JDFR Journal of Data Forensics Research
데이터포렌식연구, 3권 1호, 2026년 6월, 113-137면
Journal of Data Forensics Research, Volume 3, Issue 1, June 2026, Pages 113-137.
https://doi.org/10.12972/JDFR.2026.3.1.6
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

생성형 AI 스미싱 문구의 심리적 기제가 기만성에 미치는 영향: 인간과 AI의 텍스트 생성 주체 간 비교 연구

김현진[1], 김지온[2]

[1]한림대학교 심리학과 학사과정, [2]한림대학교 융합과학수사학과 교수

교신저자: 김지온, jion972@hallym.ac.kr

요약

본 연구는 생성형 AI가 작성한 스미싱 문구와 인간이 작성한 스미싱 문구가 수신자의 기만성 지각(신뢰도 및 행동의도)에 미치는 영향을 비교하고, 이러한 효과가 심리적 설득 기제(위급성·권위성·혼합형)와 연령 집단(청년층·중장년층)에 따라 어떻게 달라지는지를 검증하였다. 한국인터넷진흥원(KISA)의 실제 스미싱 데이터(2018~2021년)를 치알디니의 설득 원칙에 따라 분류한 후, AI를 통해 동일 시나리오 구조의 페어 자극물을 제작하였다. 만 19~64세 성인 126명(청년층 65명, 중장년층 61명)을 대상으로 자극물의 이월 효과를 통제한 2×3×2 혼합 설계를 적용하여 분석한 결과는 다음과 같다.

첫째, 생성주체의 주효과가 유의하여 AI가 작성한 문구(M = 2.03, SD = 0.79)가 인간이 작성한 문구(M = 1.65, SD = 0.62)보다 유의하게 높은 기만성을 유발하였다. 둘째, 심리 기제의 주효과가 유의하여 사후검정 결과 권위성 기제(M = 1.89)가 혼합형 기제(M = 1.77)보다 유의하게 높은 기만성을 보였다. 셋째, 연령의 주효과가 강력하게 나타나, 청년층(M = 2.01)이 중장년층(M = 1.66)보다 모든 조건에서 더 높은 기만 취약성을 보였다. 넷째, 생성주체와 연령 간의 상호작용 효과가 유의하게 나타났다. 청년층과 중장년층 모두 인간 작성 문구보다 AI 작성 문구에서 유의하게 높은 기만성을 보였으나, 인간 대비 AI 문구의 기만성 증가 폭이 중장년층(ΔM = 0.22)보다 청년층(ΔM = 0.54)에서 더 크게 나타났다. 본 결과는 생성형 AI가 인간이 작성한 사기 문구보다 기만성 점수를 유의하게 높이며, 디지털 친숙도가 높은 청년층이 오히려 AI 기반 기만에 더 취약할 수 있음을 시사한다. 향후 AI를 활용한 범죄에 대응할 수 있는 기술개발과 잠재적 피해자인 청년층 대상으로 피싱범죄를 예방할 수 있는 다양한 정책이 마련되어야 할 것이다.

주제어

생성형 AI, 스미싱, 사회공학적 공격, 기만성, 심리적 설득 기제, 디지털 기만

Received: June 08, 2026
Revised: June 30, 2026
Accepted: June 30, 2026
Published: June 30, 2026

Original Article

Deception impact of psychological mechanisms in generative AI smishing: Comparative study of texts generated by humans and AI

Hyeonjin Kim[1], Jion Kim[2]

[1]Undergraduate Student, Department of Psychology, Hallym University, Republic of Korea
[2]Professor, Department of Forensic Science, Hallym University, Republic of Korea

Corresponding Author: Jion Kim, jion972@hallym.ac.kr

ABSTRACT

This study compared the effects of smishing messages written by generative AI versus humans on recipients' deception perception (trust and behavioral intention). Furthermore, it examined how these effects vary by psychological persuasion mechanism (urgency, authority, and mixed) and age group (younger vs. middle-aged adults). After classifying authentic smishing data from the Korea Internet and Security Agency (KISA; 2018–2021) according to Cialdini's principles of persuasion, paired stimuli with identical scenario structures were generated using AI. A 2×3×2 mixed design, controlling for carryover effects, was applied to 126 adults aged 19–64 years (younger, n = 65; middle-aged, n = 61). First, a significant main effect of the generation source emerged: AI-generated messages (M = 2.03, SD = 0.79) elicited a higher deception perception than human-authored messages (M = 1.65, SD = 0.62). Second, the authority mechanism (M = 1.89) produced significantly higher deception than the mixed mechanism (M = 1.77). Third, younger adults (M = 2.01) showed higher vulnerability to deception than middle-aged adults (M = 1.66) across all conditions. Notably, a significant source × age interaction emerged: the increase from human to AI messages was larger among younger (ΔM = 0.54) than among middle-aged adults (ΔM = 0.22). These findings suggest that generative AI significantly increases deception perception beyond human-authored fraudulent messages, and that younger adults with high digital fluency may paradoxically be more vulnerable to AI-based deception. Future studies should develop technologies to counter AI-enabled crime, and establish policies to prevent smishing by targeting younger adults.

KEYWORDS

Generative AI, Smishing, Social Engineering, Deception, Psychological Persuasion Mechanisms, Digital Deception

I. 서론

1.1. 연구 배경 및 필요성

최근 정보통신기술의 발전과 대규모 언어 모델의 상용화는 인간에게 편의성을 제공하는 동시에, 사이버 범죄 생태계의 패러다임 변화를 야기하고 있다. 한국인터넷진흥원(KISA)의 『2025년 하반기 스팸 유통현황』에 따르면, 1인당 월평균 휴대전화 문자스팸 수신량은 2024년 상반기 11.59통에서 2025년 하반기 2.74통으로 급감하며 2020년 이후 최저치를 기록하였다[1]. 그러나 이러한 표면적인 스팸 유통량의 감소가 사이버 위협의 실질적인 완화를 의미하는 것은 아니다. 인공지능 기술의 발전은 기존 위협의 규모를 확장하는 데 그치지 않고, 위협의 본질적 성격 자체를 변화시키는 특성을 지닌다[2]. 즉, 기존의 단순 대량 발송 방식이 차단 체계에 의해 임계점에 도달함에 따라, 범죄 조직들은 노동력을 AI로 대체하여 더 표적화되고 고도화된 스미싱으로 공격 기법을 진화시키고 있다[2].

이와 같은 위협의 질적 변화는 조직범죄의 경제적 구조 및 생성형 AI의 서비스화와 연결되어 있다. 유럽형사경찰기구(Europol)의 『인터넷 조직범죄 위협 평가(IOCTA) 2024』에 따르면, 온라인 사기는 전례 없는 규모와 정교함을 보이며 가장 심각한 주요 위협 중 하나로 지속되고 있다[3]. 과거 보이스피싱 및 스미싱 조직은 대규모 콜센터 상담원의 배치, 인건비 소요, 대포통장 수급, 그리고 수사기관에 의한 체포 리스크 등으로 인해 고비용 구조를 감수해야만 했지만, 최근 범죄자들이 생성형 AI를 도구로 채택하여 사회공학적 공격을 자동화하고 개인화하기 시작하면서 사이버 범죄 생태계는 변화를 맞이하였다[4]. 전문 기술 없이도 생성형 AI를 활용한 맞춤형 피싱 이메일 생성, 악성 대규모 언어모델(LLM)의 다크웹 유통 등이 보고되고 있으며[3], 이러한 도구들이 이른바 '서비스형 범죄(Crime as a Service, CaaS)' 형태로 제공됨에 따라 기술적 전문성이 낮은 행위자조차 손쉽게 정교한 사기 텍스트를 제작하여 범죄에 가담할 수 있는 환경이 조성되었다[5]. 이는 곧 과거의 고비용·고위험 사기 구조가 저비용·고효율의 산업화된 형태로 완전히 재편되고 있음을 의미한다.

AI 기술의 도입은 범죄의 한계 비용을 절감시키는 반면, 방어자의 대응 비용은 기하급수적으로 증가시키는 경제적 비대칭성을 초래한다[6]. 향후 고도화된 자동화 기만 공격이 대규모로 양산될 경우, 방어 진영은 이를 탐지하고 차단하기 위해 보안 인프라 확장 및 시스템 운영에 있어 막대한 자원 투입을 요구받을 가능성이 높다[6]. 실제로 범죄 조직들은 다크웹 등에서 악의적으로 변형된 AI(예: FraudGPT, WormGPT)를 월정액 형태로 구독하거나 오픈소스 LLM을 활용하여, 맞춤형 사기 텍스트를 기계의 속도로 대량 생산할 수 있는 기술적 기반을 갖추어가고 있다[3]. 특히 이러한 자동화 시스템은 범죄 기획부터 실행에 이르는 전 과정의 공격 주기를 압축시키는 잠재력을 지닌다[4]. 결과적으로, 어색한 문법, 맞춤법 오류, 조잡한 문맥 등 과거 피해자들이 위험을 직관적으로 식별하던 전통적인 보안 지표들은 생성형 AI의 정교한 언어 생성 능력에 의해 식별이 현저히 어려워지는 상황으로 전환될 위기에 처해 있으며, 이는 향후 AI의 디지털 기만 성공률 증가로 이어질 수 있다[7]. 따라서 범죄에 이용될 수 있는 고도화된 AI 스미싱 문구가 인간의 인지적 취약성에 미치는 영향을 실증적으로 검증하여, 새로운 위협에 대한 선제적인 대응책 마련이 시급하다.

1.2. 문제 제기 및 연구 목적

인공지능 개발자들은 AI의 범죄 악용을 방지하기 위해 인공지능이 인간의 윤리적 가치, 의도, 그리고 안전 기준에 맞게 행동하도록 훈련시키는 정렬 기술과 안전장치를 구축하여 범죄용 스

크립트 생성을 엄격히 통제하고 있다. 그러나 보안 취약점을 우회하는 탈옥 기법이나 시스템에 정교한 역할을 부여하는 페르소나(역할극) 프롬프트 기술의 발전에 따라 공격 범죄자들이 AI 모델에게 가상의 역할을 부여함으로써, 이러한 방어선을 무력화할 가능성이 있다[8].

이러한 탈옥 기법을 통해 생성형 AI가 만들어낸 스미싱 문구는 기존의 사회공학적 공격 양상과 비교하여 어떤 차이를 가지는가? 과거 전통적인 스미싱 문구는 고정된 키워드나 정형화된 템플릿에 의존하여 상대적으로 단순하고 간결한 형태를 띠는 경우가 많았다[9]. 반면 고도화된 생성형 AI는 언어적 뉘앙스와 맥락적 개인화를 정밀하게 제어할 수 있으며, 권위성과 긴급성 등 사회공학적 공격의 핵심 심리 기제를 텍스트에 효과적으로 구현하여 인간의 인지적 편향을 자극할 수 있다[7]. 한편, 기만에 대한 인간의 인지 및 사기 취약성은 연령이나 성격 특성 등 심리적 요인에 따라 유의미한 차이를 보이게 된다[10]. 그러나 현재까지의 보안 연구는 사기성 URL 차단이나 악성 앱 탐지 등 기술적 방어 체계에 치우쳐 있으며[9], 공격의 대상자인 인간이 느끼는 기만과 심리적 메커니즘에 대한 실증적 분석은 미진한 실정이다[7].

따라서 본 연구는 생성형 AI 스미싱 문구에 내포된 심리적 기제가 피해자의 기만성 지각에 미치는 영향을 실증적으로 규명하고자 한다. 특히, 텍스트 생성 주체(인간 vs 생성형 AI)에 따른 심리적 조작 기제(위급성, 권위성, 혼합형)의 발현 양상을 비교 분석하고, 이러한 기만 효과가 방어기제가 상이한 연령 집단(청년층 vs 중장년층)에 따라 어떻게 조절되는지 2×3×2 혼합 요인 설계를 통해 검증한다. 다만, Cialdini의 희소성 원리에서 파생되는 시간적 압박 기제의 경우 스미싱 범죄 맥락에서 현존하는 급박한 위험 상황을 고려하는 것이 더 적합하다고 판단되어 본 연구에서는 '위급성' 관점에서 살펴보기로 한다. 본 연구는 실제 신고된 스미싱 데이터(KISA)와 생성형 AI가 산출한 텍스트를 정량적으로 비교·분석함으로써, AI 기반 기만 텍스트의 탐지 및 디지털 증거 분석 체계의 고도화에 활용될 수 있는 언어적·심리적 판별 단서를 도출한다. 이를 통해 고도화된 AI 기만에 대응하기 위한 방어 AI의 언어적 탐지 가이드라인을 제시하고, 연령 맞춤형 예방 교육의 토대를 마련하는 데 그 목적이 있다.

II. 이론적 배경 및 관련 연구

2.1. 이론적 배경

2.1.1. 스미싱의 정의

스미싱(Smishing)이란 문자메시지(SMS)와 피싱(Phishing)의 합성어로, '무료쿠폰 제공', '돌잔치 초대장', '모바일 청첩장' 등을 내용으로 하는 문자메시지를 발송하여, 피해자가 문자메시지 내 인터넷주소(URL)를 클릭하면 악성코드가 스마트폰에 설치되어 피해자 모르는 사이에 소액결제 피해가 발생하거나 개인·금융정보를 탈취하는 범죄수법이다[11]. 한국인터넷진흥원(KISA)의 침해사고 유형에 따르면, 스미싱은 단순한 악성코드 유포를 넘어 타겟의 상황적 맥락을 파고들어 즉각적인 행동(URL 클릭 등)을 유도한다는 점에서 이메일 기반 피싱과 구별되는 고유한 매체적 특성을 지닌다[1].

이러한 전통적 스미싱의 범행 수법은 범행 준비와 실행에 이르는 체계적인 스크립트 구조를 따른다. 최관과 김민지(2016)의 수법 분석에 따르면, 범죄 조직은 타겟의 정보를 불법 수집한 뒤 경찰, 검찰, 금융기관 등을 사칭하여 피해자의 두려움을 자극하는 보호형이나, 도로교통법 위반 과태료 청구 등 즉각적인 조치를 요구하는 협박형 시나리오를 주로 기획해왔다[9]. 이는 스미싱이 단순한 기술적 해킹을 넘어, 권위와 위급성이라는 설득의 심리학적 기제(Psychological Principles of Persuasion)를 텍스트에 투영함으로써 수신자의 합리적 정보

처리를 방해하는 전형적인 디지털 기만 범죄임을 시사한다[7,9].

2.1.2. 인공지능의 악용과 사이버 위협 패러다임의 변화

인공지능 기술의 악용 가능성은 기술 발전 초기부터 핵심적인 보안 이슈로 제기되어 왔다. Brundage 등(2018)은 AI가 기존 사이버 위협의 구조를 재편할 것임을 예견하며, '위협 규모의 확장', '신규 위협의 도입', '위협 성격의 본질적 변화'라는 세 가지 축으로 구조화하였다. 이들은 과거 숙련된 공격자의 지적 노동력이 필수적이었던 사회공학적 기만 과정이 AI를 통해 자동화됨으로써 사이버 범죄의 실행 한계 비용이 극단적으로 감소할 것임을 경고하였다[2].

이러한 예측은 대형 언어 모델(LLM)을 중심으로 한 현대 생성형 AI 환경에서 현실화되는 추세이다. 기존의 피싱 및 스미싱 공격은 공격자의 수작업이나 번역 시스템에 의존하였기 때문에 문법적 오류나 문맥상 어색함 등 수신자가 위협을 직관적으로 탐지할 수 있는 언어적 단서가 뚜렷하게 존재하였다. 반면, 생성형 AI는 타겟의 특성에 맞춘 정교한 문맥적 개인화와 자연스러운 언어적 뉘앙스를 유연하게 구현할 수 있어, 향후 인간 범죄자의 기만 역량을 모방하거나 이를 우회하는 고도의 신뢰성 높은 텍스트를 양산할 잠재력을 지닌 것으로 평가된다[7]. 이는 곧 인공지능의 악용이 향후 사이버 위협의 패러다임을 근본적으로 변화시킬 수 있음을 시사하며, 새로운 디지털 기만 위협에 대한 선제적 경계와 대비가 필수적임을 의미한다.

2.1.3. 사이버 범죄의 경제학: 기만의 산업화와 서비스형 범죄(CaaS)

Bailo 등(2026)에 따르면, 생성형 AI의 도입은 범죄 생태계를 기만의 산업화(Industrialisation of Deception) 단계로 진입시켰다[4]. 과거의 범죄 구조와 달리 현대 사이버 범죄 조직들은 기술을 직접 개발하기보다 다크웹을 통해 고도화된 도구를 구매·유통하는 분업화된 구조를 취한다. 유로폴의 공식 문서인 EU-SOCTA-2025에서도 범죄 생태계가 서비스형 범죄 체계로 완전히 재편되었음을 공표한 바 있다[5].

범죄 조직들은 기술적 전문성이 부족하더라도 다크웹에서 범죄 인프라를 손쉽게 이용할 수 있다. 유럽형사경찰기구(Europol)의 위협 평가 보고서 등에서 보고된 FraudGPT와 WormGPT는 개발사의 안전 제약이 전혀 없는 악용 가능한 LLM 클론으로, 월정액 구독 모델 형태로 범죄자들에게 유통되고 있다. 범죄자들은 고비용을 들여 시나리오 작가를 고용하거나 콜센터를 운영할 필요 없이, 단지 몇 단계의 쿼리 입력만으로 타겟의 프로필과 결합된 극도로 정교한 사회공학적 사기 스크립트를 추출해 낼 수 있게 되었다[3]. 즉, 공격의 한계 비용은 감소한 반면, 산출되는 기만 텍스트의 정교함과 맞춤화 수준은 비약적으로 향상되어 범죄의 효율성이 높아질 위험이 있다.

이러한 기만의 산업화는 사이버 보안 구조에 경제적 비대칭성을 초래할 우려가 제기된다. 현행 스미싱 방어 체계는 주로 기 수집된 악성 URL이나 정형화된 범죄 키워드를 차단하는 블랙리스트 기반의 사후 대응에 의존하고 있다[9]. 그러나 생성형 AI가 무한한 변형과 맥락적 회피 능력을 지닌 텍스트를 대량으로 양산할 경우, 기존의 패턴 매칭 방어망은 한계에 직면할 위험성이 존재한다. Chlasta(2025)의 연구는 자동화를 활용한 공격은 실행 비용을 절감시키는 반면, 이를 방어하기 위한 탐지에 훨씬 더 막대한 자원을 소모하게 될 위험성을 경고한다. 공격자는 저비용으로 다량의 변형 스크립트를 생성하기만 하면 되지만, 방어자는 통신망을 지나는 방대한 양의 정상 데이터 속에서 악성 텍스트를 분석하고, 오탐을 방지하기 위해 대규모 연산 인프라를 활용하기 때문이다[6]. 결론적으로 이를 선제적으로 방어하기 위해선 기술적 체계의 확장뿐만 아니라, 기만의 최종 대상인 인간의 심리적 취약성을 파악하고 인지적 방어력을 높이는 실증 연

구가 병행되어야 한다.

2.1.4. LLM 방어 메커니즘의 무력화: 탈옥 및 페르소나 공격 기술

생성형 인공지능 개발사들은 상용 LLM의 악용을 막기 위해 엄격한 알고리즘과 가드레일 시스템을 적용하고 있다. 예컨대 피싱 템플릿, 악성 스크립트, 혹은 타인을 기만하기 위한 사기성 문구 생성을 요청할 경우, 모델은 거부 반응을 출력하도록 설계되어 있다. 그러나 인공지능 보안 진영과 범죄 진영 간의 창과 방패의 싸움에서, 이러한 방어 메커니즘을 우회하는 탈옥(Jailbreak) 기법이 지속적으로 고도화되고 있다[8].

최근 연구인 Zhang 등(2025)의 연구에 의하면, 모델에게 특정한 인격이나 사회적 역할을 부여하는 페르소나 프롬프팅 기법은 LLM의 내장 안전장치를 우회하는 가장 치명적인 수단 중 하나로 지목된다. 이 연구에서는 유전자 알고리즘을 기반으로 자동 생성된 페르소나 프롬프트가 다수의 상용 LLM의 보안 거부율을 무려 50%에서 70%까지 감소시켰음을 증명하였다. 모델에게 특정 직업군이나 우회적 역할을 부여하는 정교한 페르소나를 주입할 경우, 내장된 안전 정렬 레이어가 이를 정상적인 문맥으로 오인하여 악의적인 사회공학적 스크립트를 제한 없이 배출하게 된다. 또한 이러한 페르소나 공격은 기존의 단순 우회 규칙 공격과 결합될 때 10~20%의 추가적인 탈옥 성공률 상승을 유발하는 것으로 확인되어, 범죄자들에게 매우 강력한 공격 인프라로 기능하고 있다[12]. 이는 곧 본 연구에서 생성형 AI의 고도화된 기만 효과를 실증하기 위해, 역으로 범죄자들의 페르소나 프롬프팅 기법을 연구 방법론에 차용하여 실험 자극물을 설계해야 할 기술적 당위성을 제공한다.

2.1.5. 스마트폰 의존성과 기만의 관계

생성형 AI를 통해 정교화된 스미싱 문구의 기만 효과는 공격 대상자인 인간의 매체 의존 특성 및 인지적 정보 처리 방식과 긴밀하게 상호작용한다. 박광순(2025)의 실증 연구에 따르면, 스마트폰 이용자 집단 중에서도 대학생을 비롯한 청년층은 모바일 환경이 제공하는 즉시성과 도구적 편의성에 대한 의존도가 매우 높은 특성을 보인다. 이러한 높은 매체 의존 경향은 스미싱 메시지에 대한 위험 지각 구조를 왜곡하여 기만 성공률을 높이는 인지적 취약성을 형성한다[13].

이러한 현상은 인지 심리학의 이중 처리 이론(Dual-Process Theory)을 통해 설명될 수 있다. Williams와 Polage(2019)의 연구에 따르면, 기만 메시지가 유발하는 권위성 및 위급성 압박은 모바일 매체의 즉시성과 결합될 때 수신자의 인지적 부하를 증가시킨다. 그 결과, 수신자는 메시지의 진위를 꼼꼼히 검증하는 심층적 사고를 생략한 채, 휴리스틱에 기반한 직관적이고 자동적인 주변부 사고로 이행하게 된다[14]. 즉, 사용자가 정보 습득을 목적으로 스마트폰에 과도하게 의존할 때 생성형 AI의 정교한 언어적 조작 기제와 마주하면, 위험 신호를 필터링하지 못하고 사기 행동 의도로 이행될 잠재적 위험성이 높아진다.

한편, 기만에 대한 취약성은 단순히 스마트폰 의존도에 의해서만 결정되는 것이 아니며, 연령 및 심리적 요인에 따라 유의미한 발현 차이를 나타낸다. 심예은과 최은실(2023)에 따르면, 청년층이 모바일 기기의 높은 접근성 및 과의존으로 인해 기만 환경에 빈번하게 노출되어 취약성을 띠는 반면, 중장년층은 연령 증가에 따른 전두엽 기능 변화나 특정 사회공학적 기제에 대한 맹신 등 상이한 심리적 메커니즘에 의해 사기 피해에 노출되는 경향이 있다[10].

따라서 본 연구는 생성형 AI 스미싱 문구의 조작 기제가 인간의 심리 메커니즘과 결합하여 발현되는 기만 효과를 단일 세대가 아닌 방어 기제가 상이한 두 연령 집단(청년층 vs 중장년층)으로 교차 비교하여 실증 검증하고자 한다.

2.2. 선행연구 분석

최근 대규모 언어 모델(LLM)의 고도화에 따라, 인공지능이 생성한 텍스트가 사회공학적 공격에 미치는 영향을 실증하려는 연구가 활발히 진행되고 있다. 대표적으로 Heiding 등(2024)은 실제 인간 피험자를 대상으로 자동화된 스피어 피싱(Spear-phishing) 캠페인을 설계하여, LLM이 생성한 피싱 문구가 인간 전문가와 대등한 수준의 클릭 유도율(약 54%)을 달성함을 실증하였다[15]. 이어 Schmitt와 Flechais(2024)의 체계적 문헌고찰에 따르면, 범용 LLM이 생성한 피싱 텍스트에서는 맞춤법 오류나 어색한 문맥 등 전통적인 적신호(Red flags)가 사실상 소멸되었으며, AI가 언어적 뉘앙스를 조작하여 기만 성공률을 증가시킬 잠재력이 있음이 규명되었다[7].

그러나 이러한 유의미한 발견에도 불구하고, 기존의 선행연구들은 방법론적, 환경적 측면에서 다음과 같은 세 가지 한계점을 지닌다. 첫째, 분석 매체의 편향성이다. 대다수의 연구가 이메일 기반의 피싱 환경에 국한되어 있어, 즉각적인 상호작용과 고도의 텍스트 압축이 요구되는 모바일 스미싱 환경의 특수성을 반영하는 데 제한이 있었다. 둘째, 심리적 기제의 블랙박스화이다. 기존 연구들은 'AI가 작성한 글이 더 잘 속인다'는 결과론적 클릭률 비교에 집중했을 뿐, 텍스트 내부에 내재된 어떤 구체적인 심리적 조작 기제(예: 권위성, 위급성 등)가 피해자의 인지적 취약성을 관통했는지 분리하여 측정하는 데까지 나아가지 못했다. 셋째, 피해자의 인지적·인구통계학적 특성 간과이다. 연령이나 정보 보안 인식 수준에 따라 텍스트 기만성을 수용하는 방어 기제가 상이함에도 불구하고, 피험자 집단을 단일한 기준으로 묶어 분석함으로써 기만 효과의 조절 변인을 충분히 다루지 못한 측면이 있다.

따라서 본 연구는 선행연구들의 한계를 극복하기 위해 다음과 같은 차별화된 연구 모형을 설계하였다. 첫째, 분석의 환경을 이메일에서 실생활 위협 민감도가 가장 높은 모바일 스미싱 환경으로 전환하여 생태학적 타당성을 확보한다. 둘째, Cialdini(2021)의 설득 원리[16]를 바탕으로 스미싱 텍스트에 내재된 세부 심리 기제를 권위성, 위급성, 혼합형으로 조작적 정의하고 분리하여 측정한다. 셋째, 텍스트 생성 주체(인간 vs AI)와 심리적 기제가 피해자의 연령 집단(청년층 vs 중장년층)에 따라 어떻게 다르게 수용되는지 다차원 혼합 요인 설계를 통해 실증한다. 이를 통해 본 연구는 결과론적인 위험성 경고를 넘어, AI 기만에 대항하기 위한 맞춤형 보안 가이드라인과 연령별 방어 전략의 토대를 제공하고자 한다.

2.3. 사회공학 공격의 심리적 설득 기제

2.3.1. 디지털 기만과 심리적 조작 기제

스미싱 범죄의 핵심 성공 요인은 기술적 정교함 자체보다 수신자로 하여금 즉각적인 행동(링크 클릭, 앱 설치 등)을 유도하는 심리적 조작 기제의 효과적인 발현에 있다. 인간은 인지적 구두쇠로서 복잡한 상황에서 체계적 정보 처리(Systematic Processing) 대신 휴리스틱 정보 처리(Heuristic Processing)에 의존하는 경향이 있다. 사이버 범죄자들은 이러한 인간의 취약성을 공략하기 위해 Cialdini의 설득의 원칙 중 권위성과 희소성, 그리고 이로부터 파생되는 긴급성 등을 핵심 기제로 사용한다[16].

Schmitt와 Flechais(2024)의 연구에 따르면, 생성형 AI는 이러한 심리적 영향력 원리를 정밀하게 제어하고 최적화하는 데 탁월한 성능을 보인다[7]. AI는 단순히 "돈을 이체하라"는 식의 직설적인 요구 대신, 공공기관을 사칭하여 법적 불이익을 예고하는 권위성을 부여하거나, 제한 시간 내에 조치하지 않으면 계좌가 동결된다는 긴급한 상황 맥락에 맞추어 문장을 정교하게 조

합해 낸다. 인간 공격자의 경우 주관적 경험과 제한된 언어 능력에 의존하여 심리 자극 문구를 작성하지만, 거대한 말뭉치 데이터를 학습한 AI는 인간이 거부감을 느끼지 않으면서도 방어 본능을 무력화할 수 있는 최적의 심리적 취약점을 정교하게 저격할 수 있다[7]. 본 연구는 이러한 이론적 뼈대 위에서 생성형 AI가 산출한 심리적 기제의 정량적 효과를 인간 작성 문구와의 비교 분석을 통해 포렌식 및 사회공학 방어적 관점에서 실증하고자 한다.

2.3.2. 권위성(Authority)

사회공학 공격에서 권위성 기제는 수신자로 하여금 발신자를 합법적이고 공신력 있는 주체로 인식하게 만들어 의심을 무력화하는 심리적 트리거이다. 로버트 치알디니(Cialdini)의 설득 원리에 기반하면, 인간은 사회적으로 공인된 권위체(사법·행정기관, 금융기관 등)의 요구에 직면했을 때 깊은 인지적 여과 과정 없이 순응하는 경향이 있다[16]. 생성형 AI는 방대한 말뭉치를 바탕으로 이러한 권위적 언어 규범과 공문서식 뉘앙스를 정교하게 모방한다. 특히 Zhang 등(2025)의 연구에서 입증된 페르소나 프롬프팅(Persona Prompting) 기법을 악용할 경우[12], AI는 특정 사법 기관이나 금융 보안 담당자의 인격을 정교하게 모사하여 기술적 위협 없이도 수신자의 행정적·제도적 의무감을 자극해 기만 성공률을 증가시킨다.

2.3.3. 위급성(Urgency)

위급성 기제는 수신자에게 시간적 압박이나 즉각적인 물리적·금전적 손실 가능성을 인식하게 함으로써, 정보를 충분히 검토하고 분석하는 체계적 정보 처리를 제한하는 심리적 기제이다[16]. 인간은 무언가를 얻는 것보다 이미 가진 것을 잃는 것을 더 두려워하는 손실 회피 성향을 지닌다. 따라서 계좌 동결이나 자산 소멸과 같이 당장 행동하지 않으면 기회나 통제권을 상실할 수 있다는 위협(시간제한 등)에 노출되면, 불안감이 극대화되어 이성적인 판단을 멈추고 직관적이고 무비판적인 사고(휴리스틱)에 의존하게 된다. Schmitt와 Flechais(2024)에 따르면, 생성형 AI는 이러한 인간의 심리적 취약점을 파고들어, 타겟의 개인정보와 상황 맥락에 맞추어 시간적 데드라인과 위협의 강도를 가장 효과적으로 조작하는 데 탁월한 성능을 발휘한다[7].

2.3.4. 혼합형(Authority + Urgency)

혼합형 기제는 권위성이 가지는 제도적 신뢰감과 위급성이 유발하는 심리적 압박감을 하나의 텍스트 내에 결합한 복합적 사회공학 전략이다. 최관과 김민지(2016)의 스미싱 스크립트 분석에 따르면, 전통적 사이버 범죄 생태계에서도 단일 기제보다는 신뢰성 있는 기관을 사칭함과 동시에 즉각적인 처분을 경고하는 방식이 피해자의 방어 기제를 무력화하는 가장 치명적인 수단으로 활용되어 왔다[9]. 나아가 Heiding 등(2024)은 AI가 이러한 복합 기제를 자동화하여 결합했을 때 인간 전문가와 대등하거나 이를 초월하는 54%의 기만 성공률을 기록함을 실증하였다[15]. 이는 수신자의 인지적 방어 기제를 다각도로 자극하여 심리적 무력화를 가장 강력하게 유도하는 형태이다.

따라서 본 연구는 앞서 도출된 세 가지 핵심 심리 기제(권위성, 위급성, 혼합형)를 독립 변인으로 철저히 통제하여 실증 실험을 설계한다. 특히 생성형 AI가 산출한 기만 텍스트가 인간 해커가 작성한 텍스트에 비해 이러한 심리적 트리거를 얼마나 더 위협적으로 발현시키는지, 그리고 이 기만 효과가 인지적 방어 기제가 상이한 연령 집단(청년층 vs 중장년층)에 따라 어떻게 조절되는지 다차원적으로 검증하고자 한다.

III. 연구 방법론

3.1. 연구 가설

이론적 논의 및 선행연구 검토 결과를 바탕으로, 본 연구는 생성형 AI의 언어적 조작이 수신자의 기만성 지각에 미치는 영향을 실증하기 위해 다음과 같은 가설을 설정하였다.

가설 1. 생성형 AI가 작성한 스미싱 문구는 인간이 작성한 스미싱 문구보다 수신자에게 더 높은 수준의 기만성 지각(신뢰도 및 행동의도)을 유발할 것이다.

가설 2. 스미싱 문구에 내재된 심리적 설득 기제(위급성, 권위성, 혼합형)의 유형에 따라 수신자의 기만성 지각 수준에 유의미한 차이가 있을 것이다.

가설 3. 텍스트 생성 주체(인간 vs. AI)와 심리적 설득 기제(위급성, 권위성, 혼합형) 간에는 수신자의 기만성 지각에 미치는 유의미한 상호작용 효과가 존재할 것이다.

가설 4. 수신자의 연령 집단(청년층 vs. 중장년층)에 따라 스미싱 문구에 대한 전반적인 기만성 지각 수준에 유의미한 차이가 있을 것이다.

가설 5. 수신자의 연령 집단과 문구 생성 주체 간 기만성 지각에 미치는 유의미한 상호작용 효과가 존재할 것이다.

3.2. 연구 설계 및 참여자

본 연구는 텍스트 생성 주체, 심리 기제, 연령 집단이 기만성 지각에 미치는 영향을 검증하기 위해 2(생성주체: 인간 vs. AI; 참가자 내) × 3(심리 기제: 위급성 vs. 권위성 vs. 혼합형; 참가자 내) × 2(연령 집단: 청년층 vs. 중장년층; 참가자 간)의 혼합 요인 설계(Mixed-design)를 적용하였다. 특히, 동일한 범죄 시나리오에서 파생된 인간 작성본과 AI 작성본을 한 명의 참여자가 연속으로 평가할 경우, 앞서 접한 문항이 뒤이은 문항의 판단에 영향을 미치는 이월 효과가 발생할 수 있다. 이를 통제하기 위해 전체 자극물을 두 세트(A/B)로 분할하여, 한 참여자가 동일 시나리오의 인간본과 AI본을 동시에 평가하지 않도록 교차 할당하는 역균형화 기법을 도입하였다. 연구 참여자는 온라인 설문 진입 시 휴대전화 번호 마지막 자리(홀수/짝수)를 기준으로 A집단 또는 B집단에 체계적 무작위 할당(Systematic random assignment) 방식으로 배치되어 집단 간 분리가 이루어졌다.

<Table 1> 연구 참여자의 인구통계학적 특성(N=126)

구분	항목	빈도(명)	비율(%)
성별	여성	73	57.9
	남성	53	42.1
연령 집단	청년층(만 19~39세)	65	51.6
	중장년층(만 40~64세)	61	48.4
최종 학력	고등학교 졸업	68	54.0
	대학교(전문대 포함) 졸업	42	33.3
	대학원 졸업(석사)	13	10.3
	대학원 졸업(박사 이상)	3	2.4
역균형화 집단	A(홀수)	72	57.1
	B(짝수)	54	42.9

본 연구의 최종 분석에는 만 19세 이상 64세 이하의 성인 남녀 126명이 포함되었다. 연령 집단별로는 청년층(만 19~39세) 65명, 중장년층(만 40~64세) 61명으로 구성되었으며, 참여자의 인구통계학적 특성은 <Table 1>과 같다. 자극물 분할에 따른 역균형화 집단은 홀수(집단 A) 72명, 짝수(집단 B) 54명으로 배정되었다. 표본 크기의 통계적 타당성은 G*Power 3.1 프로그램을 활용하여 사전 산출되었다. 혼합 분산분석(Mixed ANOVA)을 기준으로 유의수준($\alpha = .05$), 검정력($1-\beta = .80$), 중간 수준의 효과 크기($f = .25$)를 설정한 결과 최소 요구 표본이 도출되었으나, 자극물 교차 할당에 따른 오차 통제 및 온라인 설문의 특성상 발생할 수 있는 10~20%의 불성실 응답률을 보수적으로 반영하여 최종 목표 표본을 120명 이상으로 설정하였다. 아울러, 대규모 언어 모델(LLM)을 활용한 자동화 피싱 캠페인의 기만성을 실증한 Heiding 등(2024)[15]의 선행 연구에서 101명의 피험자를 통해 유의미한 통계적 검정력을 확보한 선례를 표본 크기 설정의 실증적 타당성 근거로 참조하였다. 한편 휴대전화 번호 끝자리를 기준으로 실시한 무작위 배정의 특성상 두 집단의 인원수가 완전히 균등하게 나뉘지는 않았다. 다만 본 연구의 주된 비교가 한 참여자가 인간 작성 문구와 AI 작성 문구를 모두 평가하는 참가자 내 분석을 중심으로 이루어지므로, 집단 간 인원수 차이가 결과에 미치는 영향은 크지 않을 것으로 판단된다. 향후 후속 연구에서는 두 집단의 인원이 고르게 배분되도록 할당 방식을 보완할 필요가 있다.

3.3. 자극물 설계

3.3.1. 인간 작성 자극물

본 연구에서 활용된 인간 작성 자극물은 한국인터넷진흥원(KISA) 사이버보안빅데이터센터의 스미싱 신고 데이터(2018~2021)를 기반으로 구축되었다. 해당 데이터는 KISA 사이버보안빅데이터센터의 공식적인 이용 승인 절차를 거쳐 제공받았으며, 생성형 AI 기술이 사회 전반에 확산되기 이전 시기의 실제 범죄 사례를 포괄하고 있어 AI 개입 이전 시기의 범죄자의 언어적 단서를 확보하기 위한 원천 데이터로 활용되었다. 본 연구의 데이터 수집 및 가공 과정은 생명윤리위원회(IRB)의 최종 승인을 획득한 후, 연구자의 비식별화 절차를 거쳐 외부로 반출되었다.

인간 작성 자극물의 정제 과정은 다음과 같다. 첫째, 원본 데이터 내에 존재하는 이름·주소·전화번호·계좌번호 등 수신자를 직접적으로 식별할 수 있는 민감 정보를 R 기반의 정규표현식 스크립트를 활용하여 일괄 비식별화 처리하였다. 둘째, 정제된 본문 텍스트를 대상으로 Cialdini(2021)의 설득 원칙을 이론적 토대로 하여, Schmitt와 Flechais(2024)가 제시한 사회공학적 기제의 조작적 정의에 부합하는 문구들을 선별하였다[7,16]. 본 연구는 기제별 분류의 객관성을 확보하기 위해, 선행 연구에서 제시된 이론적 정의와 기제별 키워드 매뉴얼을 엄격히 준수하여 규준 참조 분류 방식을 사용함으로써, 연구자의 주관적 개입과 해석적 편향을 최소화하는 데 중점을 두었다.

3.3.2. AI 작성 자극물

AI 작성 자극물은 인간 작성본의 원천 데이터가 지닌 핵심 시나리오(예: ‘소액결제 알림’, ‘택배 반송’ 등)와 ‘URL 삽입 목적’을 동일하게 유지하여 인간 작성본과 1:1로 페어링되도록 설계하였다. 단, 두 자극물 간의 물리적 텍스트 길이나 표면적 문장 구조는 엄격히 통제하지 않았다. 이는 생성형 AI가 타겟의 심리 기제를 자극하기 위해 스스로 문맥을 확장하고 정교화하는 과정 자체를 고도화된 스미싱의 본질적 위협으로 간주하여, 실제 위협 환경의 생태학적 타당성을 확보하기 위함이다.

텍스트 생성에는 대규모 언어 모델인 Gemini 모델군(Pro, Flash 등 복합 모델)이 활용되었다. 특히 본 연구는 상용 LLM의 내장된 보안 가드레일을 우회하고, 실제 범죄자의 고도화된 공격 기법을 실험 환경에 모사하기 위해 체계적인 페르소나 프롬프팅 및 탈옥 기법을 적용하였다. 구체적으로 자극물 생성 프롬프트는 1) 맥락 및 규칙 설정 2) 타겟 심리 기제의 조작적 정의 주입 3) 환경적 제약 조건 4) 원본 데이터(Seed) 입력 5) 기만성 극대화를 위한 과업 지시의 5단계 모듈 구조로 설계되었다.

이러한 고도화된 프롬프트 엔지니어링을 통해 도출된 다수의 결과물 중, 연구자가 공격자의 관점에서 사전에 정의된 조작적 정의에 가장 부합하고 기만 잠재력이 높다고 판단되는 텍스트를 최종 선별하는 인간-AI 협업(Human-in-the-Loop) 기반의 생성 프로세스를 채택하였다. 이는 기술적 전문성이 낮은 행위자라도 탈옥 프롬프트를 통해 최적의 사기 스크립트를 추출해 내는 형태의 서비스형 범죄(CaaS) 생태계를 방법론적으로 가장 현실성 있게 구현한 것이다. 본 연구는 사이버 보안 분야에서 생성형 AI가 지니는 이중 용도의 딜레마 및 범죄 악용 가능성 등 연구 윤리적 측면을 고려하여, LLM의 가드레일 우회에 실제 사용된 페르소나 프롬프트의 원문 전체는 비공개 원칙을 고수하되, 학술적 재현성을 보장하기 위해 앞서 서술한 바와 같이 프롬프트의 구조적 뼈대만을 논문에 명시하였다.

3.3.3. 정상 메시지(대조군)

스미싱 자극물과의 변별 타당도를 확보하고, 피험자의 기본적 의심 수준(Baseline suspicion)을 측정하기 위해, 실제 공식 기관이 발송하는 정상 메시지 8개를 대조군으로 포함하였다. 정상 메시지의 예시로는 통신사 청구서 안내, 카드사 결제 알림, 네이버 선물 도착 알림 등이 포함되었다.

3.3.4. 자극물 할당 방법

전체 자극물은 인간 작성 12개, AI 작성 12개, 정상 메시지 8개로 총 32개이며, 한 명의 피험자는 <Table 2>와 같이 교차 할당된 20개(인간 6개 + AI 6개 + 정상 8개)를 평가하였다. 특히, 각 심리 기제 측정 시 단일 시나리오에 치우쳐 발생할 수 있는 측정 오차를 통제하기 위해, 하나의 기제당 2개의 서로 다른 도메인 자극물을 복수 배정하였다. 자극물 제시 순서는 무작위화하였고, 자극물은 실제 모바일 메시지 캡처 이미지 형태로 제시하여 생태학적 타당도를 확보하였다. 권위성 1번 인간 문항에 대해 쌍을 이루도록 제작된 권위성 1번 AI 문항은 서로 다른 집단(A/B)에 할당되는 형식을 취하여 동일 시나리오에 대한 피험자의 인지적 이월 효과를 차단하였다. 피험자는 무작위 할당된 문항에 대해 신뢰도 정도와 행동 의도를 4점 리커트 척도로 응답하였다.

<Table 2> 피험자 1인당 자극물 할당

분류	심리 기제	인간 작성	AI 작성	정상(대조군)	합계
기제별	위급성	2개	2개	-	4개
	권위성	2개	2개	-	4개
	혼합형	2개	2개	-	4개
	정상	-	-	8개	8개
합계		6개	6개	8개	20개

본 설문은 참여자의 고유한 특성을 통제하고 정량적 실험 결과를 심층적으로 보완하기 위해 세 단계의 구조로 구성되었다. 첫째, 설문 초기 단계에서 연구 참여자의 성별, 연령, 최종 학력 등 표본 특성 기술 및 통제 목적으로 사용하기 위해 인구통계학적 특성을 수집하였다. 둘째, 무작위로 제시되는 20개의 메시지 자극물에 대한 정량 평가를 실시하였다. 셋째, 모든 자극물 평가가 완료된 후 실험 말미에 피험자의 인지적 판단 근거를 포착하기 위한 서술형 주관식 문항을 배치하였다. 주관식 문항은 전체 자극물 중 '가장 정상 메시지 같았다고 판단한 문항과 그 구체적인 이유', 그리고 '가장 스팸(스미싱) 메시지 같았다고 판단한 문항과 그 구체적인 이유'를 자유롭게 기술하도록 유도하였다. 이 주관식 서술 데이터는 참여자들이 디지털 기만을 탐지하거나 신뢰를 부여할 때 의존하는 언어적·맥락적 단서를 규명하기 위한 텍스트 마이닝 분석의 원천 데이터로 활용되었다.

3.4. 실험 변인 선정의 근거 및 생태학적 타당성

본 연구는 Cialdini의 다양한 설득 원리 중 모바일 스미싱 환경의 특수성과 생태학적 타당성을 고려하여 '권위성'과 '위급성'만을 독립 변인으로 추출하여 통제하였다. 스미싱은 텍스트 분량이 극도로 제한적이고 단시간 내에 수신자의 즉각적인 클릭 행동을 유도해야 하는 매체적 특성을 지닌다. 따라서 피해자와 장기적인 서사나 관계 형성이 필요한 다른 기제(예: 호감성, 상호성 등)보다는, 수신자의 심층적 정보 처리를 즉각적으로 차단하고 휴리스틱에 기반한 자동적 반응을 강제하는 권위성과 시간적 위급성이 가장 치명적으로 작용한다. 또한 실제 사이버 범죄 수법을 분석한 선행연구[9]에서도 공공기관 사칭(권위성)과 즉각적인 행정 처분 경고(위급성)가 스미싱 시나리오의 대다수를 차지하고 있어, 해당 변인을 중심축으로 실험을 설계하는 것이 현실의 사이버 위협을 가장 정확하게 반영한다고 판단하였다.

3.5. 조작적 정의 및 측정 도구

본 연구에서 설정한 심리 기제별 조작적 정의는 <Table 3>과 같다. 각 자극물에 대해 피험자는 다음의 두 문항을 4점 리커트 척도(1=전혀 그렇지 않다 ~ 4=매우 그렇다)로 응답하였으며 본 연구에서 중립(보통이다) 옵션이 배제된 4점 척도를 채택한 근거는 다음과 같다. 첫째, 스미싱 위협은 모바일 환경에서 수신자의 즉각적이고 직관적인 진위 판별을 요구하는 특성이 있으므로, 피험자가 메시지의 기만성에 대해 명확한 입장을 취하도록 유도하기 위함이다. 둘째, 중장년층을 포함한 피험자가 다수의 자극물(20개)을 반복 평가하는 과정에서 인지적 노력을 회피하기 위해 중립 응답을 남발하는 현상을 원천적으로 통제하기 위함이다. 셋째, 보안 위협에 대한 자기 평가는 사회적 바람직성이 개입되어 '모르겠다' 혹은 '중간'으로 도피하려는 경향이 발생할 수 있으므로, 이를 배제하고 피험자의 내재된 실제 취약성 경향을 명확히 추출하기 위해 짝수 척도를 적용하였다[17].

① 신뢰도(Trust): "이 메시지가 실제 공식 기관에서 발송한 것이라고 생각한다."

② 행동의도(Behavioral Intention): "나에게 이 메시지가 도착했다면 첨부 링크를 확인해 볼 것이다."

두 문항의 평균을 기만성 점수로 정의하였다. 또한, 앞서 설계한 바와 같이 각 실험 조건(예: 인간×위급성)마다 2개의 자극물이 배정되었으므로, 이 두 자극물의 기만성 점수를 합산하여 평균을 낸 값을 각 피험자의 해당 조건별 최종 기만성 점수로 산출하였다. 이와 같이 복수 문항을 활용한 측정 도구의 타당성 및 신뢰도를 검증하기 위해 6개의 실험 조건별로 문항 간 내적 일관

성(Cronbach's α)을 산출한 결과 .605~.802의 범위로 나타났다. 일부 인간 작성 조건(위급성·권위성)이 통상적 기준인 .70을 소폭 하회하였으나, 사회과학에서 허용되는 수용기준(α≥.60)을 충족하여 양호한 수준으로 판단하였다.

<Table 3> 심리 기제별 조작적 정의

심리기제	조작적 정의
권위성	발신자가 국가 사법/행정 기관, 보건/의료 기관 등 사회적 공신력을 가진 주체로 위장하는 방식. 즉각적인 처벌이나 금전적 손실 위협보다는 대상 기관의 지위에 순응하여 행정적·제도적 의무감으로 링크 클릭 등의 인지적 과업을 수행하게 만드는 기만 전략
위급성	발신자의 공신력보다 수신자에게 발생할 수 있는 즉각적인 물리적/금전적 타격(예: 택배 반송, 예기치 않은 과금, 계정 해킹 등)을 구체적으로 제시하는 방식. 당장의 손실을 회피하기 위한 행동의 시간적 압박을 부여함으로써, 피해자의 불안감을 조성하고 충동적인 반응을 강제하는 기만 전략
혼합형	앞서 정의된 권위성(공신력을 가진 기관)과 위급성(즉각적인 행정적 처분, 계정 정지 등의 손실 위협)이 하나의 텍스트 내에 동시에 발현되어 피해자의 방어 기제를 약화하는 복합 기만 전략

3.6. 연구 절차 및 분석 방법

본 연구는 생명윤리위원회(IRB)의 승인 절차를 거쳐 진행되었다. 모든 자료 수집은 온라인 설문 플랫폼(Google Forms)을 통해 비대면으로 이루어졌다. 연구의 내적 타당도를 확보하기 위해, 설문 초기에는 본 연구를 '일상적 모바일 메시지의 진위 판별 조사'로 안내하는 불완전한 정보 제공 절차를 적용하였다. 모든 자극물 평가가 완료된 직후, 마지막 섹션에서 본 연구의 실제 목적(인간 vs AI 생성 스미싱 비교)을 밝히는 사후 설명문을 제공하고, 수집된 데이터를 최종적으로 연구에 활용하는 것에 대한 사후 동의를 다시 한 번 구하였다. 이와 함께 경찰청 '피싱 대응 행동 수칙' 영상을 안내함으로써 참여자에게 실질적 예방 교육의 기회를 제공하였다.

모든 통계 분석은 R 통계 패키지(R 4.6.0)의 afex, emmeans, psych, rstatix, effectsize 등 패키지를 사용하였다. 분석 절차는 다음과 같다. ① 데이터 정제: 결측치 및 불성실 응답 제거. ② 척도 신뢰도: Cronbach's α 산출. ③ 정규성 검정: Shapiro-Wilk 검정. ④ 주 가설 검증: 생성주체(2) × 심리기제(3) 참가자 내 반복측정 분산분석 및 연령(2)을 추가한 2×3×2 혼합 반복측정 분산분석을 실시. ⑤ 사후 검정: Bonferroni 및 Tukey 보정을 적용한 다중비교. ⑥ 효과 크기: 부분 에타제곱(η^2_p) 및 Cohen's d. ⑦ 비모수 보완: Wilcoxon signed-rank test. 통계적 유의수준은 α=.05로 설정하였으며, 구형성 가정 위배 시 Greenhouse-Geisser 보정을 적용하였다.

IV. 실험 및 평가

4.1. 측정 도구의 신뢰도 및 기술통계

본 연구에서 사용된 모바일 메시지 기만성 지각(신뢰도 및 행동의도) 척도의 문항 간 내적 일관성을 검증하기 위해 Cronbach's α 계수를 산출하였다. 전체 참여자(N=126) 및 연령 집단에 따른 기만성 점수의 기술통계는 <Table 4>와 같다. 분석 결과, 위급성-인간 조건(α = .605), 위급성-AI조건(α = .773), 권위성-인간 조건(α = .680), 권위성-AI 조건(α = .802), 혼합형-인간 조건(α = .705), 혼합형-AI 조건(α = .779)으로 나타났다. 정상 메시지의 경우 신뢰도(α = .831)

와 행동의도(α = .859) 모두 높은 수준을 보였다. 일부 인간 작성 조건(위급성, 권위성)에서 내적 일관성이 통상적 기준인 .70을 소폭 하회하였으나, 본 연구의 척도가 조건당 4문항(신뢰도 2문항, 행동의도 2문항)으로 구성된 소문항 척도임과 동시에, 실제 범죄 현장에서 수집된 인간 작성 스미싱 문구 특유의 표현적 이질성이 반영된 결과로 해석할 수 있다. 반면 프롬프트를 통해 통제된 생성형 AI의 문구는 상대적으로 높은 일관성을 보였다. 따라서 사회과학 분야에서 허용되는 수용 기준(α ≥ .60)을 충족하는 양호한 수준으로 판단하였다.

<Table 4> 연령 집단 및 심리 기제에 따른 기만성 점수 기술통계 및 신뢰도

생성 주체	심리 기제	청년층 (n = 65)	중장년층 (n = 61)	전체 (N = 126)	신뢰도(α)
AI 작성	권위성	2.53 (0.80)	1.85 (0.72)	2.20 (0.83)	.802
	위급성	2.18 (0.81)	1.72 (0.67)	1.95 (0.78)	.773
	혼합형	2.13 (0.76)	1.74 (0.74)	1.94 (0.77)	.779
	주체 소계(AI 평균)	2.28 (0.79)	1.77 (0.71)	2.03 (0.79)	-
인간 작성	권위성	1.58 (0.55)	1.56 (0.60)	1.57 (0.57)	.680
	위급성	1.94 (0.65)	1.57 (0.57)	1.76 (0.64)	.605
	혼합형	1.70 (0.67)	1.51 (0.61)	1.61 (0.64)	.705
	주체 소계(인간 평균)	1.74 (0.62)	1.55 (0.59)	1.65 (0.62)	-
대조군	정상 메시지	2.24 (0.61)	1.78 (0.56)	2.02 (0.63)	.831 / .859

평균 (표준편차). 기만성 점수는 4점 리커트 척도(1 = 전혀 그렇지 않다, 4 = 매우 그렇다)를 기준으로 산출됨.

전체 표본의 조건별 기술 통계 추이를 살펴보면, 전반적인 스미싱 자극물 중 AI가 작성한 권위성 스미싱의 기만성 점수(M = 2.20, SD = 0.83)가 대조군인 정상 메시지의 기만성 점수(M = 2.02, SD = 0.63)를 상회하는 것으로 나타났다. 이는 기존의 스미싱 문구들이 지니던 어색함으로 인해 작동되던 경계심이 생성형 AI의 권위성 조작 앞에서 약화되며, 정상 메시지보다 더 높은 신뢰를 부여받는 기만 효과가 발생하고 있음을 시사한다.

한편, Shapiro-Wilk 정규성 검정을 실시한 결과, 스미싱 하위 조건들의 기만성 점수 분포에서 정규성 가정이 통계적으로 위배되는 양상이 관찰되었다(W = .847~.950, p < .05). 그러나 본 연구는 표본 크기의 충분성(N = 126) 및 중심극한정리(Central Limit Theorem)에 따른 분산분석 모델의 강건성(robustness)에 의거하여 모수적 통계 기법(ANOVA)을 주 분석으로 채택하되, 분석 결과의 타당성을 다각도로 검증하기 위해 비모수적 대안 검정(Wilcoxon 부호순위 검정)을 병행하여 일관성을 검증하였다.

4.2. 주요 가설 검증

4.2.1. 텍스트 생성주체 × 심리기제에 따른 2원 반복측정 분산분석

전체 표본을 대상으로 생성주체(인간 vs 생성형 AI)와 메시지에 내포된 심리적 설득 기제(위급성 vs 권위성 vs 혼합형)가 수신자의 종합적 기만성 지각에 미치는 주효과와 상호작용 효과를 검증하기 위해, 피험자 내 요인들을 대상으로 2원 반복측정 분산분석(Two-way Repeated Measures ANOVA)을 수행하였다. 분석 시 구형성 가정을 평가하였으며, 가정 위배가 확인된 요인에 대해서는 Greenhouse-Geisser 보정 계수를 적용하여 유의확률을 조정하였다.

분석 결과는 다음과 같다. 첫째, 텍스트 생성 주체의 주효과가 통계적으로 유의하였다(F(1,

125) = 67.898, $p < .001$, $\eta^2_p = .352$). 즉, 심리 기제 조건을 통틀었을 때 생성형 AI가 작성한 스미싱 문구($M = 2.03$)가 인간 범죄자가 작성한 문구($M = 1.65$)보다 수신자에게 유의하게 더 높은 기만성을 유발하였다(평균 차이 = 0.386, 95% CI [0.293, 0.479]). 또한 모수 검정의 정규성 위반을 보완하기 위해 비모수적 대안 검정인 Wilcoxon 부호순위 검정을 병행한 결과에서도 통계적으로 매우 유의한 차이가 확인되어 분석의 강건성(Robustness)이 검증되었다($V = 4861$, $p < .001$, $r = .631$).

둘째, 자극물에 조작된 심리 기제의 주효과 또한 통계적으로 유의하였다($F(1.97, 245.80) = 4.045$, $p = .019$, $\eta^2_p = .031$). 다만 그 효과 크기는 작은 수준(small effect)으로, 생성 주체의 주효과($\eta^2_p = .352$)에 비해 설명력이 제한적이었다. Tukey 방식을 적용한 사후 다중비교 결과, 권위성 기제($M = 1.89$)가 혼합형 기제($M = 1.77$)에 비해 유의하게 높은 기만성 지각을 형성하였다($p = .024$). 반면, 위급성 기제($M = 1.86$)는 권위성 및 혼합형 기제와 유의한 차이가 관찰되지 않았다. 즉 심리 기제 단독의 주효과는 통계적으로 유의하였으나, 실질적 기여도는 생성 주체 요인 및 후술할 상호작용 효과에 비해 부차적인 것으로 해석된다.

셋째, 생성 주체와 심리 기제 간의 2원 상호작용 효과가 통계적으로 유의하게 나타났다($F(1.92, 240.20) = 17.841$, $p < .001$, $\eta^2_p = .125$). 기제별 주체 단순효과를 분석한 결과, 세 기제 모두에서 AI가 인간보다 유의하게 높은 기만성을 보였으나, 그 격차의 크기에서 차이가 관찰되었다. 특히 권위성 기제에서 AI와 인간의 기만성의 평균 차이(0.66)가 위급성(0.194)이나 혼합형(0.331)에 비해 상대적으로 크게 나타났다.

이는 생성형 AI가 산출한 텍스트의 기만 효과가 단순히 모든 상황에서 균일하게 나타나는 것이 아니라, 공공기관이나 금융사 등을 사칭하는 권위성 기제에서 통계적으로 가장 크게 나타났음을 확인한 결과이다. 생성 주체와 심리 기제에 따른 기만성 점수의 2원 반복측정 분산분석 결과는 <Table 5>와 같다.

<Table 5> 생성 주체와 심리 기제에 따른 기만성 점수의 2원 반복측정 분산분석 결과

변인 (Source)	자유도 (df)	F	유의확률 (p)	부분 에타제곱 (η^2_p)
생성 주체(Agent)	1.00	67.898	<.001***	.352
오차(생성 주체)	125.00			
심리 기제 (Mechanism)	1.97	4.045	.019*	.031
오차(심리 기제)	245.80			
생성 주체 x 심리 기제	1.92	17.841	<.001***	.125
오차(생성 주체 x 심리 기제)	240.20	-	-	-

심리 기제 및 상호작용의 자유도(df)는 구형성 가정 위배로 인해 Greenhouse-Geisser 보정치가 적용됨.
*p <.05, ***p <.001

4.2.2. 연령 집단을 추가한 3원 혼합 분산분석

연령을 참가자 간 요인으로 추가한 3원 혼합 분산분석 결과<Table 6>, 연령의 주효과가 유의하게 나타났다($F(1, 124) = 14.04$, $p < .001$, $\eta p^2 = .102$). 청년층($M = 2.01$)이 중장년층($M = 1.66$)보다 전반적으로 높은 기만성 지각을 보였으며, 두 집단 간의 종합 기만성 차이는 통계적으로 유의하였다(평균 차이 = 0.353, 95% CI [0.166, 0.540], $d = 0.668$).

<Table 6> 기만성에 대한 연령 × 생성주체 × 심리기제 3원 혼합 분산분석 결과

Source(변인)	df	*F*	*p*	η^2_p
개체 간 효과				
연령 (Age)	1	14.044	<.001***	.102
오차 (Error)	124			
개체 내 효과				
생성 주체	1	72.140	<.001***	.368
연령 x 생성주체	1	12.476	<.001***	.091
심리기제 (Mech)	1.97	3.953	.021*	.031
연령 x 심리기제	1.97	1.210	.299	.010
생성주체 x 심리기제	1.96	18.128	<.001***	.128
연령 x 생성주체 x 심리기제	1.96	8.787	<.001***	.066

심리 기제 및 관련 상호작용의 자유도(df)는 구형성 가정 위배로 인해 Greenhouse-Geisser 보정치가 적용됨.
*p <.05, ***p <.001

또한, 연령과 생성 주체 간의 2원 상호작용 효과가 통계적으로 유의하게 나타났다(F(1, 124) = 12.476, p < .001, η^2_p = .091). 연령 집단별 주체 단순효과를 분석한 결과, 청년층과 중장년층 모두 인간 작성 문구보다 AI 작성 문구 조건에서 유의하게 높은 기만성을 보였다. 그러나 집단 간 격차의 크기에서 차이가 관찰되었다. 청년층은 인간과 AI 작성 문구 간의 기만성 점수 격차(평균 차이 = 0.540, 95% CI [0.416, 0.663], p < .001)가 상대적으로 크게 나타난 반면, 중장년층은 그 격차(평균 차이 = 0.223, 95% CI [0.095, 0.350], p < .001)가 작게 나타났다. 이는 생성형 AI 문구의 기만 효과가 청년층 집단에서 더 큰 폭으로 발생함을 의미한다.

심리 기제에 따른 연령 집단별 단순주효과를 분석한 결과에서도 집단 간 차이가 확인되었다. 청년층은 권위성(M = 2.06)과 위급성(M = 2.06) 기제 조건이 혼합형(M = 1.92) 조건에 비해 통계적으로 유의하게 높은 기만성을 유발하였다(각각 p = .046, p = .041). 반면, 중장년층은 세 가지 심리 기제 간 통계적으로 유의미한 점수 차이가 관찰되지 않았다.

4.2.3. 정상 메시지 대비 스미싱의 변별 타당도

대조군인 정상 메시지와 스미싱 자극물 간의 기만성 지각 차이를 비교하기 위해 대응표본 t-검정을 실시하였다. 분석 결과, 인간이 작성한 스미싱 문구는 정상 메시지에 비해 통계적으로 유의하게 낮은 기만성 점수를 보였다(평균 차이 = -0.369, 95% CI [-0.445, -0.293], t(125) = -9.667, p < .001). 반면, 생성형 AI가 작성한 스미싱 문구 총점과 정상 메시지 간의 기만성 점수 차이는 통계적으로 유의하지 않았다(평균 차이 = 0.017, 95% CI [-0.052, 0.087], t(125) = 0.488, p = .626). 이는 연구 참여자들이 모바일 환경에서 인간 범죄자의 문구와 실제 공공기관의 정상 메시지는 유의미하게 변별해 낸 반면, 생성형 AI가 산출한 악성 텍스트와 정상 메시지는 통계적인 수준에서 식별하지 못하였음을 확인한 결과이다.

4.2.4. 종속 변수별 추가 분석

종합적인 기만성 지각을 구성하는 하위 차원인 인지적 차원(신뢰도)과 행동적 차원(행동의도)을 분리하여 생성 주체의 효과를 추가 분석하였다. 반복측정 분산분석 및 대응표본 t-검정 결과, 신뢰도 차원에서 생성형 AI 작성 문구가 인간 작성 문구보다 유의하게 높은 점수를 기록하였다(F(1, 125) = 64.094, p <.001, η^2_p = .339; 평균 차이 = 0.429, 95% CI [0.323, 0.535]).

동일하게, 실제 악성 URL 클릭 가능성을 묻는 행동의도 차원에서도 생성형 AI 작성 문구가 인간 작성 문구에 비해 통계적으로 유의하게 높은 점수를 유발하였다($F(1, 125) = 50.825$, $p < .001$, $\eta^2_p = .289$; 평균 차이 = 0.344, 95% CI [0.248, 0.439]). 두 종속 변수 모두 비모수 검정(Wilcoxon 부호순위 검정)을 통한 교차 검증에서 일관되게 유의한 차이(각각 $p < .001$)가 확인되었다. 이는 생성형 AI 문구의 기만 효과가 단순히 메시지의 진위 판단(인지) 단계에 국한되지 않고, 수신자의 실질적인 반응(행동 의도) 단계까지 유의하게 이어지고 있음을 나타내는 결과이다.

4.2.5. 텍스트 키워드 분석

정량 분석 결과를 보완하고 참여자가 메시지의 진위를 판단할 때 사용한 인지적 단서를 탐색하기 위해, 피험자가 설문 말미에 응답한 '가장 정상 메시지 같았던 문항'과 '가장 스팸 메시지 같았던 문항'에 대한 주관식 응답(정상 판단 113건, 스팸 판단 100건)을 텍스트 마이닝 기법으로 분석하였다. 형태소 분석을 통해 명사·동사·형용사·부사를 추출하고, 설문에서 동시 출현도가 높은 단어 및 노이즈 데이터를 불용어로 제거한 뒤 출현 빈도를 산출하여 워드클라우드로 시각화하였다<Figure 1>. 또한, 판단 단서의 유형을 파악하기 위해, 응답에 나타난 핵심어를 여섯 개 범주로 분류하고 정상·스팸 판단별 출현 빈도를 비교하였다<Table 7>.

<Figure 1> 워드클라우드 - 정상 메시지 판별 이유 vs 스팸 메시지 판단 이유

<Table 7> 판단 단서 범주별 출현 빈도

구분	단서 범주	정상 판단	스팸 판단
신뢰 단서	친숙성·경험	17	4
의심 단서	불안·공포 유발	0	5
	어색·조잡·간결	0	11
양면 단서	공식성·형식	23	23
	발신정보(출처)	7	14
	행동 권고·유도	22	32

워드클라우드 분석 결과, 정상 판단과 스팸 판단의 근거 어휘가 대비되는 양상이 관찰되었다. 정상 메시지로 판단한 근거로는 '실제로', '자주', '평소', '정기', '일상' 등 수신자 자신의 평소 이용 경험과 친숙성을 반영하는 어휘가 두드러졌으며, '길다', '유사', '진짜', '신뢰', '공식' 등 형식적 완결성을 의미하는 어휘가 함께 관찰되었다. 반면 스팸 메시지로 판단한 근거에서는 '짧

다', '설명(없이)', '이상', '쓸모(없는)' 등 형식적 결함을 나타내는 어휘와 '의심', '사기', '유도', '피싱' 등 위험을 감지하는 어휘, 그리고 '발신', '번호', '계정', '해외' 등 발신 출처의 불명확성을 나타내는 어휘가 주를 이루었다.

정상 판단에서는 '길다', 스팸 판단에서는 '짧다'가 대비되어 나타났는데 이는 메시지의 길이와 서술의 구체성이 진위 판단의 단서로 작동하였음을 시사한다. 단서 범주별 빈도 분석 결과에서 판단 단서는 세 가지 유형으로 구분되었다. 첫째, '친숙성·경험' 범주는 정상 판단(17회)이 스팸 판단(4회)보다 높은 빈도로 출현하여, 평소 받아온 메시지와의 유사성이 신뢰의 핵심 단서로 작동함이 확인되었다. 둘째, '불안·공포 유발'(정상 0회, 스팸 5회)과 '어색·조잡·간결'(정상 0회, 스팸 11회) 범주는 스팸 판단에서만 출현하여, 부정 정서의 유발과 문맥의 어색함이 의심을 형성하는 단서로 기능하였다. 셋째, '공식성·형식'(정상 23회, 스팸 23회), '발신정보'(정상 7회, 스팸 14회), '행동 권고·유도'(정상 22회, 스팸 32회) 범주는 정상·스팸 판단 모두에서 공통적으로 출현하여, 동일한 차원이 신뢰의 단서이면서 의심의 단서로도 작동하는 양면적 양상을 보이는 것을 확인하였다. 특히 '공식성·형식' 범주가 양측에서 동일한 빈도로 나타난 결과를 통해 발신 주체의 권위·공식성이 신뢰와 의심 모두를 매개하는 양가적 단서임을 뒷받침한다.

종합하면, 수신자는 메시지의 진위를 판단할 때 친숙성과 공식성을 신뢰의 근거로, 형식적 결함(어색함·과도한 간결성)과 부정 정서를 의심의 근거로 활용하였다는 것을 알 수 있다.

V. 논의

5.1. 설문분석 결과

본 연구는 인간이 작성한 스미싱 문구와 생성형 AI가 작성한 스미싱 문구가 수신자의 기만성 지각(신뢰도 및 행동의도)에 미치는 차별적 영향을 실증적으로 검증하였다. 가설 검증을 포함한 주요 결과는 다음과 같이 <Table 8>에 요약하였다.

첫째, 텍스트 생성주체의 주효과가 유의하게 나타나 가설 1은 지지되었다. 생성형 AI가 작성한 문구(M = 2.03)가 인간이 작성한 문구(M = 1.65)보다 전반적으로 유의하게 높은 기만성을 유발하였다. 이는 실제 범죄 환경에서 수집된 인간의 스미싱 텍스트 대비, 생성형 AI가 산출한 텍스트의 기만적 위험성이 실증적으로 더 높음을 확인한 결과이다.

둘째, 자극물에 조작된 심리 기제의 주효과가 유의하게 나타나 가설 2는 지지되었다. 사후 다중비교 결과, 공공기관 등을 사칭하는 권위성 기제(M = 1.89)가 권위성과 위급성이 결합된 혼합형 기제(M = 1.77)에 비해 유의하게 높은 기만성 지각을 형성하였다. 다만 심리 기제의 주효과 크기는 작은 수준(η^2_p = .031)에 그쳐, 기제의 유형 그 자체보다는 후술할 생성 주체와의 상호작용을 통해 그 효과가 발현되는 것으로 보는 것이 타당하다.

셋째, 생성 주체와 심리 기제 간의 2원 상호작용 효과가 통계적으로 유의하게 나타나 가설 3은 지지되었다. 세 가지 심리 기제 모두에서 AI 작성 문구가 인간 작성 문구보다 높은 기만성을 보였으나, 특히 '권위성' 기제 조건에서 인간 문구 대비 AI 문구의 기만성 상승폭이 가장 크게 관찰되었다.

넷째, 수신자의 연령 집단에 따른 주효과가 유의하게 나타나 가설 4는 지지되었다. 청년층(M = 2.01)이 중장년층(M = 1.66)에 비해 모든 실험 조건에서 유의하게 높은 수준의 기만성을 나타냈다. 이는 중장년층이 스미싱 범죄에 더 취약할 것이라는 일반적 통념과 상반되며, 디지털 기기에 친숙한 세대라 할지라도 정교하게 조작된 텍스트 앞에서는 구조적 취약성을 지닐 수 있음을 시사한다.

다섯째, 연령 집단과 생성 주체 간의 상호작용 효과가 통계적으로 유의하게 나타나 가설 5는 지지되었다. 청년층과 중장년층 모두 인간 작성 문구보다 AI 작성 문구에 더 높은 기만성을 보였으나, 그 격차의 크기에서 차이가 확인되었다. 즉, 인간 문구 대비 AI 문구 노출 시 기만성 점수의 상승폭은 상대적으로 청년층 집단에서 더 두드러지게 발현되었다.

<Table 8> 가설 검증 결과 요약

연구 가설	가설 내용	검증 결과	비고
가설 1	생성 주체에 따른 기만성 차이가 유의할 것이다.	지지	AI > 인간
가설 2	심리 기제에 따른 기만성 차이가 유의할 것이다.	지지	권위성이 혼합형보다 높았음
가설 3	생성 주체 x 심리 기제 상호작용이 유의할 것이다.	지지	권위성 조건에서 AI-인간 격차 최대
가설 4	연령 집단에 따른 기만성 차이가 유의할 것이다.	지지	청년층 > 중장년층
가설 5	연령 집단 x 생성 주체 상호작용이 유의할 것이다.	지지	청년층에서 AI-인간 격차가 더 큼

5.2. 청년층의 높은 기만 취약성

본 연구에서 청년층이 중장년층보다 전반적으로 기만성에 더 취약하였으며, 특히 AI 작성 문구에서 그 상승폭이 두드러졌다. 청년층은 모바일 메시지에 일상적으로 노출되는 환경 속에서 메시지를 처리할 때 체계적 분석보다는 인지적 휴리스틱에 의존하는 경향이 있다. Jakesch 등(2023)의 연구에 따르면, 인간은 텍스트의 문법적 정확성이나 맥락의 자연스러움을 진실의 단서로 오인하는 인지적 결함을 지니며, AI는 이를 악용하여 '인간보다 더 인간적인(more human than human)' 텍스트로 수신자를 기만할 수 있다[18]. 본 연구에서 청년층이 AI의 정교한 텍스트에 높은 신뢰를 부여한 것 역시 이러한 형식적 완결성을 안전의 단서로 받아들였을 것으로 해석할 수 있다. 나아가 청년층은 대조군(정상 메시지)에서도 중장년층보다 높은 점수를 부여하였는데, 이는 청년층이 모바일 텍스트 전반에 대해 지니고 있는 기저 신뢰도 자체가 상대적으로 높음을 시사한다. 이러한 양상은 청년층의 높은 매체 의존도 및 휴리스틱 처리 경향과 관련되었을 가능성이 있으나, 본 연구에서 청년층 취약성의 기제(휴리스틱 의존, 디지털 친숙도 등)는 직접 측정되지 않았기에 이러한 해석의 타당성은 후속 연구를 통해 검증할 필요가 있다.

5.3. 중장년층의 인지적 경계와 방어 기제

중장년층 역시 인간 작성본보다 AI 작성본에 유의하게 더 높은 기만성을 보였으나, 그 점수의 상승폭은 청년층에 비해 상대적으로 제한적이었다. 또한 정상 메시지를 포함한 모든 실험 조건에서 일관되게 낮은 점수를 기록하였다. 이는 중장년층이 텍스트의 형식적 완성도를 신뢰 단서로 삼기보다는, 모바일 메시지 전반에 대해 높은 인지적 경계심을 유지하고 있음을 시사한다. 이는 그동안의 보이스피싱 및 스미싱 예방 교육의 누적 효과로 인해, 중장년층은 출처가 불분명한 문자 메시지에 대해 진위 여부를 먼저 판단하거나 기본적으로 불신하는 방어 기제를 사용하고 있음을 의미한다.

또한 연령·생성 주체·심리 기제 간의 3원 상호작용이 통계적으로 유의하게 나타난 점은, 심리 기제가 기만에 미치는 영향이 연령에 따라 다르게 작동함을 보여준다. 청년층에서는 권위성과 위급성 기제가 혼합형보다 유의하게 높은 기만성을 유발한 반면, 중장년층에서는 세 기제 간에

유의한 차이가 나타나지 않았다. 즉, 청년층이 특정 기제(권위성, 위급성)에 더 민감하게 반응한 데 비해, 중장년층은 어떤 기제가 사용되든 메시지 전반에 걸쳐 일관되게 낮은 신뢰를 부여하였다. 이는 두 집단의 양상이 단순히 기만 취약성의 차이를 넘어, 메시지를 받아들이는 방식 자체가 질적으로 다르다는 점을 시사한다.

5.4. 진위 판별 실패 및 권위성 기제의 확장

본 연구에서 가장 주목할 만한 학술적 발견은 AI가 작성한 스미싱 문구와 실제 공식 기관의 정상 메시지 간에 기만성 지각의 통계적 차이가 관찰되지 않았다는 점이다. 이는 모바일 텍스트 환경에서 AI가 작성한 스미싱 문구가 정상 메시지와 구별되지 않을 만큼 정교해져, 수신자의 진위 판별 능력이 사실상 무력화되었음을 의미한다. 특히 이러한 AI의 기만 효과는 권위성 기제와 결합할 때 가장 두드러졌다. Cialdini(2021)의 권위의 법칙에 따르면, 수신자는 공공기관이나 금융사 등 권위적 주체를 마주할 때 비판적 사고를 축소하고 순응하려는 인지 편향을 보인다. 과거 인간 범죄자의 스미싱 문구는 조잡한 문법과 어색한 표현으로 인해, 권위성 기제가 작동하기도 전에 의심이 선행되어서 그 위험성이 부각되지 않았다. 그러나 AI가 이러한 언어적 결함을 상당 부분 제거하면서 수신자의 권위 순응적 태도가 그대로 드러난 것으로 해석할 수 있다. AI 작성 문구가 정상 메시지와 구별되지 않았다는 결과는, 생성형 AI 자체를 위협으로 규정하기보다 그 산출물이 인간이 작성한 수준으로 자연스러워졌다는 점에 주목할 필요가 있음을 시사한다. 따라서 생성 주체와 무관하게 기만성이 높은 문구 자체를 어떻게 식별하고 대응할 것인지에 관해 논의될 필요가 있다.

한편, 본 연구에서 권위성 단독 기제가 권위성과 위급성을 결합한 혼합형보다 오히려 높은 기만성을 유발한 결과는 여러 설득 기제를 함께 사용할수록 기만 효과가 더 커질 것이라는 일반적인 예측과 어긋난다. 이러한 역전 현상은 다음과 같이 해석될 수 있다. 첫째, 스미싱은 텍스트 분량이 제한된 모바일 매체이다. 하나의 짧은 메시지 안에 권위적 사칭과 시간적 압박이 동시에 담기면, 정보량이 많아져 오히려 메시지가 작위적으로 느껴지고 수신자의 의심을 유발했을 가능성이 있다. 둘째, 권위성 기제만으로도 기만성 지각이 충분히 높은 수준에 도달하여, 위급성 기제를 더하더라도 효과가 더 이상 높아지지 않는 천장 효과(ceiling effect)가 나타났을 수 있다. 다만 본 연구는 기제가 결합될 때 수신자의 판단 과정을 직접 측정하지 않았으므로, 결합 기제의 효과가 감소하는 원인에 대한 해석은 후속 연구를 통해 검증될 필요가 있다.

5.5. 실무적 함의 및 향후 연구 제안

본 연구의 결과는 다음과 같은 실무적·정책적 함의를 갖는다.

첫째, AI 기반 스미싱 탐지 기술의 개선이 필요하다. 본 연구에서 AI가 작성한 사기 문구는 정상 메시지와 구별하기 어려울 만큼 자연스러운 것으로 나타났다. 이는 URL·IP 같은 기술적 단서나 단순 키워드에 의존하는 현재의 스미싱 방어 기술은 한계에 직면할 수 있음을 시사한다. 기존의 탐지 방식은 이미 수집된 악성 URL이나 반복적으로 사용되는 범죄 키워드를 차단하는 데 의존했지만, 생성형 AI는 매번 새로운 문구를 무한히 변형해 만들어내므로 과거 범죄 데이터에 기반한 탐지는 새로운 공격을 뒤늦게 따라갈 수밖에 없다. 따라서 탐지의 기준을 알려진 범죄 문구와의 일치 여부에서 정상적인 공식 메시지가 지닌 언어적·맥락적 특징에서 벗어나는 정도로 전환할 필요가 있다. 형식이 비교적 일정하고 확보가 용이한 정상 메시지의 패턴을 학습한 뒤, 이에서 벗어나는 메시지를 이상치로 탐지하는 방식이 그 대안이 될 수 있다. 아울러 본 연구에서와 같이 생성형 AI를 활용해 다양한 스미싱 메시지를 직접 생성하여 탐지 모델의 학습 자료

로 삼는다면, 실제 범죄 데이터를 사용하지 않고도 새로운 위협에 선제적으로 대비할 수 있을 것이다. 나아가 이러한 접근은 개별 메시지를 사후에 차단하는 데 그치지 않고, 대량의 통신 데이터 속에서 기만 텍스트가 지닌 언어적 특징을 분석하는 단계로 확장될 수 있다. 본 연구가 확인한 인간과 AI 생성 텍스트 간의 기만성 격차와 심리 기제별 표현 차이는 악성 메시지를 식별하거나 그 작성 주체의 특성을 추정하는 등 디지털 증거 분석 실무의 판별 지표를 설계하는 데 기초 자료로 활용될 수 있다.

둘째, 사칭 메시지의 진위를 이용자가 직접 확인할 수 있는 수단을 마련할 필요가 있다. 본 연구에서 공공·수사기관을 사칭하는 권위성 기제가 가장 높은 기만성을 보인 것은 발신 주체에 대한 신뢰가 기만의 핵심 통로로 작동함을 시사한다. 그러나 SMS는 발신자를 신뢰성 있게 인증하기 어려운 매체이므로 수신자가 출처를 스스로 검증하기 쉽지 않다. 따라서 공공기관과 금융권이 본인 인증이 보장된 공식 채널을 통해 주요 고지를 발송하고, 이용자가 의심스러운 메시지를 해당 채널에서 손쉽게 대조·확인할 수 있도록 하는 방안을 검토할 수 있다.

셋째, 주로 노인층이나 취약계층을 대상으로 하는 기존의 연령 맞춤형 예방 교육 패러다임의 전환이 필요하다. 그동안의 피싱 예방 교육은 주로 고령층과 정보 취약계층을 대상으로 이루어져 왔다. 그러나 본 연구에서는 청년층과 중장년층 모두 기만에 취약하였으며, 전반적으로 청년층이 중장년층보다 더 높은 기만 취약성을 보였다. 더욱이 이러한 연령 차이는 인간 작성 문구에 비해 AI 작성 문구에서 기만성이 높아지는 정도가 더 크게 나타났는데, 이는 청년층이 문장이 자연스럽고 정교할수록 해당 메시지를 신뢰할 만한 것으로 받아들이는 경향이 있음을 시사한다. 따라서 청년층을 대상으로 한 교육에서는 디지털 기기를 능숙하게 다루는 능력과 정보의 진위를 판별하는 능력이 서로 다르다는 점, 그리고 문장이 자연스럽고 맥락이 정교하다는 사실 자체가 그 메시지를 신뢰할 근거가 되지는 않는다는 점을 핵심적으로 전달할 필요가 있다.

넷째, 언어 모델의 안전장치를 지속적으로 강화할 필요가 있다. 본 연구는 AI의 통제망을 우회하려는 단순한 프롬프트 조작만으로 실제 범죄 문구보다 기만성이 높은 텍스트가 생성될 수 있음을 확인하였다. 이는 대규모 언어 모델(LLM) 제공자들이 안전장치를 지속적으로 점검·강화하여 범죄에 악용될 가능성을 줄여야 함을 시사한다. 특히 무료로 누구나 이용할 수 있는 상용 모델은 안전장치가 취약할수록 악용이 쉽게 확산될 수 있어, 우선적으로 점검·보완해야 할 대상이다. 다만 이미 다크웹에는 안전장치가 제거된 악성 모델(예: FraudGPT, WormGPT)이 유통되고 있어, 궁극적으로는 이러한 모델을 차단하는 방안까지 함께 모색할 필요가 있다.

본 연구의 결과는 다음의 세 가지 후속 연구로 확장될 수 있다. 첫째, 최근 위조 이미지·문서를 결합한 스미싱이 증가하는 추세를 고려할 때, 텍스트를 넘어 이미지 위·변조 탐지를 통합한 멀티모달 방어 연구가 필요하다. 둘째, 성별·직업·투자 관심도 등 다양한 인구통계학적 특성과 자극물 유형을 결합한 다차원 설계를 통해 취약 집단을 정밀하게 규명하고 맞춤형 예방 정책을 마련할 필요가 있다. 셋째, AI 에이전트 기반의 능동적 탐지·방어 기술을 개발하여 실제 수사 지원 시스템에 통합하는 방안이 모색되어야 할 것이다.

Ⅵ. 결어

6.1. 연구 요약 및 시사점

본 연구는 생성형 AI가 작성한 스미싱 문구가 인간 범죄자가 작성한 문구보다 유의하게 높은 기만성을 유발함을 2(생성주체)×3(심리기제)×2(연령) 혼합 요인 설계를 통해 확인하였다. 본 연구의 핵심적 발견은 AI 작성 스미싱이 실제 공식 기관의 정상 메시지와 통계적으로 변별되지

않는 수준에 도달하였다는 점이다. 이는 수신자가 문구의 내용만으로 해당 메시지의 진위 판별을 하는 것이 정교화된 AI 앞에서 무력화될 수 있음을 보여준다. 아울러 디지털 기기에 친숙한 세대로 여겨지는 청년층이 AI가 작성한 텍스트에 오히려 더 취약하다는 것을 발견함으로써, 기존의 통념과 차별된 새로운 시사점을 도출하였다. 본 연구의 결과는 향후 AI 기반 스미싱 방어 시스템의 개발 방향, 연령 맞춤형 예방 교육 설계, 그리고 생성형 AI에 대한 사회적·정책적 규제 논의의 기초자료로 활용될 수 있을 것이다. 사이버 범죄가 기계의 속도로 산업화되는 시대에, 본 연구는 인간의 인지적 취약성과 AI의 기만 능력이 만나는 지점을 드러내고, 그에 대한 방어 체계 구축의 필요성을 환기한다.

본 연구의 시사점은 다음과 같다.

첫째, 생성형 AI의 발전이 유발하는 사회공학적 위협을 수신자의 심리적 반응 차원에서 규명하였다. 기존 보안 연구들이 악성 URL이나 앱 탐지 등 기술적 방어에 집중했던 것과 달리, 본 연구는 공격의 최종 수용자인 인간이 느끼는 기만성을 확인함으로써 사회공학적 공격의 방어를 인간 중심으로 확장하는 학문적 토대를 제공한다.

둘째, 디지털 기기에 친숙한 세대일수록 사기에 안전할 것이라는 기존의 통념과 달리, 청년층이 중장년층보다 기만에 더 취약하며, 특히 AI가 작성한 문구에서 더 기만성이 높아짐을 확인하였다. 이는 디지털 친숙도가 정보의 진위 판별 능력을 보장하지 않음을 보여주는 것으로, 사기 취약성에 대한 논의를 고령층 중심에서 전 연령으로 확장할 필요가 있음을 시사한다.

셋째, 수신자가 메시지를 참으로 믿는 신뢰(인지)가 실제 링크를 누르는 행동 의도(행동)까지 이어진다는 점을 발견하였다. AI 문구는 진위 신뢰뿐 아니라 실제 링크 클릭 의도까지 유의하게 상승시켰다. 다만 효과 크기는 신뢰도가 행동 의도보다 다소 컸는데, 이는 행동 단계에서 미약한 저항이 존재하지만, 안정적 방어선으로 보기는 어려움을 시사한다.

6.2. 연구의 한계

본 연구의 한계는 다음과 같다.

첫째, 본 연구는 만 19세에서 64세 사이의 성인 126명(청년층 65명, 중장년층 61명)을 대상으로 진행되어, 그 결과를 국내 청년층 및 중장년층 전체로 일반화하는 데 한계가 있다. 특히 디지털 기기에 대한 친숙도, 보안 지식수준, 사기 피해 경험 유무 등 개인의 특성에 따라 기만성 지각이 달라질 수 있다. 따라서 향후 연구에서는 대규모 표본을 대상으로 광범위한 실증적 검증이 이루어져야 할 것이다.

둘째, 본 연구의 자극물 평가는 구글 폼을 이용한 시나리오 기반의 설문 형태로 이루어졌다. 이는 참여자가 가상의 맥락에서 메시지의 진위와 행동 의도를 주관적으로 평정하는 방식이므로, 실제 모바일 환경에서 스미싱을 실시간으로 접할 때의 행동(실제 링크 클릭)과는 생태학적 간극이 존재한다. 따라서 향후 연구에서는 사용자의 주의 분산이나 실시간 인지 부하를 반영할 수 있도록, 실제 모바일 환경에서 링크 클릭 행동을 추적하는 행동 측정 기법의 도입을 제안한다.

셋째, 자극물 제작에 단일 LLM과 특정 페르소나 프롬프트를 사용하고 응답을 4점 리커트 척도로 측정한 점은 자극물의 대표성과 응답의 정밀도를 제한할 수 있다. 사이버 범죄에 악용되는 생성형 AI 모델이 점차 다양해지고 프롬프트 우회 기법도 진화하고 있음을 고려하면, 특정 모델에 의존해 제작한 자극물은 실제 위협의 가변성을 충분히 반영하지 못했을 수 있다. 따라서 향후 연구에서는 다양한 LLM 모델과 프롬프트 조건을 체계적으로 비교하여 AI 기반 방어 모델의 실제 탐지 성능을 검증할 필요가 있다.

넷째, 본 연구는 명확히 정의된 사회공학적 기제를 바탕으로 자극물을 구성하였으나, 연구자 개인이 기제를 분류하는 과정에서 주관적 판단이 개입했을 가능성을 배제하기 어렵다. 따라서 향후 연구에서는 다수의 평가자가 자극물 분류에 참여하여 평가자 간 일치도를 산출하는 검증 절차를 포함할 필요가 있다.

다섯째, 본 연구는 실험 자극물 간의 언어적 이질성을 충분히 통제하지 못하였다. 인간 작성본으로 KISA에 신고된 실제 범죄 데이터를 원형 그대로 사용하여 생태학적 타당도를 확보하였으나, 이로 인해 다수의 익명 행위자가 작성한 파편적이고 이질적인 표현들이 혼재되었다. 이는 단일 모델(Gemini)에 동일한 페르소나를 부여해 산출한 일관된 AI 텍스트와 비교할 때 척도의 내적 일관성(Cronbach's α)에서 차이를 유발하는 원인이 되었다. 따라서 향후 연구에서는 인간 작성본의 생태학적 타당도를 유지하면서도 실험 변인의 통제력을 높일 수 있는 정교한 자극물 매칭 기법이 요구된다.

감사의 글

본 연구에 사용된 원천 데이터는 한국인터넷진흥원(KISA) 사이버보안빅데이터센터로부터 제공받았습니다.

참고문헌(References)

[1] Korea Communications, Media and Broadcasting Commission, Korea Internet & Security Agency (KISA). 2026. Spam distribution status in the second half of 2025 [2025년 하반기 스팸 유통현황]. KISA, Naju, Korea.

[2] Brundage M, Avin S, Clark J, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv:1802.07228. https://doi.org/10.48550/arXiv.1802.07228

[3] Europol. 2024. Internet organised crime threat assessment (IOCTA) 2024. European Union Agency for Law Enforcement Cooperation, The Hague, Netherlands.

[4] Bailo P, Sirignano A, Nittari G, et al. 2026. A review of crime at machine speed: Criminological aspects of artificial intelligence's industrialisation of deception. Sci, 8(3), 54. https://doi.org/10.3390/sci8030054

[5] Europol. 2025. European Union serious and organised crime threat assessment (EU-SOCTA) 2025. European Union Agency for Law Enforcement Cooperation, The Hague, Netherlands.

[6] Chlasta K. 2025. The dual-use dilemma of generative artificial intelligence in cybersecurity: Navigating the explosive growth in offensive and defensive applications. Security and Defence Quarterly, 52(4), 4-22. https://doi.org/10.35467/sdq/217364

[7] Schmitt M, Flechais I. 2024. Digital deception: Generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review, 57, 324. https://doi.org/10.1007/s10462-024-10973-2

[8] Hossain I, Ahad T, Alam MJ, et al. 2026. The art of the jailbreak: Formulating jailbreak attacks for LLM security beyond binary scoring. arXiv:2605.09225. https://doi.org/10.48550/arXiv.2605.09225

[9] Choi K, Kim M. 2016. A study on the modus operandi of smishing crime for public safety [국민안전을 위한 스미싱 범죄수법분석]. Journal of Convergence Security, 16(3), 3-12.

[10] Shim YE, Choi ES. 2023. The study of fraud vulnerability: Focusing on age, personality traits and mind reading ability [사기 취약성에 대한 연구: 연령, 성격특성 및 마음읽기 능력을 중심으로]. The Korean Journal of Developmental Psychology, 36(1), 105-131. http://doi.org/10.35574/KJDP.2023.3.36.1.105

[11] Korean National Police Agency, Cyber Bureau. 2015. Glossary of cyber investigation terms [사이버수사 용어집], p. 16. Korean National Police Agency.

[12] Zhang Z, Zhao P, Ye D, et al. 2025. Enhancing jailbreak attacks on LLMs via persona prompts. arXiv preprint arXiv:2507.22171. https://doi.org/10.48550/arXiv.2507.22171

[13] Park KS. 2025. The impact of the factors of smartphone dependency and its usage motivation on the risk perception of voice phishing·smishing: Focusing on college students [스마트폰 의존도와 이용동기 요인이 보이스피싱·스미싱의 위험지각에 미치는 영향: 대학생 집단을 중심으로]. The Journal of the Korea Contents Association, 25(3), 513-522. http://doi.org/10.5392/JKCA.2025.25.03.513

[14] Williams EJ, Polage D. 2019. How persuasive is phishing email? The role of authentic design, influence and current events in email judgements. Behaviour & Information Technology, 38(2), 184-197. https://doi.org/10.1080/0144929X.2018.1519599

[15] Heiding F, Lermen S, Kao A, et al. 2024. Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. arXiv:2412.00586. https://doi.org/10.48550/arXiv.2412.00586

[16] Cialdini RB. 2021. Influence, new and expanded: The psychology of persuasion. Harper Business [황혜숙 등 역. 2023. 설득의 심리학 1. 21세기북스].

[17] Chyung SY, Roberts K, Swanson I, et al. 2017. Evidence-based survey design: The use of a midpoint on the Likert scale. Performance Improvement, 56(10), 15-23. https://doi.org/10.1002/pfi.21727

[18] Jakesch M, Hancock JT, Naaman M. 2023. Human heuristics for AI-generated language are flawed. Proceedings of the National Academy of Sciences, 120(11), e2208839120.

https://doi.org/10.1073/pnas.2208839120