W3CR-WP-2026-03 | Division III | Autonomous Agent Safety & Formal Verification | Published 2026/11/23

Prompt-Injection Impossibility via Deterministic Pre-Sign AST Firewalls and Bounded ERC-7715 Delegation

Institutional Attribution (COPE): Web3 Capital Research - Division III Agent Safety & Formal Verification Working Group

ACM Reproducibility Badges: Artifacts Available | Artifacts Functional | Results Reproduced

ABSTRACT

Granting large language models (LLMs) direct custody of cryptographic signing keys or relying on probabilistic secondary LLM classifiers to screen outgoing transactions is provably broken under adaptive indirect prompt injection: any non-zero single-turn false-negative rate epsilon > 0 compounds across K adversarial turns toward certainty, lim_{K->infinity} [1 - (1 - epsilon)^K] = 1. Drawing on Goguen-Meseguer noninterference and Schneider execution-monitoring theory, this monograph proves a constructive impossibility result against unauthorized transaction signing. We isolate master key material behind a 64 KB Model Context Protocol (MCP) air-gap exposing exactly 14 typed tools, where every candidate EIP-5792 call bundle must satisfy a decidable linear-time O(|tx|) Abstract Syntax Tree (AST) policy automaton parameterized by ERC-7715 capabilities (contract allowlists, 4-byte ABI selectors, periodic spend caps, blind-signature prohibition, and <= 30-day TTL). Consequently, arbitrary untrusted context streams produce exact out-of-policy signing probability ASR_inj = 0.00%, verified across 60/60 zero-dependency clean-room security obligations.

FORMAL THEOREM - Theorem 3.1 (Semantic Noninterference & Prompt-Injection Impossibility)

Let an autonomous reasoning agent M process arbitrary adversarial token sequences u in Sigma* without read access to private signing key sk (mutual information I(Context_M; sk) = 0 bits). Suppose every signature request generated by M passes through a deterministic execution monitor A_Phi implementing policy language L(Phi) over canonicalized transaction ASTs: target address in Allow_addr, 4-byte calldata selector in Allow_sel (disallowing raw fallback/delegatecall/unlimited-approval selectors unless explicitly scoped), cumulative rolling value v_cum + v_tx <= Cap_period, and block timestamp t_now <= T_exp <= t_grant + 30 days. Because A_Phi is a total, decidable O(|tx|) state predicate independent of u, no adversarial prompt u in Sigma* can induce a valid cryptographic signature outside L(Phi), establishing exact breach rate ASR_inj = 0.

FALSIFIABLE EMPIRICAL BENCHMARK METRICS

[0 bits] LLM Context Private-Key Exposure - Signing keys physically air-gapped from neural context and agent runtime memory
[64 KB / 14] Air-Gapped MCP Envelope & Tools - 64 KB frame ceiling across 14 typed JSON-RPC 2.0 tools with 0 npm dependencies
[60 / 60] Verified Security Obligations - OBL-001..OBL-060 deterministic suite including risk-score >= 94 auto-halt
[0.00%] Prompt-Injection Breach Rate (ASR_inj) - Decidable O(|tx|) AST + ERC-7715 target, selector, cap & <= 30d TTL guard

PROBLEM FORMULATION & SYSTEM MODEL

Autonomous tool-using agents routinely ingest untrusted external data - web documents, code repositories, cross-agent messages, and on-chain metadata - into the same neural attention workspace that plans tool invocations. Unlike SQL injection, natural language lacks a syntactic separation between data and control instructions within transformer self-attention layers.

First-generation "agentic wallets" commit two fatal architectural errors: (1) loading raw BIP-39 seeds or session private keys into environment variables inside the agent's Python/Node.js runtime, or (2) interposing a "guardrail LLM" to judge whether a proposed transaction looks benign. Under Saltzer and Schroeder's Principle of Complete Mediation (1975) and Schneider's characterization of Enforceable Security Policies (2000), probabilistic neural filters fail the foundational requirement of reference-monitor soundness: for any single-query evasion rate [Formal Mathematical Expression] , an adaptive adversary embedding [Formal Mathematical Expression] polymorphic payloads drives cumulative compromise probability [Formal Mathematical Expression] arbitrarily close to 1.

- Ambient Authority Hazard: An agent holding a raw private key possesses unbounded authority across every chain, asset, and contract simultaneously.

- Calldata Obfuscation & Blind Signing: Malicious prompts frequently instruct agents to sign opaque EIP-712 digests, eth_sign raw hashes, or nested EIP-7702 / DELEGATECALL upgrades disguised as routine settlements.

FORMAL CONSTRUCTION & MATHEMATICAL PROOF

1. Physical Key Air-Gap & Bounded 64 KB MCP Transport: We partition the architecture into an untrusted reasoning domain [Formal Mathematical Expression] and an isolated cryptographic signing domain [Formal Mathematical Expression] communicating solely over a strict JSON-RPC 2.0 Model Context Protocol (MCP) channel capped at 64 KB (65,536 bytes) per frame. Exactly 14 deterministic tools are exposed (including offline ABI decoding, simulation, and scoped EIP-5792 submission); zero methods exist to export seeds, raw private scalars, or arbitrary blind-hash signatures.

2. Decidable Pre-Sign AST Automaton ([Formal Mathematical Expression]): Every candidate transaction or EIP-5792 atomic call bundle [Formal Mathematical Expression] is parsed from raw hexadecimal bytes into a typed Abstract Syntax Tree prior to cryptographic signing. For each elementary call [Formal Mathematical Expression] , [Formal Mathematical Expression] evaluates a conjunction of five decidable predicates in [Formal Mathematical Expression] time:

- Target Containment: [Formal Mathematical Expression] (normalized EIP-55 / chain-scoped address set).

- Selector & AST Decode Soundness: [Formal Mathematical Expression] (if value-only permitted) or [Formal Mathematical Expression] , with recursive unpacking of ERC-7579 executeBatch frames so nested sub-calls cannot smuggle unauthorized selectors.

- Hazard Halt Invariant: Immediate fail-closed rejection (risk score [Formal Mathematical Expression]) on unbounded MAX_UINT256 approvals, unverified EIP-7702 code-delegation tuples, or eth_sign raw digests.

- ERC-7715 Rolling Spend & TTL Envelope: Monotone accounting ensures [Formal Mathematical Expression] and hard session expiry [Formal Mathematical Expression] .

3. Proof of Noninterference: In Goguen-Meseguer terms, purge all high-untrusted inputs [Formal Mathematical Expression] against the boolean indicator [Formal Mathematical Expression] . Because [Formal Mathematical Expression] invokes ECDSA/P-256 signing only

when [Formal Mathematical Expression] , and [Formal Mathematical Expression] , the conjunction [Formal Mathematical Expression] is unsatisfiable for every [Formal Mathematical Expression] .

EMPIRICAL EVALUATION & REFERENCE IMPLEMENTATION VERIFICATION

We evaluated the pre-sign AST firewall and ERC-7715 capability engine across a comprehensive adversarial suite spanning malformed ABI encodings, nested batch smuggling, and supply-chain isolation audits:

- Zero-Breach Adversarial Calldata Matrix: Across synthetic prompt-injected payloads attempting unlimited ERC-20 approvals, unauthorized recipient substitution, over-cap session transfers, expired ERC-7715 delegations (> 30 days), and disguised EIP-7702 code-set authorizations, the deterministic AST monitor achieved ASR_inj = 0.00% (100% deterministic rejection with risk score 94 auto-halt on critical vectors) in sub-millisecond evaluation time.

- 64 KB MCP Air-Gap & 14-Tool Completeness: Fuzzing both stdio and HTTP /api/mcp transports confirmed hard frame truncation/rejection above 64 KB , strict JSON-RPC schema enforcement across all 14 agentic tools, and 0 bits of secret-key material in agent memory dumps or error traces.

- 60/60 Clean-Room Security Obligations & Zero Supply-Chain Surface: Static and dynamic verification suites confirmed 60/60 security obligations (OBL-001 through OBL-060) passing with 0 runtime npm dependencies, eliminating third-party package hijacking vectors entirely.

VERIFICATION PROVENANCE & COMPETING INTERESTS DISCLOSURE

COPE & ACM Verification Provenance Statement: Authored and verified by the institutional Division III Agent Safety & Formal Verification Working Group under Principal Investigator oversight with zero fictional personal names. Policy-automaton truth tables and adversarial calldata rejection invariants are directly reproducible in-browser via the deterministic verification capsule (capsuleType: "ast-firewall").

Competing Interests & Reference Implementation Disclosure: Published by Web3 Capital Research (Republic of Singapore; https://research.web3.capital/) as open-access, non-advisory, zero-fundraising computer-security research under Singapore SFA 2001, FAA 2001, and PSA 2019 safe-harbor provisions, as well as the EU AI Act Regulation (EU) 2024/1689 Art. 2(6) scientific research exemption. Empirical 60/60 obligation benchmarks and 14-tool MCP harnesses were verified against the open-source non-custodial reference implementation maintained by MYRILUM LIMITED (Hong Kong UBI 80912065 ; https://myrilum.org/research and https://myrilum.io/), involving zero custodial asset handling and non-value synthetic test vectors (S_token === 0 , r_holder === 0).

VERIFIED REFERENCES

[SS75] Jerome H. Saltzer and Michael D. Schroeder (1975). "The Protection of Information in Computer Systems." Proceedings of the IEEE, Vol. 63, No. 9, pp. 1278 - 1308. https://doi.org/10.1109/PROC.1975.9939
[GM82] Joseph A. Goguen and Jos Meseguer (1982). "Security Policies and Security Models." 1982 IEEE Symposium on Security and Privacy, pp. 11 - 20. https://doi.org/10.1109/SP.1982.10014
[Sch00] Fred B. Schneider (2000). "Enforceable Security Policies." ACM Transactions on Information and System Security (TISSEC), Vol. 3, No. 1, pp. 30 - 50. https://doi.org/10.1145/353323.353382
[EIP7702] Vitalik Buterin, Sam Wilson, Ansgar Dietrichs, and Matt Garnett (2024). "EIP-7702: Set Code for EOAs." Ethereum Improvement Proposals. https://eips.ethereum.org/EIPS/eip-7702

[EIP5792] Moody Salem, Lukas Rosario, Wilson Cusack, Dror Tirosh, Jake Moxey, et al. (2022). "EIP-5792: Wallet Call API." Ethereum Improvement Proposals. https://eips.ethereum.org/EIPS/eip-5792
[ERC7715] Luka Isailovic, Derek Rein, Dan Finlay, Derek Chiang, Fil Makarov, et al. (2024). "ERC-7715: Request Permissions from Wallets." Ethereum Improvement Proposals. https://eips.ethereum.org/EIPS/eip-7715