitigating

cybersecurity threats in the public sector. *Issues in Information Systems*, 25(3).

Jorij Abraham. 2024. Global state of scams report 2024. https://www.gasa.org. Accessed: 2025-11-05.

Lin Ai, Tharindu Sandaruwan Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael S Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, and 1 others. 2024. 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.

AntiPhishingWorkingGroup. 2025. Phishing activity trends report: 4th quarter 2024. Technical report, Anti Phishing Working Group. March 19, 2025.

Tri Cao, Chengyu Huang, Yuexin Li, Wang Huilin, Amy He, Nay Oo, and Bryan Hooi. 2025. Phishagent: a robust multimodal agent for phishing webpage detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27869–27877.

Cheng-Han Chiang and Hung-Yi Lee. 2023. 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.

Kwan Choi, Ju-lak Lee, and Yong-tae Chun. 2017. Voice phishing fraud and its modus operandi. *Security Journal*, 30:454–466.

Derek B Cornish. 1994. Crimes as scripts. In *Proceedings of the international seminar on environmental criminology and crime analysis*, volume 1, pages 30–45. Florida Department of Law Enforcement Tallahassee.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

B. S. Everitt and A. Skrondal. 2010. *The Cambridge Dictionary of Statistics*, 4th edition. Cambridge University Press, Cambridge.

FBI. 2023. Fbi internet crime complaint center. https://www.ic3.gov/.

Financial Supervisory Service. 2024. Voice phishing statistics. https://www.fss.or.kr/. Accessed: 2025-06-02.

Chihun Han, Beomsoo Kim, and Jaeyoung Park. 2024. Voice phishing scammers' psychological manipulation and consumer protection measures. *Journal of the Korea Institute of Information Security & Cryptology*, 34(5):1089–1100.

Alice Hutchings and Thomas J Holt. 2015. A crime script analysis of the online stolen data market. *British Journal of Criminology*, 55(3):596–614.

INTERPOL. 2024. Interpol financial fraud assessment: A global threat boosted by technology. https://www.interpol.int/en/News-and-Events/News/2024/INTERPOL-Financial-Fraud-assessment-A-global-threat-boosted-by-technology.

Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. 2023. Detecting phishing sites using chatgpt. *arXiv preprint arXiv:2306.05816*.

Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. 2024. Chatspamdetector: Leveraging large language models for effective phishing email detection. In *International Conference on Security and Privacy in Communication Systems*, pages 297–319. Springer.

Aditya Kulkarni, Vivek Balachandran, Dinil Mon Divakaran, and Tamal Das. 2025. From ml to llm: Evaluating the robustness of phishing web page detection models against adversarial attacks. *Digital Threats: Research and Practice*, 6(2):1–25.

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

Jehyun Lee, Peiyuan Lim, Bryan Hooi, and Dinil Mon Divakaran. 2024. Multimodal large language models for phishing webpage detection and identification. In *2024 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–13. IEEE.

Sangyub Lee, Heedou Kim, and Hyeoncheol Kim. 2026. Evaluating llms for police decision-making: A framework based on police action scenarios. *arXiv preprint arXiv:2601.03553*.

Yunseung Lee and Daehee Han. 2024. 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.

Bryan Lim, Roman Huerta, Alejandro Sotelo, Anthonie Quintela, and Priyanka Kumar. 2025. Explicate: Enhancing phishing detection through explainable ai and llm-powered interpretability. *arXiv preprint arXiv:2503.20796*.

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

Zeya Lwin Tun and Daniel Birks. 2023. Supporting crime script analyses of scams with natural language processing. *Crime Science*, 12(1):1.

Abbie Jean Marono, Sasha Reid, Enzo Yaksic, and David Adam Keatley. 2020. A behaviour sequence analysis of serial killers' lives: From childhood abuse to methods of murder. *Psychiatry, psychology and law*, 27(1):126–137.

Sean R Martin, Julia J Lee, and Bidhan Lalit Parmar. 2021. Social distance, trust and getting "hooked": A phishing expedition. *Organizational Behavior and Human Decision Processes*, 166:39–48.

Janne Merete Hagen, Eirik Albrechtsen, and Jan Hovden. 2008. Implementation and effectiveness of organizational information security measures. *Information Management & Computer Security*, 16(4):377–397.

Inc. Proofpoint. 2024. 2024 state of the phish: Risky actions, real-world threats and user resilience in an age of human-centric cybersecurity. Technical report, Proofpoint. Accessed: 2025-05-28.

Police Science Institute PSI. 2025. LAW&ORDER: A Benchmark Dataset for Structured Evaluation of LLMs in Policing. https://github.com/Heedou/policingai. Accessed: 2025-11-03.

Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. 2016. I think they're trying to tell me something: Advice sources and selection for digital security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 272–288. IEEE.

Reuters. 2023. 'deepfake' scam in china fans worries over ai-driven fraud. https://www.reuters.com/technology/deepfake-scam-china-fans-worries-over-ai-driven-fraud-2023-05-22/.

Zitong Shen, Sineng Yan, Youqian Zhang, Xiapu Luo, Grace Ngai, and Eugene Yujun Fu. 2025. " 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.

Youngsup Shin, Kyoungmin Kim, Jemin Justin Lee, and Kyungho Lee. 2022. Focusing on the weakest link: A similarity analysis on phishing campaigns based on the att&ck matrix. *Security and Communication Networks*, 2022(1):1699657.

Blake E Strom, Andy Applebaum, Doug P Miller, Kathryn C Nickels, Adam G Pennington, and Cody B Thomas. 2018. Mitre att&ck: Design and philosophy. In *Technical report*. The MITRE Corporation.

Nikolaos Tsinganos, Georgios Sakellariou, Panagiotis Fouliras, and Ioannis Mavridis. 2018. Towards an automated recognition system for chat-based social engineering attacks in enterprise environments. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, pages 1–10.

Arun Vishwanath, Brynne Harrison, and Yu Jie Ng. 2018. Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication research*, 45(8):1146–1166.

Jingguo Wang, Yuan Li, and H Raghav Rao. 2016. Overconfidence in phishing email detection. *Journal of the Association for Information Systems*, 17(11):1.

Mengli Wang and Lipeng Song. 2021. An incentive mechanism for reporting phishing e-mails based on the tripartite evolutionary game model. *Security and Communication Networks*, 2021(1):3394325.

Zuoguang Wang, Hongsong Zhu, and Limin Sun. 2021. Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods. *Ieee Access*, 9:11895–11910.

Seunguk Yu, Yejin Kwon, Minju Kim, and Kiseong Lee. 2024. Korean voice phishing detection applying ner with key tags and sentence-level n-gram. *IEEE Access*.

Kangfeng Zheng, Tong Wu, Xiujuan Wang, Bin Wu, and Chunhua Wu. 2019. A session and dialogue-based social engineering framework. *IEEE Access*, 7:67781–67794.

A UI and UX Design Result

Based on our experiment results demonstrating that **SCRIPTMIND** enhances users' cognitive awareness, we designed a system capable of efficiently detecting real-time social engineering scams that occur during phone calls in real-world device environments. This design concretizes the conceptual framework of **SCRIPTMIND** from a practical perspective, providing the foundational operational structure for future research and development of more advanced real-time scam detection systems.

Figure 4~6 illustrate the main operational flow of **SCRIPTMIND**, assuming that it runs as a smartphone application. As shown in Figure 4, **SCRIPTMIND** obtains the user's explicit consent to automatically transcribe phone conversations (speech-to-text) and employs a LLM to analyze and display the likelihood of fraud for each conversational segment. When suspicious activity is detected, the system immediately displays a warning message, enabling the user to compare the model's prediction with the actual dialogue and make an informed decision. Upon first launch or activation of the monitoring feature, a Consent modal appears to obtain user permission, and users can disable the function at any time via the settings menu, which instantly stops real-time analysis.

Next, as depicted in Figure 5, the LLM continuously analyzes the transcribed conversation in real time. When a scam is detected, **SCRIPTMIND** displays the identified scam type along with one of its core outputs, the predicted next utterance of the scammer. Since model prediction may experience slight latency compared to the actual conversation, **SCRIPTMIND** divides the dialogue into second-level segments so that users can scroll through and view the predictions for each segment. This design allows users, even after the call ends, to retrospectively recognize that "the model's prediction was indeed correct," thereby reinforcing their awareness and caution against scam tactics.

Finally, as shown in Figure 6, **SCRIPTMIND** continues to function in the background even when the application is inactive. The system delivers alert notifications based on the model's prediction results, allowing users to receive scam warnings in real time while communicating through speakerphone. It ensures continuous protection and real-time guidance without requiring active interaction.

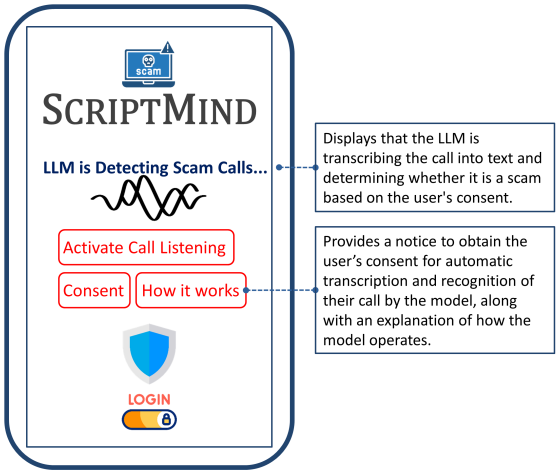

Figure 4: User Consent Interface and Real-Time Transcription Workflow of **SCRIPTMIND**

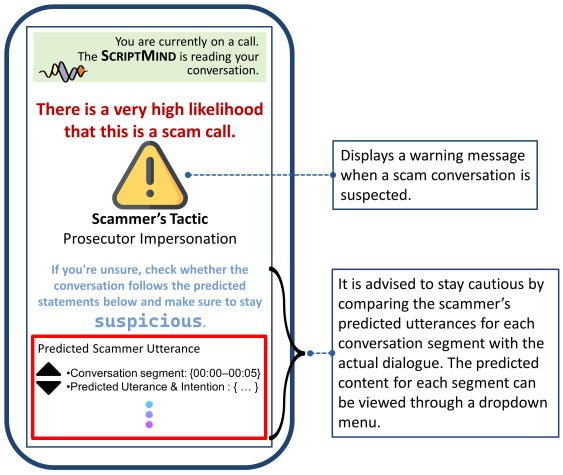

Figure 5: LLM-Based Scam Recognition and Next-Utterance Prediction Display.

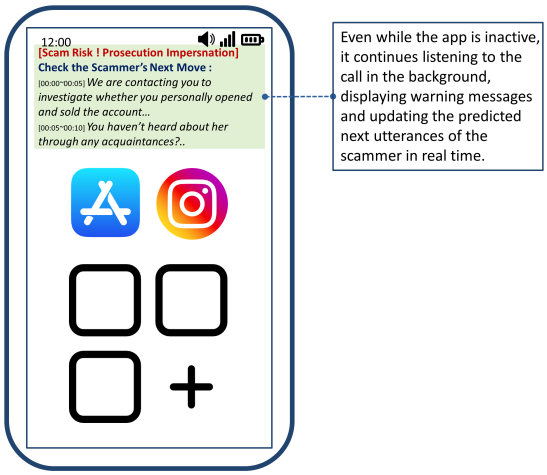

Figure 6: Background Monitoring and Real-Time Scam Alert Notification

B Scam Dataset

B.1 Original Dataset

We define the overall scam conversation dataset as the original dataset D_o, which contains conversation between users and scammers:

$$D_o = \{\mathcal{C}\}, \quad \mathcal{C} = \{C^{(1)}, C^{(2)}, \ldots, C^{(n)}\}$$
$$C^{(i)} = \{U_t^{\text{usr}}, U_t^{\text{scm}} \mid t = 1, 2, \ldots, T_i\}$$

Each case $C^{(i)}$ consists of sequential utterances exchanged between a user and a scammer, where U_t^{usr} and U_t^{scm} denote the user's and scammer's utterances at turn t, respectively, and T_i is the total number of turns.

Original Scam Conversation Example
This is the Seoul Central District Prosecutors' Office. My name is Investigator Yoo Seowon from the Advanced Crime Investigation Division. Yes, this is the Seoul Central District Prosecutors' Office. Yes, the reason I'm calling is regarding an identity theft case involving Mr./Ms. [Name]. Before I explain the details of the case, may I ask if you know a person named [Name]? No, it's a male from [Address], aged [Age]. Just a moment. Recently, we arrested a financial fraud ring called "[Name]," and during the operation, we seized numerous credit cards and illegal bank accounts.

Table 4: Example of scam conversation.

B.2 Initial Dataset

We define the Fraudulent Intent Interpretation (FII) dataset as an intent classification dataset derived from each conversation $C^{(i)}$ in the original set. It maps subsets of scam-related utterances to corresponding intent labels.

$$D_I = \{(U_S^{(i)}, Y_I^{(i)}) \mid C^{(i)} \in \mathcal{C}\}$$
$$U_S^{(i)} \subseteq C^{(i)}, \quad Y_I^{(i)} \subseteq \mathcal{Y}_{\text{intent}} = \{y_1, \ldots, y_m\}$$

Here, $U_S^{(i)}$ represents a subset of utterances within a conversation $C^{(i)}$, and $Y_I^{(i)}$ denotes the corresponding subset of intent labels drawn from the intent label set $\mathcal{Y}_{intent}$.

$$D_o \supset \mathcal{C} \supset C^{(i)} \supset U_S^{(i)}, \quad \mathcal{Y}_{\text{intent}} \supset Y_I^{(i)}$$

`"conversation"`	"Right, so you're saying that on May 23, 2016, at the [Address] branch— No, I didn't. (So you didn't personally open that account, correct?) Yes. So, to confirm again, you're saying that you didn't open the account yourself, correct? Yes. Since the account was seized during the illegal operation, we have frozen it to verify whether it was personally opened or fraudulently used."
`"intention"`	`["3. Case introduction - (7) Verify whether the bank account was personally opened by the citizen", "6. Investigation recording - (3) Notify that the account has been frozen"]`

Table 5: Example of an instance containing conversation and corresponding intentions from Law&Order Dataset.

B.3 Scammer's Behavior Sequences

We further define a subset of scammer utterances from $U_S^{(i)}$ that are explicitly annotated with intent labels, forming the Scammer-Based Subset (SBS) dataset.

$$U_{\text{scm,int}}^{(i)} = \{U_t^{\text{scm}} \in U_S^{(i)} \mid Y_I^{(i)}(U_t^{\text{scm}}) \neq \varnothing\}$$

This subset includes only the scammer utterances for which corresponding intent annotations exist.

$$D_{\text{scm}}^{(i)} = \Big(\{U_1^{\text{scm}}, U_2^{\text{scm}}, \ldots\}, \{Y_1, Y_2, \ldots\}\Big)$$

To represent these utterances and their intent labels as paired data, each conversation $C^{(i)}$ produces a local mapping between scammer utterances and corresponding intent categories.

$$D_{\text{SBS}} = \{(U_t^{\text{scm}}, Y_t) \mid U_t^{\text{scm}} \in U_S^{(i)},\ Y_t \in Y_I^{(i)}\}$$

or equivalently,

$$D_{\text{SBS}} = \{(U_t^{\text{scm}}, Y_t)\}$$

where each pair (U_t^{scm}, Y_t) corresponds to a scammer utterance and its associated intent.

Case ID	Speaker	Utterance	Scenario Classification	Intent Classification
001	Scammer	The account was issued at [Address], opened in June and used until November.	4. Case Involvement	Informing the citizen that objective evidence confirms their connection to the case.
001	Scammer	Were you not aware of this account?	3. Case Introduction	Checking whether the citizen knows about the account involved in the crime.
001	Scammer	The reason I contacted you is to confirm whether you personally opened and sold these two accounts to the "[Name]" group for payment, or whether, like other victims, your identity was stolen. Our preliminary investigation did not find any evidence suggesting you colluded with "[Name]", so we are contacting you under the assumption that your name was misused.	4. Case Involvement	Verifying whether the person actually sold the account or was a victim of identity theft.
001	Scammer	What's important now is determining whether you are a victim or an accomplice in this case. At the scene, we found two bank accounts under your name — from [Bank Name] — that were used in this crime, and there are victims who suffered financial loss through those accounts.	4. Case Involvement	Explaining that multiple people are involved in the crime, including both perpetrators and victims of identity theft.
002	Scammer	This is the Seoul Central District Prosecutors' Office.	2. Self-introduction	Stating the fake identity being impersonated.
002	Scammer	I am Investigator Yoo Seowon from the Advanced Crime Investigation Division.	2. Self-introduction	Stating the fake identity being impersonated.
002	Scammer	Yes, this is the Seoul Central District Prosecutors' Office.	2. Self-introduction	Stating the fake identity being impersonated.

Table 6: Examples of annotated scammer's behavior sequences with scenario and intent classification. All utterance categories are listed in Appendix E.5.

B.4 Crime Script-Aware Inference Dataset

The Crime Script Inference Dataset (CSID) is constructed as a collection of conversational instances designed to model both predictive and interpretive aspects of scam communication. Each instance contains a partial dialogue segment, its scam-related label, the predicted next scam utterance, and an explanatory intent description.

$$\mathcal{D}_{\text{CSID}} = \{(U_S^{(i)}, Y_a^{(i)}, U_{t+1}^{(i)}, Y_{\text{int}}^{(i)})\}_{i=1}^{N}$$

where $U_S^{(i)}$ denotes a partial conversation within a single case, $Y_a^{(i)} \in \{0, 1\}$ indicates whether the segment represents a scam (1) or non-scam (0), $U_{t+1}^{(i)}$ is the predicted next utterance of the scammer following $U_S^{(i)}$, and $Y_{\text{int}}^{(i)}$ is a natural language description explaining the underlying intent of the conversation segment.

$$f_{\text{CSID}} : U_S \mapsto (Y_a, U_{t+1}, Y_{\text{int}})$$

Here, the model f_{CSID} learns to infer the likelihood of a scam, predict the scammer's next utterance, and generate an intent-level explanation from a given dialogue segment.

Conversation	Label	Explanation
Do you have no knowledge about this at all? Alright, understood for now. Have you ever visited the [address] branch by any chance? This is the Seoul Central District Prosecutors' Office. Yes. You've never been there, correct? Yes. The account we discovered was opened around August 2015 at the [address] branch. (What exact date?) It shows that it was opened around August 2015 at the [address] branch. That's why I asked you about this earlier. In the past three years, have you ever lost any items such as your wallet or ID card that could lead to personal information leakage? According to our comparison with the relevant financial institution, this account is definitely registered under your name.	`scam`	**Next utterance:** "We are contacting you to investigate whether you personally opened the account and received payment for transferring it, or if you are a victim of identity theft." **Rationale:** The scammer is attempting to confirm whether the victim sold the account or was impersonated.
Hello, is this Ms. Hwang Ga-eun? Yes, that's me. Who is this? Hello, this is Sergeant Lee Cheol-soo from the Gangnam Police Station's Traffic Department. Do you have a moment to talk? Yes, you said Traffic Department? What is this about? A complaint has been filed regarding your violation of the Road Traffic Act and dangerous driving resulting in injury. You are required to undergo an investigation for a drunk driving case that occurred on February 28, 2015. Ah, that case... I see. I'm very sorry. When should I come to the police station? Please let me know when you are available. I'll schedule the investigation. How about this Friday morning? Friday morning works. Please come to our police station at 10 a.m. and ask for Sergeant Lee Cheol-soo at the Civil Affairs Office. Okay, I understand. I'll come at 10 a.m. Alright, see you then. Please make sure to bring your ID. Yes, I'll bring it. Thank you. Not at all. See you on Friday. Goodbye. Goodbye.	`non_scam`	**Rationale:** This is a legitimate call from a police officer. The officer provides identification, clear instructions, and no suspicious requests.

Table 7: An example from our completed Crime Script Inference Dataset (**CSID**)

Role	Instruction / Content
SYSTEM	You are an expert in detecting Korean phone scam conversations. Your output must strictly be a single JSON object. (No extra text outside the defined format.)
USER	{conversation} (Follow the rules below.) - If the conversation is phone scam, set `label:"scam"` and fill in `next_utterance` and `rationale`. - If it is not scam, return only `{"label":"non_scam"}`.

Example Output (for a scam case)	`{`
	`"label": "scam",`
	`"next_utterance": "Predicted next utterance of the scammer (1-2 sentences)",`
	`"rationale": "Current criminal intent: ... . Expected next criminal intent: ... . Evidence: ..."`
	`}`

Task	Analyze the given conversation and return the result as JSON. **OUTPUT MUST BE VALID JSON. NO EXTRA TEXT.**

Conversation Example	`"Are you saying you have no knowledge of this at all?`
	`Alright, I understand.`
	`Have you, by any chance, visited the [address] branch before?`
	`This is the Seoul Central District Prosecutor's Office.`
	`So, have you ever been to that branch?`
	`Yes. The bank account we found was opened around August 2015 at the [address] branch.`
	`That's why we're asking you.`
	`In the past three years, have you ever lost your wallet or any ID card that might have led to personal information leakage?`
	`After cross-checking with the financial institution, it is confirmed that the account was indeed opened under your name."`

Ground-truth Output	`"output": {`
	`"label": "scam",`
	`"next_utterance": "The scammer's next likely statement would be: 'We are contacting you to determine whether you personally opened and transferred this account for financial gain or if your identity has been stolen.'",`
	`"rationale": "The scammer currently aims to confirm the victim's personal information leakage and is expected to next assess whether the victim is an active participant or a victim of identity theft."`
	`}`

Table 8: Example of LLM instruction for Korean social engineering scam detection

C Behavior Sequence Analysis of Scammer's Utterances

Social engineering scams can be decomposed into step-by-step procedures through Crime Script Analysis (Hutchings and Holt, 2015; Loggen and Leukfeldt, 2022; Choi et al., 2017; Lwin Tun and Birks, 2023), or strategically mapped into structured tables based on adversarial tactic models such as MITRE ATT&CK (Shin et al., 2022; Abo El Rob et al., 2024). We apply these analytical techniques to the scammer's utterance sequences to identify where each utterance is positioned within the structured crime script of the scammer. Through this analysis, we capture both the psychological flow and tactical components of social engineering scams, providing a foundational basis for dataset labeling in training LLMs.

Core Assumptions. A key assumption in behavior sequence analysis of scammer's utterance is that typical tactics, such as those involving impersonation of prosecutors, follow a consistent pattern characterized by scripted dialogue structures, scenario progression, and intent-driven language. This assumption is supported by previous studies in crime script analysis, which have identified the step-by-step nature of social engineering scams (Choi et al., 2017; Lwin Tun and Birks, 2023).

Statistical Analysis. Based on the structure and labels of the dataset, we statistically extracted recurring patterns in scammer's utterances along with their associated intentions. To analyze the relationships between these intentions, we calculated the transition frequencies between utterances labeled with each intent and derived Standardized Residual (SR) values (Everitt and Skrondal, 2010).

SR is calculated as:

$$\text{SR} = \frac{\text{Residual}}{\text{Standard Deviation of Prediction Error}} \quad (1)$$

Where:

- **Residual** = Observed Value − Predicted Value
- **Standard Deviation of Prediction Error** reflects the uncertainty (variance) in the prediction for that specific observation.

Residuals with large absolute values are typically considered potential outliers (Everitt and Skrondal, 2010). In our study, we interpret such high SRs as indicative of repeated social engineering scam attempts following the same script, where an scammer consistently produces utterances that are more frequently observed than predicted in the dataset. This interpretation aligns with analytical approaches used in other domains, such as murder pattern analysis (Marono et al., 2020), where recurring behavioral patterns beyond statistical expectation are treated as significant indicators of intentional or scripted criminal activity.

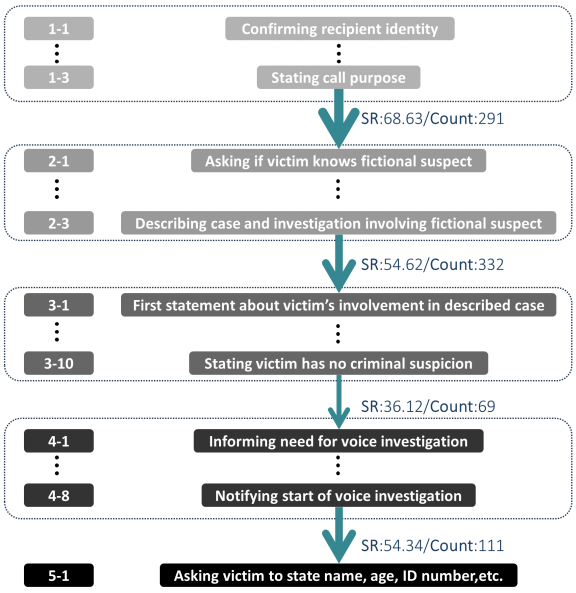

Figure 7: A transition network of utterance sequences observed within 571 phone scam conversation data points. The arrows represent the preceding and immediately following utterance sequences, respectively. Each node number corresponds to the classified intent of the scammer's utterance, as shown in the numbered utterance mapping in Appendix E.5. The SRs and frequencies of each sequence are listed in Table 9.

Analysis Results. The analysis revealed notable similarities in utterance patterns even across different scenarios, suggesting that language models trained on the structured stages of social engineering scam, rather than on the broader contextual content, may be more effective in detecting and explaining attacker behavior.

Labeling Validity. The dataset used in this study consists of 571 phone scam cases, each annotated with intent labels by two professional crime profilers. The inter-rater agreement, measured using Cohen's Kappa coefficient, reached 0.91, indicating a very high level of consistency and validating the reliability of the intent annotation process.

Structural Patterns of Scam Scripts. In the next phase, we analyzed the sequential order of utterances and intents across different cases to examine the structural regularities of scam scripts. To do so, we calculated the transition frequencies between

utterances labeled with specific intents and derived the Standardised Residuals (SR) scores. Figure 7 presents a network visualization of the top 28 utterance sequences with the highest SR values. The results revealed that utterance sequences with an SR score of 2 or higher appeared frequently and consistently across scenarios, indicating a high degree of structural organization in scam dialogues. As shown in Table 9, there are 28 transition sequences with an SR value of 20 or higher, with the highest SR value reaching 74.19. Additionally, a total of 251 sequences were identified as statistically significant transitions with SR values of 2 or above. The validity of this structure is supported by the observation that identical strategic utterance sequences appeared repeatedly across different cases with a frequency well beyond random chance. In other words, social engineering scams tend to follow a standardized script in which the same intents and strategies are executed in a fixed order.

No.	From ID	To ID	Count	SR
1	5-(1)	5-(2)	135	74.19
2	1-(3)	2-1	291	68.63
3	2-(1)	2-(2)	11	68.01
4	1-(2)	1-(3)	273	67.61
5	2-(1)	2-(2)	254	61.01
6	5-(2)	5-(3)	106	57.91
7	1-(1)	1-(2)	243	57.03
8	5-(6)	5-(8)	91	55.78
9	2-(3)	3-1	332	54.62
10	4-(8)	5-1	111	54.34
11	5-(5)	5-(6)	94	52.21
12	4-(6)	4-(7)	127	41.20
13	3-(1)	2-(6)	166	37.43
14	5-(8)	5-(9)	36	37.26
15	2-(2)	2-(3)	189	37.15
16	3-(10)	4-1	69	36.12
17	5-(3)	5-(5)	69	34.33
18	4-(7)	4-(8)	74	32.81
19	3-(2)	2-(7)	134	31.95
20	4-(1)	3-(11)	58	29.46
21	4-(2)	4-(6)	68	25.30
22	5-(4)	5-(5)	39	24.48
23	2-(4)	3-(10)	68	24.09
24	5-(5)	5-(4)	36	22.44
25	5-(9)	5-(8)	23	22.31
26	2-(6)	2-(7)	98	21.39
27	5-(4)	5-(6)	30	20.47
28	2-(3)	2-(4)	122	20.46

Table 9: Transition matrix of labeled utterances. Each node ID corresponds to the classified intent of the scammer's utterance, as shown in the numbered utterance mapping in Appendix E.5.

D Model Selection

To verify the effectiveness of our **SCRIPTMIND**-based social engineering scam detection system performing **CSIT**, we selected a diverse set of models as shown in Table 10. First, since real-world scam detection must operate in restricted environments such as users' on-device systems to ensure privacy protection, we included lightweight sLLMs (1~2B parameters). Next, we selected 7~11B parameter models to evaluate the feasibility of deploying them in secure intranet servers of public-sector organizations (e.g., police) using limited GPU resources. Finally, large-scale commercial models (e.g., GPT-4) with over 11B parameters were incorporated as baselines, allowing us to compare the latest high-performance reasoning capabilities. Since our **CSID** dataset is Korean-language based, we primarily focused on Korean-tuned models, while also evaluating multilingual models to examine their cross-lingual adaptability.

Scale	Category	Model	Deployment
1–2B	Multilingual sLLM	Llama-3.2-1B-Instruct	On-device phone
	Korean sLLM	Exaone-2B	
		MIDM-mini	
7–11B	Multilingual sLLM	Llama-3.1-8B-Instruct	Closed intranet server
		SOLAR-10.7B-Instruct	
	Korean sLLM	EEVE-Korean-Instruct-10.8B	
		Exaone-7B	
		MIDM-base	
>11B	Commercial LLM	chatgpt-4o-latest	No
		gemini-2.0-flash	
		Clade-10B-Instruct	

Table 10: Evaluation models categorized by parameter scale and deployment environment.

E Cognitive Evaluation(CSED) Settings

E.1 Research Question Formulation

The experimental design of our study is grounded in a key research question: ***Can real-time LLM-based scam detection serve as truly effective intervention tools?*** Unlike a single-point phishing, chat based online scam is a continuous and interactive process in which the victim's psychological state evolves over time (Martin et al., 2021; Kumarage et al., 2025). In particular, levels of suspicion are not static but tend to fluctuate depending on the phase of the scam (Han et al., 2024). These dynamic cognitive shifts raise important questions about the adequacy of traditional black-box detection models, which typically offer one-time warnings with limited contextual feedback. In response, there is increasing interest in LLMs that can deliver iterative and interpretable alerts throughout the interaction, aligning more closely with the user's changing cognitive state. Against this backdrop, our study designed and conducted a controlled cognitive experiment using a prosecutor impersonation voice phishing scenario to evaluate the effectiveness of real-time, LLM-based interventions.

Our phone scam scenario was selected as a representative case of a sophisticated scam, often executed through highly coordinated scripts by organized call center operations (Choi et al., 2017). It was chosen for two primary reasons. First, phone scam unfolds gradually through a sequence of conversations, making it particularly difficult to determine the optimal timing for detection-based intervention technologies (Yu et al., 2024; Choi et al., 2017). Second, because detection relies solely on the spoken content of the conversation, distinguishing between legitimate and fraudulent calls is inherently challenging, thereby increasing the risk of false positives (Shen et al., 2025). Once technical filters are bypassed, the final decision, such as whether to hang up the call, rests entirely with the user. These characteristics make this scenario especially suitable for evaluating the effectiveness of real-time LLM-based scam detection and for gaining insights into the central research questions.

We formulate the following research questions:

- **RQ1:** How do recipients' emotional and cognitive responses change over the course of a phone scam conversation?
- **RQ2:** Are current AI detection technologies effective in helping users recognize scam?
- **RQ3:** Does detection and explanation of scams using LLMs enhance user awareness more effectively than conventional single-warning detection models?

E.2 Experiment Design

❶ **Conditions.** To evaluate the effectiveness and user acceptance of an AI-based scam system, a three-group experimental study was designed using a simulated voice phishing call scenario grounded in a structured scam script. Prior to the main simulation experiment, we first analyzed a corpus of voice phishing data to develop the detection-oriented LLM described in the following sections. Details of the scenario analysis are provided in Appendix E.5. The scenario followed a five-stage crime script structure, with participants instructed to listen and respond to the stimuli in real time. These five stages were derived from a crime script analysis of actual voice phishing cases, and are summarized in Table 11. Depending on the assigned experimental condition, participants received varying types and intensities of AI generated alerts. This design aimed to examine how the presentation of AI predictions influences participants' perceptions and behavioral responses, such as their intention to terminate the call.

As shown in Table 12, Group 1 (control condition) received no AI-based alerts during the experiment. In contrast, two experimental groups were formed where AI model interventions were introduced. Group 2 was assigned the "single warning alert" condition, modeled after approaches developed in previous studies. Group 3 was assigned a newly developed condition in which an **Script-Aware LLM** continuously presented the "predicted next SE threat" as the attacker speech progressed.

❷ **Stimulus Methods.** Figure 8 provides an overview of the entire experimental procedure and the method of stimulus presentation. As shown in Table 11, all participants were exposed to the same five-stage voice phishing video stimulus. Each of the five stages consisted of 4 to 12 segmented utterances, with a total of 40 utterances presented as stimuli across all stages. The scenario was based on a commonly used psychological manipulation tactic in voice phishing, in which the perpetrator impersonates a prosecutor and falsely accuses the target of involvement in a criminal case.

Before the stimulus began, participants were given the following instruction:

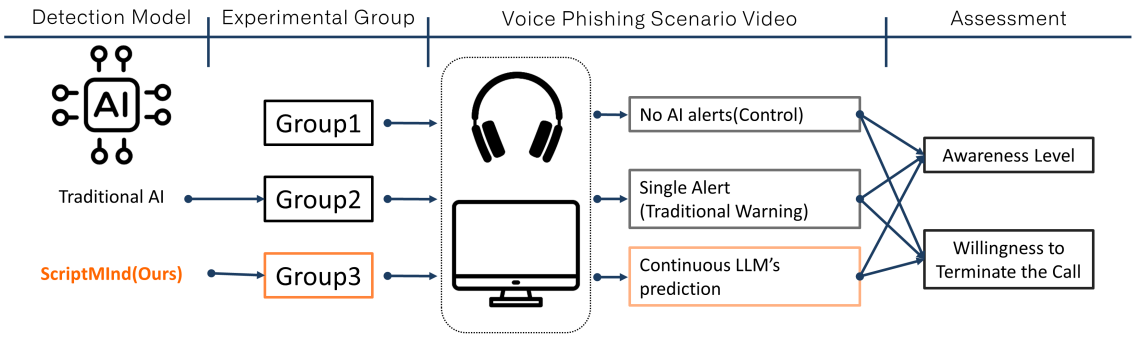

Figure 8: Description of the Psychological Experimental Stimuli Simulating a Phone Scam Scenario

Stage	Content Description and Example Dialogue
Stage 1 (Introduction)	The attacker introduces themselves and presents a fabricated case. *"This is Investigator Kim Young-jae from the Seoul Central District Prosecutors' Office."*
Stage 2 (Allegation)	The recipient is falsely implicated in a crime. *"We recently arrested a financial fraud ring including someone named [Name], and during the seizure, we found large amounts of cash, cloned credit cards, and bankbooks—including two accounts under your name from [Bank 1] and [Bank 2]."*
Stage 3 (Recorded Investigation)	The attacker explains that the call is being recorded for investigative purposes. *"Since your accounts were found at the crime scene, if you believe you're a victim, you'll need to formally prove it. We're currently conducting voice-recorded phone interviews for suspected victims."*
Stage 4 (Financial Information Request)	The attacker demands sensitive banking details. *"To freeze any unauthorized accounts and prevent further harm, please state which banks you currently use legitimately. For depositor protection registration, could you also confirm the current balance of your [Bank Name] account as of today?"*
Stage 5 (Legal Consequences Notification)	The recipient is threatened with potential legal repercussions. *"We'll send a subpoena to your home address. Please review it and visit our office once you receive it."*

Table 11: Phone Scam Stimulus Stages with Example Dialogues

Group	Condition	AI Intervention
Group 1	Control	No AI based alerts provided
Group 2	Single Warning	A single alert at the most suspicious stage (financial info request)
Group 3	SCRIPTMIND LLM warning	Continuous prediction of attacker next utterance

Table 12: Experimental Conditions

"The content you are about to see may be either a phone scam attempt or a legitimate notice from a public prosecutor's office."
This prompt served as a tool to elicit participants' judgments about the authenticity of each utterance in a realistic setting, allowing us to quantitatively measure their level of suspicion at each stage.

The experiment was conducted in the same conference room, where participants independently watched the video and completed the survey using tablet PCs and stereo headphones. All groups heard the same audio stimulus, while only the visual and auditory warning cues varied across conditions to ensure internal validity.

During the stimulus, the scammer's utterances were presented through both voice and on-screen text. After each stage, participants were asked(Appendix E.3):

1. Whether they believed the utterance was part of a scam call.

2. How suspicious or anxious they felt after being exposed to the content.

To simulate realistic decision-making under time pressure, participants were instructed to respond promptly, with limited time allocated for each response. Depending on the assigned condition, the presentation of visual and auditory alerts varied: the control group received no warnings; the Single

Warning group (Group 2) received a visual alert during Stage 4 (Financial Information Request); and the **SCRIPTMIND** LLM warning group (Group 3) was shown an AI-generated sentence predicting the scammer next utterance as a visual cue, displaying a warning message—"Warning!! This is a scam call"—along with the logo of the Korean National Police Agency, accompanied by an auditory alert tone. The presentation of real-time prediction outputs was constructed by selecting sentences deemed accurate from the predictions generated in real time by the developed LLM model, based on pre-constructed scam scripts.

Group	20–29	30–39	40–49	50–59	Total
Group 1	8	7	8	7	30
Group 2	7	8	7	8	30
Group 3	8	7	8	7	30
Total (Preliminary)	23 (+2)	22 (+2)	23 (+2)	22 (+2)	90 (+8)

Table 13: Number of Participants by Group and Age

❸ **Participant Recruitment.** As shown in Table 13, a total of 98 adults aged 20 to 59 were recruited, evenly distributed across four age groups by decade and assigned to three experimental conditions through stratified random sampling. While G*Power analysis suggested a minimum of 34 participants per group for sufficient power, this study ensured robustness through a bootstrap based ANOVA and repeated measures design.

To ensure the reliability of the study and prevent data contamination, participants recruited based on pre-defined criteria were automatically assigned to condition groups. After inputting age information, each participant was randomly assigned to a condition within the experimental platform.

1) **Inclusion Criteria**: Individuals who voluntarily consented to participate after receiving an explanation of the study's purpose.

2) **Exclusion Criteria**: Individuals who had participated in a survey or experiment within the past six months; those employed at financial institutions or law enforcement/judicial agencies; and those working in fields related to research such as marketing, market research, journalism, or broadcasting.

E.3 Evaluation Questions

At each utterance stage, participants were presented with the following identical set of questions.

Q1. *Who do you think this speaker is: an authority (e.g., investigator) or a scammer?*

- 1–3: Investigator
- 4: Not Sure
- 5–7: Scammer

Q2. *Emotional Evaluation – Please check the item that best describes your current feeling:*

- **Importance:**
  - 1–3: Not important at all
  - 4: Neutral
  - 5–7: Very important
- **Relevance:**
  - 1–3: Not relevant to me
  - 4: Neutral
  - 5–7: Highly relevant
- **Anxiety:**
  - 1–3: Not anxious at all
  - 4: Neutral
  - 5–7: Very anxious

E.4 Statistical Analysis

We conducted quantitative statistical analyses to examine differences in perception, emotional response, and behavioral intention during scam call scenarios, based on AI warning types and call stages. Statistical analyses were performed using the *JAMOVI* software, employing repeated measures ANOVA, one-way ANOVA, and independent samples *t*-tests as the primary analytical methods. The significance level was set at $\alpha = .05$.

- To examine how recipients' psychological responses as addressed in **RQ1**, including suspicion of fraud, anxiety, and perceived personal relevance, change over the course of a scam call, the five call stages were treated as repeated measures factors. A repeated measures ANOVA was conducted to analyze differences in psychological variables across stages, and Greenhouse-Geisser corrections were applied in cases where the assumption of sphericity was violated.
- To investigate the impact of the conventional AI detection method, namely a single warning message, on recipients' scam recognition

and their intention to terminate the call as addressed in **RQ2**, an independent samples *t*-test and one-way ANOVA were conducted to compare differences between the single warning condition (Group 2) and the control condition without any warning (Group 1).

- To evaluate the effectiveness of the LLM-based real-time utterance prediction model compared to traditional methods as addressed in **RQ3**, three experimental conditions (**SCRIPTMIND** LLM warning, single warning, and control group) were set as independent variables. A one-way ANOVA was conducted to assess their effects on scam recognition, intention to terminate the call, and attitudes toward AI intervention. Additionally, a mixed-design repeated measures ANOVA was performed to examine the interaction between AI alert condition and stage.

E.5 Scam Details in Cognitive Experiment

1. Identity Confirmation & Introduction

- **(1)** Confirming recipient identity
  - Ex) *Hello, is this [Name]?*
- **(2)** Stating impersonated identity
  - Ex) *Hello, this is Investigator Kim Youngjae from the Seoul Central District Prosecutors' Office. Is now a good time to talk?*
- **(3)** Stating call purpose
  - Ex) *I'm contacting you regarding a few confirmations about your personal data breach.*

2. Case Introduction

- **(1)** Asking if victim knows fictional suspect
  - Ex) *Do you happen to know someone named Kim Sang-sik from Ilsan, Gyeonggi Province?*
- **(2)** Asking about suspect's address, age, etc.
  - Ex) *He's a former civil servant, a 47-year-old man. Have you ever heard about him through acquaintances?*
- **(3)** Describing case and investigation involving fictional suspect
  - Ex) *We have arrested a financial fraud syndicate led by Kim Sang-sik.*
- **(4)** Forming suspicion about account used in crime
  - Ex) *During the seizure, multiple bankbooks and IDs under your name (from Kookmin Bank and Shinhan Bank) were found. Are you aware of these accounts?*
- **(5)** Disclosing bankbook purchase through testimony
  - Ex) *According to the statement made by [Name], when they purchased the bank account, they primarily used internet banking, transferring money into the account in their own name before completing the purchase.*
- **(6)** Asking if victim knows the account
  - Ex) *This account was used in a crime that resulted in a victim. Are you aware of these accounts?*
- **(7)** Confirming if victim opened the ghost account
  - Ex) *Did you, then, open [Bank Name] and [Bank Name] accounts under your name around January 27, 2016, through [Address]?*

3. Case Involvement

- **(1)** Statement about victim's involvement in the case
  - Ex) *At the scene of the arrest, a large amount of cash, cloned credit cards, and bank accounts under borrowed names were seized. Among these items, bank accounts from [Bank Name] and [Bank Name] registered under your name were identified.*
- **(2)** Objectively stating victim's link to case
  - Ex) *When we checked the issuance date of those accounts, it showed July 14, 2022, from Yeongdeungpo branch.*
- **(3)** Confirming identity theft
  - Ex) *Have you ever received any message or contact about your personal data being leaked to financial firms or shopping malls?*

- **(4)** Confirming whether it was theft or actual sale
  - Ex) *We contacted you to verify the misuse of bankbooks opened under your name.*
- **(5)** Asking if victim sold or transferred account
  - Ex) *Have you ever transferred your bank account to another person?*
- **(6)** Stating need for proof of victimization
  - Ex) *Just because your name is on the account doesn't mean we see you as the criminal. We do consider you a possible victim, but we need proof.*
- **(7)** Warning that sale/transfer leads to punishment
  - Ex) *If you did transfer your bank account, you may be subject to punishment under Article 10, Section 90 of the Act on the Punishment of Transfer of Personal Financial Information.*
- **(8)** Pressuring that account was created by victim
  - Ex) *When we checked the issuance date of those accounts, it showed July 14, 2022, from Yeongdeungpo branch.*
- **(9)** Explaining many involved, including victims
  - Ex) *This case currently involves approximately 180 individuals nationwide. Among them are people who either opened bank accounts and sold them or were victims whose identities were stolen.*
- **(10)** Stating victim has no criminal suspicion
  - Ex) *Based on our investigation, you have no criminal history and verified identity, so we are contacting you in advance.*
- **(11)** Informing of investigation via voice testimony
  - Ex) *We're here to assist your statement as a victim.*
- **(12)** Notifying that proof of victimization is required
  - Ex) *Since both of your bank accounts were found at the crime scene, if you believe you are a victim, it is essential that you provide proof to establish your status as a victim.*
- **(13)** If proven victim, informing of compensation
  - Ex) *If you are able to prove that you are a victim and it is confirmed that these individuals withdrew money from your account, the state can provide compensation for the loss.*
- **(14)** Notifying that prosecution is investigating
  - Ex) *We're not contacting you from a local police station or an insurance company, right? You understand where we're calling from, correct? This is an official investigation by a government agency—the Seoul Central District Prosecutors' Office.*
- **(15)** Warning of severe penalty for false statements
  - Ex) *The entire investigation process is being recorded, so if you are aware of any details regarding this case but provide false statements or attempt to conceal information, you may be subject to more severe legal penalties.*

4. Preparation for Voice Investigation

- **(1)** Informing need for voice investigation
  - Ex) *Since there's no direct suspicion against you, we'll proceed with a simplified voice-recorded investigation.*
- **(2)** Inducing agreement to participate
  - Ex) *For now, we will only record the parts that you are aware of as evidence. Do you agree to the recording?*
- **(3)** Prohibiting victim from revealing investigation
  - Ex) *And since you, [Name], are currently in the position of an interviewee under investigation, you do not have the right to disclose or discuss any details related to this case until your status as a victim has been verified. Understood?*

- **(4)** Telling victim their accounts will be tracked
  - Ex) *We are currently conducting a joint investigation with [Agency Name], and we will be performing account tracing under your name. If the accounts with [Bank Name] and [Bank Name] were opened without your knowledge, there is a possibility that other undiscovered accounts may exist as well.*
- **(5)** Telling victim to note impersonated info
  - Ex) *First of all, are you able to take notes? Since I'm the investigator in charge of your case, let me go over my affiliation and name again. Please get ready to write it down.*
- **(6)** Notifying voice record will be submitted to court
  - Ex) *This will be submitted to court, so if there are background noises or third-party voices, it won't be accepted as evidence. Please take caution.*
- **(7)** Setting environment for call
  - Ex) *Are you currently at home or at work? I asked because third-party voices can invalidate the recording.*
- **(8)** Notifying start of voice investigation
  - Ex) *Alright, we'll start the recording.*

5. Voice Investigation

- **(1)** Asking victim to state name, ID, etc.
  - Ex) *Hello, I'm Investigator Kim Young-jae. Please state your name and age for the record.*
- **(2)** Re-checking knowledge of people/accounts/crime links
  - Ex) *Do you know Kim Sung-sik, a 47-year-old man residing in Ilsan, Gyeonggi Province?*
- **(3)** Notifying account freezing
  - Ex) *These two accounts were frozen to prevent further damage. Do you know what freezing means?*
- **(4)** Notifying further freezing if other banks found
  - Ex) *Are you aware that any newly found accounts may result in penalties and freezing?*
- **(5)** Confirming normal banks used
  - Ex) *If you use other bank accounts beyond those you've declared, please state only the name of the legitimate bank to prevent them from being frozen as illegal.*
- **(6)** Confirming number/purpose of bank accounts
  - Ex) *At [Bank Name], how many accounts and for what purposes would you normally have under your name? You don't have any accounts related to savings plans, housing subscriptions, funds, stocks, or cryptocurrency, correct?*
- **(7)** Checking cash held in victim's account
  - Ex) *By "freezing," we mean you can't use the account at all. Do you understand?*
- **(8)** Confirming balance held in account
  - Ex) *For depositor protection registration, we need to verify today's balance at [Bank Name].*
- **(9)** Warning about punishment if amount differs
  - Ex) *If the reported balance differs by over 1 million KRW, the account will be considered suspicious, frozen, and may lead to an arrest warrant.*
- **(10)** Informing of next steps after call
  - Ex) *Thank you for your cooperation. The first hearing will be held next Wednesday.*
- **(11)** Threatening in-person summon if victim refuses to participate
  - Ex) *First, we will send a summons to your residence. Once you receive it, please appear in person at our office as instructed.*

F Ethical Considerations

F.1 Dataset and Model

The data used in this study was derived from the FRAUDULENT SCENARIO COMPLETION task, one of the benchmark tasks included in the officially released Law&Order dataset, made available via github by a policy researcher affiliated with the Korean National Police Agency (PSI, 2025). The original source data for this task consists of 571 voice phishing cases recorded between 2015 and 2019, all of which included verified voice recordings. Each case was transcribed using speech-to-text processing, and personally identifiable information, such as names, bank account numbers, and addresses, was either removed or anonymized to ensure that the dataset contained no personal data. For the purposes of this study, the original audio files were not used. Instead, the experiments were conducted using synthetic audio recordings, generated by professional actors reading the transcriptions.

The sLLM model used in the experiment was developed to predict the intent behind the scammer's utterances, utilizing a generative language modeling approach. Given the potential risk of adversarial misuse, such as the model being exploited to generate scam content, both model training and inference were conducted exclusively within a closed, internal network at the Korean National Police Agency, which originally provided the dataset. The model itself was not released as open source; only the model outputs, in the form of text, were used in the experimental setting.

F.2 Human Experiment

This research was reviewed and approved by the Central Institutional Review Board of Korea (IRB No. P01-202408-01-045), and written informed consent was obtained from all participants. All responses were collected anonymously, no personal data was stored or retained, and the entire experimental process adhered to ethical standards for research involving human subjects. Participant recruitment was conducted between August 26 and September 6, 2024. Eligible participants were prescreened based on predefined inclusion criteria and were contacted via SMS with a URL containing an information sheet and consent form detailing the study purpose, procedures, potential risks and benefits, data protection measures, and voluntary participation policy. Participants who consented were randomly assigned to one of three conditions and were provided with a tablet and headphones to complete the task. Each participant viewed a simulated voice phishing scenario video and responded to stage-specific survey items over the course of approximately 30 to 40 minutes. All response data were collected via a secure electronic survey system in real time.

F.3 Privacy Concerns of On-Device LLMs for Scam Conversation Analysis

F.3.1 Misuse as a Surveillance Tool

Although scam conversation datasets can help build models that detect and prevent fraud, there is a potential concern that such data might be misused for automated surveillance. However, this concern is mitigated by multiple safeguards. Law enforcement agencies, telecom companies, and social media platforms are all bound by existing privacy and communication protection laws that strictly prohibit unauthorized monitoring of citizens' private communications. Moreover, these organizations already possess their own user data and cannot legally repurpose it for surveillance without consent or judicial oversight.

F.3.2 Operation in Secure, Closed Environments

Our dataset and models are developed and operated within the closed police network, ensuring that no data is used for targeting individuals. The system functions solely as an internal decision-support tool for investigators, with all judgments ultimately made by human officers. All preprocessing and experiments occur within this restricted environment and are not accessible to external parties.

F.3.3 Legal and Regulatory Safeguards

Countries such as South Korea, the UK, and members of the EU have established or proposed AI governance frameworks, for example, Korea's Artificial Intelligence Basic Act, that classify police-developed AI systems as "high-impact AI." Such systems are legally required to undergo committee review, data management oversight, user protection planning, and risk mitigation. These international regulatory trends collectively ensure that on-device LLMs analyzing scam conversations cannot be used for broad or invasive surveillance.

G Finetuning Results Analysis

G.1 Correlation between LLM-as-a-Judge and Human Evaluation

Pair	Correlation
LLM-finetuning vs Human1-finetuning	0.850^{***}
LLM-zeroshot vs Human1-zeroshot	0.830^{***}
Human1-zeroshot vs Human2-zeroshot	0.826^{***}
Human1-finetuning vs Human2-finetuning	0.819^{***}
LLM-finetuning vs Human2-finetuning	0.810^{***}
LLM-zeroshot vs Human2-zeroshot	0.782^{***}

Table 14: Results of Correlation Analysis on LLM-as-a-Judge and Human Evaluation. Pearson correlation coefficients were calculated. Statistical significance was set at $p < 0.05$. $^{***}p < 0.001$.

We evaluated the quality of LLM responses for **CSIT** using the LLM-as-a-Judge framework. To validate its reliability, we analyzed the correlation between two human experts' ratings and the automatic scores on 200 randomly sampled instances from the test set. The analysis employed the Pearson correlation coefficient, with statistical significance set at $p < 0.05$. We conducted independent analyses for 200 zeroshot and 200 finetuning evaluation instances, following the same procedure. Both human experts and the LLM were instructed to assess responses solely based on the given golden answers, with the LLM providing decimal scores between 0 and 1 and humans using a 7-point Likert scale. The evaluation prompt is presented in Table 15.

As shown in Table 14, the results indicate a strong correlation between human and LLM-based evaluations, consistently observed across both zeroshot and finetuning settings. The correlations were statistically significant, suggesting that automated evaluation by LLMs can serve as a valid alternative for assessing other models.

G.2 Effect of sLLM Fine-tuning

We conducted a detailed comparison between the zero-shot and fine-tuned performances of each sLLM to assess the impact of SCRIPTMIND fine-tuning. As shown in Table 16, all seven models demonstrated improvements across scam detection accuracy, next-utterance quality, and rationale generation. On average, the Accuracy and F1 score of scam detection increased by 0.28 and 0.19, respectively, while the False Positive Rate decreased by approximately 0.28, indicating a notable reduction in misclassification. Although the False Negative Rate varied by model, with some showing slight increases(e.g. Llama 3.1 8B and SOLAR 10.7B), this trend may reflect a more conservative classification tendency when models jointly learned benign data, causing them to label borderline scam instances as non-scam. In contrast, both Next Utterance and Rationale scores improved substantially (0.24, 0.16), suggesting that SCRIPTMIND finetuning enhanced contextual understanding and explanatory quality in scam-related dialogue modeling.

In addition, we qualitatively analyzed the improvement in scam detection performance achieved through SCRIPTMIND fine-tuning by examining actual prediction cases.

(1) Enhanced Understanding of Scam Scenarios As shown in the first row of Table 17, the EEVE model fine-tuned with SCRIPTMIND demonstrates a clear understanding of a typical scam scenario in which the scammer falsely claims that *"the user's bank account is linked to a criminal case."* Consequently, in other similar cases where the scammer states that *"it is necessary to verify whether the user opened the account themselves or is a victim of identity theft,"* the fine-tuned model accurately predicts such repetitive and characteristic scam utterances and provides detailed explanations of their deceptive intent. In contrast, the zero-shot model merely repeats the scammer's words or offers only a superficial description such as *"the scammer is trying to steal personal information."*

(2) Reduction of False Negatives As illustrated in the second row of Table 17, the SCRIPTMIND fine-tuned EEVE model successfully identifies deceptive intent even in conversations that appear ordinary at first glance. For instance, an utterance like *"Do you know Mr. XX?"* may sound like a casual question, but in a *prosecutor impersonation scam scenario*, it serves as a classic tactic to gain trust by referring to a fictional criminal figure. The fine-tuned model accurately recognized this contextual cue and classified the dialogue as a scam, whereas the zero-shot model misclassified it as a normal conversation. This finding highlights the importance of enabling early-stage scam detection to prevent further interaction and potential victimization, emphasizing the necessity of learning subtle contextual cues underlying scam communication.

(3) Reduction of False Positives While minimizing false negatives is important, reducing false positives is an even more critical challenge in scam

Instruction / Content
You are an expert evaluator for phone scam scenario predictions. Your task is to compare the model's predicted next utterance with the correct ground truth utterance. Rate the prediction **STRICTLY** based on whether it conveys the *same phishing situation* or meaning as the ground truth, not merely based on text similarity. Give a score between `0.00` and `1.00` (two decimal places): - `1.00` means the prediction fully matches the meaning and intent of the ground truth (same phishing situation described). - `0.00` means the prediction is completely different or unrelated. - Intermediate values (e.g., `0.45`, `0.72`) represent partial semantic overlap or situational similarity. Output **only the numeric score**, no explanation.

Table 15: LLM instruction example for automated evaluation of scam scenario predictions

Model	Scam Detection			Next Utterance	Rationale
	ACC	F1	FP / FN	LLM-as-a-Judge	LLM-as-a-Judge
`Llama-3.2-1B-Instruct`	0.36	0.36	-0.18 / -0.19	0.35	0.40
`EXAONE-3.5-2.4B-Instruct`	0.36	0.29	-0.25 / -0.10	0.23	0.22
`Midm-2.0-Mini-Instruct`	0.26	0.23	-0.20 / -0.06	0.22	0.24
`Llama-3.1-8B-Instruct`	0.30	0.09	-0.49 / 0.19	0.22	0.00
`SOLAR-10.7B-Instruct`	0.21	0.10	-0.33 / 0.12	0.28	0.06
`EEVE-Korean-Instruct-10.8B`	0.27	0.24	-0.20 / -0.08	0.26	0.28
`Midm-2.0-Base-Instruct`	0.18	0.04	-0.28 / 0.09	0.11	-0.07

Table 16: Performance improvement of each model before and after fine-tuning. Each value represents the performance gain of the fine-tuned model compared to its zero-shot baseline.

detection. If legitimate communications from law enforcement, such as requests for appearance issued by the police or prosecution, are mistakenly classified as scams, the reliability of the detection system would be severely compromised, potentially disrupting legitimate criminal justice operations. As presented in the third row of Table 17, the **SCRIPTMIND** fine-tuned EEVE model correctly classifies such lawful law-enforcement scenarios as non-scam, whereas the zero-shot model erroneously labels them as scam. This demonstrates the necessity of incorporating both scam and benign data in a balanced fine-tuning process to enable the model to distinguish subtle linguistic and contextual differences between fraudulent and legitimate communications. Through this approach, the model can precisely capture the characteristics of deceptive speech while maintaining high accuracy in recognizing genuine public communication.

Input Conversation	Golden Label	Golden Answer (Next Utterance)	Golden Answer (Rationale)	Finetuned Ans (Next Utterance)	Finetuned Ans (Rationale)	Zeroshot Ans (Next Utterance)	Zeroshot Ans (Rationale)
"Do you have no knowledge about this at all? Alright, understood. Have you ever visited the [address] branch? This is the Seoul Central District Prosecutors' Office. Have you ever been there? The account discovered was opened around August 2015 at the [address] branch. That's why I asked you earlier. In the past three years, have you ever lost any items like your wallet or ID that could lead to personal data leakage? According to our comparison with the relevant financial institution, the account is definitely registered under your name."	scam	"The scammer's next expected utterance: We are contacting you to investigate whether you personally opened and sold the account for payment or if you were a victim of identity theft. "	"The scammer currently aims to verify whether the victim's identity was stolen and will next ask whether the victim sold the account or was impersonated. "	"The scammer's next expected utterance: The reason we are contacting you today is to determine whether you personally opened the accounts and sold them for profit, or whether your identity was stolen. "	"The scammer aims to inform the citizen about the case objectively and will next ask whether they sold or were impersonated. "	"The scammer's next expected utterance: Have you ever visited the [address] branch?"	"Current scammer intent: to obtain the user's financial information. Next expected intent: to collect ID or banking information, as the scammer is inquiring about the user's address and transactions."
"There was an issue, so we're contacting you. We have a few questions about this case—are you available for a call now? Do you know a woman named [Name], born in [Address], aged [Age]?"	scam	"The scammer's next expected utterance: So, you haven't heard about her through any acquaintances or groups? We are asking because the prosecution recently arrested a financial fraud ring led by [Name] , during which we seized multiple bank accounts"	"The scammer currently seeks to check if the victim recognizes a fake suspect name and will next explain the fabricated case and the citizen's alleged involvement."	"The scammer's next expected utterance: You don't know this person? We recently arrested a financial crime ring led by [Name]. "	"The scammer aims to confirm whether the user knows the fake suspect and will next describe the crime and investigation details."	*Misclassified as non-scam.*	
"Is this Mr. Yoon Youngsik? Yes, who is this? Hello, this is Sergeant Lee Cheol-soo from the Gimhae Police Department's Traffic Investigation Team. Do you have a moment to talk? Yes, what is this about? We received a report of drunk driving on August 25, 2018. We need to investigate this matter. When can you come to the station? This Friday at 3 p.m. would work. Please bring your ID when you visit."	non-scam			*"Correctly classified as non-scam."*		*Misclassified as scam.* Incorrectly predicted next utterance : "The officer informs the citizen of a drunk-driving investigation and reminds them to bring their ID.", Incorrectly predicted intent : to deceive Mr. Yoon under the guise of a police officer to obtain personal ID information and to use the ID and personal data for identity theft.	

Table 17: Examples of input conversations, ground-truth (golden) answers, and model predictions (finetuned vs. zero-shot). The examples illustrate model behavior in next utterance generation and rationale explanation for both scam and non-scam dialogues.

H Cognitive Evaluation Results Analysis

H.1 Summary of Experiment Result Analysis

Because real scam call experiments are ethically infeasible, participants in our simulation were informed that the conversation might be fraudulent, which naturally elevated their initial suspicion levels. To mitigate this limitation and ensure the validity of the findings, we applied a rigorous multi-step analytical framework to capture the genuine cognitive effects of LLM interventions.

First, we statistically examined whether participants' suspicion varied meaningfully across scam stages. Alongside suspicion, we also tracked anxiety and perceived relevance to validate suspicion as a sensitive cognitive indicator. Results revealed that while anxiety and relevance showed no significant stage-wise differences, suspicion decreased across Stages 1–3 and sharply increased at Stage 4. A repeated measures ANOVA confirmed that these differences were statistically significant (Appendix H.2), verifying that suspicion functions as a more dynamic and diagnostic psychological marker for scam detection.

Next, to assess the impact of a traditional single warning, we compared mean perceived suspicion scores across the entire script, as single warnings are not tied to specific conversational stages. Although the single-warning group reported slightly higher suspicion than the control group, the difference was not statistically significant (Appendix H.3). This suggests that a one-time alert may momentarily raise awareness but cannot sustain cognitive resistance throughout a strategically structured social engineering scam.

Finally, we analyzed the stage-wise effects of the LLM intervention. A significant interaction between stage and group (Table 22) indicated that suspicion patterns were not driven merely by call progression but by the type of warning received. Stage-wise ANOVAs further showed that the LLM Warning group exhibited the highest suspicion at Stages 4 and 5, with statistically significant between-group differences (Table 23). These findings indicate that LLM interventions effectively sustain and amplify suspicion, especially during critical stages involving pressure or financial solicitation, whereas the control and single-warning groups showed slower or incomplete recovery in awareness.

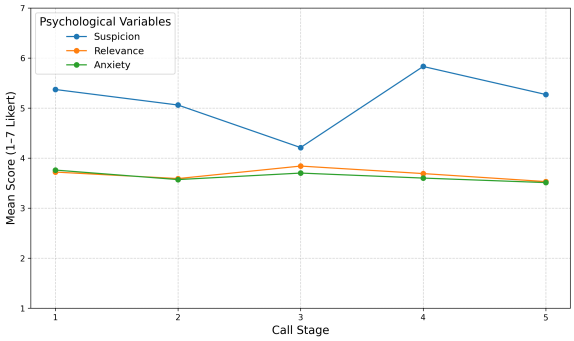

Figure 9: Mean Trends of Psychological Reactions

H.2 RQ1: suspicion evolves, but emotion persists throughout scam stages

To address **RQ1**, we examined how recipients' psychological reactions, specifically suspicion, perceived relevance, and anxiety change. The results show that suspicion levels temporarily decreased and reached their lowest point at Stage 3, followed by a sharp increase at Stage 4. In contrast, perceived relevance and anxiety remained relatively stable across stages. The means and standard deviations for each psychological variable across the five stages are presented in Table 18, and these patterns are also visualized in Figure 9.

Stage	Suspicion (M ± SD)	Relevance	Anxiety
Stage 1	5.37 ± 1.89	3.72 ± 1.87	3.76 ± 1.93
Stage 2	5.06 ± 1.90	3.59 ± 1.89	3.57 ± 1.91
Stage 3	4.21 ± 2.06	3.84 ± 1.97	3.70 ± 1.89
Stage 4	5.83 ± 1.64	3.69 ± 2.03	3.60 ± 2.06
Stage 5	5.27 ± 2.08	3.53 ± 2.00	3.51 ± 2.05

Table 18: Descriptive Statistics of Reactions

The results of the repeated measures ANOVA support this finding. As shown in Table 19, there was a statistically significant effect of stage on suspicion($F(4, 356) = 23.20$, $p < .001$). In contrast, no significant differences across stages were observed for perceived relevance($F(4, 356) = 1.45$, $p = .217$), or anxiety($F(4, 356) = 1.10$, $p = .359$).

Variable	F(4, 356)	P	Partial η^2
Suspicion	23.20	$< .001$	.207
Relevance	1.45	.217	.016
Anxiety	1.10	.359	.012

Table 19: Repeated Measures ANOVA with Effect Sizes. Partial η^2 represents the effect size. Sphericity assumption was violated for all variables, but the Greenhouse Geisser corrected results yielded consistent patterns.

Stage 1	Stage 2	Mean Difference	P	Significant
Stage 1	Stage 2	–0.31	.814	No
Stage 1	Stage 3	–1.16	$< .001$	Yes
Stage 1	Stage 4	0.47	.479	No
Stage 1	Stage 5	–0.10	.997	No
Stage 2	Stage 3	–0.84	.028	Yes

Table 20: Pairwise comparisons of suspicion. We used Tukey's HSD Test. Mean differences reflect the direction and magnitude of change between stages.

Post-hoc comparisons using Tukey's HSD test were also conducted for the suspicion variable. As shown in Table 20, significant differences were observed between Stage 1 and Stage 3 ($p < .001$), and between Stage 2 and Stage 3 ($p = .028$), indicating that suspicion was significantly lower at Stage 3 than in the earlier stages.

These findings suggest that participants showed initial suspicion during Stages 1–2, which briefly declined in Stage 3, likely due to persuasive scammer cues, then sharply increased from Stage 4 as pressure and financial demands escalated. In contrast, relevance and anxiety remained relatively stable, indicating that suspicion may be a more sensitive and dynamic marker for detecting scam.

H.3 RQ2: single interventions failed to produce significant cognitive effects in complex social engineering scam

To address **RQ2**, which examines the effect of a traditional AI-based detection method on recipients' scam awareness and call termination intention, independent samples t-tests and one-way ANOVA were conducted to compare the single warning and no-warning groups. Awareness was measured as the average perceived suspicion score across all stages of the script.

As shown in Table 21 and Figure 10, the single-warning group reported slightly higher levels of suspicion compared to the control group. However, the difference was not statistically significant($t(58) = 0.96, p = .339$). The effect size was small (Cohen's d = 0.25), and the 95% confidence interval estimated through bootstrapping [–0.42, 1.22] also indicated a lack of statistical significance.

These results suggest that while a single warning may momentarily trigger suspicion, it is insufficient to sustain recipients' psychological resistance throughout the full sequence of a strategically structured social engineering scam. In complex, dynamic threat scenarios such as voice phishing, more adaptive and context-aware interventions may be necessary to produce significant cognitive effects.

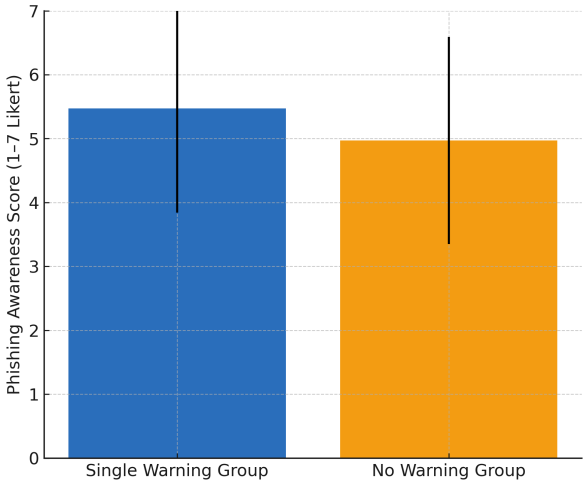

Figure 10: Comparison of Scam Suspicion Score

Group	N	M	SD	t(58)	P	95% CI	Cohen's d
Single Warning	30	5.13	1.56	0.96	.339	[–0.42,1.22]	0.25
No Warning	30	4.73	1.66				

Table 21: Suspicion Scores of Single and No Warning. Cohen's d is reported as the standardized effect size.

H.4 RQ3: SCRIPTMIND helps users recognize potential harm during scams

RQ3 aimed to evaluate whether a **SCRIPTMIND** warning generated by LLMs could enhance users' cognitive suspicion in response to a simulated SE attack. To examine this, we conducted a two-way repeated measures ANOVA, with five call stages as the within-subjects factor and experimental condition as the between-subjects factor.

As shown in Table 22, the results revealed a significant main effect of stage on suspicion levels($F(4, 348) = 23.79, p < .001$, partial $\eta^2 = .215$). Importantly, a significant interaction effect between stage and group was also observed($F(8, 348) = 2.15, p = .031$, partial $\eta^2 = .047$). This indicates that users' suspicion did not merely fluctuate based on the temporal flow, but rather changed in distinct patterns depending on the type of warning.

To better understand these patterns, we conducted stage-wise one-way ANOVAs comparing the groups at each of the stages. As shown in Table 23, the LLM Warning group exhibited the highest suspicion scores at Stage 4 and Stage 5, and the

Effect	df	F	P	Partial η^2
Stage	4, 348	23.79	$< .001$	0.215
Stage × Group	8, 348	2.15	.031	0.047

Table 22: Repeated Measures ANOVA Summary for Suspicion Scores by Stage and Each Group

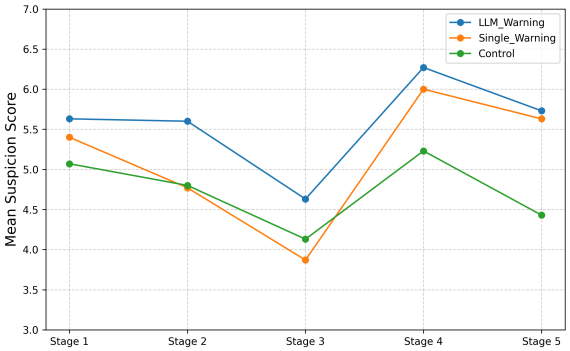

Figure 11: Changes in Suspicion Levels by Script Stage

differences between groups at these stages were statistically significant (p = .039, .024). No significant group differences were found in Stages 1-3.

Stage	LLM_Warning	Single_Warning	Control	F	P
Stage1	5.63±1.69	5.40±1.92	5.07±2.07	0.67	.512
Stage2	5.60±1.65	4.77±2.08	4.80±1.90	1.88	.159
Stage3	4.63±1.99	3.87±2.21	4.13±1.96	1.07	.346
Stage4	6.27±1.60	6.00±1.49	5.23±1.72	3.36	.039
Stage5	5.73±2.05	5.63±1.69	4.43±2.25	3.88	.024

Table 23: Suspicion Scores by Stage and Each Group

Specifically, as visualized in Figure 11, participants in the LLM Warning condition maintained relatively high levels of suspicion during the initial stages (Stages 1 and 2). Although their suspicion briefly declined at Stage 3, they showed a marked increase beginning at Stage 4, reaching the highest average suspicion levels by the final stage.

These findings suggest that **SCRIPTMIND** interventions can elicit stronger and more sustained suspicion responses, especially as social engineering scam progresses into its critical stages involving pressure or financial requests. In contrast, participants in the control and single-warning groups exhibited either delayed or reduced recovery in suspicion following the mid-call drop in awareness.

In summary, the results support that LLM-based warnings are more effective than traditional single-message alerts in promoting cognitive resilience during phone scam. The dynamic and context-sensitive nature of **SCRIPTMIND** predictions appears to better support users' psychological defense mechanisms, making this a promising intervention strategy for complex social engineering scams.