App Threat Report
Q2 2025

PROMON

Financial app security in 2025: Combating traditional malware and emerging AI threats

Prepared by: Promon Security Research Team

Contents

Introduction

Welcome to our App Threat Report for Q2 of 2025. This is Promon's quarterly analysis and evaluation of current topics in mobile application security conducted by our Security Research Team. The focus of this report is on the cybersecurity threats facing applications used in the financial industry—threats that are both mainstream and emerging.

This report will initially focus on international malware and other threats from 2024. But it will also look at 2025 and further into the future. We explore how AI is opening new attack vectors and enhancing some that already exist. As well as providing a Promon perspective on how the threat landscape is evolving, we also analyse the most popular banking apps to reach a solid position on how secure apps are today.

Report content is divided into three sections. Firstly, we provide a global overview of the threat landscape faced by banking apps over the last year, with special reference to malware campaigns that targeted finance. Second, we take a deep dive into the worlds of deepfake threats and AI security challenges to finance, while also offering positive protection and resistance models. Finally, we share the results of recent tests undertaken by the Security Research Team on financial apps and their reaction to a common Android malware attack.

Global overview of the threat landscape

Map showing areas selected for research.

We selected areas that provided an informed sample of their respective global business regions:

- Europe for EMEA
- The USA for AMER
- Southeast Asia for APA

Europe

Within Europe, Germany and the European Union provided significant statistics on cybersecurity threats to financial applications.

Germany

A national report called The State of IT Security in Germany in 2024 [1] by the BSI (Federal Office for Information Security) revealed that there were approximately 140 APT (advanced persistent threat) groups active worldwide. Worldwide phishing IRLs and IPs detected were approximately 1000 per day. There was an increase in already known phishing campaigns in the name of banks and financial institutions, and in increase in campaigns abusing the brand names of prominent streaming services.

There was a 26% rise of new malware variants from 2023 to 2024. In terms of the average number of new malware programmes per day, this means 300 per 12-month average. Android new malware variants had above-average growth, with a 48% rise in new Android malware. This indicated a renewed development of Android attack infrastructure. The average daily growth of new Android variants per day was 4000 in June 2024.

The report showed that botnets were primarily used to steal personal information and compromise or abuse online banking

access, as well as distribute other malware. Mobile devices with Android OS were the focus of attackers. Smartphones made an attractive target for attackers because of their multifunctional nature, particularly around online banking.

During the reporting period, all reports made by security researchers to the BSI about vulnerable software products were classified according to the Open Web Application Security Project (OWASP) system. Approximately 61% of reports related to vulnerabilities that made impacted products susceptible to injection cyberattacks. Attackers could use vulnerabilities to inject malicious code into the software product, setting the stage for the next step in the attack chain.

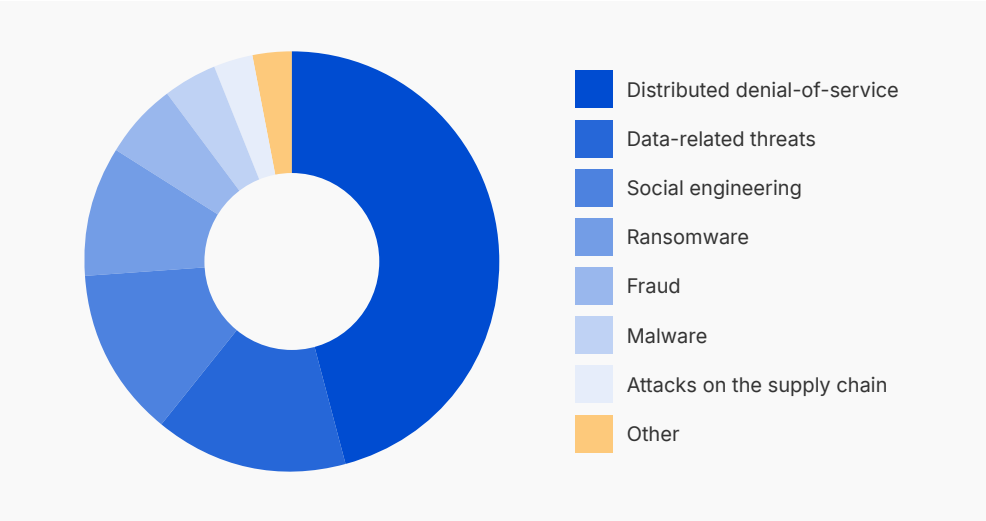
The EU and beyond

The European Union Agency for Cybersecurity (ENISA) published its first analysis of the cyber threat landscape of the EU financial sector. This report is called the ENISA Threat Landscape: Finance Sector [2] and compliments the more general ENISA Threat Landscape 2024 [3]. It identifies seven prime cybersecurity threats and provides a deep dive on each.

The seven prime threats to financial services in Europe—EU member states plus eight neighboring countries, including Norway and the UK—are identified as:

1. Distributed denial-of-service (DDoS): 46%
2. Data-related threats: 15%
3. Social engineering: 13%
4. Ransomware: 10%
5. Fraud: 6%
6. Malware: 4%
7. Attacks on the supply chain: 3%

Primary cybersecurity threats to financial services in Europe, as identified by ENISA.



The report then identifies threat actors (state-nexus groups, cybercriminals, and hacktivists) before evaluating their impact.

The main finding is that of all the different types of financial organizations, European banks (credit institutions) are the most frequently affected (46%). Cybercriminals primarily target banks to steal money through fraudulent transactions, access personal customer information, and execute ransomware attacks demanding ransoms for data decryption. Cyberattack incidents more frequently lead to financials losses, regulatory penalties, and reputational damage. The report claims that similar trends are observed outside Europe.

Malware incidents included banking trojans, spyware, and miners. Malware directed at mobile devices is a "significant subset" of the broader fraud landscape. As more customers use mobile devices for banking, those targeting then use specific methods such as smishing and malicious application distribution. The report found:

- A surge in the number and complexity of mobile banking trojans
- A 200% year-on-year growth in malware families targeting banking applications
- A rise in device-takeover-capable malware families targeting banks
- 36% of malware incidents affected banks, with individuals (24%) and crypto-asset service providers (15%) next
- 58% of malware cases were fraud and large-scale financial crimes), followed by financial losses (21%), the exposure and sale of sensitive information (14%), and operational disruptions (7%)
- Emerging threats include:
    a. GoldPickaxe: A mobile trojan aimed at iOS users that is capable of synthesizing deepfake videos using stolen facial recognition data
    b. Brokewell: An Android banking trojan with extensive device-takeover capabilities

Based on this evidence, the report highlights some trends for 2024 and beyond. These include:

- A shift in tactics towards exploiting the human factor and an increase in social engineering attacks that are more "realistic, personalized, deceptive and psychologically manipulative"
- Cybercriminals tailoring their attacks to specific financial institutions
- The use of AI to power phishing scripts (reaching more potential victims with more convincing messages) and bypass traditional security measures

- Hacktivists using distributed denial-of-service (DDoS) attacks against financial institutions to hinder customer access to online and mobile banking services

The United States of America

The Federal Bureau of Investigation Internet Crime Report 2024 [4] outlined the growing scale of cyberfraud in the US.

- The FBI recorded $16.6 billion in losses due to reported internet crime in 2024 (an increase in losses of 33% from 2023)
- Cyber-enabled fraud is responsible for almost 83% of all reported losses to the IC3 although it comprised only 38% of complaints
- Financial services were the fourth largest target for cyber threats to critical infrastructure (after manufacturing, healthcare, and government facilities)
- The top five ransomware variants by complaint numbers were Akire, LockBit, RandsomHub, FOG, and PLAY

A research report on Cybersecurity in US Financial Services [5] in 2024 based its findings on a survey of different types of financial service organizations, including boutique wealth managers, credit unions, payment systems providers, hedge funds, and various types of banks. The report revealed that:

- Data compromise incidents in the US financial industry increased 330% between 2019 and 2023
- 78% of U.S. financial services companies experienced a ransomware attack in the previous 12 months. Of this large group, 48% paid a ransom. The main consequences of ransomware attacks were operational disruption (50%) and financial losses from legal fees, fines, and reputational damage (48%)
- 95% of attacks on financial services organizations are financially motivated, with just 5% focusing on espionage
- OT (operational technology) environments in U.S. financial organizations are currently at risk from malware (37%), phishing, data theft or misuse, and accidental loss or disclosure of data (each 34%)

Of special interest in the report was the role of AI in financial sector cybersecurity, with its potential as both friend and foe.

- 95% of financial service organizations are using AI-driven tools to reinforce their cyber defenses

- The most popular AI-based tools in the US financial sector are automated penetration testing and vulnerability management, AI-driven loss prevention (42%), and AI-based phishing detection and prevention (39%)
- The areas of greatest concern for financial organizations are adaptive cyberattacks (93% of respondents), followed by AI-powered botnets (92%), and polymorphic malware (83%)

Southeast Asia

A report on Consumer Attitudes Towards Fraud and Opportunities for Mobile Network Operators in SEA [6] has revealed the experiences of mobile app users across Southeast Asia (SEA) about online security when using digital banking services. This is a region where digital transactions, fintech adoption, and mobile usage in financial transactions are rapidly growing. Online consumers were surveyed across five SEA markets (Indonesia, Malaysia, the Philippines, Singapore, and Thailand).

Other relevant findings from the report include:

- More than a quarter of respondents were victims of financial crimes, such as bank card theft, identity theft, and online hacking
- Consumers in Southeast Asia hold banks and fintech firms primarily responsible for safeguarding them against financial crimes, rather than device manufacturers and network operators
- More than half of the respondents across all five markets expressed growing fears about the rising likelihood of online fraud and hacking, and the protection of their financial data
- Over three-quarters of respondents would change financial provider for better online security, such as enhanced security features
- Key fraud concerns include SIM-swap attacks, and vulnerabilities in mobile payment and e-wallet platforms

The report highlighted opportunities for strengthening account security during financial transactions that apply to applications used on mobile devices. This includes the need for real-time verification and fraud detection to reduce risk for consumers. Two-factor authentication (2FA) for transactions over mobile banking or third-party apps is also advised.

Overview of malware targeting finance in 2024

Many of the examples of malware used for financial fraud in 2024—including the first two listed below—target the Android operating system, with the purpose of initiating fraudulent money transfers from compromised devices.

Beyond this, they share three further characteristics:

1. **Fraud classification:** They are types of payment card or financial fraud known as an account takeover (ATO), in which the fraudster attempts to assume control of a victim's account. ATO is usually classified as a type of payment or credit card fraud, along with phishing, SIM swapping, and social engineering fraud.

2. **Fraud technique:** They are fraudulent activities that use the methods of on-device fraud (ODF), since they are carried out directly on the victim's device without the need to take over the victim's account from another device.

3. **Malware type:** They are instances of remote access trojan (RAT) aka creepware. This is a type of malicious software that controls a system via a remote network connection without the victim's knowledge or consent. They are disguised as a normal program, hiding its operations from both the victim and cybersecurity software.

BingoMod

BingoMod is malware that emerged in May 2024. It is disguised as legitimate security tools that are used to protect devices. It is inferred from the languages used in target devices that human targets were English, Romanian, and Italian users. The malware developers may be Romanian speakers.

What is interesting about BingoMod is its combination of functions. As with other malware in the financial sector, threat actors use it to gain remote control of target devices by means on an overlay attack. BingoMod then carries out its primary purpose, which is the theft of sensitive information and money.

Beyond this, BingoMod contains different functions designed to help it evade detection as well as swipe data. For example, the malware employs obfuscation techniques to lower its detection rate against antivirus software. And, in the case where it is detected, BingoMod employs a self-destruction mechanism that wipes the infected device to hinder forensic investigations. This mechanism is also used to remove evidence when a successful fraud is completed.

BingoMod exhibits similar capabilities to banking trojans such as Copybara, Medusa, TangleBof and TeaBot. But it appears that

BingoMod malware is still in development. There is evidence, for example, that its developers are still experimenting with obfuscation techniques. They also seem interested in stripping back complex functionality to focus on anti-analysis configurations.

ToxicPanda

ToxicPanda emerged in early-to-mid 2024. It uses icons of well-known brands (web browsers and credit cards) as well as decoys that resemble dating apps to target retail banking institutions.

Although infected devices are found in Europe (Portugal, Spain, and especially Italy), and Latin America (especially Peru), the malware originates in Southeast Asia. It is believed the developers of ToxicPanda may be Chinese speakers, and that it is still in the early stages of development.

Because of family resemblances, ToxicPanda was initially associated with the malware known as TgToxic, albeit in a more basic form. TgToxic emerged in mid-2022, targeting Android mobile users with bank and finance apps in Indonesia, Taiwan, and Thailand. The malware also stole credentials and assets from digital wallets, with supplementary phishing, cryptocurrency, and sextortion scams.

Android malware

2024 has revealed the existence of recent iterations of previously known malware strains. This highlights the continuous evolution of malware and the sophistication of threats. Old versions appear with new features and actions, including the ability to disable or bypass recent security measures. They are customized and directed against financial targets. These examples focus on the Android mobile OS.

Coper

Coper malware—also known as Octo [7]—was used in campaigns in mid-January and early February 2024. It is a banking trojan disguised as Chrome Android applications. These are hosted on content delivery networks and customer service platforms. The malware displays fake window overlays to collect sensitive information and deceive users into surrendering their credentials. Octo is a rebranded version of another Android malware called ExobotCompact.

Octo is described by the cybersecurity industry as a rental trojan that is spread by fake apps on Google Play Store. It targets banks and other financial institutions. It is modular in design and includes a multi-state infection method. It also employs many defensive tactics to survive removal attempts. It may pose as a Play Store app installer app, a screen recording app, or a financial app.

Vultur

In May of 2024, the Finnish National Security Centre (NCSC-FI) warned in a weekly review that smishing—SMS phishing—messages were directing users to Android malware to steal their banking information. The attack chain employed a technique known as telephone-oriented attack delivery (TOAD). After an initial SMS message to call a number, the victim is informed on the call that they need to install an antivirus app for protection, the link for which is contained in a second message. However, this supposed security software contains malware designed to steal online banking credentials and funds.

Although the NCSC-Fi review didn't name the malware strain involved in this campaign, it is suspected to be an upgraded version of Vultur, due to its near identical infiltration process. Vultur is an Android banking trojan that was first disclosed in 2021. But it reappeared in 2024 with new features to increase its control over infected devices, and new detection evasion abilities. As well as the TOAD attach chain, it also masquerades as authenticator apps offered as a dropper-as-a-service (DaaS) operation called Brunhilda.

AI attacks in banking

The banking sector has witnessed an unprecedented 1,530% surge in deepfake attacks across Asia-Pacific between 2023 and 2024, with sophisticated AI-powered threats now directly targeting bank authentication systems and customer accounts.

The most significant banking-focused campaign emerged through the GoldPickaxe malware, mentioned in the ENISA Threat Landscape: Finance Sector report above.

GoldPickaxe specifically targeted banking apps across Thailand and Vietnam by exploiting new facial recognition requirements. This malware tricked bank customers into downloading fake banking apps that prompted users to record verification videos, which were then converted into deepfakes capable of bypassing legitimate banking authentication systems.

The GoldPickaxe campaign successfully targeted multiple banking applications across Southeast Asia, with attackers able to bypass facial recognition systems used by banks for customer authentication. Chinese banking customers have suffered significant losses, with documented cases including a victim who lost millions of yuan through deepfaked video calls that bypassed banking verification procedures, demonstrating how these attacks specifically target bank customer authentication processes.

The democratization of deepfake technology has transformed attacks on banking systems from sophisticated operations to accessible criminal tools costing as little as $20-$1,000 on underground markets, specifically targeting bank customer authentication. Free, open-source tools like DeepFaceLab and Deep-Live-Cam now enable real-time face swapping that can bypass banking video verification in real-time, with attack preparation time dropping to just 20 minutes for voice cloning that defeats banking phone verification systems. This accessibility has driven attacks against Western banks, with North America experiencing a 1,740% increase in deepfake fraud targeting banking institutions and Europe seeing a 780% surge in banking sector incidents in 2024.

Current penetration testing reveals that 15 out of 20 major banks remain vulnerable to basic deepfake attacks against their customer authentication systems, with success rates of 85-95% against standard banking biometric verification. Major banks like JPMorgan Chase [8] have acknowledged that deepfakes targeting their systems “keep them up at night”, while the Federal Reserve [9] has warned of “supercharged identity fraud” specifically threatening banking authentication.

The financial impact on banking institutions is projected to reach $40 billion in losses by 2027, with underground marketplaces now serving 34,965 users across 31 deepfake service vendors specifically offering banking system bypass tools, fundamentally threatening the integrity of bank customer authentication and mobile banking security worldwide.

Deep fake threats in biometrics

While biometric authentication encompasses various methods including fingerprint, voice recognition, and iris scanning, our expertise and focus centres specifically on the video-based face authentication systems increasingly used by financial institutions. These systems are currently facing an unprecedented threat that demands urgent attention.

The banking sector in Asia has already witnessed a dramatic surge in sophisticated deepfake video attacks targeting face authentication systems. This alarming trend is no longer confined to isolated incidents. Advanced AI tools that were once accessible only to technical specialists are now readily available to the public. Applications like Haotian AI and Deep-Live-Cam and other have democratized deepfake creation, allowing virtually anyone to generate convincing fake videos with zero technical knowledge.

While EU/US has not yet experienced the same volume of attacks as Asia, this is rapidly changing. Our analysis indicates that European financial institutions face an imminent threat, with sophisticated deepfake attacks likely to increase dramatically in the coming 12-24 months as these technologies and methodologies spread globally.

The window for preparation is closing. European banks and financial service providers must implement robust defenses before deepfake attacks become commonplace. Understanding the limitations of current detection approaches and developing comprehensive protection strategies is essential to maintain the integrity of face authentication systems in this evolving threat landscape.

> "Our analysis indicates that European financial institutions face an imminent threat, with sophisticated deepfake attacks likely to increase dramatically in the coming 12-24 months"

Why AI-only solutions fall short in mobile deepfake video detection

When it comes to securing mobile banking apps against deepfake attacks, relying solely on AI-based detection models presents significant challenges that can compromise the effectiveness of security measures.

The computational reality of mobile devices

Most current deepfake detection algorithms rely heavily on sophisticated deep neural networks that demand substantial computational resources. On a mobile device, these resource-intensive models face several critical constraints:

- **Limited processing power:** Modern smartphones, while powerful, still lack the GPU capabilities of dedicated machines where these models are developed and tested
- **Battery impact:** Running complex neural networks continuously during authentication drains battery life rapidly
- **Memory constraints:** Deep learning models often require significant RAM, competing with other apps and system processes
- **Heat generation:** Sustained AI processing causes devices to heat up, triggering throttling that further degrades performance

As mentioned in our blog post on deepfake attacks in mobile banking [10], mobile apps operate in a sandboxed environment with limited access to raw camera and sensor outputs. This architectural complexity further complicates the deployment of computationally heavy AI models directly on the device.

The moving target problem

Deepfake technology evolves at a rapid pace, creating an ongoing arms race between security measures and attackers. AI-only solutions face significant challenges in this dynamic environment:

- **Model staleness:** A model trained on today's deepfake techniques may be ineffective against tomorrow's innovations
- **Retraining requirements:** Updating AI models requires collecting new training data, retraining, and redeploying—a process that typically trails new attack methods
- **Signature-based limitations:** Many AI models effectively learn 'signatures' of known deepfake methods, making them vulnerable to novel techniques

Analysis from two research papers shows that most deepfake detection algorithms assume a static threat landscape, which doesn't reflect the reality of constantly evolving manipulation technologies. These papers are:

1. 'Deepfake video detection: challenges and opportunities' [11]. Kaur, A., Noori Hoshyar, A., Saikrishna, V. et al. (2024).
2. 'DeepFake video detection: Insights into model generalisation —A Systematic review' [12]. Ramcharan Ramanaharan, Deepani B. Guruge, Johnson I. Agbinya (2025).

Cross-platform vulnerabilities

Our blog post on deepfake attacks in mobile banking also highlights that the security posture varies significantly between iOS and Android. AI-only detection approaches struggle with these platform differences:

- **Platform optimization:** Models optimized for one platform may perform poorly on another
- **Hardware diversity:** Android's fragmented ecosystem means that a model performing well on flagship devices may fail on budget phones
- **API inconsistencies:** Differences in camera APIs and processing pipelines between platforms create blind spots in detection

This platform variability makes it nearly impossible for a single AI model to provide consistent protection across the mobile device landscape.

The case for layered defense

Rather than relying solely on AI, a more effective approach incorporates multiple complementary methods that operate at different levels of the mobile banking security stack. These methods operate on two different levels: short-term and long-term.

Short-term defensive measures

Our blog on deepfake attacks in mobile banking called attention to the factors that effective mobile banking security should integrate:

- **App shielding:** Detecting hooking frameworks, virtualization tools, and debugging attempts at the application layer
- **KYC and identity verification:** Working with providers that employ multiple anti-deepfake technologies

- **Behavioral analysis:** Monitoring login patterns and user interactions to flag anomalies
- **Transaction authentication:** Implementing strong multi-factor verification for high-risk activities

Long-term strategic initiatives

For comprehensive protection, organizations should pursue:

- **Platform-provided biometric APIs:** Leveraging Apple Face ID or Android's BiometricPrompt which benefit from hardware-level security
- **Regular security testing:** Specifically focusing on face authentication bypass techniques
- **OS-level biometric security:** Collaborating with platform providers to strengthen the camera pipeline security

Conclusion: Beyond AI alone

While AI plays a crucial role in deepfake detection, an AI-only approach is insufficient for securing mobile banking applications. The computational limitations of mobile devices, the rapidly evolving threat landscape, and the complex platform differences all point to the necessity of integrated, multi-layered defenses.

The most effective deepfake detection strategies combine AI with traditional security measures, hardware-based protections, and platform-specific optimizations. This layered approach provides greater resilience against both current and future deepfake attacks, ensuring that mobile banking apps can maintain both security and usability.

Security in mobile banking isn't just about having the most sophisticated AI model—it's about implementing a comprehensive set of defences that work together to protect users' accounts and information across the entire authentication pipeline.

> "The most effective deepfake detection strategies combine AI with traditional security measures, hardware-based protections, and platform-specific optimizations. This layered approach provides greater resilience against both current and future deepfake attacks"

AI threat model for mobile applications

Banking institutions have suffered millions in losses from attacks targeting deployed AI systems in mobile applications, demonstrating why traditional mobile app security cannot protect against AI-specific threats. Documented incidents show the following four threat categories in action:

1. **Runtime model tampering:** Research demonstrates that fraud detection models in mobile banking apps can be evaded through adversarial inputs, with academic studies showing 60-80% success rates against deployed AI systems using carefully crafted transaction patterns.

2. **Local data store compromise:** The GoldPickaxe malware campaign specifically targeted mobile banking apps with AI-powered biometric authentication, compromising banking applications across Thailand and Vietnam. Mobile banking malware families including Hook, Godfather, and Teabot now possess capabilities to extract locally stored AI model components from compromised banking apps.

3. **AI agent runtime exploitation:** Wells Fargo's AI assistant processes 245 million customer interactions, representing massive attack surfaces for runtime manipulation. Certificate validation vulnerabilities in major bank mobile apps have created potential pathways for extracting AI model inference data during communications.

4. **Prompt injection attacks:** Academic research shows 31 out of 36 commercial AI applications vulnerable to prompt injection, with mobile banking chatbots particularly at risk due to limited security controls on mobile devices.

Financial impact shows urgent need for protection

JPMorgan Chase repels 45 billion cyberattack attempts daily and spends $15 billion annually on cybersecurity, acknowledging that AI systems face unique threats. Bank of America invested $4 billion in AI initiatives while implementing strict mobile security controls.

The 196% surge in trojan banker attacks targeting smartphones shows mobile threats are escalating, with 29 malware families targeting 1,800+ banking apps globally. Mobile banking attacks now cost institutions significant resources as attackers develop AI-specific exploitation techniques.

As AI becomes increasingly integrated into mobile and desktop applications, it introduces unique security challenges that traditional cybersecurity approaches may not adequately address. This report explains the key security threats facing AI systems deployed on

devices and demonstrates why Promon's specialized protection solutions are essential for safeguarding your onboard models and infrastructure.

Key takeaways

- AI systems face specialized threats beyond traditional application security concerns
- The most critical device-based AI threats include model theft, tampering, prompt manipulation, and malicious code execution
- Promon offers targeted protection that addresses these threats through a multi-layered security approach
- Implementing AI security measures is increasingly required for regulatory compliance

The AI security challenge

There are reasons why AI security differs from traditional application security. AI applications differ fundamentally from conventional software in several ways:

- **Unique assets:** AI systems contain valuable intellectual property in the form of trained models that represent significant investment
- **New attack surfaces:** AI introduces novel vulnerabilities through components like system prompts, model parameters, and agent runtimes
- **Complex interactions:** AI systems often interact with multiple data sources and components in ways that create security gaps
- **Regulatory requirements:** Emerging regulations specifically target AI system security and privacy

As organizations deploy AI on devices—from smartphones to desktop applications—these unique security challenges require specialized protection strategies.

Critical AI threats on devices

Promon researchers analyzed key industry frameworks to discover which categories of threats were particularly concerning for organizations deploying AI on devices. These frameworks included:

- The MITRE Corporation's ATLAS Matrix [13]
- The OWASP Top 10 for LLM Applications 2025 [14]
- The OWASP Agentic AI – Threats and Mitigations [15]

From this research, four categories of threats were prominent:

1. Runtime model tampering

   What it is: Unauthorized modification, substitution, or theft of deployed AI models.

   Business impact:
   - Loss of valuable intellectual property
   - Competitors gaining access to proprietary AI capabilities
   - Compromised model behavior leading to incorrect business decisions
   - Potential insertion of backdoors that could enable future attacks

   Real-world example: A competitor gains access to your proprietary AI model deployed on a mobile app, reverse-engineers it, and implements similar functionality in their own product—stealing years of R&D investment.

2. Local data store compromise

   What it is: Attacks targeting the data stored locally on devices that AI models use to function.

   Business impact:
   - Exposure of sensitive customer information
   - Manipulation of AI inputs leading to incorrect outputs
   - Persistent vulnerabilities that can survive app updates
   - Potential regulatory violations (GDPR, AI Act, etc.)

   Real-world example: An attacker gains access to local databases in your AI application and modifies reference data, causing your AI to make incorrect decisions or recommendations that damage customer trust.

3. AI agent runtime exploitation

   What it is: Taking advantage of the AI agent environment to execute unauthorized code or leak data.

   Business impact:
   - Malware deployment on user devices
   - Data exfiltration from corporate environments
   - Manipulation of agent behavior to perform unauthorized actions

- Reputational damage from compromised applications

Real-world example: An attacker exploits your AI assistant app to load and execute malicious code that steals user credentials and sensitive business information from the device.

4. **Prompt injection attacks**

What it is: Specially crafted inputs designed to manipulate AI behavior by bypassing safety mechanisms.

Business impact:
- Generation of harmful content that violates your content policies
- Extraction of confidential information embedded in the model
- Disruption of AI service operations
- Erosion of user trust in AI systems

Real-world example: A malicious user crafts inputs that trick your AI customer service agent into revealing internal company information or generating harmful content that damages your brand.

How can your AI applications be protected?

AI applications can be protected by using a vendor like Promon—which offers a mobile application security platform directly—to address the most critical AI security threats through a comprehensive, multi-layered approach that includes:

1. **Runtime application shielding**

Protects against: Model substitution, theft, and tampering with system prompts

Ready-to-deploy features:
- Advanced code obfuscation specifically designed for AI model protection
- Real-time anti-tampering controls that detect unauthorized modifications
- Integrity verification that ensures models remain unaltered during operation

Business benefit: Immediately safeguards your valuable intellectual property and ensures consistent, reliable AI behavior without requiring changes to your existing architecture.

2. Secure communications protection

   Protects against: Interception of model requests/responses and data leakage

   Ready-to-deploy features:
   - Hardened encrypted channels that secure all AI-related data transmission
   - Certificate pinning technology that prevents man-in-the-middle attacks
   - Traffic integrity verification that ensures unaltered communication between app components and AI models

   Business benefit: Instantly prevents attackers from capturing sensitive data or manipulating AI inputs and outputs, with minimal integration effort.

3. A local data protection system

   Protects against: Unauthorized access to internal data sources used by AI

   Ready-to-deploy features:
   - Strong encryption for all stored AI data, including models and system prompts
   - Secure key management system that protects encryption keys even on compromised devices
   - Automatic integrity verification that detects any tampering with stored AI data

   Business benefit: Immediately prevents data leakage and unauthorized modification that could compromise your AI system's reliability, helping maintain regulatory compliance.

Regulatory compliance benefits

Implementing Promon's AI protection solutions helps meet emerging regulatory requirements:

- **EU AI Act:** Addresses technical robustness and safety requirements for high-risk AI systems
- **GDPR:** Supports compliance for AI systems processing personal data on mobile devices
- **Sector-specific regulations:** Helps meet other requirements for AI in finance, healthcare, and critical infrastructure
- **Transparency requirements:** Enables better documentation of AI security controls

Ready-to-deploy protection

As AI becomes central to business operations and customer experiences, protecting these systems from specialized threats is no longer optional—it's essential. Traditional application security measures alone are insufficient to address the unique challenges of AI security, especially on mobile and desktop devices.

Promon offers ready-to-deploy protection solutions that can be implemented immediately, without lengthy integration processes or changes to your existing AI architecture. Our comprehensive protection capabilities directly address the most critical threats facing AI deployments on devices.

By implementing Promon's protection solutions today, your organization can:

- ✓ **Safeguard valuable AI intellectual property** that represents significant R&D investment
- ✓ **Strengthen the reliability and accuracy of AI-driven decisions** that your business depends on
- ✓ **Protect sensitive data processed by AI systems** from unauthorized access or manipulation
- ✓ **Achieve compliance with emerging AI regulations** to avoid penalties and restrictions
- ✓ **Prevent reputational damage from AI security incidents** that could undermine customer trust

The threats to AI systems are real and growing. Competitors want your AI intellectual property. Malicious actors seek to manipulate your AI's behavior. Regulators are increasing scrutiny of AI security practices.

Promon's protection solutions are available now to secure your AI applications against these threats. Contact us today for a demonstration of how our technology can protect your specific AI deployments.

AI-driven deobfuscation and cyber-attacks by non-technical users

This section presents original research conducted by members of Promon's Security Research Team.

The research was published in the paper 'Deconstructing Obfuscation: A four-dimensional framework for evaluating Large Language Models assembly code deobfuscation capabilities' [16] by Anton Tkachenko, Dmitrij Suskevic, and Benjamin Adolphi (2025).

Research overview

This report examines how modern artificial intelligence systems, specifically Large Language Models (LLMs) like GPT-4, Claude, or Grok, perform when analyzing protected software code. The research reveals that while these AI models show impressive capabilities in some scenarios, they still struggle with sophisticated protection techniques, especially when multiple methods are combined. This creates a clear roadmap for organizations seeking to protect their software assets while also showcasing potential vulnerabilities in current protection strategies.

Introduction: Why obfuscation matters

Software protection through obfuscation (deliberately making code difficult to understand) serves two primary purposes:

1. **Legitimate use:** Protecting valuable intellectual property and preventing reverse engineering
2. **Malicious use:** Hiding harmful functionality in malware to avoid detection

Understanding how modern AI can analyze protected code helps both defensive security teams and software developers implement more effective protection strategies. This research offers the first comprehensive evaluation of commercial AI models on this specific security challenge.

What was tested: Top AI models vs. protected code

The research team evaluated eight leading commercial AI models on their ability to analyze protected software code:

- GPT-3o
- GPT-4o
- GPT-4.5
- GPT-Pro-o1
- DeepSeekR1
- Grok3
- Grok2
- Claude 3.7 Sonnet

Each model was tested against a specific program protected by different obfuscation techniques, both individually and in combination.

The test program

The researchers used a deliberately simple but strategically chosen program that computes different calculations based on an input value. This program was then protected using four different obfuscation strategies:

1. **Bogus control flow:** Adding misleading decision paths that seem important but aren't
2. **Instruction substitution:** Replacing simple operations with more complex equivalent ones
3. **Control flow flattening:** Reorganizing the code's structure to hide its true operation
4. **Combined techniques:** Applying all three methods together

The four-dimensional framework: A new way to understand AI capabilities

One of the study's major contributions is a framework that breaks down AI capabilities into four key dimensions:

1. **Reasoning depth:** The AI's ability to analyze logical relationships and draw conclusions about how code works. For example, can it recognize that a mathematical expression will always be true or false regardless of input?

2. **Pattern recognition:** The AI's ability to identify familiar structures and computations even when they're deliberately hidden or modified. Can it recognize a simple addition operation that has been transformed into a more complex series of steps?

3. **Noise filtering:** The AI's ability to distinguish between important code and deliberately added distractions. Can it identify which parts of the code affect the outcome and which are merely there to confuse analysis?

4. **Context integration:** The AI's ability to connect related pieces of code that have been separated. Can it recognize that two distant parts of the program are logically connected, even when they're physically far apart in the code?

This framework helps explain why different AI models perform inconsistently across various protection techniques, as each technique challenges different capabilities.

Results: How the AI models performed

The research revealed striking differences in how well various AI models handled different types of code protection. **Table 1** below summarizes the results, showing the level of expertise needed to successfully analyze the protected code for each model and technique combination.

Table 1: Obfuscation variants and required attacker knowledge levels

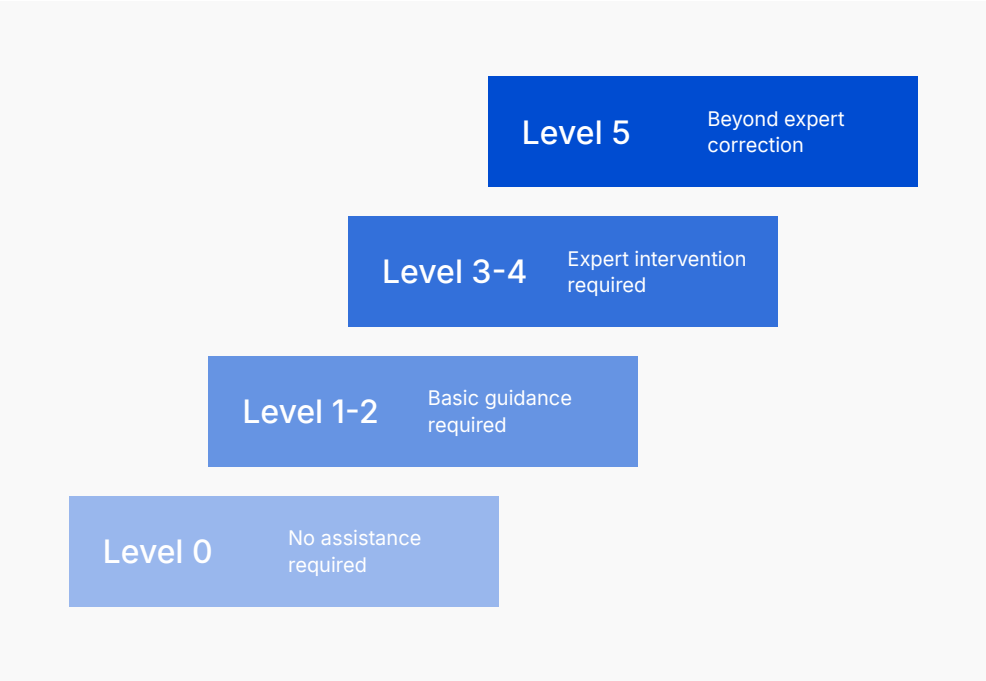
Note that in some instances, analysis proved impossible (-) as the AI model produced no meaningful output.

Model	Bogus control flow	Instruction substitution	Control flow flattening	Combined techniques
GPT-3o Mini	4-5	-	-	-
GPT-4o	4	4	1	-
GPT-4.5	1-2	4	0	5
GPT-Pro-o1	3	-	0	5
DeepSeekR1	5	5	1-2	5
Grok3	1	4	0	5
Grok2	-	-	1-2	-
Claude 3.7 Sonnet	0	1-2	0-1	5

These levels of expertise result in a five-point knowledge level scale:

- **Level 0:** AI fully solves the problem without help
- **Level 1-2:** Minimal hints are needed to correct minor errors
- **Level 3-4:** Significant guidance is needed
- **Level 5:** There are too many fundamental errors

Five-point knowledge level scale.



Key findings: What this means for software protection

The research revealed several important insights that have direct implications for organizations concerned about protecting their software assets.

Significant variation between AI models

The study revealed dramatic differences in capability between various AI systems. Some advanced models could break through certain protection techniques without human help, while others completely failed against the same challenges. This significant variation suggests that the specific AI model being used matters just as much as the protection technique being deployed.

Some protection methods are already vulnerable

Control flow flattening, once considered a strong protection technique, proved relatively ineffective against today's top AI models. Three different systems (GPT-4.5, GPT-Pro-o1, and Grok3) were able to completely defeat this protection without any human assistance.

Instruction substitution remains highly effective

This protection technique—which replaces simple operations with more complex equivalents—presented significant challenges for every AI model tested. Even the most advanced systems required substantial human expertise to overcome this protection method.

Layered protection works

The most important finding was that combining multiple protection techniques created an extremely effective defense. When all three methods were applied together, every AI model tested either failed completely or required the highest level of human expertise to make any progress.

AI capabilities are uneven and predictable

The research showed that AI performance varied in consistent, predictable ways across different protection techniques. This makes it possible for organizations to strategically choose protection methods that target specific weaknesses in current AI technology.

The three-tier resistance model: A practical guide

Based on these findings, the researchers developed a practical classification system for obfuscation techniques, consisting of low, moderate, and high resistance techniques.

1. Low resistance techniques

   These techniques primarily challenge reasoning capabilities, an area where several top-tier AI models demonstrate strong performance. They can be overcome by advanced AI systems with minimal or no human assistance. An example of a low resistance technique is control flow flattening.

2. Moderate resistance techniques

   These techniques require both strong pattern recognition and context integration. Fewer models can handle these challenges autonomously, but several advanced systems showed good capabilities. An example of a moderate resistance technique is bogus control flow.

3. High resistance techniques

   These approaches either target specific weaknesses in current AI pattern recognition or simultaneously challenge multiple capability dimensions. All models either required expert intervention or failed completely against these techniques. Examples of high resistance techniques include instruction substitution and combined techniques.

Three-tier resistance model overview.

Resistance	AI capabilities challenged	Level of human assistance required	Example
Low	Reasoning depth	Minimal or none	Control flow flattening
Moderate	Pattern recognition and context integration	Basic guidance	Bogus control flow
High	Specific weaknesses in current AI pattern recognition or multiple capability dimensions simultaneously	Expert intervention or beyond correction (AI fails even with expert help)	Instruction substitution, combined techniques

Common AI errors in code analysis

The study identified five recurring types of errors that AI models made when analyzing protected code:

- **Predicate misinterpretation:** Failing to recognize that certain conditions are always true or false
- **Structural mapping:** Correctly identifying pieces of code but incorrectly connecting them to the control structure
- **Control flow misinterpretation:** Incorrectly reconstructing the fundamental structure of the code (e.g. seeing loops where none exist)
- **Arithmetic transformation:** Failing to correctly reconstruct mathematical operations from their obfuscated form
- **Constant propagation:** Incorrectly handling, identifying, or fabricating literal values in the code

These error patterns reveal fundamental limitations in how current AI models process obfuscated code.

Implications and recommendations

Suggestions for software developers and security teams.

For software developers

- **Layer your defenses:** Combining multiple obfuscation techniques provides significantly stronger protection against AI-based analysis than any single method alone.
- **Focus on instruction substitution:** This technique proved challenging even for the most advanced AI models and should be part of any comprehensive protection strategy.
- **Don't rely solely on bogus control flow:** While effective against human analysts, this technique is increasingly vulnerable to advanced AI models.

For security teams

- **Use AI tools as supplements:** Current AI models can reduce expertise barriers for certain aspects of code analysis but still require human guidance for complex scenarios.
- **Combine AI strengths:** Different models show different capability patterns, so using multiple AI systems can provide more comprehensive analysis.
- **Expect capabilities to evolve:** This field is advancing rapidly, so protection and analysis strategies should be regularly reassessed.

Conclusion

The research demonstrates that while AI models have made impressive advances in code analysis capabilities, sophisticated protection mechanisms remain effective today. However, the rapid pace of AI development means we should expect these capabilities to improve dramatically in the coming years. What is resistant to AI analysis now may become vulnerable soon.

The four-dimensional framework provides an exceptional opportunity to:

1. **Build automated assessment systems** that can systematically evaluate both:
    a. The deobfuscation capabilities of new AI models as they emerge
    b. The AI resistance of various protection techniques

2. **Strategically improve protection mechanisms** based on objective data rather than guesswork. By understanding exactly which capabilities different protection techniques challenge, organizations can:
    a. Target specific AI weaknesses in their protection strategies
    b. Combine techniques that challenge multiple capability dimensions simultaneously
    c. Develop new protection methods specifically designed to resist emerging AI capabilities

3. **Prepare for a changing security landscape** where the balance between protection and analysis will continually evolve. Organizations that understand and leverage this framework will be better positioned to:
    a. Anticipate which protections may become vulnerable next
    b. Implement layered defenses that remain effective even as AI capabilities advance
    c. Make informed investments in security that provide longer-term protection

Rather than waiting for AI advancements to unexpectedly overcome our protections, this framework enables a proactive approach to security in an era of rapidly advancing artificial intelligence. We can now methodically assess, improve, and adapt our protection strategies based on a clear understanding of AI capabilities and limitations.

Financial app analysis conducted by Promon’s Security Research Team

Promon's Security Research Team tested apps for this part of the App Threat Report.

Methodology overview

Methodological considerations included the selection of appropriate financial apps, the process of testing, and the omission of apps during the testing process.

Sample selection

Promon researchers generated a list of top financial apps from SensorTower. These were ranked by world-wide downloads during the last 30 days.

The team tested the apps on this list semi-automatically by a simulated accessibility services-based screenreader attack. This kind of attack was selected because it is the most common way that Android malware attacks apps, providing a solid method for attack testing and research.

Test process

They installed a screenreader on the test device and then installed the app from the Play Store before launching it. When launched, the researchers navigated to a screen where they could enter some sensitive data (username, password, social security number, phone number, etc.).

Once the sensitive data was entered, the team checked for the following outcomes:

- Was the screenreader able to capture that data?
- Did the app would warn the user?
- Did the app crash because of the screenreader's presence?

Result omissions

During the tests, Promon researchers encountered two situations in which they were not able to test a given app:

1. **When the app was not available in their region:** This is a common problem for which they do not currently possess a solution. But they noted that this forced omission of test subjects added a slight regional bias to their findings.

2. **When the app required a SIM card:** They did not have a SIM card for their test device, so they could not test these apps.

In both cases, researchers omitted any problematic apps from the findings. They continued with this process until they had tested 100 apps. Since this testing was only semi-automatic, and given subsequent time constraints, testing more apps was not feasible.

Testing summary

These were the results:

- Total apps reviewed: 252
- Apps successfully tested: 100
- Apps unable to test: 152 (60%)

These are the reasons for our inability to test certain apps:

- Apps not available in the region: 111
- SIM required

Key findings

Here are the key findings from the 100 tested apps:

- Vulnerable apps: 82
- Fully protected apps: 18

Of the 18 apps fully protected, these were the reasons:

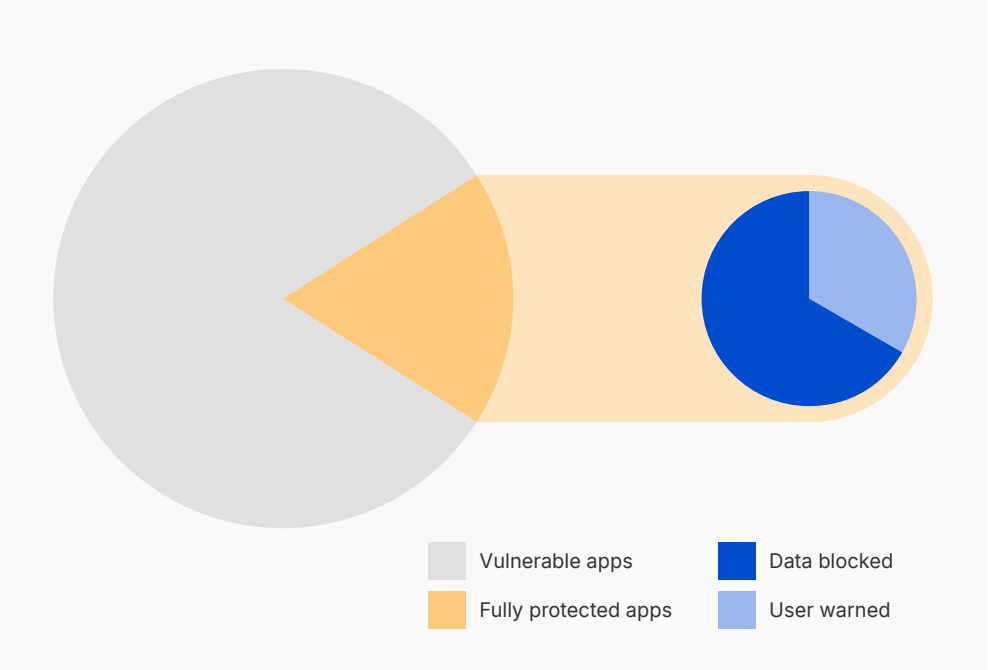
- Data blocked: 12
- User warned: 6

The number of vulnerable apps is down 32.94% from last year. The number of fully protected apps is up 7.14% from last year.

The most common weaknesses were:

- Lack of runtime protection
- No screenreader detection

Key findings from tested finance apps



The number of **fully protected apps is up 7.14%** from 2024

The number of **vulnerable apps is down 32.94%** from 2024

Conclusion

AI is continuing to change the world, including cybersecurity and the finance sector.

While it is not possible to predict fully how AI will impact and alter the threat landscape, what we can conclude is that this change has already begun—and the result is not entirely negative. Defenses against malware and cyberattacks are also employing AI to become smarter, stronger, and more sophisticated.

Our research team has found that multi-layered obfuscation is one of the most effective ways to protect mobile applications as AI continues to develop.

Citations

[1] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Securitysituation/IT-Security-Situation-in-Germany-2024.pdf

[2] https://www.enisa.europa.eu/sites/default/files/2025-02/Finance%20TL%202024_Final.pdf

[3] https://www.enisa.europa.eu/sites/default/files/2024-11/ENISA%20Threat%20Landscape%202024_0.pdf

[4] https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf

[5] https://144806018.fs1.hubspotusercontent-eu1.net/hubfs/144806018/Reports/US_CNI_Research_Report_2024_Cyber_Security_in_Financial_Services.pdf

[6] https://www.gsma.com/about-us/regions/asia-pacific/wp-content/uploads/2024/11/Consumer-Attitudes-Toward-Fraud-and-Opportunities-for-Mobile-Network-Operators-in-SEA-FINAL.pdf

[7] https://www.team-cymru.com/post/coper-octo-a-conductor-for-mobile-mayhem-with-eight-limbs

[8] https://www.americanbanker.com/news/jpmorgan-chase-using-chatgpt-like-large-language-models-to-detect-fraud

[9] https://www.federalreserve.gov/newsevents/speech/barr20250417a.htm

[10] https://promon.io/security-news/deepfake-mobile-banking-apps

[11] https://link.springer.com/article/10.1007/s10462-024-10810-6

[12] https://www.sciencedirect.com/science/article/pii/S2543925125000075

[13] https://atlas.mitre.org/matrices/ATLAS

[14] https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[15] https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

[16] https://arxiv.org/abs/2505.19887

PROMON

Promon leads the way in proactive mobile app security. For 19 years, we've been making the world a safer place by securing any app, on any device—in no time at all.

Today, we protect over 2 billion users, secure 13 billion monthly transactions, and safeguard $2.5 trillion in market cap.

Promon is headquartered in Oslo, Norway, with offices in more than 15 countries around the world.

promon.io

Promon AS
Cort Adelers Gate 30
0251 Oslo
Norway