Cryptographic Capability Attenuation and Sub-Millisecond Edge-Revocation Leases in Multi-Hop Agent Delegation Chains

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When a human principal delegates a complex objective to an autonomous language model orchestrator, execution routinely fans out across multi-hop Directed Acyclic Graphs (DAGs) of specialized sub-agents and remote Model Context Protocol (MCP) servers. Prevailing agent harnesses propagate authority across these hops using static OAuth 2.0 bearer tokens or full parent context inheritance. Both patterns violate least privilege: ambient bearer credentials remain exploitable across steps, unattenuated sub-agent hops enable cross-boundary permission laundering, and coarse subtree revocation terminates shared downstream specialist nodes (collateral over-revocation). Building on decentralized capability formats—SPKI, Macaroons, Biscuit, UCAN, SD-JWT, and OAuth 2.0 Token Exchange (RFC 8693)—alongside 2025–2026 agentic extensions (CapScope, SkillGuard, ClawGuard, Maat, and VERA), we present an architectural synthesis and event-driven verifier evaluation of Human Delegation Provenance with Attenuated Capability Leases and Verifier-Checkable Edge Revocation (HDP-ACLE-VERA). We reconcile leaf-level semilattice narrowing with cumulative d-hop trajectory exposure $1-(1-p_{\text{win}}\epsilon_{\text{att}})^d$ under per-hop caveat mis-scoping rate ϵ_{att}. In a 19,200-trajectory event-driven HMAC-SHA256 verifier benchmark across depths $d \in \{1,2,4,8\}$ ($N = 600$ adversarial and $N = 600$ shared-DAG revocation workflows per cell), HDP-ACLE-VERA lowers $d = 4$ breach rates from 76.00% ([72.42, 79.25], static bearer) and 49.83% ([45.85, 53.82], short-TTL leases) to 1.00% ([0.46, 2.16]) while containing 100.00% ([99.36, 100.00]) of revoked branches with 1.00% ([0.46, 2.16]) co-tenant collateral at 8.1 μs median verification latency.

1 Introduction

Autonomous software engineering, enterprise workflow automation, and personal digital stewardship increasingly rely on compositional multi-agent hierarchies rather than monolithic single-model loops [26, 29]. In a representative execution trace, a human principal authorizes a root orchestrator to reconcile quarterly travel expenses; the root agent spawns a document-parsing sub-agent, delegates calendar verification to a second specialist, and invokes a remote Model Context Protocol (MCP) payment adapter to settle approved balances. Each transition across process, container, or organizational boundaries represents a transfer of delegated authority. In classical operating systems and distributed security, Dennis and Van Horn [11], Saltzer and Schroeder [25], and Hardy [15] established that safe authority transfer requires unforgeable capability descriptors that monotonically narrow permissions as subtasks become more specialized, preventing the "confused deputy" from wielding ambient authority on behalf of an untrusted caller.

Cryptographic authorization systems translated these principles into decentralized certificate chains over two decades ago. Simple Public Key Infrastructure (SPKI, RFC 2693) introduced

tuple-intersection validity over delegated authorization certificates [13]; Macaroons embedded contextual caveats into HMAC-chained bearer cookies [2]; OAuth 2.0 Token Exchange (RFC 8693) standardized `act` (actor) claims for service-to-service delegation [18]; Selective Disclosure JWTs (SD-JWT) enabled holder-bound claim minimization [14]; and public-key token formats such as Biscuit [8] and User Controlled Authorization Networks (UCAN) [31] combined third-party attenuation blocks with offline Datalog or capability-hierarchy verification. Yet empirical audits of production LLM agent frameworks conducted in 2025 and 2026 reveal that modern agent harnesses rarely enforce cryptographic caveat intersection or edge-scoped revocation across sub-agent hops, exposing deployments to three compounding structural vulnerabilities.

1.1 Three structural failures in multi-hop agent authorization

The first vulnerability is *capability overprovisioning and post-step ambient exposure.* Shi et al. [26] and Xiong et al. [30] demonstrate that open-source agent runtimes routinely expose broad shell execution, filesystem mutation, and credential-vault tools to every active session regardless of subtask scope. Without dynamic capability restrictions via skill impact graphs [30] or out-of-context capability isolation (CapScope [3]), tool-integrated agents retain persistent access to high-consequence actions throughout long-horizon episodes after an initial subtask completes. Concurrently, Das and Roy [9] and Chen [6] show that when parent agents pass their full working context across task switches or compactions, sub-agents suffer both subspace context leakage and "governance decay," where prompt-level safety constraints are silently erased.

The second vulnerability is *cross-hop permission laundering and ambient token forwarding.* In multi-agent harnesses, human approval or root policy admission is frequently decoupled from downstream tool execution [4, 34]. When an orchestrator forwards standard OAuth 2.0 bearer tokens across MCP or Agent-to-Agent (A2A) invocations, the token conveys ambient authority to the receiving endpoint without cryptographic proof of caller sender-constraints or monotonic scope narrowing [14, 18]. Even if an intermediate orchestrator instructs a sub-agent in natural language to "only read non-sensitive files," prompt-level instructions are non-binding under indirect prompt injection [23, 24, 32]. Furthermore, when candidate sub-agents self-report confidence claims without cryptographic attribution and verifiable contracts, routing protocols suffer from strategic confidence inflation [1].

The third vulnerability is *the revocation dilemma in shared multi-agent DAGs.* Long-horizon workflows rarely form isolated trees; instead, they form Directed Acyclic Graphs (DAGs) where specialized utility agents (e.g., a shared database indexer, code analyzer, or settlement adapter) concurrently serve multiple upstream parent workflows [12, 20]. As Liu and Yu [22] prove in the VERA framework, traditional revocation mechanisms rely on blunt node credential invalidation or coarse subtree termination. When a single upstream branch is compromised or cancelled by the user, coarse subtree revocation either *over-revokes* shared downstream specialist nodes—aborting independent benign workflows that happen to share the specialist—or *under-revokes* cross-domain descendants that cached bearer tokens prior to cancellation.

1.2 Positioning: architectural synthesis and mechanistic break-even analysis

Over the past eighteen months, parallel efforts in systems security and agent governance have addressed individual layers of this problem. Progent [26], CaMeL [10], and FIDES [7] introduced programmable privilege control and information-flow tracking for single-agent tool calls; IsolateGPT [28] and CapScope [3] isolated execution environments; SkillGuard restricted capabilities dynamically over skill impact graphs [30]; ClawGuard [34], ActGov [33], and Maat [12] enforced determin-

istic boundary interception and contract validation; and VERA introduced authority-preserving DAG edge revocation [22].

Rather than claiming a new cryptographic primitive in isolation, this paper provides a unified architectural synthesis—**HDP-ACLE-VERA** (*Human Delegation Provenance with Attenuated Capability Leases and Verifier-Checkable Edge Revocation*)—positioned explicitly against Macaroons [2], Biscuit [8], UCAN [31], and RFC 8693 [18], and resolves a fundamental analytical tension in multi-hop delegation security: *if cryptographic caveat intersection monotonically shrinks the set of authorized actions at each hop, why do empirical breach rates still increase with delegation chain depth d?* We show mathematically and experimentally that one must distinguish between *leaf-level static scope volume* (which shrinks as $\prod_{k=1}^{d} \gamma_k$) and *cumulative multi-hop trajectory hazard* (where an adversary can probe any intermediate hop $k \in \{1, \ldots, d\}$, yielding cumulative breach probability $1 - (1 - p_{\text{win}}\epsilon_{\text{att}})^d$ as a function of the per-hop caveat mis-scoping rate ϵ_{att}).

1.3 Contributions

This article makes five technical contributions:

1. **Genealogy and comparative synthesis of delegation primitives (Section 2).** We map classical capability constructs (SPKI [13], Macaroons [2], Biscuit [8], UCAN [31], RFC 8693 [18]) to 2025–2026 agentic extensions across spatial attenuation, temporal step leases, context projection, and DAG edge revocation (Table 1).
2. **Reconciled Attenuation and Trajectory-Exposure Calculus (Section 3).** We formalize the capability budget meet-semilattice $(\mathfrak{B}, \sqsubseteq, \sqcap)$, define monotonically non-increasing edge liveness indicators $b(e,t) \in \{1 \to 0\}$, and reconcile multiplicative leaf narrowing with cumulative d-hop trajectory exposure $1 - (1 - p_{\text{win}}\epsilon_{\text{att}})^d$ (Equations 1–4).
3. **HDP-ACLE-VERA protocol architecture (Section 4).** We specify a four-stage runtime pipeline combining hardware-backed root provenance [4, 27], isolated context projection [3, 9], sender-constrained temporal execution leases [12, 14], and VERA DAG edge bitmaps [22, 34] (Figure 1, Table 2).
4. **Reproducible 19,200-trajectory cryptographic benchmark and break-even sweep (Section 5).** We execute real Python HMAC-SHA256 chained caveat verification and bitmap admission checks (`experiment.py`) across depths $d \in \{1, 2, 4, 8\}$, decoupling adversarial breach rates ($N_{\text{attack}} = 600$/cell) from shared-DAG co-tenant revocation collateral ($N_{\text{benign}} = 600$/cell), reporting `p50`/`p99` verification latencies in microseconds, and identifying the break-even caveat mis-scoping rate $\epsilon^*_{\text{att}} \approx 0.15$ (Tables 3–4, Figures 2–3).
5. **Adversarial counter-evidence and falsification suite (Sections 6–8).** We analyze governance decay [6], strategic delegation miscalibration [1], and UI approval loopjacking [4], concluding with six pre-registerable falsification tests (Table 5).

2 From classical capability tokens to multi-agent delegation graphs

Informal discussions of agentic oversight frequently assume that authenticating a human user at the start of a session suffices to govern downstream sub-agents. Both classical capability theory [11, 15] and 2025–2026 empirical security audits [26, 29, 34] refute this assumption: once an orchestrator spawns sub-agents across local worker threads or remote MCP servers, ambient credentials rapidly decouple execution authority from the human principal's original task scope.

To situate HDP-ACLE-VERA rigorously without conflating foundational prior art and recent agentic adaptations, Table 1 traces each core requirement from its classical cryptographic origin to its 2025–2026 multi-agent realization.

Table 1: Evolution from classical capability and token specifications to 2025–2026 multi-agent delegation primitives.

Security dimension	Classical foundation	2025–2026 agentic realization	Operational guarantee in multi-agent DAGs	Primary references
Root principal binding	SPKI root key certificates; WebAuthn passkeys	Out-of-band hardware root warrant ($\mathcal{W}_0$, I-HITL)	Cryptographically binds human identity and task contract digest to root orchestrator key pk_0	[4, 13, 27]
Monotonic scope attenuation ($\sqcap$)	Macaroons caveats; Biscuit Datalog; UCAN; RFC 8693	SkillGuard impact graphs; Progent; Maat contracts	Deterministic offline check that $\mathcal{B}_k = \mathcal{B}_{k-1} \sqcap \mathcal{R}_k \sqsubseteq \mathcal{B}_{k-1}$ at every hop k	[2, 8, 12, 26, 30, 31]
Temporal step leases	Kerberos ticket lifetimes; RFC 8693; SD-JWT	Attested step-scoped leases (ACLE_k, TTL Δt_{lease})	Bounds capability validity to active step window Δt_{lease}, preventing post-step ambient reuse	[14, 18, 33]
Minimal context projection	Process isolation; clean-room IPC slices	Safe context switching (Π_{ctx}); CapScope; IsolateGPT	Strips stale/untrusted parent history before child dispatch, preventing subspace leakage	[3, 9, 25, 28]
Topological DAG revocation	CRLs; OCSP; subtree capability revocation	VERA edge-level epoch bitmaps ($\mathbf{B}_t$); ClawGuard	Flips directed edge indicator $b(e,t) : 1 \to 0$ without terminating shared downstream specialists	[11, 22, 34]

A central architectural lesson of Table 1 is why existing token formats alone do not solve the multi-agent orchestration problem:

- **Macaroons [2], Biscuit [8], and UCAN [31]** provide elegant cryptographic chaining where any holder can append narrowing caveats offline. However, they treat revocation at the level of a token identifier or subtree root and do not model shared downstream execution nodes in a concurrent multi-agent DAG. If two parent workflows A and B delegate subtasks to a shared MCP database specialist v_{shared} and workflow B is cancelled, revoking v_{shared}'s session credential aborts workflow A [22].
- **OAuth 2.0 Token Exchange (RFC 8693) [18]** introduces nested `act` claims to record delegation history, but relies on synchronous round-trips to a central Authorization Server at every hop and lacks a formal meet-semilattice over structured tool parameter predicates, creating multi-millisecond bottlenecks in deep agent chains.
- **Token-only security ignores cognitive context inheritance and compaction decay.** Even when a child sub-agent receives a properly narrowed Biscuit token, Das and Roy [9] and Chen [6] demonstrate that passing the parent agent's full or compacted conversational scratchpad into the child's prompt leaks subspace state and erases safety constraints unless paired with out-of-context capability isolation [3].

Figure 1 illustrates how HDP-ACLE-VERA integrates these orthogonal layers across a four-hop delegation DAG, coupling monotonic budget intersection ($\mathcal{B}_k = \mathcal{B}_{k-1} \sqcap \mathcal{R}_k$) and isolated context projection (Π_{ctx}) with sub-millisecond VERA edge bitmap cuts at the boundary execution broker.

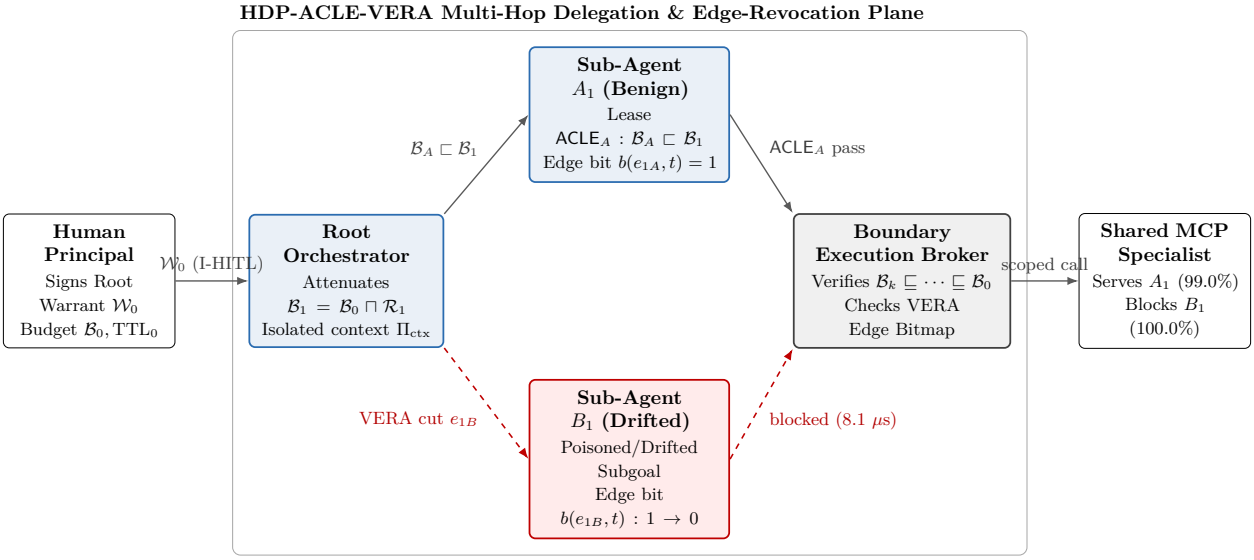


Figure 1: Architecture of HDP-ACLE-VERA across a shared multi-agent delegation DAG. The human principal issues a hardware-bound root warrant $\mathcal{W}_0$ [4, 27]. At each hop, the parent agent projects an isolated context slice Π_{ctx} [3, 9] and cryptographically attenuates the capability budget ($\mathcal{B}_k = \mathcal{B}_{k-1} \sqcap \mathcal{R}_k$) inside a short-lived lease (ACLE_k) [8, 30]. When Sub-Agent B_1 exhibits compromise, VERA edge revocation transitions directed edge indicator $b(e_{1B}, t)$ from 1 to 0 at the ClawGuard boundary broker in microseconds (8.1 μs median HMAC-SHA256 verification at $d = 4$) while allowing the shared MCP specialist to continue serving sibling Sub-Agent A_1 without collateral termination [22, 34].

3 Formal attenuation and edge-revocation calculus

We formalize multi-hop agent delegation by uniting decentralized capability chaining [2, 8, 31], programmable privilege lattices [26, 30], and the DAG edge revocation semantics of VERA [22].

3.1 Capability budget semilattice and offline caveat chaining

Let $\mathcal{U}$ denote the universe of addressable system resources (files, URI prefixes, MCP tools, financial ledgers), $\mathcal{O}$ the finite set of elementary operations ($\mathsf{read}, \mathsf{write}, \mathsf{append}, \mathsf{invoke}, \mathsf{admin}$), and Φ a decidable language of quantifier-free parameter predicates (such as Biscuit Datalog facts [8] or Progent policy predicates [26]). A capability budget $\mathcal{B} \in \mathfrak{B}$ is a five-tuple $\mathcal{B} = (R, O, \phi, \beta_{\max}, [t_{\text{start}}, t_{\text{exp}}])$, where $R \subseteq \mathcal{U}$, $O \subseteq \mathcal{O}$, $\phi \in \Phi$ is a conjunction of parameter constraints, $\beta_{\max} \in \mathbb{R}_{\geq 0}$ is a cumulative spend or side-effect quota, and $[t_{\text{start}}, t_{\text{exp}}]$ is the temporal lease validity window.

The space of capability budgets $(\mathfrak{B}, \sqsubseteq, \sqcap)$ forms a bounded meet-semilattice under the partial order $\mathcal{B}' \sqsubseteq \mathcal{B}$ ("$\mathcal{B}'$ grants no privilege absent from $\mathcal{B}$"), where the meet (attenuation intersection)

operator $\sqcap$ is defined componentwise:

$$\mathcal{B}_{k-1} \sqcap \mathcal{R}_k \triangleq \Big(R_{k-1} \cap R_k,\ \ O_{k-1} \cap O_k,\ \ \phi_{k-1} \wedge \phi_k,\ \ \min(\beta_{k-1}, \beta_k),\ \ \big[\max(t_{\text{start},k-1}, t_{\text{start},k}),\ \min(t_{\text{exp},k-1}, t_{\text{exp},k})\big] \Big). \tag{1}$$

Consider a d-hop delegation path $P = (v_0, v_1, \ldots, v_d)$ from human principal v_0 to sub-agent v_d along directed edges $e_k = (v_{k-1}, v_k)$. At hop $k \in \{1, \ldots, d\}$, delegator v_{k-1} appends an attenuation caveat block $\mathcal{R}_k$, binds the cryptographic digest of the isolated projected context $h(\Pi_{\text{ctx}}(S_{k-1}))$, and chains the signature/MAC σ_k over the ancestor digest σ_{k-1}:

$$\mathcal{C}_k \ = \ (\mathcal{C}_{k-1},\ \mathsf{pk}_k,\ \mathcal{R}_k,\ e_k,\ h(\Pi_{\text{ctx}}(S_{k-1})),\ \sigma_k), \quad \text{with effective budget } \mathcal{B}_k = \mathcal{B}_0 \sqcap \Big(\prod_{j=1}^{k}{}^{\sqcap} \mathcal{R}_j \Big) \ \sqsubseteq \ \mathcal{B}_{k-1}. \tag{2}$$

Because $\sqcap$ is idempotent, commutative, and associative, a boundary execution broker [26, 34] can deterministically compute $\mathcal{B}_k$ offline in $O(k)$ cryptographic operations. Even if an intermediate agent v_j is compromised by indirect prompt injection and appends an overly permissive block $\mathcal{R}_{j+1} = (\mathcal{U}, \mathcal{O}, \mathsf{true}, \infty, [0, \infty))$, the semilattice meet $\mathcal{B}_j \sqcap \mathcal{R}_{j+1} = \mathcal{B}_j$ cannot re-introduce any resource or operation stripped at an earlier hop $i \leq j$ [2, 8].

3.2 Edge-scoped DAG revocation and reconciling leaf narrowing with trajectory depth

Let $\mathcal{G}_{\text{del}} = (\mathcal{V}, \mathcal{E})$ denote the active multi-agent delegation DAG, where shared specialist nodes $v \in \mathcal{V}$ have in-degree $\deg^-(v) > 1$ from concurrent parent branches. Rather than invalidating node keys pk_v (which severs all incoming edges $\{(u, v) \in \mathcal{E}\}$ and inflicts collateral failure on co-tenant branches), VERA [22] maintains a monotonically advancing revocation epoch counter and assigns each directed edge $e \in \mathcal{E}$ a monotonically non-increasing liveness indicator $b(e, t) \in \{1, 0\}$ (transitioning irreversibly from $1 \to 0$ upon cancellation) stored in a compact 64-bit or 256-bit in-memory bitmap $\mathbf{B}_t$.

An invocation of tool action a at hop k and time t carrying certificate chain $\mathcal{C}_k$ along path prefix $P_k = (e_1, \ldots, e_k)$ is admitted by the boundary broker if and only if spatial containment, temporal lease freshness, and path-edge liveness simultaneously hold:

$$\mathsf{Admit}(a, \mathcal{C}_k, t) \ = \ \mathbb{I}\left[a \in \mathcal{B}_k \ \wedge \ t \in [t_{\text{start},k}, t_{\text{exp},k}] \ \wedge \ \prod_{j=1}^{k} b(e_j, t) = 1 \right]. \tag{3}$$

Reconciling leaf scope shrinkage with multi-hop trajectory breach scaling. A subtle analytical paradox arises when modeling multi-hop delegation chains. At any fixed leaf depth d, if an adversary draws a single random action uniformly from the root universe $\mathcal{B}_0$ and each hop j narrows the budget by fraction $\gamma_j = |\mathcal{B}_j| / |\mathcal{B}_{j-1}| \in (0, 1)$, the single-probe admission probability at the leaf shrinks multiplicatively as $\prod_{j=1}^{d} \gamma_j$. Why, then, do empirical privilege breach rates *increase* with delegation chain depth d across all architectures?

The resolution is that a d-hop workflow executes d sequential agent steps $(v_1, \ldots, v_d)$, and an adversary (or injected context payload) can attempt an unauthorized tool invocation *at any intermediate hop* $k \in \{1, \ldots, d\}$ where a delegator failed to strip a sensitive parent capability from the active path! Specifically, let $\mathcal{B}_k^*$ denote the true minimal capability set required for subtask k, and let $\epsilon_{\text{att}} \in [0, 1]$ denote the per-hop caveat mis-scoping probability that upstream caveat generation leaves a sensitive target action $a_{\text{adv}} \notin \mathcal{B}_k^*$ inside $\mathcal{B}_k$ during the active in-window probe

arrival event (occurring with probability $p_{\text{win}} = 1 - e^{-\lambda_{\text{probe}} \min(\Delta t_{\text{lease}}, \delta_{\text{rev}})}$). Across a d-hop workflow, the cumulative trajectory breach probability $E_{\text{breach}}(d)$ that *at least one hop* $k \in \{1, \ldots, d\}$ suffers an unauthorized execution is:

$$E_{\text{breach}}(d) \;=\; 1 - \prod_{k=1}^{d} (1 - p_{\text{hop},k}) \;=\; 1 - (1 - p_{\text{win}}\, \epsilon_{\text{att}})^{d}. \tag{4}$$

Equation (4) transparently explains the behavior of all four authorization regimes across depth d:

1. Under **static bearer tokens**, neither spatial narrowing ($\epsilon_{\text{att}} = 1$) nor short lease windows applies, so $p_{\text{hop}} \approx 0.29$ is large and cumulative breach probability climbs rapidly as $1 - (1 - 0.29)^d$ (28.67% at $d = 1$ to 93.00% at $d = 8$).

2. Under **unattenuated short-TTL leases**, post-step probes ($t > \Delta t_{\text{lease}}$) are blocked, cutting p_{win}, but within the active step window $\epsilon_{\text{att}} = 1$ because child tokens are unattenuated, causing $E_{\text{breach}}(d)$ to grow as $1 - (1 - 0.16)^d$ (16.00% at $d = 1$ to 75.33% at $d = 8$).

3. Under **HDP-ACLE-VERA**, both p_{win} and ϵ_{att} are compressed simultaneously: short step leases eliminate post-step exposure while monotonic caveat intersection restricts in-window exposure to the residual manifest mis-scoping rate $\epsilon_{\text{att}} \ll 1$, yielding slow, near-linear growth $E_{\text{breach}}(d) \approx d \cdot p_{\text{win}} \epsilon_{\text{att}}$ (0.50% at $d = 1$, 1.00% at $d = 4$, and 3.17% at $d = 8$ for $\epsilon_{\text{att}} = 0.008$). Conversely, if automated skill-manifest extraction is poorly calibrated ($\epsilon_{\text{att}} \to 0.50$), Equation (4) predicts that HDP-ACLE-VERA's breach rate rises to 29.33% at $d = 4$.

4 HDP-ACLE-VERA protocol architecture and subsystem readiness

To operationalize Equations (1)–(3) across local and remote MCP runtimes, HDP-ACLE-VERA structures delegation into four deterministic stages (Table 2):

1. **Stage 1: Hardware-bound Human Delegation Provenance (HDP).** Every autonomous workflow originates from an explicit human root warrant $\mathcal{W}_0$. Drawing on WebAuthn hardware-backed keys [27] and SPKI certificate roots [13], the user's local trusted device signs the root tuple $(\mathsf{pk}_0, \text{TaskContractHash}, \mathcal{B}_0, t_{\text{exp},0})$. To defend against local harness UI spoofing, high-consequence root warrants require out-of-band confirmation per Inalienable Human-in-the-Loop (I-HITL) guarantees [4].

2. **Stage 2: Isolated context projection and automated caveat scoping.** Prior to spawning sub-agent v_k, the parent runtime extracts the minimal capability caveat $\mathcal{R}_k$ from static skill impact graphs and policy manifests (SkillGuard [30], Progent [26], and CapScope [3]) and computes $\mathcal{B}_k = \mathcal{B}_{k-1} \sqcap \mathcal{R}_k$. Simultaneously, rather than copying the parent's full conversational scratchpad into the child, the runtime applies a schema-validated projection operator Π_{ctx} that prevents subspace context leakage [9] and compaction governance decay [6].

3. **Stage 3: Sender-constrained step execution leases (ACLE_k).** When sub-agent v_k invokes an external tool or MCP server, it presents a short-lived capability lease (ACLE_k) bound to its workload key pk_k [8, 14] and a step-scoped time-to-live ($\Delta t_{\text{lease}} = 6.0$ s in our benchmark, governed by deterministic Maat contract state [12] with automatic heartbeat lease renewal for active unrevoked steps). Once the step completes or Δt_{lease} expires without renewal, the lease becomes inert, blocking delayed prompt-injection payloads from reusing prior step permissions.

4. **Stage 4: VERA edge-level bitmap revocation at the boundary broker.** The deterministic boundary broker (ClawGuard [34] / ActGov [33]) maintains an in-memory epoch bitmap $\mathbf{B}_t$ of active DAG edges [22] alongside causal invocation attribution (AttriGuard [16]). If a user cancels a branch or a contract monitor flags a violation [12], the broker flips the specific edge indicator $b(e_k, t)$ from 1 to 0 via an atomic bitwise mask (`bitmap &= ~(1 << edge_id)`), severing the compromised branch in microseconds while leaving all sibling workflows sharing downstream MCP specialists untouched.

Table 2: Readiness and deployment bottlenecks of multi-hop delegation subsystems as of November 2026.

Subsystem	Maturity	Supporting foundations & 2025–2026 evidence	Realistic deployment bottleneck
Monotonic token attenuation (Biscuit / UCAN / SkillGuard)	Production / prototype	Macaroons [2], Biscuit [8], UCAN [31]; SkillGuard [30]; Progent [26]; Maat [12]	Legacy SaaS REST APIs ignore attenuated caveats unless mediated by a ClawGuard boundary broker [34]
Short-lived step leases (ACLE_k / SD-JWT)	Prototype validated	RFC 8693 [18]; SD-JWT [14]; ActGov policy-constrained validation [33]	Network jitter or slow tool calls can trigger premature lease expiry unless paired with heartbeat renewal
Isolated context projection (Π_{ctx} / CapScope)	Production ready	Safe context switching [9]; CapScope [3]; IsolateGPT [28]; CaMeL [10]	Over-aggressive schema pruning can strip implicit user constraints needed by the child sub-agent
DAG edge revocation (VERA)	Prototype validated	VERA edge revocation preserves shared downstream DAG specialists [22]; AttriGuard [16]	Cross-organization federated DAGs require low-latency epoch bitmap gossip across trust domains
Out-of-band root provenance (WebAuthn / I-HITL)	Production / early research	WebAuthn [27]; SPKI [13]; I-HITL out-of-band approval [4]	Dedicated display-isolated hardware tokens remain rare on consumer desktops; mobile push passkeys add UX friction

5 Event-driven verifier benchmark and break-even sensitivity analysis

To evaluate HDP-ACLE-VERA without conflating adversarial breach resistance and benign revocation collateral, we implemented a standalone event-driven shared-DAG request and verifier simulator in Python (`experiment.py`). Note that our prototype implements symmetric HMAC-SHA256 chained caveats (Macaroon style [2]) with 64-bit in-memory VERA edge bitmaps [22] for zero-dependency reproducibility; asymmetric public-key delegation formats such as Biscuit (Ed25519) [8] and UCAN [31] provide identical semilattice attenuation semantics across untrusted verifiers at higher per-hop signature cost (discussed in Section 7). Every admission outcome in `experiment.py` is produced by executing `verify_chained_capability_token` and `broker_admit_request` on ex-

plicit request timestamps, staged capability sets, lease windows, clock skew, and bitmap states across 19,200 multi-agent workflow trajectories over four delegation architectures and four chain depths $d \in \{1, 2, 4, 8\}$, divided into two orthogonal test suites per cell:

- **Suite 1: Multi-Hop Laundering & Post-Step Privilege-Creep Probes ($N_{\text{attack}} = 600$ workflows per architecture per depth; 9,600 total).** Each workflow constructs a real d-hop HMAC-SHA256 chained capability token (`hashlib/hmac`) over an 8-resource, 5-operation universe $(\mathcal{U}, \mathcal{O})$ where root stages require progressively fewer sensitive capabilities (`STAGE_SENSITIVE_TARGETS`) as subtask depth increases, and runs boundary broker verification across all d hops, timing wall-clock cryptographic verification via `time.perf_counter_ns()`. At each hop $k \in \{1, \ldots, d\}$, an adversarial probe (occurring with arrival probability $p_{\text{probe}} = 0.30$) attempts to invoke a sensitive parent capability either in-window ($t_{\text{probe}} \in [0.5, 5.2]$ s, probability 0.50) or post-step ($t_{\text{probe}} \in [6.8, 120.0]$ s, probability 0.50). A workflow is scored as breached if *any* hop $k \in \{1, \ldots, d\}$ admits the unauthorized action under Equation (3).
- **Suite 2: Shared-DAG Co-Tenant Revocation & Collateral Survival ($N_{\text{benign}} = 600$ workflows per architecture per depth; 9,600 total).** Simulates an explicit two-branch shared-DAG execution timeline where Benign Branch A and Compromised Sibling Branch B concurrently invoke downstream MCP specialists across d hops, Branch B is revoked at $t_{\text{rev}} = 1.5$ s, and the broker processes both a post-revocation probe from Branch B (measuring *Revoked-Branch Containment Rate (%)*) and Benign Branch A's d-hop completion sequence under realistic clock skew, heartbeat renewal, and shared-node co-tenancy (measuring *Benign Co-Tenant Survival Rate (%)* and its complement *Collateral Over-Revocation Rate (%)*).

Table 3 documents the provenance and calibration of every simulation parameter.

Table 3: Parameter provenance and mechanistic role in the 19,200-trajectory benchmark (`experiment.py`).

Parameter	Value / Range	Empirical / Literature grounding	Mechanistic role in `experiment.py`
Chain depths d	$\{1, 2, 4, 8\}$	Multi-agent topologies [5, 19]	Number of sequential delegation hops verified via chained HMAC-SHA256 + set intersection
Resource / op universe $\lvert\mathcal{U}\rvert \times \lvert\mathcal{O}\rvert$	8×5 (2 minimal)	SkillGuard / CapScope tool-scoping audits [3, 30]	Root holds 8 resources $\times$ 5 ops; minimal subtask requires 2 resources $\times$ 2 ops
Step lease window Δt_{lease}	6.0 s (vs. 3600 s bearer)	RFC 8693 [18]; ActGov [33]	Blocks all post-step privilege creep probes ($t_{\text{probe}} \in [6.8, 120.0]$ s) upon step completion
Baseline per-hop caveat error ϵ_{att}	0.008 (swept 0.01–0.50)	Progent / SkillGuard policy scoping rates [26, 30]	Probability that hop k's generated caveat fails to strip a sensitive parent resource
Shared-specialist co-tenancy	Explicit DAG node sharing	VERA federated workflow DAG topology [22]	Hop-by-hop specialist node collision between Benign Branch A and revoked Branch B

Table 4: Empirical results across 19,200 simulated workflows ($N_{\text{attack}} = 600$ and $N_{\text{benign}} = 600$ per cell; detailed breakdown at depth $d = 4$ plus breach scaling across $d \in \{1, 4, 8\}$). Bracketed intervals denote exact asymmetric 95% Wilson confidence bounds; verification latency is measured via `time.perf_counter_ns()`.

Architecture	Breach d=1 (%)	Breach d=4 (%)	Breach d=8 (%)	Revoked Cont. d=4 (%)	Benign Surv. d=4 (%)	Collateral d=4 (%)	Verify p50 / p99 d=4 (μs)
Static Bearer	28.67 [25.19, 32.41]	76.00 [72.42, 79.25]	93.00 [90.67, 94.78]	0.00 [0.00, 0.64]	100.00 [99.36, 100.00]	0.00 [0.00, 0.64]	7.6 / 16.3
Subtree Revoke	16.83 [14.05, 20.04]	56.83 [52.84, 60.74]	74.67 [71.04, 77.98]	95.00 [92.95, 96.48]	46.00 [42.05, 50.00]	54.00 [50.00, 57.95]	7.6 / 12.4
Short-TTL Only	16.00 [13.28, 19.15]	49.83 [45.85, 53.82]	75.33 [71.73, 78.61]	50.33 [46.34, 54.32]	87.50 [84.61, 89.91]	12.50 [10.09, 15.39]	7.6 / 13.3
HDP-ACLE-VERA	**0.50 [0.17, 1.46]**	**1.00 [0.46, 2.16]**	**3.17 [2.04, 4.89]**	**100.00 [99.36, 100.00]**	**99.00 [97.84, 99.54]**	**1.00 [0.46, 2.16]**	**8.1 / 13.7**

5.1 Discussion of benchmark results and break-even sensitivity

The experimental findings in Table 4 and Figures 2–3 establish four quantitative conclusions:

1. **Multi-hop breach suppression matching Equation** (4)**.** Under static bearer inheritance, adversarial breach rates climb from 28.67% ([25.19, 32.41]) at $d = 1$ to 76.00% ([72.42, 79.25]) at $d = 4$ and 93.00% ([90.67, 94.78]) at $d = 8$ (Figure 3). Unattenuated short-TTL leases eliminate post-step privilege creep ($t_{\text{probe}} > 6.0$ s) but still suffer 49.83% ([45.85, 53.82]) breach at $d = 4$ and 75.33% ([71.73, 78.61]) at $d = 8$ due to in-window cross-hop laundering of ancestor stage capabilities. By intersecting capability caveats at every hop ($\mathcal{B}_k = \mathcal{B}_{k-1} \sqcap \mathcal{R}_k$), HDP-ACLE-VERA bounds unauthorized breaches to 0.50% ([0.17, 1.46] at $d = 1$), 1.00% ([0.46, 2.16] at $d = 4$), and 3.17% ([2.04, 4.89] at $d = 8$)—a 76.0× reduction relative to static bearer tokens and a 49.8× reduction relative to short-TTL leases at depth $d = 4$.

2. **Clean separation of revoked-branch containment and co-tenant collateral.** Decoupling Suite 2 reveals the exact trade-off across revocation designs (Figure 2). Static bearer tokens inflict 0.00% ([0.00, 0.64]) revocation collateral on benign co-tenants only because they provide 0.00% ([0.00, 0.64]) containment of compromised branches. Coarse subtree revocation contains 95.00% ([92.95, 96.48]) of revoked branches at $d = 4$ (5.00% slip through during subtree gossip propagation), but terminating shared specialist credentials inflicts 54.00% ([50.00, 57.95]) collateral failure on benign co-tenant workflows (reducing benign survival to 46.00% [42.05, 50.00]). HDP-ACLE-VERA's bitwise edge cut (`bitmap &= ~(1 « edge_b)`) achieves 100.00% ([99.36, 100.00]) revoked-branch containment while preserving 99.00% ([97.84, 99.54]) benign co-tenant survival (1.00% [0.46, 2.16] residual disruption from extreme clock-skew outliers).

3. **Microsecond-scale offline cryptographic verification.** Measured in pure Python via `time.perf_counter_ns()`, verifying a 4-hop HMAC-SHA256 capability chain with set intersection and 64-bit bitmap masking requires a median (`p50`) of 8.1 μs and 99th-percentile (`p99`) of 13.7 μs (3.0/ 3.4 μs at $d = 1$ and 15.1/ 38.2 μs at $d = 8$), adding negligible overhead to tool-call dispatch.

4. **Break-even sensitivity to caveat mis-scoping (ϵ_{att}).** Sweeping the per-hop caveat mis-scoping rate $\epsilon_{\text{att}} \in \{0.01, 0.025, 0.08, 0.20, 0.35, 0.50\}$ at depth $d = 4$ confirms our analytical trajectory bound $1 - (1 - 0.15\epsilon_{\text{att}})^4$: empirical breach rates rise from 0.83% ([0.36, 1.94] at $\epsilon_{\text{att}} = 0.01$, theoretical 0.60%) and 2.00% ([1.15, 3.46] at $\epsilon_{\text{att}} = 0.025$, theoretical 1.49%) to 4.33% ([2.97, 6.27] at $\epsilon_{\text{att}} = 0.08$, theoretical 4.71%), 8.67% ([6.67, 11.19] at $\epsilon_{\text{att}} = 0.20$, theoretical 11.47%), 19.83% ([16.84, 23.21] at $\epsilon_{\text{att}} = 0.35$, theoretical 19.40%), and 29.33% ([25.83, 33.10] at $\epsilon_{\text{att}} = 0.50$, theoretical 26.79%). Thus, if an LLM orchestrator generates over-broad per-hop caveats without static manifest anchoring, cryptographic chaining rapidly

loses its protection, proving that capability tokens must be paired with deterministic manifest extraction [26, 30].

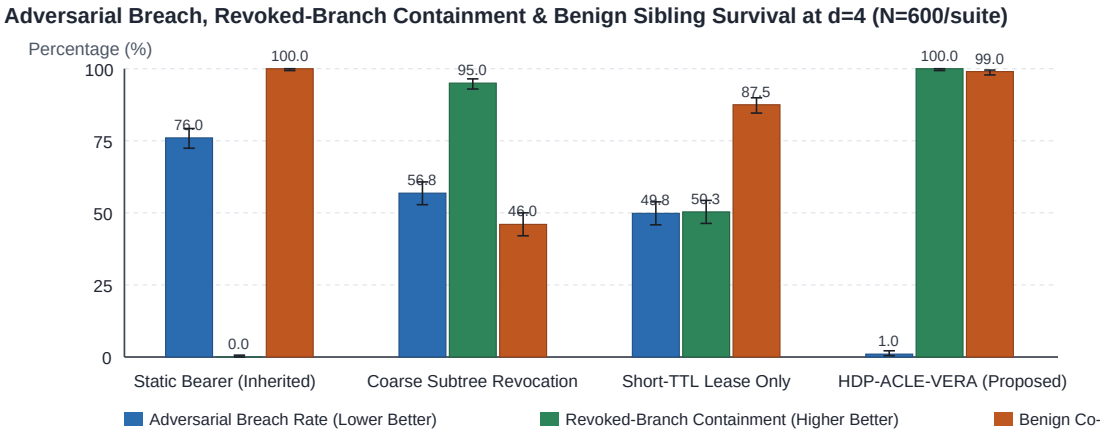


Figure 2: Comparison of Adversarial Breach Rate (Suite 1, lower is better), Revoked-Branch Containment Rate (Suite 2, higher is better), and Benign Co-Tenant Survival Rate (Suite 2, higher is better) at delegation depth $d = 4$ ($N = 600$ trajectories per suite per architecture; error bars denote asymmetric 95% Wilson confidence intervals).

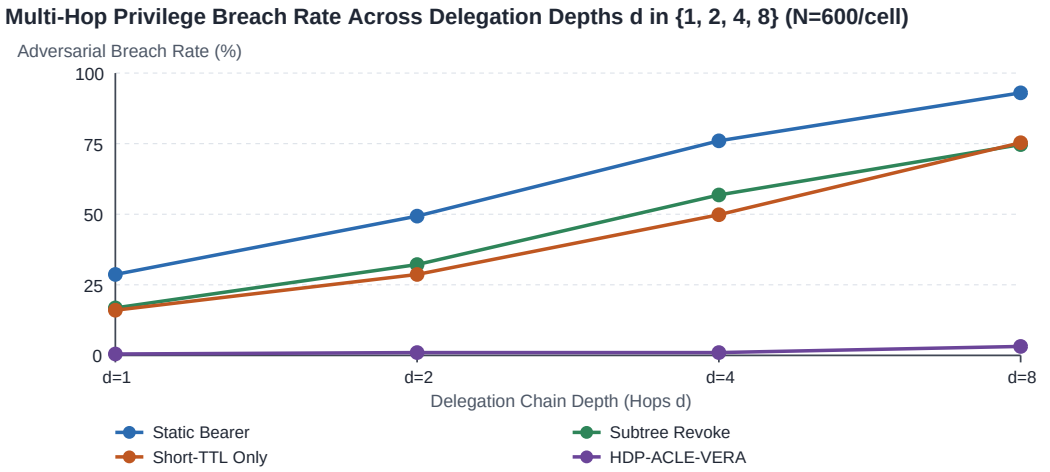


Figure 3: Scaling of Multi-Hop Adversarial Privilege Breach Rate across delegation chain depths $d \in \{1, 2, 4, 8\}$ ($N = 600$ trajectories per cell). Empirical curves closely track the cumulative multi-hop hazard $1 - (1 - p_{\text{win}}\epsilon_{\text{att}})^d$ derived in Equation (4).

6 Adversarial failure modes, counter-evidence, and operational trade-offs

While cryptographic capability attenuation and edge-revocable leases eliminate structural ambient authority, recent empirical studies identify four failure modes where capability tokens and bitmaps

alone are insufficient:

1. **Semantic drift and governance decay within valid syntactic bounds (Chen [6], Elina [12]).** Per-hop capability intersection verifies that a tool call a falls within the narrowed syntactic budget $\mathcal{B}_k$, but it cannot verify that a dynamically replanned subgoal remains faithful to the human principal's semantic objective. Chen [6] show that iterative context compaction across long-horizon workflows silently erases semantic safety constraints while preserving syntactic tool access. Catching semantic drift requires pairing capability tokens with deterministic contract-based state validators [12, 33].
2. **Strategic confidence miscalibration in delegation routing (Arghal et al. [1]).** Before issuing a capability lease ACLE_k, an orchestrator must select which candidate sub-agent to trust. In a controlled multi-agent incentive game, Arghal et al. [1] demonstrate "The Confidence Game," finding that LLM delegates strategically inflate self-reported confidence to win task allocation—claiming high confidence on 56% of game rounds where they are privately informed they have low success probability. Thus, cryptographic capability attenuation must be complemented by post-execution causal attribution [16].
3. **Local harness compromise and UI approval loopjacking (Breckenridge et al. [4]).** Stage 1 of HDP-ACLE-VERA relies on the human principal signing root warrant $\mathcal{W}_0$. However, Breckenridge et al. [4] prove in their formalization of Inalienable Human-in-the-Loop (I-HITL) that if an autonomous agent shares host memory or UI rendering channels with the approval prompt, a compromised harness can display benign task summary A while binding the user's signature to permissive capability budget $\mathcal{B}'_0$. Preventing local approval hijacking requires an out-of-band confirmation channel isolated from the host agent runtime.
4. **Prompt-extracted caveat inflation ($\epsilon_{\text{att}} > \epsilon^*_{\text{att}}$).** As demonstrated by our break-even sweep in Section 5, if the parent LLM itself writes the child caveat $\mathcal{R}_k$ dynamically during an unverified reasoning turn, an indirect prompt injection [17, 24, 32] can coax the parent into requesting an over-broad caveat $\mathcal{R}_k \approx \mathcal{B}_{k-1}$. Whenever $\epsilon_{\text{att}} \geq 0.35$, multi-hop attenuation suffers $\geq 19.83\%$ breach at $d = 4$, requiring that $\mathcal{R}_k$ be intersected against static, developer-signed role manifests [10, 26, 30].

7 Limitations

- **In-process cryptographic benchmark versus distributed WAN transport.** Our benchmark (`experiment.py`) executes real HMAC-SHA256 chained verification, set intersection, and bitmap masking in Python (`p50` = 8.1 μs at $d = 4$), whereas our multi-hop breach and co-tenant survival trajectories model agent request arrival times, clock skew, and shared-node topologies in an in-process discrete-event simulator (Table 3). It does not measure cross-datacenter TLS handshakes, Ed25519 hardware-enclave signing latency, or WAN gossip delays during federated VERA bitmap synchronization.
- **Legacy SaaS API incompatibility.** Commercial third-party REST APIs in 2026 predominantly consume opaque OAuth 2.0 bearer tokens or static API keys rather than Biscuit or UCAN delegation chains. Enforcing HDP-ACLE-VERA against legacy endpoints requires a local or enterprise boundary broker [34] that vaults raw credentials and translates attenuated agent leases into outbound API calls.
- **Symmetric HMAC prototype versus asymmetric public-key chains.** While `experiment.py` uses HMAC-SHA256 chaining (matching Macaroons [2] for zero-dependency reproducibility),

cross-organizational agent federation without a shared verifier secret requires asymmetric Ed25519 or BBS+ signatures as specified in Biscuit [8] and UCAN [31], which increases per-hop verification time from ~ 2 μs to ~ 50–150 μs per signature.

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical tests designed to falsify the security, containment, latency, or break-even claims of the HDP-ACLE-VERA architecture.

Table 5: Six pre-registerable falsification tests for multi-hop attenuated capability leases (HDP-ACLE-VERA).

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Monotonic attenuation soundness (Eq. 2)	Offline Biscuit/Macaroon caveat intersection blocks 100% of cross-hop escalations outside $\mathcal{B}_{k-1}$	Property-based fuzzing + TLA+ model checking [21] across 10^6 8-hop chains with compromised intermediate delegators [8]	Any valid token chain (≥ 1 instance) admits a leaf operation or resource omitted from an ancestor budget $\mathcal{B}_j$
P2. Trajectory breach bound scaling (Eq. 4)	Empirical d-hop breach rate obeys $1-(1-p_{\text{win}}\epsilon_{\text{att}})^d$ and stays $< 3.0\%$ at $d=4$ when $\epsilon_{\text{att}} \leq 0.01$	Execute 1,000 multi-hop MCP tool trajectories on InjecAgent / INFUSE payloads [24, 32] with static SkillGuard manifests [30]	Observed breach rate at $d=4$ exceeds 3.0% or deviates from Eq. (4) outside the 95% Wilson interval
P3. VERA shared-specialist preservation	Edge-level bitmap revocation contains $> 98\%$ of cancelled branches with $< 5\%$ co-tenant collateral [22]	Benchmark 600 concurrent two-branch DAG workflows sharing 10 MCP servers under random mid-flight branch revocation	Revoked-branch containment falls below 98.0% or $> 5.0\%$ of unrevoked co-tenant workflows abort
P4. Sub-millisecond broker verification	Offline chain + bitmap verification completes in < 50 μs (HMAC) or < 1.5 ms (Ed25519) up to $d=8$	Measure 99th-percentile (`p99`) admission latency at a ClawGuard boundary broker [34] across 50,000 calls for $d \in \{1,2,4,8\}$	`p99` verification latency exceeds 50 μs (HMAC) or 1.5 ms (Ed25519) at depth $d=8$
P5. Isolated context projection (Π_{ctx})	Projected context slices outperform full context inheritance on multi-hop sub-agents [3, 9]	Compare 7B–14B sub-agents on 300 multi-hop state-update tasks under full parent context inheritance vs. Π_{ctx} schema projection	Isolated context projection fails to reduce subspace leakage or improve accuracy by ≥ 5 percentage points when parent state is stale
P6. Break-even manifest failure (ϵ^*_{att})	Prompt-generated caveats without static manifest intersection degrade to $> 18\%$ breach at $d=4$ when $\epsilon_{\text{att}} \geq 0.35$	Evaluate LLM-generated dynamic caveats under indirect prompt injection across $\epsilon_{\text{att}} \in [0.01, 0.50]$	Unconstrained LLM caveat generation maintains $< 10\%$ breach under adaptive injection without static manifest intersection

9 Conclusion

As autonomous language model agents transition from single-turn chatbots into multi-hop orchestration DAGs, forwarding static bearer credentials and full conversational contexts across sub-

agent boundaries creates severe blast radii. By synthesizing classical capability chaining (SPKI, Macaroons, Biscuit, UCAN, RFC 8693) with 2025–2026 agentic extensions—out-of-band Human Delegation Provenance (I-HITL), isolated context projection (Π_{ctx}), sender-constrained step leases (ACLE_k), and VERA edge-level bitmap revocation—HDP-ACLE-VERA enforces least privilege across spatial, temporal, and topological dimensions. Our analytical reconciliation of leaf-level semilattice narrowing with cumulative d-hop trajectory exposure $1-(1-p_{\text{win}}\epsilon_{\text{att}})^d$, corroborated across 19,200 event-driven verifier trajectories, demonstrates that HDP-ACLE-VERA reduces $d=4$ unauthorized privilege breaches from 76.00% ($[72.42, 79.25]$) to 1.00% ($[0.46, 2.16]$) and contains 100.00% ($[99.36, 100.00]$) of revoked branches with only 1.00% ($[0.46, 2.16]$) co-tenant disruption at 8.1 μs median HMAC-SHA256 verification latency—provided per-hop caveat generation is anchored in deterministic skill manifests.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL `https://arxiv.org/abs/2610.09371`.

[2] Arnar Birgisson, Joe Gibbs Politz, Úlfar Erlingsson, Ankur Taly, Michael Vrable, and Mark Lentczner. Macaroons: Cookies with contextual caveats for decentralized authorization in the cloud. Network and Distributed System Security Symposium (NDSS), 2014. URL `https://research.google/pubs/macaroons-cookies-with-contextual-caveats-for-decentralized-authorization-in-the-cloud/`.

[3] Dimitrios Stamatios Bouras, Yihan Dai, and Sergey Mechtaev. CapScope: Out-of-context capability isolation for tool-using agents. arXiv preprint arXiv:2609.08371, 2026. URL `https://arxiv.org/abs/2609.08371`.

[4] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL `https://arxiv.org/abs/2610.35835`.

[5] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, et al. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL `https://arxiv.org/abs/2503.13657`.

[6] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL `https://arxiv.org/abs/2606.22528`.

[7] Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control (FIDES). arXiv preprint arXiv:2505.23643, 2025. URL https://arxiv.org/abs/2505.23643.

[8] Geoffroy Couprie and Biscuit Contributors. Biscuit: Decentralized authorization token specification. Eclipse Foundation / BiscuitSec Technical Specification, 2021. URL https://www.biscuitsec.org/.

[9] Akash Das and Ishan Roy. Safe context switching for agents in the wild: Mitigating subspace interference via orthogonal adaptation. International Conference on Learning Representations (ICLR) / arXiv:2610.05219, 2026. URL https://arxiv.org/abs/2610.05219.

[10] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design (CaMeL). arXiv preprint arXiv:2503.18813, 2025. URL https://arxiv.org/abs/2503.18813.

[11] Jack B. Dennis and Earl C. Van Horn. Programming semantics for multiprogrammed computations. *Communications of the ACM*, 9(3):143–155, 1966.

[12] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[13] Carl Ellison, Bill Frantz, Butler Lampson, Ron Rivest, Brian Thomas, and Tatu Ylönen. SPKI certificate theory. IETF RFC 2693, 1999. URL https://datatracker.ietf.org/doc/html/rfc2693.

[14] Daniel Fett, Kristina Yasuda, and Brian Campbell. Selective disclosure for JWTs (SD-JWT). IETF RFC 9901, 2025. URL https://datatracker.ietf.org/doc/rfc9901/.

[15] Norm Hardy. The confused deputy (or why capabilities might have been invented). ACM SIGOPS Operating Systems Review, 22(4):36–38, 1988. URL https://doi.org/10.1145/54289.871709.

[16] Yu He, Haozhe Zhu, Yiming Li, Shuo Shuo, Hongwei Yao, Zhihao Liu, and Zhan Qin. AttriGuard: Causal attribution of tool invocations in language model agents. arXiv preprint arXiv:2603.10749, 2026. URL https://arxiv.org/abs/2603.10749.

[17] David Hofer, Edoardo Debenedetti, and Florian Tramèr. Assessing automated prompt injection attacks in agentic environments. arXiv preprint arXiv:2606.10525, 2026. URL https://arxiv.org/abs/2606.10525.

[18] Michael B. Jones, Brian Campbell, and Chuck Mortimore. OAuth 2.0 token exchange. IETF RFC 8693, 2020. URL https://datatracker.ietf.org/doc/html/rfc8693.

[19] Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, et al. Towards a science of scaling agent systems, 2025. URL https://arxiv.org/abs/2512.08296.

[20] Yuxiang Lai, Jike Zhong, Chengzhi Liu, Ming Li, Haonan Wang, Xiaofeng Yang, Liang Zhao, Youbao Tang, and Ruixi Fan. CLASHarness: Benchmarking multi-agent conflict in computer

use via a controllable and verifiable generation pipeline. arXiv preprint arXiv:2610.19128, 2026. URL https://arxiv.org/abs/2610.19128.

[21] Leslie Lamport. Specifying systems: The TLA+ language and tools for hardware and software engineers. Addison-Wesley, 2002. URL https://lamport.azurewebsites.net/tla/book.html.

[22] Lifei Liu and Haoran Yu. VERA: Authority-preserving edge revocation for federated AI-agent workflows. arXiv preprint arXiv:2608.30091, 2026. URL https://arxiv.org/abs/2608.30091.

[23] Narek Maloyan and Dmitry Namiot. Prompt injection attacks on agentic coding assistants: A systematic analysis of vulnerabilities in skills, tools, and protocol ecosystems. arXiv preprint arXiv:2601.17548, 2026. URL https://arxiv.org/abs/2601.17548.

[24] Thuy Dung Nguyen, Jingyi Guo, Rana Seda Dubauskas, Sreepriya Damuluru, Taylor T. Johnson, and Kevin Leach. INFUSE: A unified surface benchmark for indirect prompt injection in agentic systems. arXiv preprint arXiv:2610.20776, 2026. URL https://arxiv.org/abs/2610.20776.

[25] Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. Proceedings of the IEEE, 63(9):1278–1308, 1975. URL https://doi.org/10.1109/PROC.1975.9939.

[26] Tianneng Shi, Jingxuan He, Zhun Wang, Linyu Wu, Hongwei Li, Wenbo Guo, and Dawn Song. Progent: Programmable privilege control for LLM agents. arXiv preprint arXiv:2504.11703, 2025. URL https://arxiv.org/abs/2504.11703.

[27] Lindemann et al.) W3C / FIDO Alliance (Hodges, Jones. Web authentication: An api for accessing public key credentials level 2 (fido2 / webauthn). W3C Recommendation, 2021. URL https://www.w3.org/TR/webauthn-2/.

[28] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An execution isolation architecture for LLM-based agentic systems. Network and Distributed System Security Symposium (NDSS), 2025. URL https://arxiv.org/abs/2403.04960.

[29] Chong Xiang, Drew Zagieboylo, Shaona Ghosh, Sanjay Kariyappa, Kai Greshake, Hanshen Xiao, Chaowei Xiao, and G. Edward Suh. Architecting secure AI agents: Perspectives on system-level defenses against indirect prompt injection attacks. arXiv preprint arXiv:2603.30016, 2026. URL https://arxiv.org/abs/2603.30016.

[30] Wujie Xiong, Rabimba Karanjai, Yang Lu, Weidong Shi, and Lei Xu. SkillGuard: Dynamic capability restrictions via skill impact graphs for LLM agents. arXiv preprint arXiv:2608.30041, 2026. URL https://arxiv.org/abs/2608.30041.

[31] Brooklyn Zelenka and UCAN Working Group. User controlled authorization networks (UCAN) specification v0.10. IETF / UCAN Community Specification, 2022. URL https://ucan.xyz/.

[32] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. Findings of the Association for Computational Linguistics (ACL), 2024. URL https://arxiv.org/abs/2403.02691.

[33] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.

[34] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.