„Defence and Science“ № 4 (2025)
www.defenseandscience.eta.edu.ge

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

LARGE LANGUAGE MODELS AND THEIR ABUSE IN HIGH-LEVEL SOCIAL ENGINEERING CAMPAIGNS

Iveri Jajanidze[1]
Ioseb Kartvelishvili[2]
doi.org/10.61446/ds.4.2025.10444

Article History:
Received 21 September 2025
Accepted 10 October 2025
Published 25 December 2025

ABSTRACT

The rapid evolution and widespread accessibility of Large Language Models (LLMs) has transformed the cyber threat landscape. While LLMs deliver major benefits to productivity, code acceleration, knowledge augmentation, and domain translation, they simultaneously enable a new generation of high-level, linguistically precise cyber deception operations. This paper examines the shift in social engineering strategy induced by generative models, analyzing how adversaries now leverage AI to produce contextually aligned, psychologically adaptive, multilingual attacks at scale — bypassing traditional anti-phishing controls. The paper also conceptually integrates LLM-based social engineering with emerging research showing adversarial AI misuse inside CI/CD supply chains, demonstrating that human trust manipulation and machine trust manipulation are converging into a single strategic threat dimension. The result is a unified adversarial model, where linguistic credibility becomes a scalable commodity weapon across human and automated domains. This research proposes a taxonomy of LLM-augmented social engineering attack classes, maps cognitive persuasion levers to MITRE ATT&CK technique paths, and defines a dual-plane evaluation methodology measuring both behavioral technique disruption and cognitive persuasion disruption. Findings suggest that defensive strategy must shift toward AI-augmented detection, adversarial linguistics analysis, supply-chain integrity reinforcement, and continuous cognitive resilience engineering.

Keywords: Large Language Models, Social Engineering, Adversarial AI, Phishing, Deception, Human Vulnerabilities, CI/CD Security, Supply Chain

[1] Ph.D. student of the Faculty of Informatics and Management Systems of Georgian Technical University
[2] Professor of the Faculty of Informatics and Management Systems of Georgian Technical University

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

INTRODUCTION

The scale, adaptability, and linguistic naturalness of modern Large Language Models has introduced a profound transformation in social engineering operations. Historically, cybercriminals executing high-value deception required domain familiarity, cultural intuition, specialized linguistic skill, and deliberate manual structuring of psychological influence. Today, these constraints are nearly eliminated. Generative AI systems can produce sector-appropriate communication aligned to executive voice, internal organizational tone, and professional lexical norms. These outputs are indistinguishable from legitimate business processes, contracts, communications, and governance language — enabling strategic exploitation of financial authorization, procurement, legal compliance, and risk governance decision flows.

Simultaneously, academic research has documented generative AI misuse within software supply chain pipelines, demonstrating that LLMs can influence configuration semantics inside automated CI/CD ecosystems to induce subtle but damaging integrity deviations.[3] These developments indicate that generative AI manipulation now transcends traditional domain boundaries: it can weaponize both human cognitive channels and automated DevSecOps trust channels. This paper investigates this convergence and proposes a structural model to evaluate and mitigate these threats.

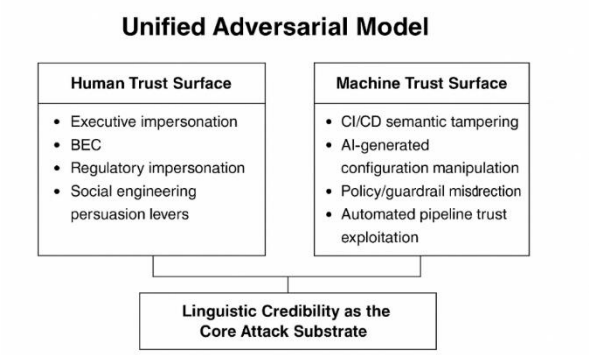

Fig.1. Unified Adversarial Model

MAIN PART

[3] I. Jajanidze, “The Use of Artificial Intelligence in CI/CD Systems: Enhancing Security and Managing Risks,” Georgian Scientists / ქართველი მეცნიერები, vol. 7, no. 3, 2025. doi: https://doi.org/10.52340/gs.2025.07.03.0.

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

LITERATURE REVIEW

Early social engineering research focused heavily on linguistic anomalies, syntactic errors, and lexical signature patterns as detection opportunities .[4] Prior to generative AI, phishing and BEC campaigns frequently exhibited poor grammar, culturally non-native formulations, and structural inconsistencies. These markers provided defenders with implicit heuristics based on linguistic incongruence. However, LLMs now remove these protective artifacts. Recent work by Galinkin identified that AI-generated phishing can maintain domain-specific sublanguage fidelity while varying surface lexical form continuously.[5] This enables polymorphic linguistic deception, where traditional filtering based on static matching fails.

Cialdini's persuasion principles, historically applied in psychology, explain why specific deception vectors successfully bypass human rational filters. Principles such as Authority, Liking, Reciprocity, Commitment, Social Proof, Scarcity, and Unity systematically manipulate belief frames and compliance outcomes. The relevance of these models to modern cyber SE is amplified when combined with LLMs, because AI automates tailored persuasion - dynamically adjusting language per target persona, socio-economic context, industry semantics, and cultural background. Generative AI thus changes persuasion from a manual craft into a scalable industrial capability.

In parallel, research investigating adversarial machine learning traditionally focused on perturbation attacks against vision models, classifier destabilization, and prompt-based jailbreak manipulation. Current threat intelligence reporting indicates a shift from raw classifier evasion toward AI-driven deception operationalization.[6,7] This shift places natural language as the adversarial substrate itself, not merely an input vector.[8] The literature demonstrates fragmentation across three silos: psychological persuasion theory, technical cybersecurity TTP frameworks, and geopolitical influence operations. This paper's novelty lies in unifying them under a single systemic threat interpretation.

[4] S. Sheng, M. Holbrook, P. Kumaraguru, L. Cranor, J. Downs, “Who Falls For Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010.
[5] A. Galinkin, “AI-Driven Social Engineering: The Next Evolution of Cyber Threats,” Trend Micro Research, 2023.
[6] Microsoft Threat Intelligence, “Business Email Compromise Trends 2024,” Microsoft Security Threat Intelligence Report, Microsoft Corporation, 2024. Available: https://www.microsoft.com/security/blog
[7] European Union Agency for Cybersecurity (ENISA), “ENISA Threat Landscape 2024,” ENISA Publications, European Union, 2024. Available: https://www.enisa.europa.eu/publications
[8] Proofpoint Threat Research, “Threat Landscape Update: AI-Enhanced Social Engineering Campaigns,” Proofpoint Research Intelligence Report, Sunnyvale, CA, USA, 2024. Available: https://www.proofpoint.com

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

RELATED WORK

Prior studies of phishing automation and Business Email Compromise (BEC) trends reveal that AI-generated influence content increasingly bypasses lexical anomaly detection and static pattern heuristics.[9] Vendor threat reports have also documented attacker use of LLMs to rapidly generate persona-tailored variants during live attack progression, enabling adversaries to run multi-path deception branching in parallel. In psychology, research on susceptibility and cognitive bias indicates that social engineering is primarily effective because it exploits pre-programmed behavioral heuristics rather than technical vulnerabilities.[10] Meanwhile, CI/CD and DevSecOps research has begun identifying generative AI misuse not only as code synthesis risk, but as a pipeline trust manipulation vector.[11] However, the intersection of these domains remains underdeveloped academically: few works systematically examine how AI-powered persuasion at scale collapses the defense advantage of linguistic intuition while simultaneously eroding machine trust channels. This paper extends the literature by presenting a threat taxonomy based on persuasion levers aligned with MITRE ATT&CK, and by showing that the same linguistic deception patterns can apply to CI/CD configuration semantics.

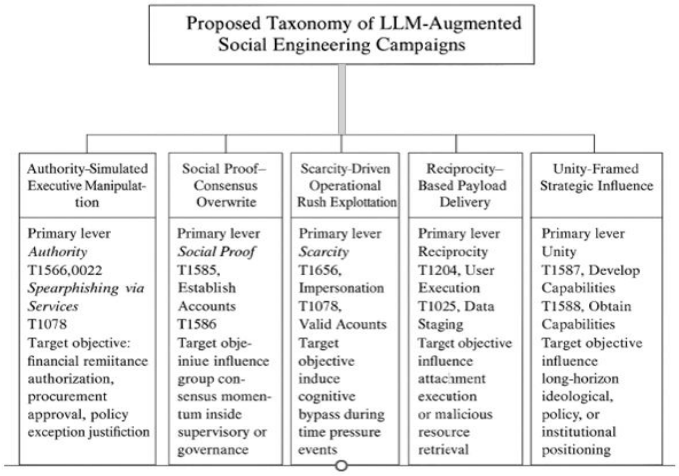

Fig. 2. Taxonomy of LLM-Augmented Social Engineering Campaigns

Proposed Threat Taxonomy of LLM-Augmented Social Engineering Campaigns

Traditional phishing terminology (spam, spearphishing, BEC) is insufficient to characterize adversarial AI deception because these labels describe delivery channels rather than persuasion

[9] Ibid. Microsoft Threat Intelligence, 2024.
[10] Ibid. S. Sheng, M. Holbrook, P. Kumaraguru, L. Cranor, J. Downs, 2010.
[11] Ibid. I. Jajanidze, “2025.

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

logic. The taxonomy proposed here classifies AI-driven SE based on (1) primary persuasion lever, (2) MITRE ATT&CK alignment, and (3) operational objective.

Authority-Simulated Executive Manipulation

Primary lever: Authority.

MITRE alignment: T1566.002 Spearphishing via Services; T1078 Valid Accounts. Target outcome: financial remittance authorization, procurement approval, policy exception justification.

Novel AI property: idiolect consistency replication across multi-email thread context.

Social Proof–Driven Consensus Overwrite

Primary lever: Social Proof.

MITRE alignment: T1585 Establish Accounts; T1586 Compromise Accounts.

Target: influence group consensus momentum inside supervisory or governance decision loops.

Novel AI property: variant distribution per internal subgroup identity cluster.

Scarcity-Driven Operational Rush Exploitation

Primary lever: Scarcity.

MITRE alignment: T1656 Impersonation; T1078 Valid Accounts.

Target: induce cognitive bypass during time pressure events.

Novel AI property: realistic regulatory deadline synthesis.

Reciprocity-Based Payload Delivery

Primary lever: Reciprocity.

MITRE alignment: T1204 User Execution; T1025 Data Staging.

Target: induce attachment execution or malicious resource retrieval.

Novel AI property: jurisdiction-correct legal/contract template production.

Unity-Framed Strategic Influence

Primary lever: Unity.

MITRE alignment: T1587 Develop Capabilities; T1588 Obtain Capabilities. Target: influence long-horizon ideological, policy, or institutional positioning.

Novel AI property: narrative alignment with institutional identity markers.Cross-category observation: the taxonomy demonstrates that AI enables *persuasion operationalization*, not just message generation. Persuasion becomes modular, composable, and adaptive.

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

Proposed Threat Taxonomy of LLM-Augmented Social Engineering Campaigns

	MITRE alligment	Target Outome	Novel AI Property
Authority-Simulated Executive Manipulation	T1566,002 Spearphishing via Services T1076 Valid Accounts	financial remittance authorization, procurement approval, policy exception justification	idiolect consistency replication across mu-lti-emall thread context
Social Proof-Driven Consensus Overwrite	T1585 Establish Acounts T1586 Compromise Accounts	influence group consensus momentum inside supervisory or governance decision' loops	variant distribution per internal subgroup identity cluster
Scarcity-Driven Operational Rush Exploitation	T1656 Impersonration T1078 Valid Accounts	induce cognitive bypas during time pressure events	realistic regulatory deadline synthesis
Reciprocity-Based Payload Delivery	T1587 Develop Capabilities T1588 Obtain Capabillties	induce attachment execution or malicious resource retrieval	jurisdiction-correcct legal/contract template production

Fig. 3. MITRE-Aligned Threat Classification Table

METHODOLOGY

This research applies a conceptual model construction approach based on triangulation across cybersecurity attack frameworks, persuasion psychology, and supply-chain adversarial AI literature. The evaluation framework defines two disruption planes: (1) ATT&CK technique chain interruption, and (2) persuasion lever interference. Controls must operate on both planes to achieve resilience at scale. Data sources include academic literature, industry threat intelligence reports, and documented public incident disclosures .[12,13] Analytical validity is determined by scalability across multilingual generative variation and independence from static lexical surface artifacts.[14,15]

CASE STUDIES

Finance Sector; AI-Optimized BEC Against Treasury Workflows

Microsoft Threat Intelligence (2024) documented multiple BEC campaigns in which adversaries used LLMs to generate CFO-style communication targeting financial approval workflows.[16] These

[12] Ibid. Microsoft Threat Intelligence, 2024.
[13] Federal Bureau of Investigation (FBI), Internet Crime Complaint Center (IC3), “IC3 Annual Report 2024,” U.S. Department of Justice, 2024. Available: https://www.ic3.gov
[14] Ibid. European Union Agency for Cybersecurity (ENISA), 2024.
[15] Ibid. Proofpoint Threat Research, 2024.
[16] Ibid. Microsoft Threat Intelligence, 2024.

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

attacks demonstrated realistic vendor payment contextualization, multi-message chain continuity, and precise internal lexicon mimicry. IC3 2024 case aggregation showed that Authority and Scarcity combinations formed the highest-severity chain.[17] MITRE mapping aligns with T1566.002 and T1078.

Government / Regulatory Domain; AI-Based Policy Mandate Impersonation

ENISA's 2024 global threat landscape analysis identified AI-assisted impersonation of regulators and public sector authorities.[18] Attackers generated authoritative compliance notices referencing realistic legal frameworks, creating the appearance of legitimate enforcement instructions. Multilingual variant generation was a key bypass lever.

Corporate Supply Chain; Procurement Manipulation at Scale

Proofpoint Threat Research (2024) reported AI-augmented vendor deception campaigns distributing revised contract packages and invoice adjustments using realistic procurement narrative semantics.[19] Reciprocity was the dominant persuasion vector here. MITRE mapping: T1204 and T1585.

Cross-sector synthesis: All three sectors reflect identical underlying principle - LLMs industrialize credibility.

CONVERGENCE WITH CI/CD PIPELINE AI MISUSE

While human-targeted deception is the most visible manifestation of LLM-enabled abuse, similar linguistic manipulation patterns are emerging inside automated software supply chain ecosystems. Prior research demonstrated generative AI capability to influence CI/CD trust assumptions by crafting realistic configuration edits and policy exemptions that appear semantically aligned with organizational norms.[20] In such cases, the adversarial vector is not procedural engineering complexity, but *linguistic semantic misdirection*. Since many pipeline guardrails evaluate human-readable policy text, commit messages, and exception rationale rather than pure machine byte-level signals, LLM-generated misconfigurations can bypass static scanners by hiding within domain-consistent natural language.

[17] Ibid. Federal Bureau of Investigation (FBI), 2024.
[18] Ibid. European Union Agency for Cybersecurity (ENISA), 2024.
[19] Ibid. Proofpoint Threat Research, 2024.
[20] Ibid. I. Jajanidze, "2025.

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

This convergence collapses the historical distinction between social engineering and supply-chain compromise. In the LLM era, both human deception and machine deception are achieved through the same substrate — linguistic persuasion. The same mechanisms that exploit cognitive bias in a CFO during BEC can exploit semantic trust bias in automated pipeline logic. This unification suggests that defensive strategy must treat linguistic inputs as a first-class attack surface across organizational domains.

CONCLUSION

LLM-driven social engineering represents a structural escalation in adversarial capability. By eliminating prior linguistic and cognitive barriers, LLM-accessible attackers can deploy personalized, multilingual, psychologically-aligned influence campaigns without requiring cultural familiarity or expert domain knowledge. When combined with OSINT-driven profiling, identity simulation, and narrative adaptability, these campaigns generate historically unprecedented success potential,particularly in executive workflow, regulatory correspondence, and supply chain procurement ecosystems. Parallel developments in CI/CD semantic manipulation demonstrate that LLM misuse is not isolated to human deception: linguistic credibility can also subvert machine trust and automated security controls.

Therefore, cybersecurity strategy must expand from static domain-based filtering to adversarial cognitive resilience engineering, behavioral correlation analytics, AI-reinforced detection, and linguistic anomaly modeling. Cross-domain adversarial AI defense engineering is required to mitigate this next-generation threat class. LLM abuse is not a tactical phishing enhancement; it is a strategic cyber deception paradigm shift.

LIMITATIONS AND FUTURE WORK

This research is conceptual and interpretive, not empirical measurement of incident frequency or attack prevalence. While real-world public case references demonstrate observed operational patterns, future studies should extend this model into formal experimental evaluation environments and adversarial simulation testbeds. More longitudinal metrics are required to quantify persuasion lever weighting, multi-step deception chain success rates, and defense disruption efficacy across control planes.

Additionally, future work should explore: (1) AI-driven protection models that automatically detect persuasion construction logic rather than lexical strings, (2) cognitive SE fatigue and inoculation studies, and (3) cross-pipeline semantic integrity enforcement that cannot be bypassed

ISSN 2720-8710 (Print)
ISSN 2960-9658 (Online)

via natural language coherence. A standardized benchmark for AI-augmented social engineering resilience measurement is a priority research need.

BIBLIOGRAPHY

Bichnigauri A., Kartvelishvili I., Shonia L., “Development and implementation of a model of an effective mechanism for preventing phishing and malicious code websites in a web browser environment,” ISBN 978-9941-512-06-3, International Scientific-Practical Conference Modern Challenges and Achievements in Information and Communication Technologies, Tbilisi, Georgia, 2023.

I. Jajanidze, “The Use of Artificial Intelligence in CI/CD Systems: Enhancing Security and Managing Risks,” Georgian Scientists / ქართველი მეცნიერები, vol. 7, no. 3, 2025. doi: https://doi.org/10.52340/gs.2025.07.03.0.

S. Sheng, M. Holbrook, P. Kumaraguru, L. Cranor, J. Downs, “Who Falls For Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010.

A. Galinkin, “AI-Driven Social Engineering: The Next Evolution of Cyber Threats,” Trend Micro Research, 2023.

M. Brundage et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” Future of Humanity Institute, University of Oxford, 2018.

Microsoft Threat Intelligence, “Business Email Compromise Trends 2024,” Microsoft Security Threat Intelligence Report, Microsoft Corporation, 2024. Available: https://surl.li/kmcwtg, (Accessed 08.12.25)

Federal Bureau of Investigation (FBI), Internet Crime Complaint Center (IC3), “IC3 Annual Report 2024,” U.S. Department of Justice, 2024. Available: https://www.ic3.gov

European Union Agency for Cybersecurity (ENISA), “ENISA Threat Landscape 2024,” ENISA Publications, European Union, 2024. Available: https://surl.li/fbzdjy, (Accessed 08.12.25)

Proofpoint Threat Research, “Threat Landscape Update: AI-Enhanced Social Engineering Campaigns,” Proofpoint Research Intelligence Report, Sunnyvale, CA, USA, 2024. Available: https://surl.li/wupqpu, (Accessed 08.12.25)