

The Impact of Artificial Intelligence on the Evolution of Cyber Scams: Emerging Techniques, Threats, and Countermeasures

Dhanashree Mankar¹, Bhoomika Bhalerao², Aditi Chaubey³, Pooja Mehta⁴

¹Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities and Sciences, Belapur, Maharashtra, India.

²Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities and Sciences, Belapur, Maharashtra, India.

³Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities and Sciences, Belapur, Maharashtra, India.

⁴Assistant Professor, D. Y. Patil Deemed to be University, School of Humanities and Sciences, Belapur, Maharashtra, India.

Abstract - Artificial Intelligence (AI) is reshaping the cybersecurity environment in two directions: it can strengthen defensive operations, while the same capabilities can also be misused to improve fraudulent activity. Generative AI, large language models, synthetic media, voice cloning, and automated content generation allow scammers to produce deceptive interactions that are more convincing, targeted, scalable, and adaptive. This paper reviews how these capabilities affect phishing, spear phishing, business email compromise, social engineering, deepfake impersonation, voice-cloning scams, automated campaigns, and related forms of fraud. The study presents a structured literature and evidence-based review of government guidance, technical research, and documented incidents, using a global scope with particular attention to India. The analysis considers changes in scale, personalization, automation, multilingual communication, impersonation, multimodal deception, and the assumptions underlying conventional detection and authentication controls. Based on these findings, the paper introduces the Layered Countermeasure Framework for AI-Enabled Cyber Scams (LCF-AICS), which brings together AI-assisted detection, identity and content verification, phishing-resistant authentication, transaction and behavioural risk analysis, human and organizational safeguards, continuous monitoring, and incident response. LCF-AICS is a conceptual framework and has not been experimentally validated. Its central proposition is that protection should not depend only on identifying AI-generated content; it should also verify identity, context, requested action, and transaction authorization before a high-impact action is completed.

Key Words: Artificial Intelligence, Cyber Scams, Generative AI, AI-Assisted Attacks, Phishing, Social Engineering, Deepfake, Voice Cloning, Cybersecurity, Fraud Detection, Countermeasures

1. INTRODUCTION

Cyber scams have traditionally relied on deception, impersonation, fraudulent messages, malicious links, and

social engineering. The expansion of digital communication has increased both the speed and reach of these activities. AI introduces another enabling capability by making deceptive material easier to generate, translate, adapt, and distribute at scale. NIST describes generative AI as a technology that can produce synthetic text, images, audio, video, and other digital content, while its risk-management guidance emphasizes the need to identify and manage risks associated with these systems [1].

The FBI has documented the use of generative AI in financial fraud across several communication and media formats. Its reported examples include AI-generated text for social engineering and spear phishing, synthetic images for fictitious identities or documents, cloned audio for impersonation, and generated video used in fraudulent promotions or authority impersonation [2]. These examples indicate that AI can function as an enabling capability that strengthens established scam methods rather than necessarily creating a completely separate class of fraud.

1.1 Problem Statement

Many conventional scam-detection practices rely on observable clues such as poor grammar, generic wording, suspicious domains, unusual formatting, or visible signs of media manipulation. AI-generated material can weaken the value of some of these clues because automated systems can produce polished, context-aware communication. Synthetic voice and video can also make informal identity checks less reliable. CERT-In identifies deepfakes as a significant risk for fraud and social engineering and recommends source verification, cross-checking, authentication, and additional detection or forensic measures [3].

The central security concern is therefore broader than deciding whether a message or media file was produced with AI. Before a consequential action is approved, the decision should independently consider the claimed

identity, the surrounding context, the requested action, and the transaction involved.

1.2 Research Gap

Existing research and security guidance address several individual safeguards, including phishing detection, synthetic-content detection, authentication, provenance, and user awareness. However, an AI-assisted scam may combine several of these capabilities within a single interaction. This study therefore examines the relationship among AI capabilities, scam evolution, detection limitations, and layered countermeasures instead of assuming that one detection mechanism can provide complete protection.

1.3 Objectives

01. Examine how AI modifies established cyber-scam methods and supports emerging techniques.
02. Analyze the use of generative AI, LLMs, synthetic media, voice cloning, and automation in scam activity.
03. Identify technical, human, and organizational threats associated with AI-enabled scams.
04. Examine limitations affecting detection, authentication, awareness, and fraud prevention.
05. Evaluate existing countermeasures and propose a layered framework combining technological, procedural, organizational, and human-centric controls.

1.4 Research Questions

- RQ1. How is AI transforming traditional cyber scams?
- RQ2. How are generative AI, LLMs, synthetic media, voice cloning, and automation being applied?
- RQ3. What technical, human, and organizational threats emerge?
- RQ4. What limitations and detection gaps exist?
- RQ5. Which combinations of controls can reduce risk?

2. METHODOLOGY

This study uses a structured literature and evidence-based review. The source material is grouped into three categories: (i) official and government guidance, (ii) academic and technical research, and (iii) documented incident evidence. The analysis follows an evidence chain that connects an AI capability with the scam technique it enables, the resulting victim interaction, the security impact, the associated detection challenge, and the countermeasure relevant to that challenge. Incident reports are treated as documented examples of techniques and are not used as statistical measures of prevalence.

The review has a global scope with a specific focus on India. Indian sources are examined for regulatory warnings, cybersecurity guidance, public-sector reporting mechanisms, and examples involving financial or impersonation risks. The proposed framework is derived from the control requirements and limitations identified during the review. It is presented as a conceptual synthesis rather than as an experimentally validated security system.

3. BACKGROUND: AI AND THE EVOLUTION OF CYBER SCAMS

The evolution of AI-enabled scams can be understood as a shift from manually prepared deception toward interactions that are increasingly automated, personalized, and multimodal. Earlier scams often required attackers to write messages and select targets themselves. Digital platforms expanded the potential audience and accelerated delivery. Generative AI can further reduce the effort needed to create or revise content; personalization can incorporate contextual information into messages; synthetic media can imitate identity or presence; and multimodal systems can combine text, audio, images, video, and conversational interaction.



Fig -1: Conceptual progression of AI-enabled cyber scams.

The FBI's 2024 advisory documents examples involving text, images, audio, and video. It explains that generative AI can support higher-volume and more convincing fraudulent communication, assist translation, create fictitious profiles and documents, imitate voices, and produce videos for fraudulent or impersonation purposes [2]. NIST's work on synthetic content similarly treats synthetic material as a broad category spanning multiple media types and identifies provenance, labeling, detection, testing, and auditing as complementary approaches [4].

3.1 Dual-Use Character of AI

AI capabilities have a dual-use nature. The same mechanisms used to generate content, classify information, translate language, or identify anomalies can

support legitimate security operations and can also be misused by attackers. The presence of AI alone does not determine whether a scam will succeed. Outcomes are influenced by attacker capability, target circumstances, communication channel, existing controls, and the decisions made by the target.

3.2 Scam Categories in Scope

The review includes phishing, spear phishing, business email compromise, social engineering, deepfake impersonation, voice-cloning and vishing, fraudulent investment promotions, fake customer-support scams, social-media impersonation, and automated scam campaigns.

4. EMERGING AI-ENABLED CYBER SCAM TECHNIQUES

Table -1: Representative AI-enabled cyber-scam techniques and their security implications

Technique	AI-enabled capability	Primary security implication
AI-assisted phishing	Rapid drafting, rewriting, and language translation	Fewer obvious language-based warning signs
Spear phishing	Context-aware personalization	More focused social-engineering attempts
BEC / impersonation	Writing-style imitation and synthetic identity cues	Greater pressure on approval and verification workflows
Deepfake impersonation	Synthetic face, image, and video creation	False authority or identity signals can appear more convincing
Voice cloning / vishing	Synthetic speech and voice imitation	A familiar voice alone is weaker evidence of identity
Investment scams	Synthetic media featuring public figures or trusted identities	Fraudulent opportunities may appear more credible
Fake support	Adaptive conversational content generation	Scammers can sustain longer, more responsive interactions
Multilingual scams	Automated translation and language generation	Campaigns can reach more language groups

Technique	AI-enabled capability	Primary security implication
Automated campaigns	Large-scale generation of profiles and content	Higher campaign volume and greater variation

5. AI-DRIVEN THREAT LANDSCAPE AND SECURITY IMPACT

The techniques examined in this review can change several operational characteristics of scam activity. Scale may increase because messages, profiles, and related material can be generated more rapidly. Personalization can become more detailed when contextual information is converted into targeted communication. Automation can reduce manual effort and support longer interactions. Synthetic media can strengthen impersonation by adding audio-visual signals that appear consistent with a claimed identity. Multilingual generation can broaden targeting across language groups, while multimodal interactions can combine several deceptive cues.

These developments place additional pressure on both technical and human safeguards. Detection systems must handle changing content patterns and previously unseen generation or manipulation techniques. Authentication mechanisms must distinguish successful credential use from the legitimacy of the action being requested. Users may also make decisions under urgency, authority pressure, or familiarity cues. Protection should therefore be treated as a layered control problem rather than as the responsibility of a single classifier or a single human decision.

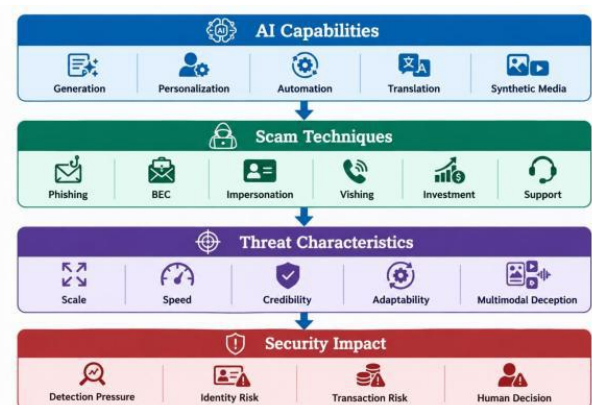


Fig -2: Threat-landscape model used in the review.

6. REAL-WORLD AND INDIA-SPECIFIC EVIDENCE

Documented advisories and incidents show that the techniques discussed above have appeared in real-world settings. In India, the Reserve Bank of India warned in

November 2024 about fake videos of the RBI Governor circulated through social media that promoted or appeared to endorse investment schemes. The RBI stated that the videos were fabricated and advised the public not to act on them [5].

CERT-In's deepfake advisory identifies realistic synthetic video, image, and audio as risks associated with misinformation, fraud, and social engineering [3]. The FBI's December 2024 advisory documents the use of AI-generated text, images, audio, and video in financial fraud and impersonation [2]. In May 2025, the FBI also reported a campaign in which malicious actors impersonated senior U.S. officials through text messages and AI-generated voice messages to build trust and encourage continued interaction [6].

Table -2: Selected documented evidence. These cases demonstrate documented techniques and should not be interpreted as prevalence estimates

Evidence	AI technique	Context	Interpretation
RBI warning, India, 2024	Deepfake video	False investment promotion using the RBI Governor's likeness	Documented risk of institutional impersonation
CERT-In advisory, 2024	Synthetic image / video / audio	Fraud and social-engineering threat	Official characterization of the threat
FBI IC3 advisory, 2024	Text / image / audio / video generation	Financial fraud and impersonation	Officially documented examples of AI-enabled fraud
FBI campaign warning, 2025	AI-generated voice + targeted messaging	Impersonation of senior officials	Documented social-engineering campaign

7. DETECTION GAPS AND LIMITATIONS OF EXISTING SECURITY MECHANISMS

AI-assisted scams expose weaknesses in controls that rely heavily on static content clues, informal identity signals, or human recognition. Language quality becomes a less dependable indicator when generated text can be polished automatically. Content detectors may also face distribution shifts as new generators, compression

conditions, or manipulation techniques appear. NIST's synthetic-content work therefore treats authentication and provenance, labeling, detection, testing, and auditing as complementary measures rather than as a single universal detection solution [4].

Identity verification presents a related challenge. NIST's Digital Identity Guidelines address identity proofing and authentication assurance [7]. This distinction matters because confirming that a claimant satisfies an identity requirement is different from deciding whether a particular action is authorized in its present context.

The resulting weakness can be understood as a trust-chain problem. A message may look legitimate, a voice may resemble a known person, and a user may authenticate successfully, yet the requested transaction can still be manipulated. Effective protection should therefore consider content, identity, context, behaviour, and action risk together.

8. EVALUATION OF EXISTING COUNTER MEASURES

Existing countermeasures remain useful, but their functions are strongest when they operate as complementary layers. AI-based detection can identify suspicious messages, URLs, behavioural patterns, or synthetic media; however, its performance should be monitored for false positives, false negatives, and changes in attack patterns. Synthetic-media detection can provide more useful evidence when it is combined, where feasible, with provenance and authentication mechanisms.

Authentication provides another important layer. NIST SP 800-63-4 and its companion volumes define requirements for identity proofing and authentication assurance [7]. For high-impact actions, successful authentication should be followed by contextual and transaction-level checks because a legitimate account holder can still be persuaded to approve a fraudulent action.

Independent callbacks or out-of-band verification can be valuable when a request is unexpected, urgent, or financially significant. Transaction monitoring can provide another layer by considering transaction amount, beneficiary changes, device or session context, timing, and behavioural anomalies. Awareness programs should emphasize verification behaviour rather than assuming that grammar problems or visible media artifacts will always expose a scam.

Table -3: Comparative evaluation of major countermeasure categories

Countermeasure	Role	Key limitation / condition
AI-based detection	Flag suspicious text, URLs, behaviour, or media	Needs continuous evaluation as content and attack patterns change
Synthetic-media detection	Identify generated or manipulated media	Performance may vary with unseen content and real-world conditions
Provenance / authentication	Provide evidence about origin and integrity	Depends on adoption and reliable signals across the ecosystem
Phishing-resistant authentication	Reduce credential-phishing risk	Does not by itself prove that a transaction is legitimate
Independent verification	Confirm the identity or request through a trusted channel	The additional channel must genuinely be independent
Transaction monitoring	Identify anomalous or high-risk actions	Intervention should be calibrated to limit unnecessary friction
Awareness / training	Strengthen verification decisions	Requires reinforcement and human responses can vary
Incident response	Contain, report, preserve evidence, and recover	Effectiveness depends on preparation and response speed

9. PROPOSED LCF-AICS FRAMEWORK

Based on the evidence and control gaps identified in the review, this paper proposes the Layered Countermeasure Framework for AI-Enabled Cyber Scams (LCF-AICS). The framework contains seven layers so that failure of a single safeguard does not automatically allow a high-impact action to proceed.

9.1 Framework Layers

Layer 1 - AI-Assisted Threat Detection: Analyze messages, URLs, media, behavioural indicators, and threat-intelligence data for suspicious characteristics.

Layer 2 - Identity and Content Verification: Validate the claimed identity and available provenance information against independent evidence.

Layer 3 - Phishing-Resistant Authentication: Use authentication appropriate to the risk level and avoid treating easily phished signals as sufficient protection.

Layer 4 - Transaction and Behaviour Risk Analysis: Examine the requested action, account context, beneficiary changes, transaction properties, and behavioural anomalies.

Layer 5 - Human and Organizational Controls: Apply independent callbacks, dual approval, trusted communication channels, and clear escalation procedures.

Layer 6 - Continuous Monitoring and Intelligence: Incorporate new indicators, incidents, false positives, and emerging techniques into detection and policy updates.

Layer 7 - Incident Response and Recovery: Enable rapid containment, reporting, evidence preservation, recovery, and lessons learned.



Fig -3: LCF-AICS: proposed seven-layer countermeasure framework.

10. OPERATIONAL RECOMMENDATIONS

Individuals: Independently verify unexpected requests. Do not treat voice, video, caller ID, or message appearance as sufficient proof of identity. Protect passwords and authentication codes, pause when urgency or secrecy is emphasized, and report suspected fraud promptly.

Organizations: Establish trusted communication channels, callback procedures, approval rules for sensitive transactions, protected employee directories, and clear escalation paths. High-risk requests should not depend on a single channel or one person's judgment.

SOC and cybersecurity teams: Correlate content, URL, identity, behavioural, and threat-intelligence signals; test detection with AI-generated samples; monitor false positives and false negatives; and maintain playbooks for AI-assisted social engineering.

Financial institutions: Combine authentication with transaction monitoring, beneficiary-change verification, anomaly detection, customer education, and rapid escalation. India's National Cyber Crime Reporting Portal provides online reporting and identifies 1930 as the helpline for immediate reporting of cyber financial fraud [9].

Awareness programs: Emphasize verification decisions instead of treating poor grammar, spelling errors, or visible media artifacts as the main warning signs.

10.1 High-Risk Decision Sequence



Fig -4: Operational verification sequence for high-risk requests.

11. DISCUSSION

The review produces five main observations. First, AI acts as an enabling layer across established scam categories. Second, generated text, synthetic media, voice cloning, and automation can collectively change the speed, personalization, adaptability, and apparent credibility of fraudulent interactions. Third, the resulting risk is socio-technical because technical detection, authentication, organizational procedures, and human decisions affect one another. Fourth, AI does not make existing controls irrelevant; it changes the assumptions and operating conditions under which those controls must be assessed. Fifth, the security objective should move beyond content classification toward verification of trust and authorization of risky actions.

11.1 Authentication Versus Transaction Authorization

Authentication and authorization represent two different decisions. Authentication determines whether a claimant satisfies an identity or credential requirement. Authorization determines whether a specific action should

be permitted in the current context. AI-assisted social engineering can target the second decision even after the first has succeeded. LCF-AICS therefore places transaction and behavioural risk analysis after authentication instead of treating successful authentication as the final safeguard.

11.2 Multimodal and Multilingual Implications

Combining text, audio, images, and video can create false corroboration, where several synthetic signals appear to support the same false identity. Multilingual and code-switched communication can further broaden the attack surface. Countermeasures should therefore use multiple independent signals and should not assume that a single media detector or language-specific rule will perform consistently in every environment.

12. PROPOSED REFERENCE ARCHITECTURE

The reference architecture converts LCF-AICS into a sequence of security decisions. Incoming communication is first examined for suspicious indicators. When necessary, identity and content are checked against independent evidence. Authentication establishes an appropriate assurance level, after which transaction and behavioural risk are evaluated. A high-risk request can then be allowed, blocked, or escalated using the combined evidence. Monitoring and incident response feed new information back into the detection and control layers.

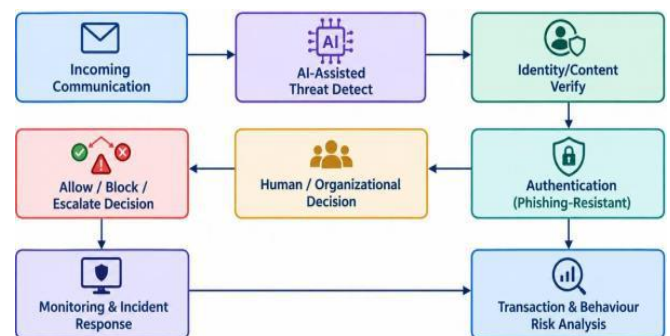


Fig -5: Proposed reference architecture for AI-enabled cyber-scam protection.

13. LIMITATIONS

This work is primarily a structured literature and evidence-based review. It does not introduce a newly collected dataset, controlled user experiment, or live organizational deployment. The findings should therefore be understood as an evidence-based synthesis rather than as a quantitative estimate of scam prevalence or a measured assessment of countermeasure effectiveness.

LCF-AICS and its reference architecture are conceptual contributions and have not been experimentally validated. Detection performance may vary with the generation

method, media quality, communication channel, language, and previously unseen content. Public incident evidence may also be incomplete or subject to later revision. India-specific observations illustrate documented risks but should not automatically be generalized to every state, language, sector, or demographic group.

The study does not assign numerical risk scores or statistically isolate the contribution of individual AI capabilities to successful scam outcomes. Human-centred recommendations are derived from the reviewed evidence and official guidance rather than from a new controlled user study.

14. FUTURE RESEARCH

Future work should evaluate LCF-AICS experimentally in controlled environments using realistic AI-generated phishing, synthetic voice, deepfake video, impersonation, and transaction scenarios. Evaluation should report detection performance, false-positive and false-negative rates, verification success, latency, intervention rates, usability, and operational overhead.

Further research should develop representative multilingual and multimodal datasets that include code-switched Indian-language scenarios and realistic communication-channel conditions. Detection systems should be tested against previously unseen generators and distribution shifts. Voice-cloning evaluations should consider noise, compression, short audio samples, and real-time interaction instead of relying only on clean benchmark recordings.

Future Research: Feedback Loop into Policies

Future studies can also investigate adaptive and privacy-preserving detection, provenance and watermarking, behavioural and transaction-level risk analysis, human decision-making, the distinction between authentication and transaction authorization, and deployment of layered controls in real organizational environments.

14.1 Suggested Evaluation Metrics

Table -4: Candidate metrics for future empirical validation

Dimension	Potential measures
Detection	Precision, recall, false-positive and false-negative rates, and robustness to unseen content
Identity verification	Verification success, escalation rate, and time to verify

Dimension	Potential measures
Authentication	Authentication success or failure, phishing resistance, and step-up frequency
Transaction risk	Intervention rate, anomaly detection, and prevented high-risk actions
Human factors	Verification compliance, decision time, and usability
Operations	Latency, computational overhead, and incident-response time

15. CONCLUSION

Artificial Intelligence is changing cyber scams by making deceptive content easier to generate, personalize, automate, and distribute. The reviewed evidence indicates that generative AI, synthetic media, voice cloning, and related capabilities can be incorporated into phishing, social engineering, impersonation, vishing, investment fraud, fake-support operations, and other scam categories. The security concern is therefore not limited to whether a piece of content was generated with AI; it also includes how AI changes an attacker's ability to create convincing and adaptive interactions.

The review points to a shift from content-focused detection toward broader trust and risk verification. Conventional indicators remain useful, but they should not be treated as sufficient on their own. Identity claims, communication context, authentication state, requested action, and transaction risk should be considered together. This is important because a legitimate user may authenticate successfully and still be persuaded to perform an unauthorized action.

To address these requirements, the study proposes LCF-AICS, a seven-layer conceptual framework that combines AI-assisted detection, identity and content verification, phishing-resistant authentication, transaction and behavioural risk analysis, human and organizational controls, continuous monitoring, and incident response. The reference architecture translates these layers into an operational sequence from incoming communication through verification, authentication, risk assessment, decision, and response.

The framework has not been experimentally validated and should therefore be treated as a structured research contribution and a basis for future evaluation. Its central principle is that verification should become stronger and more independent as the consequences of the requested action increase. Protection against AI-enabled cyber scams consequently requires defence in depth, continuous adaptation, and coordination among technical controls, organizational processes, and human decision-making.

REFERENCES

- [1] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, and K. Roberts, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST AI 600-1, National Institute of Standards and Technology, Gaithersburg, MD, USA, Jul. 2024.
- [2] Federal Bureau of Investigation, Internet Crime Complaint Center, "Criminals Use Generative Artificial Intelligence to Facilitate Financial Fraud," Public Service Announcement I-120324-PSA, Dec. 3, 2024.
- [3] Indian Computer Emergency Response Team (CERT-In), "Deepfakes – Threats and Countermeasures," Advisory CIAD-2024-0060, Nov. 27, 2024.
- [4] B. Chandra, J. Dunietz, K. Roberts, Y.-Y. Lee, P. Fontana, and G. Awad, "Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency," NIST AI 100-4, National Institute of Standards and Technology, Nov. 2024.
- [5] Reserve Bank of India, "RBI cautions public on deepfake videos of Top Management circulated over social media giving financial advice," Press Release 2024-2025/1537, Nov. 19, 2024.
- [6] Federal Bureau of Investigation, Internet Crime Complaint Center, "Senior U.S. Officials Continue To Be Impersonated in Malicious Messaging Campaign," Public Service Announcement, May 15, 2025.
- [7] D. Temoshok, D. Proud-Madruga, Y.-Y. Choong, R. Galluzzo, S. Gupta, C. LaSalle, N. Lefkovitz, and A. Regenscheid, "Digital Identity Guidelines," NIST SP 800-63-4, National Institute of Standards and Technology, Gaithersburg, MD, USA, Jul. 2025.
- [8] D. Temoshok, J. Fenton, Y.-Y. Choong, N. Lefkovitz, A. Regenscheid, R. Galluzzo, and J. Richer, "Digital Identity Guidelines: Authentication and Authenticator Management," NIST SP 800-63B-4, National Institute of Standards and Technology, Jul. 2025.
- [9] Government of India, Indian Cybercrime Coordination Centre (I4C), "National Cyber Crime Reporting Portal," Ministry of Home Affairs, Government of India, accessed 2026.

Source-verification note: The references were retained from the supplied paper and checked against the cited official sources where available. The LCF-AICS framework, tables, diagrams, and reference architecture are the authors' proposed synthesis and are not presented as established external frameworks.