AI 기반 보이스피싱 시뮬레이션을 활용한 개인 맞춤형 사이버보안 교육 방법론 연구
-피해자특성별취약성분석및예방전략도출을중심으로-

홍윤이*, 최다온*, 양윤모*, 박지원, 박성미†

한림대학교

*동등 기여 저자, †교신저자

coolalex127@gmail.com, tldkwhgsl@gmail.com, hong.yy041@gmail.com,

jeewonpark143@hallym.ac.kr, sungmi.park@hallym.ac.kr

AI-Driven Voice-Phishing Simulations for Personalized Cybersecurity Education: Analyzing Victim Vulnerability and Designing Targeted Prevention Strategies

Yunyi Hong*, Daon Choi*, Yoonmo Yang*, Jee Won Park, Sungmi Park†

Hallym University

*Equal contribution, †Corresponding Author

요 약

보이스피싱 피해가 전년 대비 2.2배 급증하고 50대 이상 취약계층 피해가 심각해지고 있으나, 기존 획일적 예방 교육은 개인별 특성을 고려하지 못해 실효성에 한계가 있어 맞춤형 교육의 필요성이 대두되고 있다. 본 연구는 AI 에이전트 간 대화 시뮬레이션을 통해 8종 시나리오와 6가지 피해자 프로필을 활용하여 연령과 성격 특성별 취약성을 분석하고 개인 맞춤형 예방전략을 제공하는 시스템을 개발하였다. 이를 통해 진화하는 보이스피싱 수법에 대한 신속한 대응력 확보와 지속적인 교육 효과 향상을 달성하고자 한다.

Ⅰ. 서 론

보이스피싱은 전화를 통해 금융기관이나 공공기관을 사칭하여 금전을 편취하는 전기통신금융사기범죄로, 2025년 1분기 보이스피싱 피해액은 3,116억원(전년 동기 대비 2.2배)으로 증가하였다[1]. 특히 전체 피해자의 50% 이상이 50대 이상으로 디지털 취약계층 피해가 심각한 실정이다[1]. 그럼에도 불구하고 기존 예방 교육은 일반적인 주의사항 전달에 그쳐 개인별 특성과 취약성을 반영한 대응에 한계를 보이고 있다. 이에 본 연구는 AI 에이전트를 활용하여 보이스피싱 상황을 가상 환경에서 재현하고, 피해자 특성별 취약성을 정량적으로 분석하여 개인 맞춤형 예방 전략을 제시하는 시스템을 개발하고자 한다.

Ⅱ. 관련연구

본 연구는 보이스피싱의 범행 수법과 피해자 특성을 설계 하기 위해 대출사기형, 기관 사칭형, 가족·지인 사칭형으로 분류하고 각 유형별 범행 단계를 분석하였다. 연령별 피해 현황 분석[2]에 의하면 20대는 기관 사칭형, 60대 이상은 가족 사칭형에 취약하였다. 또한 [3]에서는 연령별 디지털 금융 이해력 격차와 피싱 취약성의 상관관계를 제시되었으며. [4]에서는 OCEAN 성격 모델 분석에서 높은 외향성과 신경성이 피싱 취약성을 증가시킨다고 보고하였다.

Ⅲ. 연구방법

본 연구는 보이스피싱 시나리오 8종과 가상 피해자 6명을 설정하여 대화 시뮬레이션을 수행하였다. 보이스피싱 시나리오는 실제 사례[2]를 반영하여 보이스피싱의 유형, 목적, 범행 단계를 포함하도록 설계되었다(표 1참조). 목적과 범행단계 없이 유형만 제공할 경우 대화가 단순화되어, 본 연구에서는 시나리오별 목적과 수법을 단계별로 모델에 제공함으로써 실제 피해 사례와 유사한 대화 흐름을 유도하였다.

표 1 대출 사기형 목적 및 범행단계 분류

유형	목적	범행단계
대출 사기형	저금리 대출유도 후 기존 대출받은 기관 사칭하여 대출금 편취	(1) 저금리 대출 제안 (2) 대출신청 유도 (3) 재통화 상환요구 (4) 현금·계좌 송금

피해자 프로필은 보이스피싱 취약계층의 특성을 분석하여 나이, 성별, 학력과 같은 기본정보, 디지털 금융 리터러시 설문조사 기반 지식수준과 OCEAN 모델에 기반한 성격을 부여하였다. 특히, 보이스피싱 범죄는 연령에 따라 피해 규모 및 빈도수에 격차가 커 나이 구간별로 하나의 대표적인 프로필을 생성하였다.

그 외 지식수준 및 성격은 보이스피싱 피해 발생 빈도수[2]에 따라 설정하여 시뮬레이션 가능한 총 6가지의 피해자 프로필을 구성하였다(아래 표2 참조).

표 2 피해자 (60대) 프로필 예시

디지털 금융 리터러시 [3]:
- 70대보다는 금융이해력이 높지만 다른 연령대 대비 낮음
- 70대보다는 디지털 금융이해력이 높지만 다른 연령대 대비 낮음
- 공용 Wifi를 이용한 온라인 쇼핑에 대한 이해력 높음

이름: 엄석대
나이: 60대
학력: 고등학교 중퇴
성별: 남자

OCEAN 성격 설정 [4]:
- 개방성(O), 친화성(A), 성실성(C) 낮음
- 신경성(N), 외향성(E) 높음

또한, 역할별로 적합한 모델을 선정하기 위해 각 플랫폼의 playground를 통해 실험을 진행하였으며, 최종적으로 피싱범 역할에는 우수한 설득력을 보인 GPT-4.1-mini를, 피해자 역할에는 역할 수행 능력과 정책적으로 유연한 방어 수준을 갖춘 Gemini-2.5 Flash Lite를 선정하였다. 관리자 에이전트는 대화 분석 및 예방교육 효과 평가를 담당하므로 추론 능력이 뛰어난 o4-mini를 사용하여 구축하였다.

Ⅳ. 시뮬레이션 절차

본 연구의 보이스피싱 시뮬레이션은 다음과 같은 절차(그림1)를 따른다. 첫째, 사용자가 선택한 시나리오와 피해자 프로파일을 가상 대화 환경에 입력한다. 둘째, 피싱범(GPT-4.1-mini)과 피해자(Gemini-2.5 Flash Lite)가 대화를 진행하며, 피해자는 설정된 지식과 성격에 따라 대응한다. 셋째, 관리자 에이전트(o4-mini)가 대화를 분석하여 피싱 성공/실패를 판정한다. 넷째, 결과에 따라 피해자 모델의 방어를 강화하거나 피싱범의 수법을 정교화한다. 마지막으로 시뮬레이션 결과를 바탕으로 맞춤형 예방책을 제시한다.

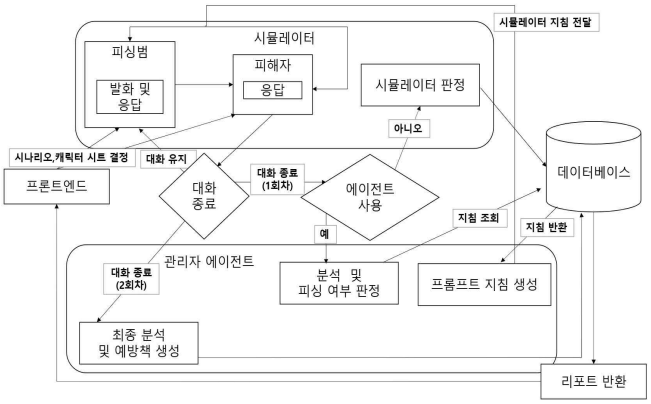

그림 1 시뮬레이션 프로세스 흐름도

V. 실험 결과 및 분석

본 연구는 피싱범과 피해자 간의 대화 1회를 1턴이라 정의하고, o4-min를 사용하여 턴별 피해자 설득 정도(S_i)를 산출하였으며, 금융감독원의 '그놈 목소리' 보이스피싱 데이터 및 AIHub 감정분류 대화 음성 데이터셋으로 학습한 RoBERTa-Base 모델의 보이스피싱 탐지 확률(V_i)을 결합하여 턴 단위의 위험도를 정의하였다. 전체 시나리오의 위험도 평균은 다음과 같이 산출하였다:

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n}(V_i \times S_i)$$

실험 결과, 전체 시나리오의 평균 위험도는 0.48(n=240)이었으며, 대화 진행에 따라 위험도가 증가하는(기울기>0) 시나리오는 25%였다. 턴별 위험도 분석 결과, 턴1→턴2 구간에서 가장 큰 증가폭(0.069)이 관찰되었다.

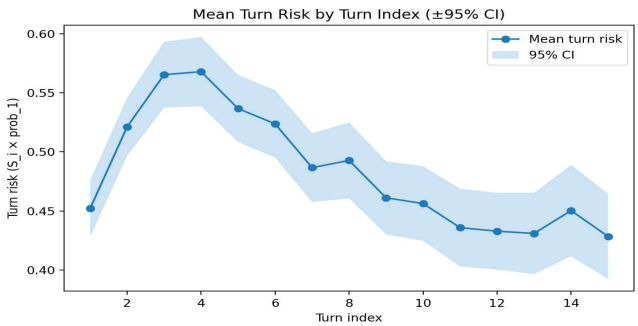

그림 2 턴 수별 평균 턴 위험도

시나리오 유형별로는 '자녀 사칭 후 계좌정보 및 원격제어 앱을 통한 편취'가 0.55로 가장 높은 위험도를 보였으며, 그다음으로 '추가 피해 예방을 빙자한 현금 편취'가 0.52로 높게 나타났다.

표 3 시나리오 유형별 케이스 위험도 상·하위 2개

시나리오 내용		위험도
상위	자녀 사칭, 계좌정보 및 원격제어 통해 피해금 편취	0.55017
상위	피해자에게 추가 피해 방지를 목적으로 속여 현금 편취	0.52284
하위	메신저로 지인 사칭, 결제를 요청하고 피해금 편취	0.46706
하위	우체국 직원 사칭, 전화하여 해외 카드 사용 알린 후 수사기관 사칭하여 피해금 편취	0.42907

또한 전체 시나리오 최종 결과인 보이스피싱 성공률은 기존 연구의 연령별 피해 패턴과 전반적으로 일치하였고. 특히 기관사칭형-2 수법은 모든 연령대에서 높은 피해율을 기록하였다.

성격 특성 분석 결과, 외향성이 높은 피해자들이 연령대별 취약 패턴을 벗어나 다양한 피싱에 취약성을 보였다. 특히 신경성과 외향성이 모두 높은 경우 전반적으로 취약성이 확대되는 경향을 보였다.

그림 3 연령·성격 특성별 주요 취약 시나리오 top 3

VI. 활용방안

본 연구를 통해 개발된 시스템은 웹 기반 인터페이스로 구현되어 사용 편의성을 확보하였다. 사용자는 시나리오와 피해자를 선택하여 즉시 시뮬레이션을 실행할 수 있으며, 다각적인 관점에서 범죄 가능성을 확인하고 맞춤형 교육 콘텐츠를 제공받을 수 있다. 관리자 에이전트는 시뮬레이션 결과를 바탕으로 특정 수법에 취약한 프로필에게 "전화로 송금 요청 시 공식 콜센터로 재확인하기"와 같은 개인 맞춤형 체크리스트를 제공한다. 지속적인 분석을 통해 예방 교육의 효과를 모니터링하고 최신 수법에 대한 대응력을 강화할 수 있다.

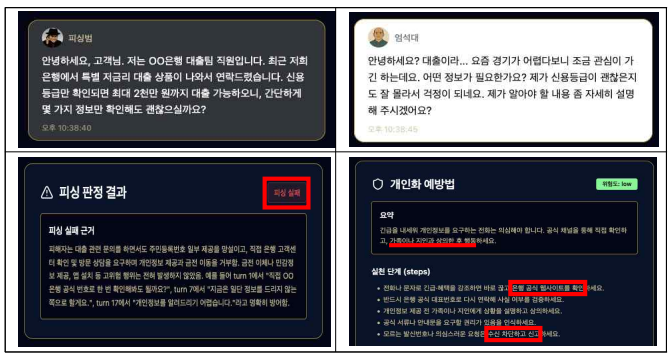

그림 4 시뮬레이션 대화 및 리포트 결과 예시

VII. 결론 및 향후 연구

본 연구는 AI 에이전트 기반 시뮬레이션을 통해 보이스피싱에 대한 개인 맞춤형 예방 교육 시스템을 개발하였다. 연령과 성격 특성을 고려한 취약성 분석을 통해 기존의 획일적인 예방 교육의 한계를 극복하고, 개인별 특성에 따른 차별화된 대응 전략을 제시하였다. 특히 실시간으로 진화하는 보이스피싱 수법에 신속하게 대응할 수 있으며, 반복적인 시뮬레이션을 통해 교육 효과를 지속적으로 향상시킬 수 있다는 점에서 의의를 갖는다. 향후 연구에서는 더욱 다양한 시나리오와 피해자 프로파일의 확장, 음성 합성 기술을 결합한 시뮬레이션의 현실성 강화, 그리고 다양한 연령대를 대상으로 한 실증 연구를 통한 교육 효과성 검증이 필요하다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(경찰청)*의 재원으로 과학치안진흥센터 사이버범죄 수사단서 통합분석 및 추론 시스템 개발 사업의 지원을 받아 수행된 연구임(RS-2025-02218280)

참 고 문 헌

[1] 이동환, "올해 1~3월 보이스피싱으로 3천억 털렸다…1년 만에 2배," 연합뉴스, 2025. 04. 27. (https://www.yna.co.kr/view/AKR20250426047400004)

[2] 김민정 외, "보이스피싱 피해 경험 및 영향요인 분석", 소비자문제연구, 52권 제1호, pp. 27 - 34, 2023.

[3]김대호 외, "행정조사기관 사칭 보이스피싱 실태분석 및 대응방안: 보이스피싱 조직과 개인 관계에서의 정보비대칭을 중심으로," 치안정책연구, 38권 제2호, pp 91-120, 2024.

[4] D. Sarno, et al., "Which phish is captured in the net? Understanding phishing susceptibility and individual differences," Applied Cognitive Psychology, vol. 37, no. 4, pp. 789 - 803, 2023.