PROTEGO TRUST BANK | Trust for Tomorrow.

Who? (are you really?)

Digital Identity for GenAI LLMs and Agents

February 4, 2025

Weiyee In, CIO Protego Trust Bank
Jim Skidmore, CISSP, PgMP, VP intiGrow
Adam McElroy, Eclypses Principal Architect
(Special Thanks to Brandon Miller, John C. Checco, and JC Vega)

Introduction

The rapidity and discontinuity of innovation and technological advancement and adoption of Generative AI (GenAI) Large Language Models (LLMs) have introduced an exponentially hyper-complex landscape of security challenges. Several of the challenges beyond more traditional data privacy and security issues came to the fore with the recent launch of DeepSeek and the challenges for transparency and reliability of GenAI LLMs and Agents. This white paper examines some of the technical aspects of these issues, focusing on identity confusion, data privacy, shadow GenAI risk for enterprise, national security concerns, cybersecurity vulnerabilities, and regulatory compliance. We will also explore the integration of digital identity and blockchain technologies for GenAI LLM agents to address some of these security challenges and industry pain points and meet growingly stringent regulatory requirements.

Identity Confusion in GenAI LLMs: "*Who are you?*"

One of the challenges for both AI governance and security came to the fore with the launch of DeepSeek's GenAI models and application in the DeepSeek R1, where the model exhibited significant instances of identity confusion, misidentifying itself as other GenAI LLMs and GenAI assistants including OpenAI's GPT-4 and Anthropic's Claude[1]. In multiple documented instances it noted "*To clarify: I'm an AI developed by Microsoft, … I'm part of Microsoft's Copilot suite (formerly Bing Chat), built on OpenAI's GPT-4 architecture.*[2]"

[1] https://opentools.ai/news/deepseeks-r1-the-open-source-ai-model-raising-eyebrows-with-identity-confusion
[2] https://www.fastcompany.com/91267647/deepseek-told-me-made-by-microsoft-r1-openai-claude-anthropic-ai-model-copilot

These and other misidentification issues surrounding DeepSeek's GenAI LLMs have exposed significant technical and security concerns that ripple through the broader AI industry.

These issues are not limited to DeepSeek, as several research efforts have identified identity issues. Shandong University researchers noted that "*We evaluated 27 LLMs and found that 25.93% exhibited identity confusion, revealing a significant vulnerability in model design and training*[3]," highlighting that GPT-4 misidentified itself as GPT-3 and GPT-3.5 during API queries, demonstrating the model's inability to accurately represent its own identity. In the case of DeepSeek, the model sometimes identifies itself as ChatGPT or other GenAI LLM systems, which has already been likened to multiple personalities in dissociative identity disorder (DID). However, this phenomenon is not a human psychological condition but rather a technical issue stemming from its training data or programming. The misidentification is primarily attributed to the model's training on datasets that include outputs from other GenAI LLM systems via scraped data, leading to confusion about its own identity.

For enterprises, especially financial institutions, the trustworthiness of GenAI LLM systems becomes fundamentally undermined when they or their agents cannot reliably confirm their own identity and provenance. This raises basic and critical questions about the integrity of the model's training data, model parameterization, model algorithms and architecture not to mention a myriad of intellectual property issues related to the use and inclusion of outputs from other GenAI LLM systems without proper attribution or permission. From a financial services perspective, the DeepSeek identity confusions highlight critical issues that intersect with regulatory compliance, cybersecurity, and third-party risk management and underscore the need for more robust Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) controls in the financial sector, particularly when adopting GenAI LLM technologies.

The governance concerns raised by DeepSeek's misrepresentations emphasize the importance of incorporating GenAI and broader AI ethics into the third-party risk management processes of financial institutions. Organizations will need to develop more comprehensive due diligence procedures for all AI / ML vendors and their third-party service and data providers, including assessments of their ethical standards and transparency in AI development and training. Identity consistency in GenAI LLMs, since the launch of DeepSeek, jumped to the fore as a critical issue in the broader field of

[3] "I'm Spartacus, No, I'm Spartacus: Measuring and Understanding LLM Identity Confusion"

artificial intelligence. Advanced GenAI LLMs that can exhibit inconsistent self-identification, potentially misrepresenting their capabilities or origin raise not only use case workflow concerns but also fundamental security issues. This inconsistency not only undermines trust in GenAI LLM systems but also poses significant compliance and regulatory risks, particularly in highly regulated industries such as financial services where accurate system identification is vital for a data-driven industry.

Regulatory Implications: "*Got to be another way*"

The rapid adoption of GenAI services without corresponding security and data governance measures increases the risk of non-compliance with data protection regulations. Financial institutions globally must now proactively consider how to integrate GenAI LLM model review, risk assessment and verification into their existing risk management frameworks for industry standards and regulatory requirements. The U.S. Federal Financial Institutions Examination Council FFIEC CAT[4], which measures a financial institution's inherent risk profile and cybersecurity maturity, has not updated specific assessments for GenAI model authenticity and performance verification, and is being phased out. During the sunset of FFIEC CAT and migration towards the National Institute of Standards and Technology (NIST) Cybersecurity Framework 2.0 and the Cybersecurity and Infrastructure Security Agency's (CISA) Cybersecurity Performance Goals the financial services industry faces a widening gap in GenAI governance.

The U.K. Financial Conduct Authority (FCA) and the European Securities and Markets Authority (ESMA) have also outlined several security requirements related broadly to AI governance and provenance. The FCA emphasizes the application of the Senior Managers and Certification Regime (SMCR) to AI governance in general making senior managers personally accountable for the use of AI in their areas of responsibility. ESMA similarly expects management bodies to have an appropriate understanding of AI technologies used within their firms and ensure proper oversight. ESMA also requires financial institutions to be transparent about the role of AI in investment decision-making processes advises conducting periodic stress tests on AI algorithms[5].

Financial institutions also seeking NIST 800-53 R5 compliance need to enhance their System and Information Integrity (SI) controls, particularly SI-7 (Software, Firmware, and Information Integrity) and SI-10 (Information Input Validation), to include measures for

[4] FFIEC has announced the sunset of its Cybersecurity Assessment Tool (CAT) effective August 31, 2025, it is being used as a baseline for principles in this discussion

[5] Artificial intelligence in EU securities markets, 1 February 2023 ESMA50-164-6247

verifying GenAI model integrity and to address the unique challenges posed by GenAI) systems, particularly in terms of digital identity and provenance. To maintain GenAI model integrity, financial institutions must implement measures that ensure the authenticity and security of GenAI systems throughout their lifecycle to ensure the model's digital identity and provenance can be authenticated throughout its lifecycle. To prevent misrepresentation of AI capabilities and protect against malicious inputs, institutions must implement robust input validation mechanisms tailored to AI systems to help prevent data poisoning attacks that could compromise the model's integrity. The risks associated with inconsistent GenAI identities can lead to potential data breaches, fraudulent activities, and misinformation campaigns or cyberattacks.

Securing Scale Matters

The convergence of GenAI LLMs, quantum computing, and the increasing ubiquity of connected devices is creating an unprecedented and complex security landscape at scale with an attack surface that is also growing in exponential scale. This confluence of technologies amplifies existing vulnerabilities and introduces new threat vectors, posing significant challenges for institutions across various sectors. Financial institutions must adopt a more proactive approach to quantum risk planning, enhance their cybersecurity measures, and develop robust strategies to address the unique challenges posed by this technological confluence. GenAI's ability to create convincing fake content raises ethical and security concerns, beyond the potential for deepfakes, synthetic identities, and counterfeit documents at scale for omnichannel phishing.

This risk extends far beyond GenAI LLMs generating massive omnichannel attack campaigns on targeted individuals or brute force attacks but goes to a deeper layer of governance for data and process. To respond, financial institutions need to not only implement quantum-resistant cryptography, implement more stringent data protection measures, but continuously adapt security protocols to address emerging threats in a digital ecosystem whose evolution is accelerating. The technical challenges lie in developing robust methods to both ensure GenAI LLMs maintain a clear and accurate sense of their own identity throughout the training process and subsequent interactions and have a digital signature to verify the integrity and provenance of GenAI LLMs and their interactions before execution. These challenges are particularly relevant for financial institutions and other regulated industries where security, trust, and accountability are paramount.

Verifying: "*Cause I really wanna know*"

The challenge of maintaining a consistent identity for GenAI LLMs is exacerbated by the increasingly hybrid nature of GenAI agents, which blur the traditional lines between human and machine identities. This new paradigm requires a rethinking of identity and access management (IAM) frameworks to more effectively address the complexities of GenAI identity management. The need to control unauthorized parties from intercepting or manipulating sensitive data and making the data itself tamper-evident come become critical. Data needs to be securely stored and transmitted in a quantum resilient zero trust model to ensure that even if the underlying infrastructure is compromised, the data remains secure. There is a critical need for a more robust framework for managing these hybrid identities by ensuring secure, transparent, and verifiable interactions.

For regulated industries these systems must ensure precise user and system identification, access control, and compliance monitoring, which becomes significantly more complex when GenAI models can be the ones not only making data or process calls but also generating unexpected or inaccurate results due to identity confusion, biases in training data, inherent limitations and misrepresentations or other complexities. Financial institutions need an integrated approach supporting regulatory compliance by providing a robust framework for tracking and verifying model updates and access, ensuring compliance with data management and AI governance regulations.

When GenAI LLMs interact with sensitive data and processes the transmission not only needs to be secure so unauthorized access is prevented but there needs to be an immutable audit trail for provenance because of inherent biases in models. This provenance tracking becomes essential for maintaining transparency and accountability in GenAI LLM governance.

Identity Consistency Across Lifecycles

As GenAI LLMs become adopted into enterprise workflows and can now make data or process calls, security and integrity require much more robust authentication mechanisms to verify the identity of both human users and AI systems and across lifecycles. This necessitates the development of GenAI-specific identity governance frameworks and robust mechanisms and advanced technology to ensure the model consistently and accurately represents itself during training, deployment, and interaction. This is essential to prevent "*identity confusion*," where models misidentify themselves or their capabilities undermining user trust as well as introducing significant risks—particularly in regulated

sectors like financial services, where compliance violations can have severe consequences.

During the training process, GenAI LLMs must be safeguarded against adversarial data poisoning attacks that could compromise their behavior or identity. Adversarial actors may attempt to inject malicious data into training datasets, altering the model's outputs, or embedding vulnerabilities. To mitigate this risk, embedding digital signatures and employing post quantum resilient data in transit solutions at the dataset and application level have become critical steps. These cryptographic signatures ensure the integrity of training data and provide verifiable provenance, enabling financial institutions to confirm that the data used in training has not been tampered with or altered.

In the deployment phase, secure practices are equally important to maintain a GenAI model's identity. Techniques such as digital watermarks or blockchain-based identity credentials can be employed to help models verify their own identity when queried or interacting with users. Digital watermarks embed unique identifiers within the model's architecture or outputs, making it possible to trace the origin and authenticity of the model. Blockchain-based credentials provide an immutable record of the model's provenance and lifecycle, ensuring that only authorized versions of the model are deployed and used in production environments. By implementing these measures across both training and deployment phases, organizations can establish a strong foundation for maintaining GenAI model identity. This not only enhances trustworthiness but also ensures compliance with regulatory requirements in sectors where transparency, accountability, and security are paramount.

Dynamic Nature of AI Models

GenAI LLM systems are inherently dynamic, often adapting to new data during fine-tuning or reinforcement learning. When GenAI LLMs are fine-tuned on new data (learn from new data), they undergo a process of adaptation by adjusting their weights to better fit the latest information. The model's weights are adjusted through backpropagation, calculating the error or difference between the model's predictions and the actual labels, and optimization algorithms, such as stochastic gradient descent (SGD) that adjust the model's weights based on the gradients calculated during backpropagation. This process effectively involves updating the model's parameters to optimize its performance off the new dataset, which can change the model behavior in ways that might not be immediately apparent.

This adaptability, while in principle beneficial for improving model performance, also increases the risk of identity drift, where updates may inadvertently alter the model's core characteristics or introduce security and data governance vulnerabilities. In the case of DeepSeek mistakenly identifying itself as ChatGPT and other "self-identification" or "identity confusion" issues among GenAI LLMs the incidents underscore the importance of data quality, integrity, and provenance in GenAI LLM training. If a GenAI LLM is trained on extensive web-scraped data that includes responses and outputs from other GenAI LLM systems, it may "learn" the perceived identity of those systems.

Verifying Integrity and Provenance: Come on, Come on

In today's rapidly evolving AI landscape, where "*identity confusion*" of GenAI LLMs converges with "*sleeper agent*" and "*alignment faking*" threat vectors, quantum computing threats, and malfeasance and misfeasance of bad actors using GenAI LLMs for advanced targeted phishing at scale, there is a critical market need for much more robust security solutions to protect GenAI LLMs and the counterparties to their interactions. The integrity and authenticity of these models become paramount, especially in regulated industries where compliance requirements are stringent.

Securing GenAI LLMs against threats from quantum computing (transitioning to a zero-trust architecture and post-quantum cryptography (PQC) etc.) and malicious actors is becoming a critical aspect of AI security and hygiene. Quantum computers pose a significant risk to current cryptographic methods by potentially enabling malicious actors to break them, thereby compromising encrypted data, including training data. Malicious actors can exploit GenAI LLM systems in several ways, from data poisoning where attackers intentionally corrupt the training data to influence the model's behavior, to targeted omnichannel phishing for access and permissions.

Securing system level integrity remains the first step to ensuring that training datasets are free from inadvertent contamination by other GenAI LLM outputs or any malicious poisoning to prevent identity confusion and hallucinations becomes feasible. Only after core security hygiene and a strong security posture with robust cryptographic methods are achieved can activities such as model verification processes, model audits and other secure data handling practices become effective in helping detect and mitigate issues.

Digital Signatures for Model Integrity

Because the industry faces significant security challenges, including quantum computi,ng to ensure the integrity and authenticity of GenAI LLMs, particularly in the context of model identity confusion, "*sleeper agents*," and "*alignment faking*," advanced technologies are

needed to provide robust alternatives to traditional PKI-based digital signatures for verifying model integrity. Implementing quantum-resistant cryptographic methods is a necessary first step to securing GenAI LLMs at a system or application level to protect against quantum computing threats and offering long-term security solutions that can adapt to emerging threats to ensure data quality, integrity, lineage, and provenance. For data in transit Micro Token Exchange (MTE)[6] offers a sophisticated approach to data substitution, replacing each byte of the model with multiple bytes of randomly generated data, making it difficult for attackers to intercept or manipulate sensitive data.

Provenance Tracking

Blockchain-based provenance tracking has also become a critical component in ensuring the integrity and transparency of GenAI LLMs through a digital identity. This approach involves documenting the origin, history, and modifications of a GenAI model throughout its lifecycle using blockchain technology.

Aligning with Standards and Regulations

Blockchain-based provenance helps ensure that GenAI LLMs are accurately identified and tracked throughout their lifecycle. This reduces the risk of identity confusion by providing a clear and verifiable history of model modifications, an approach that aligns at a high level with key frameworks like the NIST Cybersecurity Framework 2.0 (CSF). Blockchain-based provenance helps identify the origin, history, and modifications of GenAI LLMs, aligning with the "Identify" function of the NIST CSF enabling the categorizing and prioritizing assets, including GenAI LLMs, based on their risk profile. By providing an immutable record of changes, blockchain technology protects against unauthorized modifications that could lead potentially to identity confusion aligning practices with the "Protect" function focused on implementing safeguards to prevent or limit the impact of a security event. The transparent and tamper-proof nature of blockchain allows for the detection of any anomalies or unauthorized changes in the model's lifecycle, supporting the "Detect" function by enabling real-time monitoring and anomaly detection. In the event of a security incident, blockchain-based provenance provides a clear audit trail, facilitating a swift response to mitigate risks for the "Respond" function. By maintaining a verifiable history of model states, blockchain technology aids in the recovery process by ensuring that previous versions of the model can be restored if needed, supporting the "Recover" function.

[6] MicroToken Exchange is a patented solution of Eclypses

Similarly, for following NIST 800 53 r5, blockchain-based provenance helps support System and Information Integrity (SI), Audit and Accountability (AU) and Access Control (AC). Blockchain-provenance ensures that access to model modifications is controlled and auditable, aligning with Access Control requirements. Having an immutable ledger supports Audit and Accountability by maintaining a comprehensive record of all transactions related to the model. Blockchain technology also enhances System and Information Integrity by ensuring that model updates are authorized and verifiable, reducing the risk of unauthorized modifications.

From an overall Software/System Development Life Cycle (SDLC) perspective blockchain-based provenance can be integrated into the requirements definitions to ensure that security and transparency requirements are met from the outset, it can be used to track changes and ensure that the model's architecture aligns with security standards, verify that model updates are correctly implemented and authorized. During deployment, blockchain-based provenance ensures that the model operates as intended and that any changes are transparently recorded. Throughout the maintenance phase, blockchain technology provides a secure and transparent record of model updates, ensuring ongoing compliance with security standards.

Immutable Ledger for Tracking Changes

A blockchain network consists of blocks that contain a timestamp, a hash, and a set of transactions, where each block is linked to the previous one through a cryptographic hash, forming a chain that is highly secure and resistant to tampering. Blockchain further employs consensus algorithms such as Proof of Work (PoW) or Proof of Stake (PoS) to validate transactions across the network to ensure that all nodes agree on the current state of the ledger, enhancing trust in the recorded data. Configurable access rights allow different stakeholders to interact with the blockchain according to their roles through Role-Based Access Control (RBAC) to minimize risks associated with unauthorized access while maintaining accountability.

What this means for GenAI LLMs is that it adds enhanced security and integrity because the immutable nature of blockchain ensures that data integrity, lineage and provenance are tamper-proof. This provides a transparent record of all transactions, fostering accountability among stakeholders and enhancing trust and transparency of GenAI LLMs.. For regulated industries maintaining a clear and verifiable history of model changes, financial institutions comply with regulatory requirements related to data management, transparency and potentially explainability for GenAI LLM governance. By ensuring that

modifications to a GenAI LLM or its training data are transparently recorded and verified, preventing unauthorized alterations that could lead to identity confusion.

Unique Digital Identifiers: Who are you?

Once data is recorded on the blockchain, it cannot be altered or deleted without consensus from the network, ensuring that every change to a GenAI LLM or its training data is permanently logged, providing a clear and tamper-evident history of modifications. By leveraging unique digital identifiers, provenance tracking, and verification mechanisms on a blockchain GenAI LLMs can be accurately identified, and their authenticity is verifiable. A unique digital user identifier can be assigned and maintained on the blockchain for each GenAI LLM and each of its agents. These digital identifiers serve as digital fingerprints or DNA that distinguish one GenAI LLM and each of its agents from another. Leveraging Micro Token Exchange (MTE) encryption enables massively scalable encryption and non-repudiation of GenAI data in transit. By using NIST recommended Post Quantum Cryptography (PQC) to replace each packet of the GenAI model's data with multiple bytes of randomly generated data it is possible to create a dynamically hardened and tamper-evident representation of the model, enhancing security and preventing unauthorized alterations.

MTE ensures data integrity and non-repudiation in every information exchange, which is crucial in maintaining the veracity of GenAI LLMs, verifying each endpoint connection, preventing unauthorized access and ensuring that data is secured in transit, from the keyboard to the cloud. Because unauthorized access can lead to identity confusion when GenAI LLMs are trained on contaminated datasets that may include outputs from tainted models, or prompt injections, MTE provides an immutable pathway to access or modify GenAI models with authentic datasets, especially in environments where multiple stakeholders interact with genAI LLM systems; or where the genAI LLM systems and agents have access to data and automated processes. MTE encryption defends against unauthorized access to prevent data theft, destruction, or leakage, where sensitive information may be inadvertently exposed through unintended GenAI LLM outputs.

Implementing strict access controls and authentication mechanisms is crucial for preventing unauthorized access to GenAI LLMs. This includes role-based access control, attribute-based permissions, and multi-factor authentication to ensure that only authorized personnel can interact with the model. The ABAC grants access based on attributes of the subject (user human or GenAI LLM or GenAI agent) and object (data) and the micro toke exchange can be used to secure data transmission ensuring that even when access is granted based on attributes, the data itself remains secure. Because GenAI

LLMs and their agents are dynamic, an ABAC with MTE allows for dynamic access control based on changing attributes providing a flexible security framework that adapts to dynamic access conditions, ensuring that data remains secure regardless of attribute changes.

By ensuring the security of data in transit and providing endpoint verification, MTE ABAC further reduces potential attack vectors, where unauthorized access can lead to identity confusion or sleeper agent behavior in GenAI LLMs. MTE's PQC encryption and embedded non-repudiation secures data in transit and can be integrated with AI-powered threat detection systems to further enhance the identification and mitigation of potential threats. MTE further reduces unauthorized access attempts - since only MTE secured data is accepted by the relay, all unexpected or non-MTE data is rejected and therefore cannot be injected into or exfiltrated from the genAI model.

Mitigating Sleeper Agent Risks

Research into recent advancements in GenAI LLMs have also exposed significant security vulnerabilities and challenges related to identity consistency, "s*leeper agents*", and "*alignment faking*" in advanced GenAI LLM systems. These issues present complex technical hurdles for AI developers (human) and raise important concerns in GenAI LLM governance for organizations deploying AI technologies. Unauthorized or malicious modifications, inadvertent triggering or bad actor triggering GenAI LMs can introduce sleeper agent functionality, where the model appears benign initially but activates malicious or errant behaviors when triggered.

By maintaining an immutable record of model changes, blockchain technology can ensure that any modifications to GenAI LLMs are tracked, verified, and help detect unauthorized modifications that might indicate the presence of sleeper agents. Any unexpected changes in model behavior can also be traced back to specific updates or modifications recorded on the blockchain. While this may not be able to prevent sleeper agents from being introduced and launched, it makes it difficult for malicious actors to introduce sleeper agent functionality without detection. Smart contracts integrated with Blockchain provenance can automate safety protocols to help prevent sleeper agent activity by being programmed to detect and respond to specific triggers or anomalies that might activate malicious behaviors.

Formal mathematical verification provides mathematical proof of correctness, ensuring that smart contracts behave as intended in all scenarios; this is particularly important for preventing sleeper agent activity, where even small errors or trigger inputs can have significant consequences. Formal mathematics verification can help in ensuring that smart contracts, which automate safety protocols to prevent sleeper agent activity, operate as intended. By creating a formal specification of the smart contract's desired behavior against GenAI LLM behavior, defining safety protocols, anomalies and triggers, a mathematical model of the smart contract can be constructed to capture its essential components, states, and transitions as an abstract representation of the contract's behavior. Then using techniques such as model checking or theorem proving it is possible to verify that the contract's model satisfies the formal specification while model checking explores all states of the smart contract to validate whether the safety protocols are correctly implemented. Further theorem proving can be used in constructing formal proofs to demonstrate that the contract behaves as specified.

Unlike manual audits, which are fundamentally subject to human error, formal mathematical verification systematically checks the contract's logic against its desired properties. This comprehensive approach helps identify and mitigate complex vulnerabilities that might otherwise go undetected or resulting from human error. Formal mathematical verification can be applied even to complex smart contracts where manual review is often physically impractical. It thereby provides a scalable solution for ensuring the security and reliability of smart contracts in various applications, including those designed to prevent sleeper agent activity.

Formally Mitigating Alignment Faking

Similarly smart contracts can be pre-programmed to ensure that GenAI LLM training data and objectives are transparently recorded on a blockchain. This makes it more difficult for GenAI LLMs to secretly maintain “preferences” that contradict their original intended alignment. By maintaining an immutable provenance record of model states, smart contracts can ensure that any deviations from intended behavior can be traced back to specific changes in the GenAI LLM or its training data. Formal mathematical verification ensures these mechanisms have been correctly implemented and function as intended. This includes verifying that smart contracts enforce rules about GenAI LLM updates and usage, preventing alignment faking by ensuring that models operate within specified parameters.

In a digital economy that is becoming not only hyper-competitive but also increasingly complex because of significant regulatory compliance burdens on organizations for

ensuring transparency, explainability, managing data privacy and security risks, and mitigating biases in GenAI systems the compliance aspects of the standard workflows need to be pivoted. The challenge is that with so many identity-sensitive and governance frameworks so "maturely ingested" and entrenched in corporations, auditors and regulators there will be upheaval and transformation challenges.