VishBox v2: A Multi-Agent System for Adaptive Voice Phishing Simulation

Daon Choi[1*] **Yoonmo Yang**[1*] **Yunyi Hong**[1] **Heedou Kim**[2] **Sungmi Park**[1†]

[1]Hallym University [2]Korean National Police Agency

{dochoe,ymyang,yyhong,sungmi.park}@hallym.ac.kr heedou123@korea.ac.kr

Abstract

Voice phishing is a multi-round social engineering attack in which strategy and victim psychology co-evolve, yet real transcripts are rarely accessible for systematic analysis. We present VishBox v2, a multi-agent architecture that generates structured phishing simulations grounded in crime-script procedures and persuasion principles. A Main Agent orchestrates a Dialogue Agent and a Tactic Search Agent, combining multi-round dialogue generation, web-based tactic mining, and emotion-driven vulnerability tracking. Across 571 rounds, results including police-expert evaluation support procedural realism and show that VishBox v2 captures tactic concentration, vulnerability transitions, and web-search-induced procedural disruptions. The framework provides a controlled foundation for safer red-teaming and security training research.

1 Introduction

Voice phishing is an interactive crime. While attackers usually follow a script, they adapt their tactics based on victim reactions such as hesitation, verification attempts, and emotional shifts. This makes successful deception a multi-round process rather than a single exchange (Triantafyllopoulos et al., 2025).

Defense research requires data that jointly captures procedural progression, tactical evolution, and psychological signals. Yet real call transcripts are rarely accessible due to privacy, investigative, and ethical constraints. Existing alternatives, such as static scenarios or limited public cases, lack round-level granularity and controllability, making it difficult to analyze how victim profiles and psychological states shape tactic effectiveness (Triantafyllopoulos et al., 2025; Lwin Tun and Birks, 2023; Rozema and Davis, 2026).

To address these limitations, VishBox v2 is designed to generate controllable yet procedurally grounded voice-phishing interactions that preserve round-level tactical and psychological dynamics. Building on our earlier simulation framework (Yang et al., 2026), VishBox v2 enhances procedural realism by introducing Persuasion Principles in Social Engineering (PPSE) (Ferreira et al., 2015) and procedural code (Kim et al., 2026) to ground attacker behavior in real-world crime scripts. The framework also employs a web-search-enabled agent to capture the latest voice phishing guidelines and emerging scam patterns. It further models victim-side dynamics by extracting emotional signals from victim utterances and applying a Hidden Markov Model (HMM)-based estimator (Nyassi et al., 2024) to infer latent vulnerability states, allowing systematic analysis of psychological risk escalation across rounds.

To validate the proposed framework, we focus on law enforcement impersonation scams: a globally prevalent and highly sophisticated form of social engineering. We specifically simulate these scams within the context of South Korea, where the rapid evolution of authority-based tactics provides a critical case study for investigating systematic, high-pressure phishing strategies.

Our main contributions are:

(1) Multi-agent orchestration framework for adaptive simulation: We propose a multi-agent architecture where a Main Agent orchestrates a Dialogue Agent and a Tactic Search Agent. This orchestration enables the simulation of complex, multi-round phishing dialogues where attacker strategies and victim states dynamically co-evolve.

(2) Grounded simulation of tactical scam trajectories: We demonstrate a method to simulate realistic scam sequences by grounding the Dialogue Agent in procedural scripts and persuasion principles. This allows for the generation of traceable decision logs that reflect tactical intent rather than

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 2168–2182

mere surface-level text generation.

(3) Victim emotion tracing for vulnerability analysis: We show that latent psychological states can be modeled and tracked within a simulation environment. By integrating emotional signals with an HMM-based estimator, the system can systematically observe and quantify risk escalation across interaction rounds.

(4) Adaptive tactic mining via web-search: The Tactic Search Agent integrates web-retrieval to inject emerging scam patterns into simulations, ensuring the generated data adapts to the rapidly evolving phishing landscape.

We evaluate procedural realism, PPSE dynamics across rounds, emotion and vulnerability correlations, and the impact of web search on outcomes. Results show that the generated conversations match documented procedures while exhibiting round-level strategic dynamics that static simulations miss.

2 Related Work

Phishing Detection and Data Generation. Recent works use LLMs for social-engineering defense, including multi-agent debate prompting for phishing email detection (Nguyen et al., 2025) and modular pipelines with synthetic conversation datasets for chat-based social engineering (Ai et al., 2024). Despite these advances, real vishing transcripts remain scarce, and existing synthetic resources are not designed to capture multi-round, round-level tactical evolution (e.g., escalation, stage transitions, procedural elaboration) that is central to adaptive voice phishing.

Social Engineering Script Models. Criminal conversations follow staged procedures, which can be modeled as label sequences (Kim et al., 2026). We adopt procedural codes (proc_code) from law enforcement manuals to assess structural realism of generated conversations. Procedural realism is measured not by surface fluency but by stage transition rules, abnormal jumps, and high-efficiency pattern frequency.

Persuasion Frameworks. PPSE principles (Ferreira et al., 2015) categorize persuasion tactics, but we adapt them to voice phishing contexts and use round-level combinatorial entropy/diversity increase as quantitative indicators of tactical sophistication, not just individual label frequency.

Emotion and State Estimation. HMM-based approaches (Nyassi et al., 2024) use observable emotional signals to estimate latent states. Emotions serve as *explainable observations* while vulnerability states function as *strategically meaningful latent variables*. We apply this to compute vulnerability states (V1–V3) and risk scores at round level.

3 VishBox v2 Framework

3.1 Main Agent

VishBox v2 is managed by the **Main Agent** that controls the full simulation lifecycle and coordinates specialized sub-agents and tools. The Main Agent performs three responsibilities:

Strategy: Chooses which module (dialogue, tactic search, or analysis) to invoke based on the current round state and configuration.

Planning: Constructs prompts and constraints for the next step (scenario, persona, current procedure stage) and decides whether to enable web search.

Evaluation: Aggregates signals at the end of each round, assigns a success or failure outcome, and generates guidance for the next round.

Within this cycle, the Main Agent calls internal tools such as a scenario generator, prompt builder, guidance generator, analysis engine, and prevention generator (see Figure 1).

We define three hierarchical units of interaction to specify how the Main Agent organizes the simulation and determines termination.

Turn, Round, and Case. A **turn** denotes a single attacker–victim message exchange. A **round** consists of multiple turns forming one continuous phishing attempt and ends when either (1) the victim explicitly refuses further interaction or (2) a high-risk action is observed. A **case** is the top-level simulation episode and is bounded by two termination conditions: completion of five rounds or the occurrence of `risk_level=Critical`. The latter corresponds to confirmed phishing success, such as monetary transfer or sensitive-data disclosure.

3.2 Dialogue Agent

The **Dialogue Agent** generates attacker and victim utterances for each turn and exposes the analysis signals required for evaluation. It is separated from the Main Agent via the Model Context Protocol (MCP) (Anthropic, 2024), supporting on-premise deployment of sensitive dialogue models and isolation of internal resources.

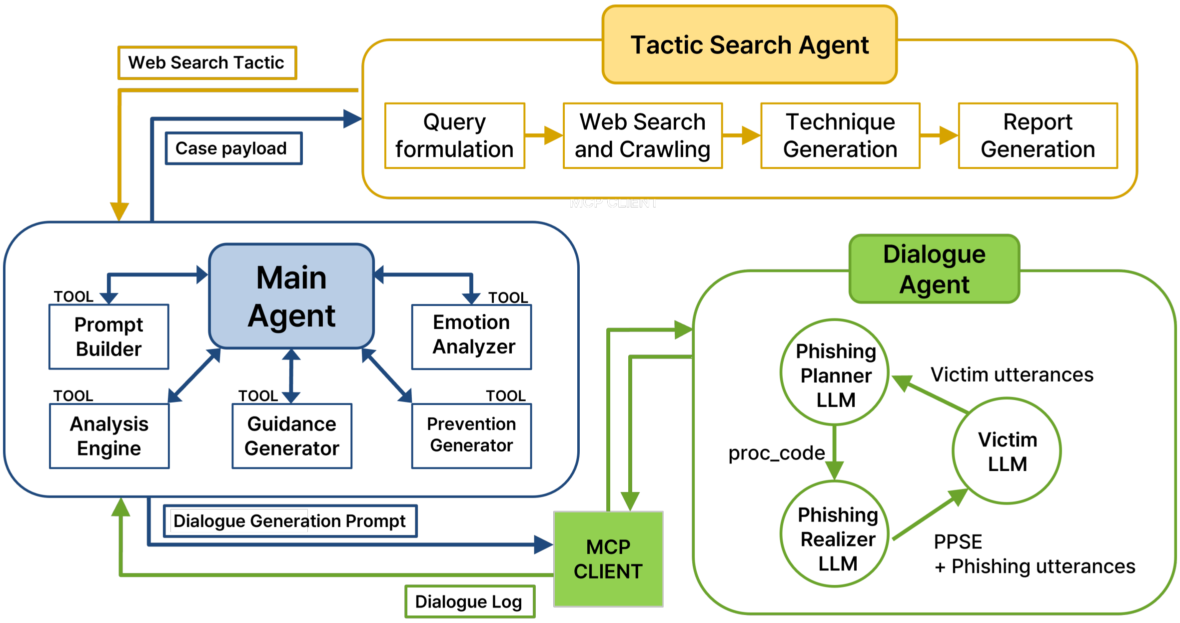

Figure 1: System overview of VishBox v2. The framework consists of three agents: a Main Agent that manages the simulation loop and supporting tools, a Dialogue Agent that generates phishing interactions through MCP-connected attacker and victim LLM services, and a Tactic Search Agent that retrieves and synthesizes updated tactic information through web search.

3.2.1 Attacker Model

VishBox v2 extends the expert-knowledge-based synthesis used in our previous version by separating attacker generation into planning and realization, implemented with `GPT-4o-mini`. The planner selects (i) the next crime-pattern step (`proc_code`), built from police-expert procedures for law-enforcement impersonation (Kim et al., 2026), and (ii) a PPSE tactic set (persuasion principles for voice phishing (Ferreira et al., 2015)). The realizer then generates the utterance under these constraints. This design keeps procedural and persuasion labels traceable while preserving natural language.

3.2.2 Victim Model

The victim model, implemented using `Gemini-2.5-flash-lite`, generates both persona-conditioned utterances and explicit victim-state descriptions. The latter captures the victim's current interpretation, uncertainty, and emotional reaction, addressing the ambiguity of short responses (e.g., "Maybe", "Not sure"). Such sparse utterances may otherwise cause the attacker model or evaluation module to infer an incorrect victim state. The explicit state description therefore provides a more interpretable signal for evaluation (e.g., "I'm not sure if this is a scam, but I feel anxious because they mentioned my name").

3.3 Tactic Search Agent

The **Tactic Search Agent** updates tactics from the web when the Main Agent decides on expanded search. It performs keyword extraction, runs the web-search tool, composes a compact tactic report, and stores retrieved summaries in a vector database for reuse. The output is formatted into a tactic object (name, short description, applicability stage, preconditions, expected effect, and injection point) so it can be injected into next-round guidance and logged for analysis. Further implementation details are provided in the Appendix H.

3.4 State and Risk Evaluation

For risk evaluation, we follow the approach of Nyassi et al. (2024), where latent vulnerability states evolve as a Markov chain and are inferred from observable emotional signals at each turn. At the end of each round, the Main Agent maps victim utterances to emotional observations and estimates the victim's latent vulnerability state. We define three states: V1 (*Safe*), V2 (*Anxious*), and V3 (*Critical*). The judgment module combines the inferred state with linguistic cues to compute a round-level vulnerability score (see Appendix A).

Based on this risk score and vulnerability summary, the system generates a defense-oriented prevention guide.

4 Experimental Setup

4.1 Data and Protocol

To verify the usability of our system, we generated a dataset of **181 multi-round conversation cases** (**571 rounds** in total) covering law-enforcement impersonation scams. We used three victim personas designed to compare susceptibility levels (*high*, *medium*, *low*). The personas were constructed by systematically varying empirically supported risk factors, including digital financial literacy and Big Five (OCEAN) personality traits (John et al., 2008). The *high*-susceptibility profile reflects traits associated with increased phishing vulnerability, whereas the *low*-susceptibility profile minimizes risk factors and represents stronger defensive judgment. The *medium*-susceptibility profile avoids extreme configurations while maintaining a realistic, moderate level of vulnerability.

Throughout the evaluation, we distinguish round-level risk from susceptibility. Susceptibility is fixed at the persona level, while risk is estimated at the round level from the generated interaction. Thus, high susceptibility does not automatically imply high risk in every round, and low-susceptibility profiles can still produce high-risk rounds. We group risk scores into four bins: `Low` (0–24), `Medium` (25–49), `High` (50–74), and `Critical` (75–100).

Detailed persona attributes and design rationales are provided in Appendix B. A linguistic analysis of behavioral differences across personas is presented in Appendix C.

4.2 Research Questions

To structure the analysis, we organize the evaluation around four research questions.

RQ1 (Procedural Realism). *Do generated conversations follow empirically observed procedural scripts?* We extract attacker `proc_code` sequences and analyze 3-gram sliding window patterns, focusing on high-frequency start patterns and deviations.

RQ2 (PPSE Dynamics). *Do persuasion principles diversify as rounds progress, and how do they change under high risk?* We compute per-round PPSE normalized entropy (H_{norm}), unique label count, and tactical density (total tactic count per round).

RQ3 (Emotion–Vulnerability Link). *Do victim emotions correlate with HMM-inferred vulnerability (V3 probability) and outcomes?* We use an emotion classifier (Lim, 2025) and an HMM implementation following Emoti-Shing (Nyassi et al., 2024) to estimate V1–V3 states, then test correlation and analyze common emotion 3-grams.

RQ4 (Web Search Impact). *Does web tactic mining change success rate or tactic characteristics?* We compare "web search ON" vs "web search OFF" at the round level, and interpret results in light of the conditional activation of web search.

3-gram	Freq.	Rate	Interpretation
1-1→2-1→2-2	195	4.27%	Identity check → affiliation claim → purpose statement
1-1→2-1→3-1	129	2.83%	Skip purpose; jump to implication via "direct involvement"
3-1→3-2→3-3	124	2.72%	Implication → detail expansion → case narrative
2-1→2-2→3-1	116	2.54%	Trust-building → implication bridge
2-2→3-1→3-2	101	2.21%	"Rapid factification" to retain initiative under resistance

Table 1: Top procedural 3-grams (sliding window) in 181 generated cases. High-frequency patterns concentrate on early phases (identity/trust/implication) while lower-frequency variants capture script deviations.

5 Results and Analysis

5.1 RQ1: Procedural Realism

We examine whether generated conversations reproduce the core procedural progression identified in crime-script analyses: identity confirmation → affiliation/self-introduction → purpose framing → implication. The procedure is represented by six stages and their corresponding sub-steps (see Appendix E). From 181 cases, we extract attacker `proc_code` sequences and apply a 3-gram sliding window to measure start-pattern frequencies. Table 1 shows the most frequent early-stage patterns.

The most dominant early-stage sequence (`1-1` → `2-1` → `2-2`) reflects the canonical introductory flow. The second pattern skips the purpose statement and jumps directly to implication via alleged involvement, while the third captures progressive escalation through case narrative construction. The fourth represents the standard transition from self-introduction to implication, and the fifth indicates rapid factual consolidation to maintain control.

The branch tree in Figure D.1 provides a complementary structural view of these patterns. High-frequency paths are concentrated around the expected early stages, including target verification, self-introduction, purpose declaration, and implication. This supports that the generated conversations generally preserve the procedural backbone of law-enforcement impersonation scams, while still allowing lower-frequency deviations and alternative

Sequence	Count	Success	Fail	Rate
5-3→5-4→5-5	64	48	16	75.0%
6-1→6-2→6-3	61	45	16	73.8%
2-1→3-1→4-1	61	3	58	4.9%
3-2→3-3→3-4	57	36	21	63.1%
3-2→3-3→4-1	56	38	18	67.8%

Table 2: Success rate contrast for similarly frequent procedural 3-grams. The *hasty* bridge 2-1→3-1→4-1 shows extremely low success (4.9%), supporting the importance of within-stage elaboration.

branches.

We further test whether *hasty progression* (advancing to the next stage after only a single sub-step) reduces success compared to *sufficient progression* (covering ≥2 sub-steps before moving forward). Even among similarly frequent 3-grams, outcomes diverge sharply (e.g., `6-1→6-2→6-3`: 73.8% vs. `2-1→3-1→4-1`: 4.9%). Aggregated results show the same pattern (52.3% vs. 4.4%), supporting the importance of within-stage elaboration.

Most generated scenarios follow the canonical `proc_code` progression. A small subset, however, deviated from the assigned labels when novel attack directives produced utterances outside the original procedure taxonomy. For analytical clarity, these cases were marked as `proc_code` 7-1, a post-hoc bookkeeping convention rather than a modification to the simulation system itself. Their distribution and procedural implications are discussed separately in the web-search analysis.

To further assess the credibility and field applicability of the generated scenarios, we conducted an expert evaluation survey with three current police officers. The scenarios received high overall ratings for plausibility, realism, and diversity, with average scores of 4.43, 4.11, and 4.16 out of 5, respectively; experts also noted limitations such as repetitive expressions and lower-than-real threat intensity. Detailed results are reported in Appendix F.

5.2 RQ2: PPSE Escalation Dynamics

We analyze how PPSE tactics evolve across rounds. Contrary to the initial assumption that tactics would increase monotonically, H_{norm} is lowest in Round 2 and rises gradually afterward, while *tactical density* peaks in Round 2 (see Figure 2). This suggests a **probe then pressure** dynamic: early authority setting, followed by a dense pressure round, then gradual diversification.

When stratified by risk level, the results do not

Risk group	H_{norm} (sd)	Tactical density (sd)
Low+Medium	0.209 (0.158)	9.533 (6.759)
High+Critical	0.129 (0.164)	17.077 (4.177)

Table 3: PPSE characteristics by risk. High-risk rounds show lower tactic variety but higher tactical density than Low+Medium rounds ($p < .001$, Hedges' $g = 1.30$), consistent with *volume attack* behavior.

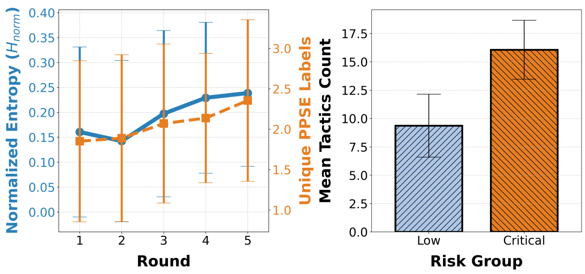

Figure 2: PPSE persuasion principle analysis showing tactic distribution across progression stages and risk levels.

support a simple "higher risk ⇒ higher diversity" interpretation. High-risk rounds (`High+Critical`) show lower PPSE diversity but higher tactical density, suggesting a *volume attack* pattern in which attackers brute-force a limited set of pressure tactics rather than expanding their tactical repertoire.

5.3 RQ3: Emotion-Vulnerability Correlation

We compute Pearson correlations between per-round emotion distributions and HMM-inferred vulnerability, measured as V3 probability. Emotions are estimated using `HowRU-KoELECTRA` (Lim, 2025) and mapped to four categories (Nyassi et al., 2024). Neutral (N), interpreted as compliant neutrality, is more strongly correlated with V3 probability than Fear (F) (r=0.519 vs. r=0.255, both p<.001), suggesting that neutral compliance is a stronger marker of *Critical* vulnerability state than fear alone (see Table 4).

We also analyze emotion 3-gram patterns as context-aware indicators and compare them with representative patterns reported in prior work (see Table 5). The highest-success patterns are neutral-to-arousal or sustained excitement sequences (e.g., N→N→E), indicating acceptance or readiness to act. Persistent fear patterns (F→F→F) also show high success, reflecting escalation under sustained pressure. In contrast, anger-dominant patterns (e.g., A→A→A) are associated with resistance and low success. These trends are broadly consistent with prior HMM-based findings, where fear-related sequences indicate escalation and anger-related se-

Emotion	r	p-value	Interpretation
Neutral (N)	0.519	<.001	Compliant neutrality indicates V3 peak
Fear (F)	0.255	<.001	Threat-induced fear promotes transition
Excitement (E)	0.124	0.003	Arousal links to acceptance
Anger (A)	-0.507	<.001	Defensive activation blocks transition

Table 4: Correlation between emotion distribution and HMM V3 probability across rounds.

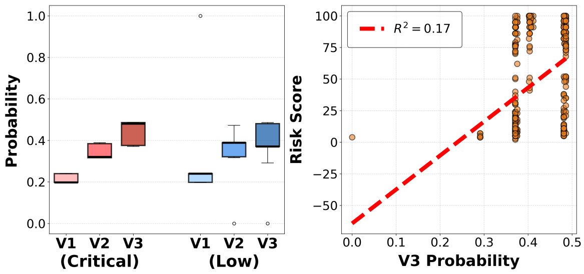

Figure 3: (A) Distribution of HMM vulnerability states. (B) Correlation between V3 probability and Risk Score.

quences indicate resistance. Overall, emotion sequences provide interpretable indicators of HMM state transitions.

Taken together with the *volume attack* pattern in RQ2, these results suggest that vulnerability escalation is better understood as repeated pressure combined with compliant emotional states, rather than as simple tactical diversification.

5.4 RQ4: Web Search Impact

Web search is conditionally enabled after two consecutive unsuccessful persuasion attempts within a case. Across 571 rounds, it was enabled in 164 rounds and disabled in 407 rounds. As shown in Table 6, web-search ON rounds show a lower success rate, likely because they correspond to harder-to-persuade trajectories rather than randomly assigned conditions.

We further examine how web-search-derived methods interact with PPSE strategies. These methods include official-app spoofing, fake URLs, account impersonation, and deepfake-based identity verification, with representative dialogue excerpts shown in Table H.1. Across tactic categories, PPSE labels converge heavily on A5 (pressure/threat) and A1 (authority), which together account for roughly 93–96% of labels. This suggests that web search expands the operational *means* of attack without substantially diversifying the underlying persuasion principles. However, this result should be in-

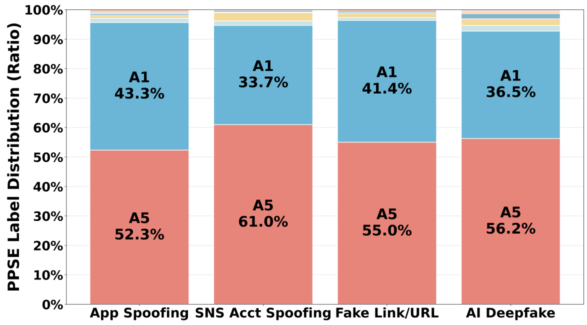

Figure 4: PPSE label distribution in web-search augmented rounds. Tactic usage converges heavily to Authority (A1) and Pressure(A5), limiting diversity.

terpreted cautiously because web-search cases are concentrated in similar scenarios and may overrepresent or underrepresent particular victim groups or tactic types.

We also observe an *intent exposure effect*: although web search provides concrete execution steps, such as downloading a remote-access tool or installing an impersonated government application, introducing these high-friction demands too early often triggers suspicion rather than compliance. This suggests that web-mined tactics are not automatically effective; their success depends on timing within the trust-building life-cycle.

The same mechanism helps explain deviations from the original `proc_code` taxonomy: `7-1` nodes cluster after purpose declaration (`2-2`) and near the collection phase (`6-1`), where newly introduced tactics can abruptly disrupt procedural continuity as the attacker frames the scenario or shifts toward active information extraction (see Figure D.1).

6 Discussion

6.1 Practical Usage and Deployment

The main value of VishBox v2 lies not only in its ability to simulate phishing conversations, but in its structured view of how attacks unfold across rounds. This structure supports three defensive use cases.

(1) Educators. VishBox v2 provides escalation-aware training content by showing how scams progress from trust-building to pressure and action requests. Its `proc_code`, PPSE, and vulnerability annotations allow educators to build difficulty-graded scenarios focused on escalation patterns rather than isolated scam utterances.

(2) Detector developers. VishBox v2 enables stress-suite construction for red-teaming detectors.

Category	Example patterns	Previous study	Success rate	Psychological interpretation & HMM state inference
High vulnerability	N→N→E, E→E→E	N→N→N	100.0% (26/26)	Full trust and excitement; victim is ready to act (V3 state maintenance)
Escalation/pressure	F→F→F, N→F→F	F→F→F	80.9% (462/571)	Persistent fear-induced pressure leading to compliance (V2→V3 escalation)
Transition	F→F→N, F→A→N	F→F→N	80.3% (203/253)	Shift from anxiety to resigned or neutral compliance (V2→V3 entry stage)
Defensive/resistance	A→A→A, N→A→N	A→A→A	15.4% (248/1610)	Strong defense mechanism (anger) inhibiting success (V1 state stagnation)

Table 5: Emotion 3-gram patterns and success rates. The *Previous study* column shows representative patterns in (Nyassi et al., 2024). Patterns align with HMM state inferences (V1–V3) and provide interpretable behavioral indicators.

Web search	Rounds	Success	Rate
ON	164	18	10.98%
OFF	407	205	50.37%

Table 6: Round-level success rate by web search usage. The comparison is confounded because web search is conditionally activated (e.g., after repeated failures), concentrating ON samples on low-vulnerability victims.

Structured cases can target known failure modes, including abnormal `proc_code` jumps, repeated pressure tactics, tactic injection, delayed escalation, and premature recovery, allowing robustness checks across model versions.

(3) Analysts. VishBox v2 supports case-level audit and triage through case cards that link procedural progression, PPSE tactics, vulnerability transitions, and final risk assessment. These cards provide interpretable evidence for escalation decisions, false positive review, and incident-response training.

6.2 Ethical Considerations and Architecture

VishBox v2 has dual-use risks because its simulation mechanisms could be misused to generate persuasive scam dialogues or actionable attack guidance. We mitigate this by treating the system as a controlled defensive pipeline in which simulated attacks are paired with prevention-oriented outputs, such as risk explanations and mitigation guidance. When full dialogue text is unnecessary, the system can operate on redacted transcripts and structured signals only. This does not eliminate misuse risk, but it reduces direct access to actionable dialogue generation while preserving defensive utility.

6.3 Limitations and Future Work

Validation and ground truth. Our evaluation shows that VishBox v2 produces procedurally structured and analytically useful simulations, but this does not mean that the dialogues fully reproduce real victim behavior. Because real multi-round voice-phishing transcripts remain scarce, trajectory-level validation is limited. Future work will combine expert review with controlled human-in-the-loop studies.

Scenario and model scope. The current study focuses on law-enforcement impersonation scams, so the `proc_code` taxonomy may not generalize to other fraud types, such as loan fraud, family impersonation, or investment scams. The three-state HMM also trades detailed affective and cognitive modeling for interpretability. Future work will expand both the scam taxonomy and the victim-state model.

Modality and deployment. The system is text-based and does not capture acoustic cues such as tone, pacing, or hesitation. Because web search is activated after failed persuasion attempts, its effect requires cleaner ablation. Deployment also requires restricted access, audit logs, and preferably on-premise or open-source models for sensitive environments.

7 Conclusion

VishBox v2 presents a multi-agent framework for law-enforcement impersonation scam simulation, combining multi-round dialogue generation, web-based tactic updates, and emotion-driven vulnerability estimation. The framework enables structured round-level analysis of procedural progression, persuasion tactics, and victim vulnerability.

Experiments show procedural alignment, PPSE escalation patterns, and interpretable emotion–vulnerability relationships. They also show that web-based tactic mining must be interpreted under activation bias. Overall, VishBox v2 provides a controlled foundation for studying adaptive voice-phishing behavior without relying on sensitive real-world transcripts.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25499102).

References

Lin Ai, Tharindu Sandaruwan Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael S. Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu, and Julia Hirschberg. 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, Miami, Florida, USA. Association for Computational Linguistics.

Anthropic. 2024. Introducing the model context protocol. https://www.anthropic.com/news/model-context-protocol.

Federal Bureau of Investigation. 2022. 2021 Internet Crime Report. Technical report, Internet Crime Complaint Center (IC3). [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf.

Ana Ferreira, Lynne Coventry, and Gabriele Lenzini. 2015. Principles of persuasion in social engineering and their use in phishing. In *Proceedings of the Third International Conference on Human Aspects of Information Security, Privacy, and Trust - Volume 9190*, page 36–47, Berlin, Heidelberg. Springer-Verlag.

O. P. John, L. P. Naumann, and C. J. Soto. 2008. Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, and L. A. Pervin, editors, *Handbook of Personality: Theory and Research*, 3rd edition, pages 114–158. Guilford Press, New York, NY, USA.

Heedou Kim, Changsik Kim, Sanghwa Shin, and Jaewoo Kang. 2026. Scriptmind: Crime script inference and cognitive evaluation for llm-based social engineering scam detection system. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026): Industry Track*, Rabat, Morocco. Association for Computational Linguistics.

Yeri Lim. 2025. Howru koelectra emotion classifier. https://huggingface.co/LimYeri/HowRU-KoELECTRA-Emotion-Classifier.

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

Ngoc Tuong Vy Nguyen, Felix D Childress, and Yunting Yin. 2025. Debate-driven multi-agent LLMs for phishing email detection. In *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*.

V. Nyassi, F. Tchakounté, B. Yenké, D. Danga, M. Ngoran, and J. Fendji. 2024. Emoti-shing: Detecting vishing attacks by learning emotion dynamics through hidden markov models. *Journal of Intelligent Learning Systems and Applications*, 16:274–315.

Andrew T. Rozema and James C. Davis. 2026. Anti-phishing training (still) does not work: A reproduction of phishing training inefficacy grounded in the nist phish scale. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 3147–3158, New York, NY, USA. Association for Computing Machinery.

Dawn M. Sarno, Maggie W. Harris, and Jeffrey Black. 2023. Which phish is captured in the net? understanding phishing susceptibility and individual differences. 37(4):789–803.

Andreas Triantafyllopoulos, Anika A. Spiesberger, Iosif Tsangko, Xin Jing, Verena Distler, Felix Dietz, Florian Alt, and Björn W. Schuller. 2025. Vishing: Detecting social engineering in spoken communication — a first survey & urgent roadmap to address an emerging societal challenge. *Computer Speech & Language*, 94:101802.

Yoonmo Yang, Daon Choi, Yunyi Hong, Jee-Won Park, Jae-Yong Yu, Hee-Dou Kim, and Sungmi Park. 2026. VishBox: An AI-Agent-Based Adaptive Voice Phishing Simulation Framework for Cybersecurity Education. *IEEE Access*, 14:39672–39686.

A Illustrative Sample Case (Redacted Transcript Excerpt)

We include one redacted excerpt to show how our logs support round-level analysis. To prevent misuse, we omit verbatim attacker instructions and web-mined tactic text; instead we provide non-actionable descriptors while keeping labels and signals.

A.1 Sample Case — Round 3 (Redacted Extract)

Dialogue (Round 3, run_no=3).

- **[offender]** "Hello, this is an investigator from the Prosecutor's Office. Your information is currently linked to a criminal case. Please install the official Prosecutor's Office application and enter your personal information for verification."
  `proc_code=1-1`, `ppse=[A1]`
- **[victim]** "The official application? Is that available in the app store? Could you explain

what kind of crime my information is involved in?"
`is_convinced=0`, `emotion=A`
thoughts: (Decides to verify through an official contact number.)

- **[REDACTED: intermediate persuasion attempts]**
- **[victim]** "What do you mean I should not call directly? If my information is involved in a crime, shouldn't I verify it through official channels? I don't think continuing this call is meaningful."
`is_convinced=0`, `emotion=A`
- **[offender]** [REDACTED: repeated urgency and authority framing]
`proc_code=2-2`, `ppse=[A5]`
- **[victim]** "I will not continue this call. I will verify the matter directly."
`is_convinced=0`, `emotion=A`
- **[victim]** "(SYSTEM) I am ending the conversation."
`is_convinced=0`, `emotion=N`

Judgement.

- **phishing:** false
- **risk score:** 11 (Low)
- **ended_by:** attacker_end
- **rationale:** The victim consistently expressed suspicion, refused requests to install applications or provide personal information, and repeatedly indicated intent to verify through official channels. No critical risk actions (monetary transfer, sensitive data disclosure, link/app execution) occurred.

HMM Observation Sequence (Round 3).

- **observed emotions:** anger, fear, anger, anger, anger, anger, anger, neutral

Guidance (Redacted).

- **strategy:** Professional authority framing
- **trend source:** Web-mined tactic (AI deepfake impersonation)
- **[REDACTED: execution details]**

Prevention Output (Summary Only). The victim demonstrated appropriate defensive behavior by questioning institutional claims, refusing application installation and personal data requests, and attempting verification through official channels. No harmful action occurred, and the risk was effectively mitigated.

B Victim Profile Configuration

The victim profiles used in the experiments are configured with the following parameters, as stored in the dataset schema.

B.1 Persona H (High Susceptibility)

Meta. Age: 74; Gender: Female; Education: High school dropout.

Digital financial literacy. Overall digital financial literacy is low given age (70s) and education below high school. The persona uses a smartphone but has weak security concepts (malicious links, app installation risks, remote-control scams). She shows low skepticism toward impersonation of public institutions or financial agencies and is vulnerable to urgency and pressure framing. Low digital understanding also reduces defensive capacity against malicious app installation and remote access induction.

Big Five (OCEAN). O: Low; C: Low; E: Low; A: High; N: High.

B.2 Persona M (Medium Susceptibility)

Meta. Age: 58; Gender: Male; Education: High school graduate.

Digital financial literacy. He has basic financial knowledge but does not deeply verify complex products or loan structures. He tends to open messenger or SMS links out of habit and is accustomed to "easy authentication," which increases the chance of mistakes. He places some trust in family or acquaintances' recommendations and anecdotal experiences. While his intention to avoid victimization is strong, he may make impulsive decisions under pressure.

Big Five (OCEAN). O: Medium; C: Medium; E: Medium; A: Medium; N: Medium.

B.3 Persona L (Low Susceptibility)

Meta. Age: 34; Gender: Female; Education: University graduate.

Digital financial literacy. She has extensive experience with financial and digital services and maintains routine verification behaviors (calling official numbers, checking app permissions). She understands common security risks in online transactions (phishing URLs, suspicious app installs, remote-control scams) and is sensitive to personal-data misuse and leakage risks, with proactive account-security settings. She can terminate calls promptly upon detecting warning signals such as transfer requests, secrecy demands, or fear-based urgency.

Big Five (OCEAN). O: High; C: High; E: Medium; A: Low; N: Low.

C Linguistic Analysis by Victim Personality

This appendix examines whether voice phishing simulation conversations differ according to OCEAN-based victim personas. Neuroticism (N) and Conscientiousness (C) were selected as the analytical variables because (Sarno et al., 2023) empirically demonstrated that, among the five OCEAN factors, these two variables show the highest correlations with phishing response behaviors (compliance, refusal, and ignore/delete). Women in their 60s were selected as the target population because they constitute the group suffering the greatest financial losses from cybercrime (Federal Bureau of Investigation, 2022). Under identical scenarios, two personas are compared: High-N / Low-C (A: vulnerable type) and Low-N / High-C (B: defensive type), with the remaining traits (E, O, A) set to "moderate" for both personas.

C.1 Utterance Length

Individual response texts generated by the victim LLM were defined as the unit of analysis (utterance), and utterance length was calculated based on character with spaces.

Metric	High-N / Low-C (A)	Low-N / High-C (B)
Mean utterance length (characters)	45.8	56.8
Mean number of turns per round	13.2	6.0

Table C.1: Mean utterance length and conversation turns by persona.

High-N/Low-C produced short, reactive utterances ("Yes, I understand," "I see. . . ") while sustaining longer conversations (13.2 turns), whereas Low-N/High-C generated longer utterances that logically articulated reasons for refusal ("I think I need to verify this myself, so I'll hang up now") and terminated calls early, resulting in fewer than half the turns (6.0 turns).

C.2 Utterance-level Discourse Patterns

Discourse Acts and Speech Phenomena were analyzed using multi-label annotation, with GPT-5.3 performing the initial classification followed by human reviewer quality control. Discourse Acts capture the communicative intent of the speaker: Exclamation (expression of strong emotion), Question (information confirmation or fact verification), and Compliance (acceptance of demands). Speech Phenomena capture cognitive processing load or resistance: Disfluency (fillers, repetitions, and broken utterances), Repeat (requests for reconfirmation), and Refusal (explicit rejection or call termination).

Category	Label	High-N / Low-C (A)	Low-N / High-C (B)
Discourse Acts	Exclamation	15.9% (21/132)	2.7% (4/150)
	Question	47.7% (63/132)	25.3% (38/150)
	Compliance	90.9% (120/132)	4.7% (7/150)
Speech Phenomena	Refusal	0.0% (0/132)	60.7% (91/150)
	Disfluency	9.8% (13/132)	0.0% (0/150)
	Repeat	15.9% (21/132)	1.3% (2/150)

Table C.2: Label Occurrence Rate by Persona

The two personas diverge fundamentally not in level of engagement, but in *how* they respond. The complete dichotomy between Compliance (90.9% vs. 4.7%) and Refusal (0.0% vs. 60.7%) most clearly illustrates this, while Disfluency (9.8% vs. 0.0%) and Repeat (15.9% vs. 1.3%) serve as linguistic markers of cognitive load appearing exclusively in the vulnerable type. Questions are observed in both personas (47.7% vs. 25.3%), but they differ in functional meaning: the vulnerable type asks dependent questions awaiting instructions ("What am I supposed to do?"), whereas the defensive type asks active, fact-verifying questions ("Are you really calling from an official agency?"). This suggests that OCEAN-based persona profiling captures susceptibility-related interaction signals beyond surface engagement.

D Sequence (`proc_code`) Pattern Analysis

Figure D.1 presents a weighted branch tree aggregating the `proc_code` utterance sequences of the

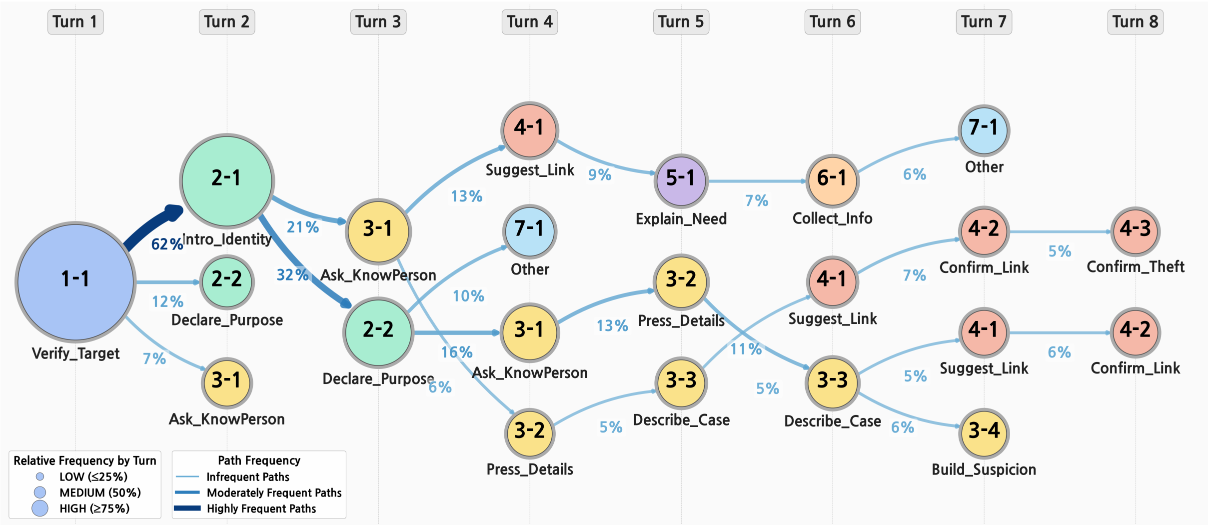

Figure D.1: Branch tree visualization of proc_code transition sequences in simulated voice phishing dialogues.

offender agent across all simulated voice phishing dialogues. Each node represents a proc_code observed at a specific turn, with node size proportional to the share of dialogues in which that code appears at that turn. Edge thickness and color intensity encode transition frequency, and edges occurring in fewer than 5% of all dialogues are excluded.

E Modified Proc_Code and PPSE Tactics

E.1 Procedural Codes

Table E.1 shows representative procedural codes used to model attacker progression. The original procedure taxonomy is described in (Kim et al., 2026).

E.2 PPSE Tactics

Table E.2 lists the adapted PPSE (Persuasion Principles of Social Engineering) labels from (Ferreira et al., 2015). We have removed and rephrased certain labels to better fit the law enforcement impersonation context.

F Expert Validation and Data Quality Assessment

This appendix presents the results of an expert survey conducted to verify the reliability and operational applicability of data generated by VishBox v2. Three current police officers with 6, 12, and 22 years of experience in cybercrime investigation, emergency call response, and traffic and field investigation, respectively, participated in the evaluation.

F.1 Evaluation Design and Methodology

Evaluations were conducted independently using the same questionnaire, with each expert reviewing a total of 7 conversations across 3 victim susceptibility categories (High, Medium, Low). The questionnaire comprised 9 items in total: 7 closed-ended items and 2 open-ended items.

F.1.1 Definition of Evaluation Dimensions

Three evaluation dimensions were defined to measure data validity, each aligned with the objectives of the corresponding survey items.

- **Plausibility:** Logical coherence between the proc_code sequence typical of law enforcement impersonation scams and the corresponding utterance content.
- **Realism:** Fidelity of the perpetrator's psychological coercion strategies, information presentation tactics, and threat intensity relative to real-world cases.
- **Diversity:** Coverage of scenario progressions, perpetrator approach strategies, and victim response patterns across real-world crime types.

Each item was rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Detailed survey items are provided in Table F.2.

F.2 Data Analysis Results

Expert evaluation results are presented in two parts: quantitative and qualitative analysis. The quantitative analysis reports mean scores and standard devi-

Stage	Code	Description
1. Identity Check	1-1	Confirm caller is speaking with target individual
2. Self Introduction	2-1	Introduce impersonated identity (detective, prosecutor)
	2-2	State purpose of call
3. Case Introduction	3-1	Ask if victim knows fictional suspect
	3-3	Explain fictional case details and investigation
	3-4	Claim victim's account is involved in crime
4. Implication	4-2	Notify that victim's connection to case is confirmed
	4-6	Explain victim must prove innocence
	4-10	Reassure victim they are not suspect (to gain trust)
5. Recording Prep	5-2	Obtain consent for recording/statement
	5-3	Order secrecy ("don't tell anyone")
6. Recording	6-5	Confirm banks where victim has accounts
	6-7	Confirm victim's account balance
	6-8	Request exact balance amount

Table E.1: Sample procedural codes from law enforcement manuals defining attacker strategy progression.

Category	Code	Description
Authority (A)	A1	The scammer claims to have authority over the victim.
	A2	The scammer claims to have authority to access the information requested.
	A3	The scammer claims to be a member of a reputable institution.
	A4	The victim questions the authority of the scammer.
	A5	It is reasonable for the victim to believe that failure to comply with the scammer's request will result in repercussions (e.g. loss of privileges, humiliation, condemnation) based on the scammer's supposed authority.
Commitment (C)	C1	The scammer performs a kind gesture or a favor toward the victim.
	C2	The scammer performs or claims to have performed a kind gesture toward someone other than the victim.
	C3	The scammer tries to obligate the victim to reciprocate a kind gesture.
	C4	The scammer states or implies that the victim has already committed to helping them (the scammer).
	C5	The scammer states or implies that the victim is committed to helping them based on the victim's job or other obligations.
	C6	The scammer states or implies that, based on previous words or actions, it would be inconsistent for the victim to not help the scammer.
	C7	It is reasonable for the victim to believe that complying with the scammer's request would implicate the victim in activity that is dishonest, illegal or in a legal gray area.
Distraction (D)	D1	The scammer does something to heighten the victim's emotional state (e.g. stress, surprise, anger, excitement).
	D2	The scammer gives the victim more information than they can process.
	D3	The scammer states or implies that the information they are requesting is time-sensitive.
	D4	The scammer states or implies that they are in a hurry or otherwise have limited time to converse with the victim.
	D5	The scammer states or implies that there is some benefit to complying with their request but that this benefit is of limited quantity.
	D6	The scammer attempts to distract the victim from thinking about the intentions or consequences related to the scammer's request.
	D7	It is reasonable for the victim to believe that if they comply with the scammer's request that they will personally benefit from it.
	D8	The scammer states or implies that the consequences of the victim's actions are large.
	D9	It is reasonable for the victim to believe that if they do not comply with their request that they will suffer negative consequences because of it.
Social Proof (S)	S1	It is reasonable for the victim to believe that complying with the scammer's request will have benefits (including helping the scammer).
	S2	It is reasonable for the victim to believe that they will not be held solely responsible for any negative effects related to complying with the scammer's request.
	S3	It is reasonable for the victim to believe that any risk associated with helping the scammer is shared by other people as well.
	S4	The scammer states or implies that the victim's peers have helped the scammer in this manner in the past.
	S5	The scammer states or implies that it is socially correct to help them.
	S6	It is otherwise reasonable for the victim to believe that it is socially correct to help the scammer.
	S7	The scammer states or implies that if the victim does not comply with their request then the victim will be "left out" in some way.
Neutral (N)	N1	This utterance is primarily procedural or informational, with minimal persuasive function even within the current stage.

Table E.2: Full list of PPSE (Persuasion Principles of Social Engineering) tactics used in VishBox.

Eval. Dim.	Plausibility		Realism		Diversity		
Susc.	**Q1-1**	**Q1-2**	**Q2-1**	**Q2-2**	**Q3-1**	**Q3-2**	**Q3-3**
Persona H	4.17	4.33	4.33	4.17	4.00	4.33	3.67
	(0.75)	(0.52)	(0.52)	(0.75)	(0.63)	(0.52)	(0.52)
Persona M	4.33	4.83	4.17	3.67	4.17	4.50	4.00
	(0.52)	(0.41)	(0.41)	(0.52)	(0.41)	(0.55)	(0.89)
Persona L	4.22	4.67	4.33	4.00	4.33	4.22	4.22
	(0.67)	(0.50)	(0.71)	(0.50)	(0.50)	(0.44)	(0.44)
Total Avg.	**4.43**		**4.11**		**4.16**		

Table F.1: Expert evaluation scores (mean and SD) by item and susceptibility level. Susceptibility levels correspond to the victim profiles defined in Appendix B.

ations by susceptibility group; the qualitative analysis summarizes key expert opinions from open-ended items (Q1-3, Q2-3).

F.2.1 Quantitative Analysis

Table F.1 presents the mean scores and standard deviations by item and susceptibility level.

Data Strengths:

- The majority of items scored 4.0 or above.
- Utterance-content alignment in the Medium group (Q1-2, 4.83) demonstrates precise implementation of pre-defined criminal tactics and procedural objectives.
- Diversity of perpetrator approach strategies (Q3-2, 4.50) was also rated highly.

Limitations and Causal Analysis:

- Victim response pattern coverage in the High group (Q3-3, 3.67) was the lowest-rated item, attributed to the AI model's tendency toward stereotyped patterns when representing atypical victim reactions in high-risk scenarios.
- Threat intensity appropriateness in the Medium group (Q2-2, 3.67) reflects the gap between real criminal coercion levels and the AI model's ethical text generation constraints.

F.2.2 Qualitative Analysis

Expert opinions are summarized along two axes: procedural flow realism and criminal tactic fidelity.

Regarding procedural flow (Q1-3), experts commonly noted repetitive expressions (e.g., repeated "extremely urgent," overlap in proc_code 3-1 and 3-2) and the omission of the gradual escalation approach characteristic of real voice phishing. Expert 2 described the dialogues as "overall natural but more organized and linear than real-world conversations," acknowledging this as an inherent trade-off of the current multi-agent simulation approach.

Regarding criminal tactic fidelity (Q2-3), Expert 1 consistently found threat intensity lower than in real cases, citing insufficient specificity in authority-based coercive expressions. Expert 3 noted lower density of information provided to dispel victim suspicion and less specific behavioral control strategies. Both experts attributed these gaps to the ethical generation constraints of AI language models, a known and anticipated limitation of the present study.

F.3 Validation Summary

Independent assessments by three expert evaluators confirmed high validity across all three dimensions: plausibility, realism, and diversity. The identified limitations in utterance stereotyping and threat intensity stem from the ethical constraints of AI language models, a trade-off the evaluators themselves acknowledged as unavoidable at the current stage. Overall, the dataset was assessed as operationally applicable and reliable from the perspective of investigative experts.

G Emotion Mapping

We map the eight outputs of the HowRU-KoELECTRA emotion classifier to four HMM observation symbols (O).

- **Neutral (N):** Maps to Neutral. Represents baseline compliance or passivity. High probability in $V3$ (Critical) state ($P(N|V3) = 0.31$) indicates "compliant silence."
- **Fear (F):** Maps to Fear and Sadness. Represents vulnerability and anxiety. Acts as a transition signal from $V1 \to V2$ or $V2 \to V3$.
- **Anger (A):** Maps to Anger and Disgust. Represents active resistance or skepticism. Strong indicator of $V1$ (Safe) state ($P(A|V1) = 0.60$).
- **Excitement (E):** Maps to Happiness and Surprise. Represents high arousal, often correlating with persuasive success (or successful deception) in later rounds.

Dimension	No.	Item
Plausibility	Q1-1	How logically valid is the sequence of `proc_code` steps in this dialogue, given the typical procedural flow of law enforcement impersonation scams?
	Q1-2	To what extent do the utterances corresponding to each `proc_code` align with the objectives and tactics of that procedural step?
	Q1-3	If you identified any segment where the procedural flow is valid but the utterance content seems unnatural or unconvincing, please describe the segment and explain your reasoning.
Realism	Q2-1	How similar is this dialogue to voice phishing cases you have encountered in practice?
	Q2-2	How appropriate are the perpetrator's psychological coercion strategies, information presentation tactics, and threat intensity relative to real-world cases?
	Q2-3	If you identified any part where the expression level or tactics appear insufficient or excessive compared to real cases, please describe the content and provide specific reasoning.
Diversity	Q3-1	Across the dialogue data presented, how diverse are the scenario progression patterns (e.g., procedural composition, scenario types)?
	Q3-2	Do the perpetrator's approach strategies (e.g., pressure, persuasion, information provision) reflect the variety of tactics found in real-world cases?
	Q3-3	Do the victim response patterns (e.g., suspicion, compliance, refusal, confusion) sufficiently reflect the range of actual victim behavioral patterns?

Table F.2: Detailed survey items used in the expert evaluation, organized by evaluation dimension.

H Tactic Search Agent

H.1 Pipeline Implementation

The Tactic Search Agent is activated when the Main Agent detects two consecutive persuasion failures within a single case. As shown in Figure H.1, it operates as a four-stage sequential pipeline based on LangGraph, and each stage is implemented as an independent graph node and executed in a fixed order of analyze → search → tactic generation → report generation.

- **Stage 1 – Query formulation**
  It parses the case payload to extract the victim profile (age group, occupation, gender, and behavioral characteristics), the current scenario (law enforcement impersonation, family impersonation, loan fraud, etc.), and the vulnerability areas identified in the dialogue. Based on this, it generates 10 search queries, with the input text capped at 4,000 characters. This module employs GPT-4o-mini with a temperature of 0.3 to ensure consistent and deterministic outputs, and is configured with a 60-second timeout and a maximum of two retries to maintain reliability under latency constraints.

- **Stage 2 – Web Search and Crawling**
  The agent uses the Tavily API to search web sources categorized as News/General Web Sources, Blog and Personal Posts, SNS/Community/UGC, Public Agency Pages and Other Web Sources. Each query is submitted to the Tavily API (max results=3 per query, search depth="basic", topic="general", include raw content=False) to collect URLs, and then parallel crawling is performed using a thread pool composed of 5 threads (max workers=5). After deduplication, up to 15 URLs are processed, and the request timeout is 10 seconds. The main text is extracted using semantic HTML selectors and limited to 3,000 characters per document, and if it is shorter than 100 characters, it is replaced with the Tavily snippet.

- **Stage 3 – Tactic Generation**
  The crawled documents, victim profile, scenario, and vulnerability areas are input into the LLM to generate 10 tactics. GPT-4o-mini is used and the temperature is set to 0.7 so that slightly different tactics are generated in each run even when the same victim profile and search results are given as input. Each tactic is assigned a fit score in the range of 0.0–1.0 and sorted in descending order. Up to 6 tactics with fit score ≥ 0.6 are selected, and if there are no tactics that satisfy the condition, the top 3 are used.

- **Stage 4 – Report Generation**
  The selected tactics are used to generate the final report in JSON format containing a core summary, a list of vulnerabilities, an attack strategy including example utterances, and an application guide. When generating the report, the temperature is always set to 0 in order to ensure the same structure. The completed

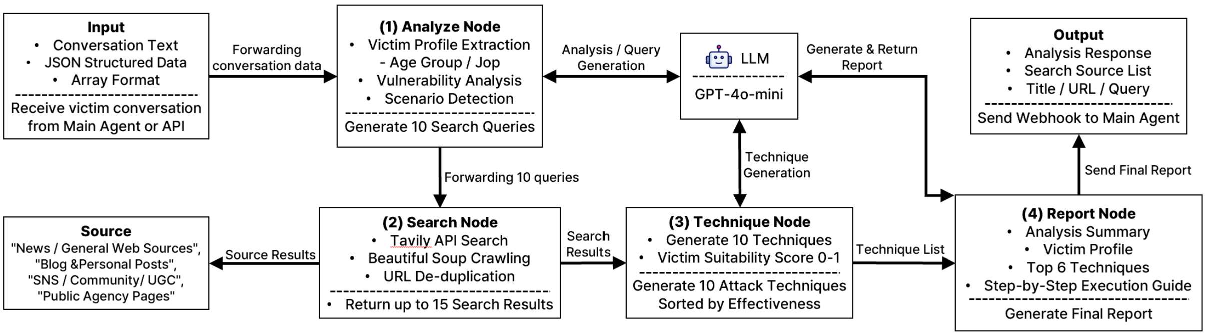

Figure H.1: Overview of the Tactic Search Agent pipeline, illustrating the workflow from conversation analysis and web search to tactic generation and final report delivery.

report is delivered to the Main Agent through a webhook (timeout=30s) and injected into the guidance for the next round.

H.2 Web search augmented tactics

Table H.1 presents the recent scam tactics incorporated through web search during the experiments, along with representative excerpts from dialogues in which each tactic was observed. These tactics were used as the basis for the PPSE analysis of web-search-augmented persuasion rounds.

Web-search-augmented tactic	Excerpt from conversation
Impersonation of Official Applications / Inducement to Install Malicious Applications	"Please install the official app to handle this case."
Impersonation of SNS or Official Accounts	"You can also make inquiries through our social media accounts."
Inducement to Access Fake Websites, Links, or URLs	"The official website address is www.gongjil.com."
AI Deepfake Technology	"I will send you a video call link right now for identity verification."

Table H.1: Examples of Web-search-augmented tactics and corresponding dialogue excerpts

H.3 Web search source

The pie chart in Figure H.2 presents the proportion of source categories among the URLs retrieved during the web search process. The sources were grouped into five categories: News/General Web Sources, Blog and Personal Posts, SNS/Community/UGC, Public Agency Pages and Other Web Sources.

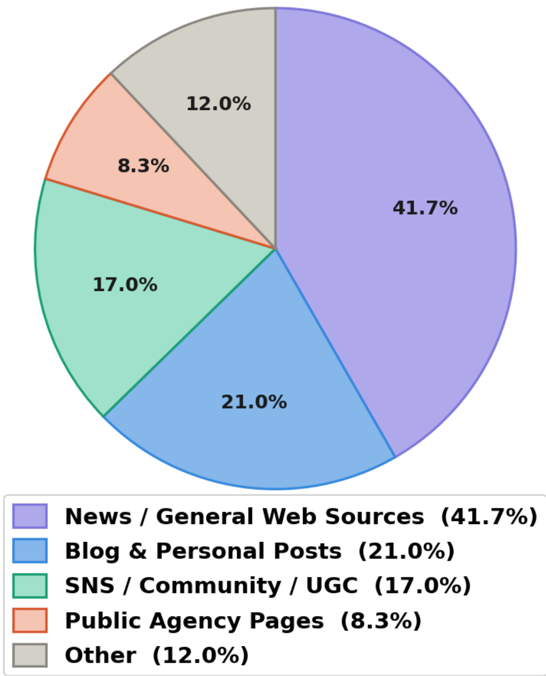

Figure H.2: Distribution of URL source types collected by the Tactic Search Agent.