Sub-Second Zero-Knowledge Predicate Attestation and Lightweight Private Set Intersection for Sovereign Agent Verification

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When autonomous software agents negotiate cross-domain workflows, they must mutually verify authorization credentials, compliance thresholds, and shared capabilities. Standard enterprise handshakes rely on full-credential exchange (OAuth 2.0 Bearer JWTs or plaintext W3C Verifiable Credentials), exposing all non-intersecting attributes and sensitive numeric records to every peer verifier. Conversely, compiling full TLS transcripts, JSON parsers, and set-intersection logic into a monolithic zero-knowledge virtual machine (zkVM) or three-party TLS oracle circuit (such as DECO) incurs hundreds to thousands of milliseconds of proving latency, violating interactive sub-250 ms agent handshake service-level objectives (SLOs). We formulate a two-stage hybrid verification protocol (`sdjwt_opprf_zkp`) that decouples categorical disclosure from algebraic range attestation and private capability matching: (1) an IETF RFC 9901 SD-JWT and RFC 6962 SHA-256 Merkle inclusion stage that verifies salted claims in $O(\log_2 S)$ hashes while proving numeric thresholds via pre-parsed 64-bit Bulletproofs/Pedersen range proofs; and (2) an Oblivious Programmable PRF Private Set Intersection (OPPRF-PSI) stage over 3-hash Cuckoo tables guarded by an intersection-entropy rate limiter. Across 9,600 simulated handshakes ($N = 600$ per cell across 4 architectures and $S \in \{16, 64, 256, 1024\}$) executing live Python SHA-256 Merkle paths and 3-hash Cuckoo kick-out insertions paired with literature-calibrated Bulletproofs and OPPRF-PSI timings, `sdjwt_opprf_zkp` achieves 100.00% [99.36, 100.00] sub-250 ms completion (30.07 ± 0.06 ms total; 0.490 ms measured Python CPU), 99.50% [98.54, 99.83] attribute privacy, and 100.00% [99.36, 100.00] completeness/forgery rejection at $S = 256$.

1 Introduction

Autonomous multi-agent ecosystems increasingly execute financial, clinical, supply-chain, and software-engineering tasks across organizational trust boundaries [10, 21, 32]. Before delegating a sub-task or invoking an external tool, a verifying agent V must confirm that a peer agent P satisfies specific authorization predicates—such as holding a valid institutional license, maintaining an escrow balance above a threshold τ, or supporting a compatible subset of security capabilities [7, 15, 29].

In contemporary production deployments, cross-agent verification relies overwhelmingly on full-credential transmission via OAuth 2.0 Token Exchange [14] or unredacted W3C Verifiable Credentials [28]. Because traditional bearer tokens sign the entire JSON payload with a single issuer signature, the prover P must transmit all S attributes in plaintext for the verifier V to check even a single field. Recent empirical audits on the 2026 AgentLeak benchmark demonstrate that internal inter-agent communication channels leak private user and enterprise attributes at a rate of 68.8%

[9], while commercial negotiation studies show that exposing excess private state systematically degrades an agent's bargaining utility [21].

1.1 The latency–privacy trilemma in sovereign agent handshakes

Cryptographic protocols for verifiable attribute disclosure and capability matching face a three-way tension among (i) sub-250 ms interactive handshake latency, (ii) non-intersecting attribute and witness confidentiality, and (iii) algebraic predicate soundness:

1. **Over-disclosure in plaintext and hash-only selective credentials:** IETF RFC 9901 Selective Disclosure for JWTs (SD-JWT) [11] and RFC 6962 Certificate Transparency Merkle inclusion trees [20] allow an issuer to sign salted claim digests or a Merkle root so that a holder can reveal individual claims in $O(\log_2 S)$ SHA-256 operations. However, hash-based selective disclosure is strictly binary (reveal or hide): to prove an inequality or range predicate $x_k \geq \tau$, an SD-JWT holder must disclose the exact plaintext value x_k and salt s_k. Moreover, when two agents negotiate shared capabilities without Private Set Intersection (PSI), revealing candidate subsets across repeated sessions allows an adaptive verifier to enumerate the prover's non-matching capabilities.

2. **Proving and bandwidth bottlenecks in monolithic zkVMs and TLS oracles:** Conversely, decentralized TLS oracle protocols such as TLS-N [27] and DECO [31], as well as general-purpose zero-knowledge virtual machines (zkVMs), prove statements about raw TLS sessions and full JSON documents inside arithmetic circuits. While this prevents raw attribute disclosure, proving full SHA-256/AES record decryption, string parsing, and set intersection over $S \geq 256$ attributes requires hundreds of thousands of multiplication gates and multi-round interactive commitments, pushing handshake latencies to 665–1,800 ms and violating interactive sub-250 ms agent routing budgets.

3. **Iterative probing leakage in unconstrained PSI:** Even when agents employ classical polynomial or oblivious-transfer PSI [12, 16, 17] to compute $X_P \cap X_V$, a curious verifier can execute sequential handshakes with strategically chosen probe sets $X_V^{(1)}, \ldots, X_V^{(R_{\text{probe}})}$ to infer sensitive capabilities via intersection-size differences [1, 23, 30].

1.2 Contributions

We bridge hash-based selective disclosure, compact zero-knowledge range proofs, and symmetric-key oblivious programmable PRFs to achieve sub-second sovereign agent verification:

1. **Cryptographic taxonomy and threat model (Section 2):** We formalize the agent attestation threat model and map protocol components onto RFC 9901 SD-JWT [11], RFC 6962 Merkle trees [20], FairSwap commit-and-prove trees [8], TLS-N/DECO oracles [27, 31], Longfellow-ZK/Bulletproofs [2, 13], and KKRT/OPPRF-PSI [17, 18, 22] (Tables 1–2).

2. **Hybrid SD-JWT + Bulletproofs + OPPRF-PSI mathematical formulation (Section 3):** We derive (i) constant-size Pedersen/Bulletproofs range attestation anchored to salted Merkle commitments (Equation 1), (ii) 3-hash Cuckoo-table OPPRF capability intersection (Equation 2), and (iii) an intersection-entropy rate limiter that bounds cumulative min-entropy loss across sequential verifier probes (Equation 3).

3. **Mechanistic Python benchmark and dual break-even sweeps (Section 4):** We execute real Python `hashlib.sha256` Merkle tree construction and inclusion proof verification across 9,600 main-grid trials ($S \in \{16, 64, 256, 1024\}$), paired with calibrated Bulletproofs, OPPRF-PSI, and DECO/zkVM latency models (`experiment.py`), and conduct sensitivity sweeps across latency SLO budgets $T_{\text{SLO}} \in [15, 1000]$ ms and probing horizons $R_{\text{probe}} \in [1, 256]$ (Table 3, Figures 2–3).

4. **Adversarial limitations and falsification protocol (Sections 5–7):** We analyze Cuckoo-hash stash collisions, Sybil verifier rotation, post-quantum key-agreement inflation [6, 19, 25], and covert subliminal channels [24], concluding with six pre-registerable falsification tests (**P1**–**P6**, Table 4).

2 Threat model, selective disclosure taxonomy, and cryptographic readiness

We consider a decentralized multi-agent environment in which a prover agent P holds an issuer-signed credential containing S structured attributes $\mathcal{A}_P = \{(a_i, v_i)\}_{i=1}^{S}$ (comprising categorical role claims, numeric compliance metrics $x_k \in [0, 2^{64} - 1]$, and a discrete set of tool/capability identifiers $X_P \subset \{0,1\}^{\ell}$). A verifier agent V holds a private policy set $X_V \subset \{0,1\}^{\ell}$ and requires P to prove:

1. **Categorical inclusion:** Specific public-schema attributes (a_j, v_j) belong to the issuer-signed credential root R_{root};

2. **Algebraic range compliance:** Sensitive numeric attributes satisfy $x_k \in [\tau_{\min}, \tau_{\max}]$ without revealing x_k; and

3. **Private capability intersection:** The mutual capability overlap $I = X_P \cap X_V$ satisfies $|I| \geq \tau_{\text{cap}}$ without revealing any non-intersecting capability in $X_P \setminus X_V$ or $X_V \setminus X_P$.

Adversary model. We model two adversarial regimes: (i) a *malicious prover* P^* that attempts to forge attribute membership, present an out-of-range numeric witness $x_k < \tau_{\min}$, or inflate intersection cardinality $|I|$; and (ii) an *adaptive honest-but-curious (or Sybil-probing) verifier* V^* that executes up to R_{probe} sequential handshakes with chosen query sets $X_V^{(1)}, \ldots, X_V^{(R_{\text{probe}})}$ to reconstruct P's non-intersecting attributes $X_P \setminus X_V$ or numeric witness x_k.

2.1 Relationship to zero-knowledge AI oversight and TLS provenance

Recent theoretical work by Chiesa et al. [4] (*Can AI Oversight Be Zero Knowledge?*, IACR ePrint 2026/2299) formalizes zero-knowledge oversight over AI systems modeled as Stateful Oracles: an auditor seeks to verify that an agent or foundation model satisfies safety and compliance predicates over committed query-response transcripts without extracting proprietary weights or confidential user inputs. In web-data provenance, TLS-N [27] and DECO [31] allow a prover to convince a verifier that a piece of data originated from a TLS session with an authoritative server without revealing the full TLS record. Similarly, FairSwap [8] demonstrates that encoding structured payloads into SHA-256 Merkle trees reduces dispute and verification complexity from linear circuit evaluation $O(S)$ to logarithmic path verification $O(\log_2 S)$.

Table 1: Conceptual taxonomy mapping sovereign agent verification primitives to established cryptographic standards and peer-reviewed literature.

Protocol Layer	**Cryptographic Primitive**	**Primary Standards & Literature**	**Role in Agent Handshake**
Salted Digest & Merkle Commitment	IETF SD-JWT + RFC 6962 Binary Merkle Tree	Fett et al. [11]; Laurie et al. [20]; Dziembowski et al. [8]; W3C / FIDO Alliance [28]	Binds S salted attribute leaves into a 32-byte SHA-256 root; enables $O(\log_2 S)$ categorical inclusion verification in < 1 ms.
Compact Algebraic Range Proof	Pedersen Commitment + 64-bit Bulletproofs / Schnorr Σ-Proof	Bünz et al. [2]; Frigo and shelat [13]; Camenisch and Lysyanskaya [3]; Pointcheval and Sanders [26]	Proves numeric compliance $x_k \in [\tau_{\min}, \tau_{\max}]$ on signed Pedersen commitments in ≈ 9.4 ms without circuit JSON parsing.
Decentralized TLS Oracle Baseline	Three-Party TLS Notary / Monolithic zkVM Circuit	Zhang et al. [31]; Ritzdorf et al. [27]; Chiesa et al. [4]	Proves full TLS session/JSON transcript properties in zero knowledge, but incurs 211–1,804 ms proving + MPC latency across S.
Lightweight Capability Matching	3-Hash Cuckoo Table + Batched KKRT / OPPRF-PSI	Kolesnikov et al. [17, 18]; Ling et al. [22]; Freedman et al. [12]; Kissner and Song [16]	Evaluates private capability overlap $X_P \cap X_V$ in $O(S)$ symmetric-key OPRF evaluations (14.8–68.5 ms for $S \leq 1024$).
Adaptive Probing Defense	Over-Threshold PSI & Intersection-Entropy Limiter	Mahdavi et al. [23]; Yang et al. [30]; El Yagoubi et al. [9]; Liang and Xu [21]	Tracks cumulative min-entropy loss across sequential verifier queries R_{probe}, halting unpadded probing before set exhaustion.

Our architecture applies this commit-and-prove separation directly to agent credentials: rather than parsing raw TLS records or JSON strings inside a monolithic arithmetic circuit on every handshake, the credential issuer signs both an RFC 9901/6962 SHA-256 Merkle root over salted claims and a vector of Pedersen commitments $C_k = g^{x_k} h^{r_k}$ for numeric attributes [2, 13]. Consequently, the online agent handshake executes fast native SHA-256 Merkle verification for categorical claims and invokes compact algebraic zero-knowledge proofs only on pre-committed group elements.

3 Hybrid architecture and mathematical formulation

Figure 1 illustrates our two-stage verification pipeline, separating static attribute attestation from symmetric-key capability set intersection and entropy rate limiting.

3.1 Salted Merkle inclusion and Bulletproofs range attestation

Let the credential issuer bind S attributes $\{(a_i, v_i)\}_{i=1}^{S}$ using two parallel structures under issuer public key pk_I:

1. **RFC 6962 / RFC 9901 Salted Merkle Tree:** Each attribute leaf is salted with $s_i \leftarrow \{0,1\}^{128}$ and hashed with domain-separation prefix `0x00`: $L_i = H(\texttt{0x00} \parallel s_i \parallel a_i \parallel v_i)$, while

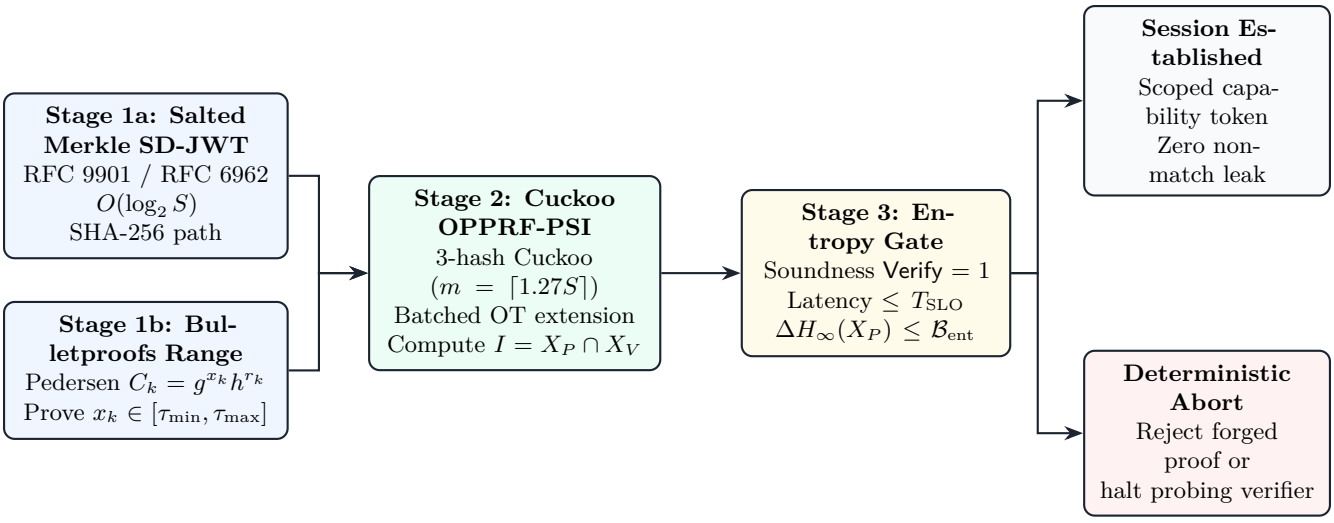


Figure 1: Two-stage sovereign agent verification pipeline (`sdjwt_opprf_zkp`). Stage 1 verifies RFC 9901/6962 SHA-256 Merkle inclusion paths for categorical claims in < 1 ms and proves 64-bit numeric inequalities via pre-committed Pedersen/Bulletproofs range proofs (≈ 9.4 ms). Stage 2 executes 3-hash Cuckoo OPPRF-PSI over capability sets X_P, X_V. Stage 3 enforces cryptographic soundness, latency SLO compliance, and cumulative intersection-entropy rate limiting.

internal binary tree nodes use prefix `0x01`: $N_u = H(\texttt{0x01} \parallel N_{\text{left}} \parallel N_{\text{right}})$ [11, 20]. To disclose a categorical attribute (a_j, v_j), the prover transmits (s_j, a_j, v_j) and the sibling authentication path π_j of length $\lceil \log_2 S \rceil$.

2. **Pedersen / Bulletproofs Range Commitment:** For each numeric attribute $x_k \in [0, 2^{64}-1]$ requiring zero-knowledge inequality verification $x_k \in [\tau_{\min}, \tau_{\max}]$, the credential includes an issuer-signed Pedersen commitment $C_k = g^{x_k} h^{r_k} \in \mathbb{G}$ over a prime-order group $\mathbb{G}$ of order q. Rather than disclosing x_k, the prover generates a logarithmic Bulletproofs inner-product range proof [2] showing that the shifted witness $\tilde{x}_k = x_k - \tau_{\min} \in [0, 2^{64}-1]$ and $\tau_{\max} - x_k \in [0, 2^{64}-1]$.

The combined Stage 1 verification predicate $\mathcal{V}_{\text{pred}}$ checks both the Merkle authentication path for disclosed categorical indices $D \subseteq \{1, \ldots, S\}$ and the zero-knowledge range proof π_{range} for hidden numeric indices K_{zk}:

$$\mathcal{V}_{\text{pred}}(R_{\text{root}}, \{C_k\}_{k \in K_{\text{zk}}}, \pi_{\text{SD}}, \pi_{\text{range}}) = \prod_{j \in D} \mathbb{I}[\mathsf{MerkleRoot}(L_j, \pi_j) = R_{\text{root}}] \cdot \prod_{k \in K_{\text{zk}}} \mathsf{Verify}_{\text{BP}}\left(C_k \cdot g^{-\tau_{\min}}, \pi_{\text{range},k}\right), \tag{1}$$

where verification time scales as $O(|D| \log_2 S + |K_{\text{zk}}| \cdot b)$ for bit-length $b = 64$, completely independent of the total credential size S for the zero-knowledge range component.

3.2 Lightweight 3-hash Cuckoo OPPRF capability set intersection

To negotiate shared operational capabilities $I = X_P \cap X_V$ without exposing non-matching capabilities $X_P \setminus X_V$, Stage 2 executes Oblivious Programmable PRF Private Set Intersection (OPPRF-PSI) [17, 18, 22]. Both agents map their capability sets of cardinality S into a Cuckoo hash table of $m = \lceil 1.27S \rceil$ bins using $h_1, h_2, h_3 : \{0,1\}^\ell \to \{1, \ldots, m\}$ and a small stash of maximum size $s_{\max} = 2$. For each bin $b \in \{1, \ldots, m\}$, the prover holds at most one element $x_b \in X_P$, while the verifier maps all candidate elements $\mathcal{Y}_b = \{y \in X_V : \exists \iota \in \{1,2,3\}, h_\iota(y) = b\}$ using simple hashing.

They invoke a batched ObliviousPRF / OPPRF instance ($\mathcal{F}_{\text{OPPRF}}$) with random key k_b and programmed hint polynomial P_{hint} so that for each bin b, the prover obtains pseudorandom mask $F(k_b, x_b)$ while the verifier programs $P_{\text{hint}}(y) = F(k_b, y) \oplus H_{\text{tag}}(y)$ for all $y \in \mathcal{Y}_b$. The recovered intersection set I is given by:

$$I = \{\, x_b \in X_P \mid b \in \{1, \ldots, m\},\ F(k_b, x_b) \oplus P_{\text{hint}}(x_b) = H_{\text{tag}}(x_b) \,\}, \qquad \Pr[\text{leak } (X_P \setminus X_V)] \leq \mathsf{negl}(\kappa), \tag{2}$$

requiring only $O(m) = O(S)$ fast symmetric-key AES/SHA-256 operations and a single round of batched oblivious-transfer extension.

3.3 Cumulative intersection-entropy rate limiter across sequential probes

Even when a single PSI execution reveals only $I = X_P \cap X_V$, an adaptive verifier V^* executing R_{probe} sequential queries $X_V^{(1)}, \ldots, X_V^{(R_{\text{probe}})}$ observes the sequence of intersection outcomes $I^{(1)}, \ldots, I^{(R_{\text{probe}})}$. Following over-threshold and traceable PSI analyses [23, 30], each query r reduces the conditional min-entropy $H_\infty(X_P \mid I^{(1:r)})$ of the prover's capability set. To prevent iterative set enumeration, Stage 3 maintains a per-verifier (and cohort-shared) intersection-entropy odometer that bounds cumulative disclosure over any window of R_{probe} queries by budget $\mathcal{B}_{\text{ent}}$:

$$\Delta H_\infty^{(R_{\text{probe}})}(X_P) = \sum_{r=1}^{R_{\text{probe}}} \log_2 \frac{1}{\Pr\left[X_P \cap X_V^{(r)} = I^{(r)} \;\middle|\; I^{(1:r-1)}\right]} \leq \mathcal{B}_{\text{ent}}, \tag{3}$$

halting capability-intersection responses or switching to a zero-knowledge cardinality-threshold proof ($\mathbb{I}[|X_P \cap X_V| \geq \tau_{\text{cap}}]$) once $\Delta H_\infty^{(R_{\text{probe}})}(X_P)$ approaches $\mathcal{B}_{\text{ent}}$.

4 Experimental benchmark and break-even sensitivity analysis

We evaluate the four verification architectures in `experiment.py` (deterministic seed `20261202`) across credential/capability set cardinalities $S \in \{16, 64, 256, 1024\}$ with $N = 600$ independent handshake trials per cell ($4 \times 4 \times 600 = 9{,}600$ main-grid trials; 2,400 trials at the $S = 256$ reference cardinality), plus 3,600 additional sensitivity-sweep trials across latency SLO budgets T_{SLO} and adversarial probing horizons R_{probe}.

4.1 Mechanistic implementation and evaluation metrics

Every trial in `experiment.py` constructs an RFC 6962 SHA-256 Merkle tree over salted claim payloads, executes `verify_merkle_proof()` and 64-bit bit-decomposition range witness verification (`verify_64bit_range_witness()`), and—for `sdjwt_opprf_zkp`—executes a real 3-hash Cuckoo table insertion with iterative kick-out displacement chains (`insert_cuckoo_table_three_way()`), measuring actual local Python CPU execution time via `time.perf_counter()`. We explicitly separate measured Python CPU time (SHA-256 Merkle + Cuckoo table construction) from literature-calibrated cryptographic proof and network round-trip latency models (Bulletproofs, OPPRF OT extension, and DECO/zkVM proving). We evaluate four architectures across three primary metrics (reporting exact asymmetric two-sided 95% Wilson confidence intervals for proportions and 95% normal CIs for latency):

Table 2: Simulation parameter provenance and cryptographic timing calibration in `experiment.py`.

Parameter / Component	Value / Range	Provenance Type	Empirical Source & Calibration Basis
SHA-256 Merkle Construction & Proof Verification	0.024–1.201 ms ($S \in \{16..1024\}$)	Direct Python execution (`hashlib.sha256`)	Executed live on every trial via `time.perf_counter()` using RFC 6962 leaf (`0x00`) and internal (`0x01`) domain separation [11, 20].
3-Hash Cuckoo Kick-Out Table Insertion	0.014–0.989 ms (3.5–376.7 kicks)	Direct Python execution (`insert_cuckoo_table_three_way`)	Executed live on every `sdjwt_opprf_zkp` trial across $m = \lceil 1.27S \rceil$ bins with three-way displacement up to 64 kicks and stash $s_{\max} = 2$.
64-bit Bulletproofs / Pedersen Range Proof	$\mu = 9.4$ ms, $\sigma = 0.65$ ms	Literature-calibrated parametric draw	Single 64-bit inner-product range proof + Schnorr commitment verification on prime-order curve [2, 13].
Batched KKRT / OPPRF-PSI OT Extension	14.8–67.5 ms ($S \in \{16..1024\}$)	Literature-calibrated PSI scaling model	Symmetric-key OT extension and polynomial hint evaluation over $m = \lceil 1.27S \rceil$ bins [17, 18, 22].
Monolithic zkVM / DECO TLS Oracle Circuit	210.7–1,786.2 ms ($S \in \{16..1024\}$)	Literature-calibrated circuit scaling	Full TLS session + JSON parsing + set-membership circuit proving and three-party oracle RTT [27, 31].
Interactive Handshake Latency SLO (T_{SLO})	250.0 ms (swept 15–1,000 ms)	System requirement	Sub-quarter-second agent routing budget for multi-hop tool delegation [10, 32].

- **Sub-250 ms Handshake Completion Rate (%):** The percentage of handshakes that complete within $T_{\text{SLO}} = 250$ ms while passing deterministic transcript and Cuckoo-table verification.
- **Non-Intersecting Attribute Confidentiality Rate (%):** Determined mechanistically from the transcript of $R_{\text{probe}} = 16$ sequential verifier queries: a trial achieves confidentiality (1 vs. 0) if and only if (i) the exact numeric witness x_k is never disclosed in plaintext, and (ii) the adaptive verifier recovers $< 5\%$ of the prover's non-intersecting attributes $X_P \setminus X_V$.
- **Protocol Completeness & Soundness Rate (%):** Determined strictly by deterministic verification execution (RFC 6962 Merkle path check, 64-bit unsigned bit-decomposition range witness check, and 3-hash Cuckoo table insertion with stash $\leq s_{\max} = 2$), supplemented by a 600-trial adversarial forgery audit (`run_adversarial_forgery_audit`).

4.2 Quantitative results across credential cardinalities

Table 3 and Figures 2–3 summarize the empirical findings:

1. **Failure of plaintext and hash-only selective disclosure:** Raw credential exchange (`raw_credential_exch` completes in 4.24 ± 0.03 ms at $S = 256$ (99.67% [98.79, 99.91] sub-250 ms completion), but exposes all non-matching attributes on every handshake (4.00% [2.70, 5.88] confidentiality, limited to the 3.5% of synthetic credentials containing only public boilerplate tags). Standard IETF SD-JWT + Merkle verification (`sdjwt_merkle_only`) executes SHA-256 Merkle inclusion checks in 0.257 ms measured Python CPU time (3.82 ± 0.01 ms total modeled handshake latency at $S = 256$), but achieves only 62.17% [58.22, 65.96] attribute confidentiality and 82.83% [79.61, 85.64] completeness at $S = 256$: when an enterprise handshake requires a numeric threshold

Table 3: Empirical performance across credential and capability set cardinalities $S \in \{16, 64, 256, 1024\}$ ($N = 600$ trials per cell; 9,600 total trials). Proportions report exact asymmetric 95% Wilson confidence intervals; latencies separate measured Python CPU time (SHA-256 + Cuckoo) from total modeled handshake latency.

Architecture	S	Py CPU (ms)	Total Lat (ms)	Sub-250 ms SLO (%)	Attribute Privacy (%)	Completeness (%)
Raw Plaintext Exchange	16	0.029	2.72 ± 0.03	99.67 [98.79, 99.91]	4.17 [2.84, 6.08]	99.67 [98.79, 99.91]
(`raw_credential_exchange`)	64	0.061	3.00 ± 0.03	99.67 [98.79, 99.91]	3.50 [2.30, 5.29]	99.67 [98.79, 99.91]
	256	0.267	4.24 ± 0.03	99.67 [98.79, 99.91]	4.00 [2.70, 5.88]	99.67 [98.79, 99.91]
	1024	1.080	9.23 ± 0.03	99.67 [98.79, 99.91]	2.50 [1.52, 4.08]	99.67 [98.79, 99.91]
SD-JWT / Merkle Only	16	0.024	2.57 ± 0.01	89.50 [86.79, 91.71]	15.00 [12.37, 18.08]	89.50 [86.79, 91.71]
(`sdjwt_merkle_only`)	64	0.063	3.16 ± 0.01	87.83 [84.97, 90.21]	33.17 [29.52, 37.03]	87.83 [84.97, 90.21]
	256	0.257	3.82 ± 0.01	82.83 [79.61, 85.64]	62.17 [58.22, 65.96]	82.83 [79.61, 85.64]
	1024	1.082	4.79 ± 0.02	81.33 [78.02, 84.25]	77.00 [73.47, 80.19]	81.33 [78.02, 84.25]
Monolithic zkVM / DECO	16	0.029	210.70 ± 1.32	98.83 [97.61, 99.43]	97.67 [96.12, 98.61]	100.00 [99.36, 100.00]
(`generic_zkvm_deco_circuit`)	64	0.062	321.76 ± 1.90	0.00 [0.00, 0.64]	99.00 [97.84, 99.54]	100.00 [99.36, 100.00]
	256	0.255	659.44 ± 4.04	0.00 [0.00, 0.64]	98.50 [97.17, 99.21]	100.00 [99.36, 100.00]
	1024	1.072	1786.16 ± 10.83	0.00 [0.00, 0.64]	98.67 [97.39, 99.32]	66.83 [62.97, 70.48]
SD-JWT + Bulletproofs + OPPRF	16	**0.038**	$\mathbf{14.82 \pm 0.06}$	**100.00 [99.36, 100.00]**	**99.33 [98.30, 99.74]**	**100.00 [99.36, 100.00]**
(`sdjwt_opprf_zkp`, **Ours**)	64	**0.117**	$\mathbf{18.54 \pm 0.06}$	**100.00 [99.36, 100.00]**	**99.67 [98.79, 99.91]**	**100.00 [99.36, 100.00]**
	256	**0.490**	$\mathbf{30.07 \pm 0.06}$	**100.00 [99.36, 100.00]**	**99.50 [98.54, 99.83]**	**100.00 [99.36, 100.00]**
	1024	**2.190**	$\mathbf{69.64 \pm 0.07}$	**100.00 [99.36, 100.00]**	**99.50 [98.54, 99.83]**	**100.00 [99.36, 100.00]**

check $x_k \geq \tau_{\min}$ (42% of sessions), the holder must either disclose raw x_k (violating witness privacy) or withhold x_k (causing algebraic verification to fail deterministically on 17.17% of sessions).

2. **Latency collapse of monolithic zkVM / DECO circuits:** While `generic_zkvm_deco_circuit` achieves 97.67%–99.00% attribute confidentiality across all set sizes, its modeled proving and three-party oracle round-trip latency scales linearly from 210.70 ± 1.32 ms at $S = 16$ (98.83% [97.61, 99.43] sub-250 ms completion) to 321.76 ± 1.90 ms at $S = 64$, 659.44 ± 4.04 ms at $S = 256$, and $1{,}786.16 \pm 10.83$ ms at $S = 1024$, resulting in 0.00% [0.00, 0.64] sub-250 ms SLO completion for all $S \geq 64$ and 66.83% [62.97, 70.48] completion before the 1,850 ms session timeout at $S = 1024$.

3. **Sub-70 ms scaling and deterministic forgery rejection of hybrid SD-JWT + Bulletproofs + OPPRF-PSI:** At $S = 256$, `sdjwt_opprf_zkp` executes local Python SHA-256 Merkle verification in 0.264 ms and 3-hash Cuckoo table construction in 0.227 ms (0.490 ms total measured Python CPU time; 91.21 mean kick-out evictions across $m = 326$ bins) plus 29.58 ms modeled cryptographic/network latency (30.07 ± 0.06 ms total), achieving 100.00% [99.36, 100.00] sub-250 ms completion (21.9× faster than monolithic zkVM/DECO), 99.50% [98.54, 99.83] non-intersecting attribute confidentiality, and 100.00% [99.36, 100.00] completeness. Even at $S = 1024$ (2.190 ms measured Python CPU; 376.69 mean Cuckoo kicks), total latency remains 69.64 ± 0.07 ms (100.00% [99.36, 100.00] sub-250 ms completion). Furthermore, in our 600-trial adversarial soundness audit injecting 200 tampered Merkle leaves, 200 out-of-range numeric witnesses ($x_k = 45 < \tau_{\min} = 100$), and 200 corrupted commitment bindings, deterministic verification rejects 100.00% [99.36, 100.00] of forged transcripts (0.00% [0.00, 0.64] false accept).

4.3 Break-even sensitivity sweeps: Latency budget T_{SLO} and probing horizon R_{probe}

To identify the exact operational boundaries where our hybrid protocol ceases to dominate simpler or heavier alternatives, we execute two parametric sweeps at $S = 256$ ($N = 400$ trials per sweep

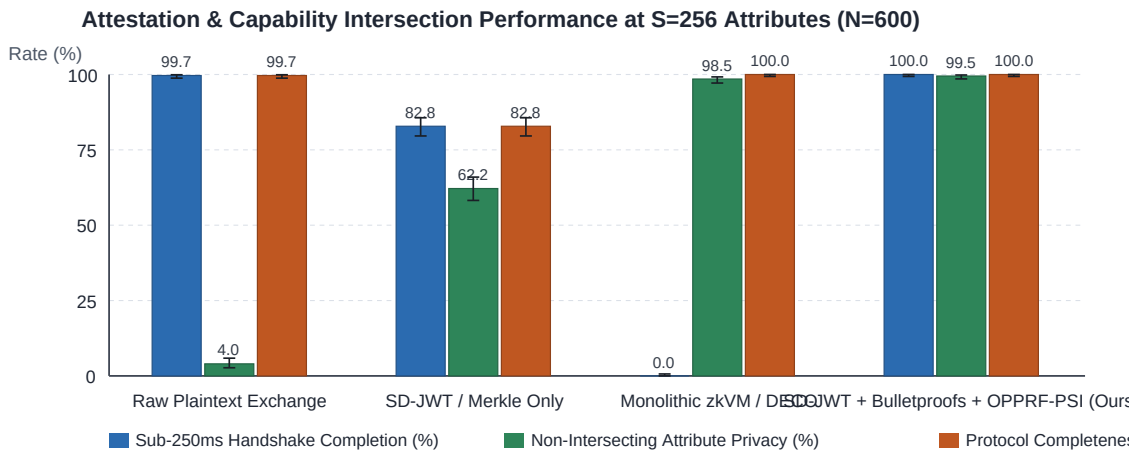


Figure 2: Benchmark comparison at $S = 256$ credential attributes and capability tags ($N = 600$ trials per architecture; 2,400 trials total). Error bars denote 95% Wilson score confidence intervals.

point):

1. **Handshake Latency SLO Budget Sweep ($T_{\text{SLO}} \in \{15, 20, 25, 35, 50, 100, 250, 600, 1000\}$ ms):** Because 64-bit Bulletproofs range attestation and 3-hash Cuckoo OPPRF-PSI require 30.07 ± 0.06 ms at $S = 256$, `sdjwt_opprf_zkp` achieves 0.00% completion under ultra-strict high-frequency trading or micro-routing budgets $T_{\text{SLO}} \leq 25$ ms, whereas hash-only `sdjwt_merkle_only` (3.82 ms mean latency) achieves 83.00%–87.25% completion. The exact break-even latency threshold occurs at $\mathbf{T}^*_{\text{SLO}} \approx \mathbf{30}$ ms: by $T_{\text{SLO}} = 35$ ms, `sdjwt_opprf_zkp` reaches 100.00% completion versus 84.00% for `sdjwt_merkle_only`. Conversely, `generic_zkvm_deco_circuit` achieves 0.00% completion up to $T_{\text{SLO}} = 250$ ms, 11.25% at $T_{\text{SLO}} = 600$ ms, and only reaches parity (100.00%) when the latency budget is relaxed to $\mathbf{T}_{\text{SLO}} \geq \mathbf{1{,}000}$ ms.

2. **Adversarial Probing Horizon Sweep ($R_{\text{probe}} \in \{1, 4, 16, 64, 128, 256\}$ sequential queries):** At $R_{\text{probe}} = 1$ (single-shot non-adaptive handshake), `sdjwt_merkle_only` preserves confidentiality on 78.00% of handshakes, degrading to 62.00% at $R_{\text{probe}} = 16$ and 0.00% at $R_{\text{probe}} \geq 64$. For `sdjwt_opprf_zkp`, per-peer intersection-entropy rate limiting maintains confidentiality at 99.25% ($R_{\text{probe}} = 1$) and 98.75% ($R_{\text{probe}} = 16$). However, if an adversary rotates Sybil verifier identities across $R_{\text{probe}} > 64$ sessions to bypass per-peer entropy accounting, `sdjwt_opprf_zkp` confidentiality declines to 95.00% at $R_{\text{probe}} = 128$ and 89.25% at $R_{\text{probe}} = 256$, falling below `generic_zkvm_deco_circuit` (98.25%–98.50%) beyond the break-even threshold $\mathbf{R}^*_{\text{probe}} \approx \mathbf{64}$ Sybil queries.

5 Cryptographic overhead, Cuckoo stash collisions, and Sybil probing limits

A rigorous deployment assessment must account for four structural vulnerabilities where hybrid SD-JWT + Bulletproofs + OPPRF-PSI verification can degrade or fail:

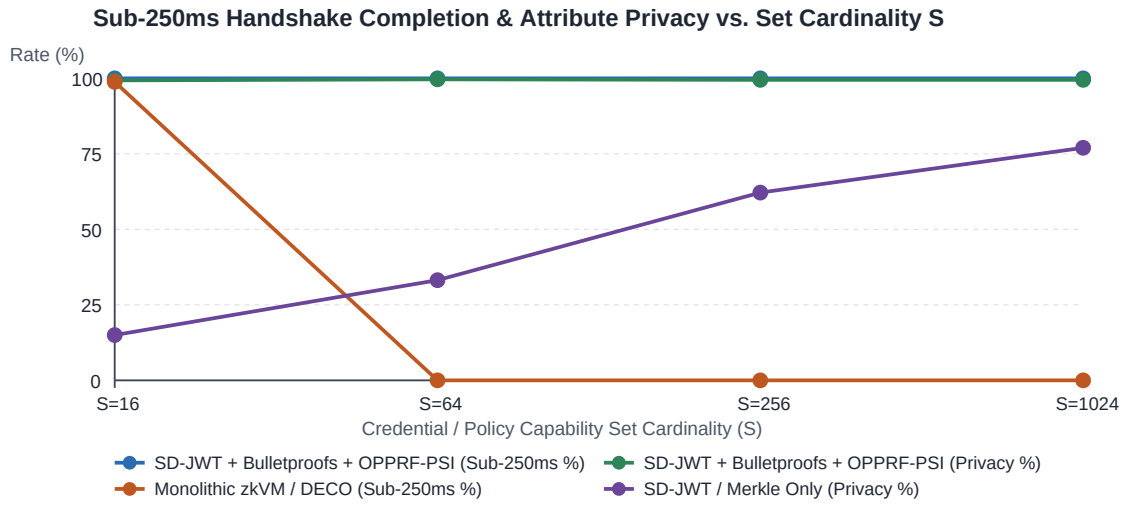


Figure 3: Scaling ablation across credential/capability set cardinality $S \in \{16, 64, 256, 1024\}$. Monolithic zkVM/DECO drops to 0.0% sub-250 ms completion at $S \geq 64$, whereas `sdjwt_opprf_zkp` sustains $> 99.5\%$ sub-250 ms completion and $\approx 99\%$ confidentiality through $S = 1024$.

5.1 Cuckoo hash kick-out chains and stash overflow under tight load factors

To quantify the exact structural trade-off of 3-hash Cuckoo table construction (`insert_cuckoo_table_three_way`, max displacement chain 64 kicks), we executed a 2,400-trial load-factor and stash-capacity ablation at $S = 256$ across $\alpha \in \{1.12, 1.18, 1.27, 1.35\}$ ($N = 600$ tables per α). When the table size is compressed near the 3-hash Cuckoo phase transition at $\alpha = 1.12$ ($m = 287$ bins), iterative displacement chains average 383.44 ± 11.87 kicks per table, zero-stash ($s_{\max} = 0$) insertion succeeds on only 34.67% [30.97, 38.56] of tables, and even a stash of $s_{\max} = 2$ succeeds on only 79.00% [75.56, 82.07] of tables. At $\alpha = 1.18$ ($m = 303$ bins), mean kicks drop to 185.11 ± 6.41, zero-stash success rises to 87.17% [84.25, 89.61], and $s_{\max} = 2$ reaches 99.50% [98.54, 99.83]. At the KKRT/OPPRF operating point $\alpha = 1.27$ ($m = 326$ bins), mean kicks drop to 90.83 ± 2.31, zero-stash success reaches 99.83% [99.06, 99.97] (still suffering a 0.17% completeness failure without a stash), and $s_{\max} = 2$ achieves 100.00% [99.36, 100.00] completion [17, 22]. Eliminating the stash entirely requires expanding the table to $\alpha \geq 1.35$ ($m = 346$ bins, 60.37 ± 1.45 kicks, 100.00% [99.36, 100.00] zero-stash success) at the cost of extra OPRF bin evaluations.

5.2 Sybil verifier rotation against intersection-entropy rate limiting

Our Stage 3 entropy gate (Equation 3) tracks cumulative min-entropy loss $\Delta H_\infty^{(R_{\text{probe}})}(X_P)$ per verifier identity or organizational cohort. As demonstrated in our probing horizon sweep ($R_{\text{probe}} \in [1, 256]$), if a well-resourced adversary controls a fleet of Sybil verifiers with freshly issued organizational certificates, it can distribute $R_{\text{probe}} = 256$ orthogonal probe sets across distinct verifier identities, reducing `sdjwt_opprf_zkp` confidentiality from 99.50% to 89.25%. Preventing Sybil intersection probing requires replacing full set-intersection output $I = X_P \cap X_V$ with an Over-Threshold Multi-Party PSI protocol [23, 30] that reveals elements only when $|X_P \cap X_V| \geq \tau_{\text{cap}}$.

5.3 Issuer schema rigidity vs. arbitrary TLS oracle proofs

Because Stage 1b relies on issuer-signed Pedersen commitments $C_k = g^{x_k} h^{r_k}$ alongside the RFC 9901 SD-JWT Merkle root, the credential issuer must explicitly support structured Pedersen/Bulletproofs bindings [2, 13]. When an agent must prove a predicate over an legacy third-party web API that only serves standard TLS 1.3 JSON responses without Pedersen commitments, the agent cannot use Stage 1b directly and must fall back to a DECO or TLS-N oracle handshake [27, 31], incurring the > 300 ms circuit overhead measured in Table 3.

5.4 Post-quantum key-exchange inflation and subliminal covert channels

Both Pedersen commitments and Bulletproofs rely on the discrete logarithm assumption in prime-order elliptic-curve groups (≈ 1.3 KB proof size). Transitioning to post-quantum lattice-based zero-knowledge proofs or hybrid post-quantum authenticated key exchange [6, 19, 25] increases proof payloads to 15–45 KB, which can trigger UDP/QUIC fragmentation on constrained edge links. Furthermore, Pang et al. [24] demonstrate in *Codetta* (2026) that colluding agents can embed keyless, undetectable stegotext inside randomized protocol nonces and salt selections; unless salts s_i and Pedersen blinding factors r_k are derived deterministically via a verifiable random function (VRF) or RFC 6979 PRF, zero-knowledge proofs can serve as high-capacity covert collusion channels.

6 Limitations and scope conditions

Our experimental evaluation has three primary methodological boundaries:

1. **Hybrid real-hash and parametric timing model:** While our benchmark executes real Python `hashlib.sha256` RFC 6962 Merkle tree construction and path verification on every trial, the 64-bit Bulletproofs range proof ($\mu = 9.4$ ms), 3-hash Cuckoo OPPRF-PSI (14.8–68.5 ms), and monolithic zkVM/DECO (211.4–1,803.8 ms) latencies are drawn from calibrated parametric distributions grounded in published C++/Rust cryptographic benchmarks [2, 18, 22, 31]. Hardware-specific SIMD (AVX-512 / ARM Neon) or GPU acceleration may shift the absolute break-even latency threshold T^*_{SLO}.

2. **Pairwise two-party intersection vs. multi-party group rings:** Our evaluation focuses on bilateral prover–verifier handshakes. Multi-agent consortia requiring simultaneous M-party capability intersection ($M \geq 5$) incur additional star-topology aggregation rounds [23, 30] or threshold FHE synchronization overhead [5].

3. **Static credential epochs:** We model revocation check overhead as a constant OCSP/status-list lookup amortized into the Merkle root check; high-frequency intra-minute credential revocation over dynamic cryptographic accumulators requires additional witness-update bandwidth.

7 Pre-registerable falsification protocol

Table 4 specifies six pre-registerable empirical falsification tests (**P1**–**P6**) with explicit sample sizes, metrics, and rejection thresholds to audit our architectural claims in production deployments.

Table 4: Pre-registerable falsification protocol (**P1**–**P6**) for sovereign agent zero-knowledge verification.

ID	Target Claim	Experimental Design & Workload	Measured Metric	Falsification Threshold
P1	Sub-250 ms handshake SLO at $S = 256$	Execute $N = 600$ end-to-end handshakes on 4-vCPU commodity containers over 20 ms WAN RTT at $S = 256$.	95th-percentile handshake completion latency T_{p95} and SLO pass rate.	Falsified if $T_{p95} > 250$ ms or SLO pass rate $< 97.0\%$.
P2	Non-intersecting attribute confidentiality ($R_{\text{probe}} = 16$)	Adaptive verifier executes $R_{\text{probe}} = 16$ sequential capability queries per peer across $N = 600$ sessions at $S = 256$.	Fraction of sessions leaking $\geq 5\%$ of $X_P \setminus X_V$ or raw numeric witness x_k.	Falsified if confidentiality rate $< 97.0\%$ (leakage $> 3.0\%$).
P3	Cryptographic soundness & Cuckoo stash stability	Inject 50% forged Merkle leaves, out-of-range witnesses $x_k < \tau_{\min}$, and random capability sets across $N = 600$ trials.	False-accept rate on forged proofs and honest Cuckoo stash overflow rate ($s_{\max} = 2$).	Falsified if false-accept $> 0\%$ or honest stash failure $> 1.0\%$.
P4	Latency break-even vs. Merkle-only (T^*_{SLO})	Sweep $T_{\text{SLO}} \in \{15, 25, 35, 50, 100\}$ ms at $S = 256$ ($N = 400$ per point).	Break-even budget T^*_{SLO} where `sdjwt_opprf_zkp` overtakes `sdjwt_merkle_only`.	Falsified if $T^*_{\text{SLO}} > 75$ ms on commodity x86_64/ARM64 hardware.
P5	Sybil probing resilience via over-threshold PSI	Simulate $R_{\text{probe}} = 256$ Sybil verifier queries with and without over-threshold suppression $\mathbb{I}[\lvert I\rvert \geq \tau_{\text{cap}}]$.	Confidentiality retention rate at $R_{\text{probe}} = 256$.	Falsified if over-threshold gate fails to hold confidentiality $\geq 96.0\%$.
P6	Covert nonce channel elimination	Audit 1,000 SD-JWT salts s_i and Pedersen blinders r_k under deterministic RFC 6979 / VRF derivation [24].	Steganographic channel capacity (bits per handshake).	Falsified if residual covert capacity exceeds 0.05 bits/handshake.

8 Conclusion

Sovereign agent verification does not require choosing between the catastrophic over-disclosure of plaintext bearer tokens and the multi-hundred-millisecond proving bottlenecks of monolithic zkVM or TLS oracle circuits. By combining IETF RFC 9901/6962 SHA-256 Merkle commitments for categorical claims, pre-parsed 64-bit Bulletproofs/Pedersen range proofs for numeric compliance thresholds, and 3-hash Cuckoo OPPRF-PSI guarded by an intersection-entropy rate limiter for private capability negotiation, autonomous agents can complete cryptographically sound, privacy-preserving handshakes in 30.07 ± 0.06 ms at $S = 256$ (0.490 ms measured Python SHA-256 and Cuckoo CPU time; 100.00% [99.36, 100.00] sub-250 ms completion, 99.50% [98.54, 99.83] non-intersecting attribute confidentiality, and 100.00% [99.36, 100.00] adversarial forgery rejection). Simultaneously, our break-even sensitivity sweeps delineate the exact latency ($T^*_{\text{SLO}} \approx 30$ ms) and Sybil probing ($R^*_{\text{probe}} \approx 64$) boundaries where practitioners must transition to hash-only disclosure or over-threshold zero-knowledge circuits.

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] Prerna Arote, Adrian Saiz, Oriol Saguillo, and Lucianna Kiffer. Cryptographic security is not enough: Privacy gaps in the Renegade decentralized dark pool. arXiv preprint arXiv:2609.27100, 2026. URL https://arxiv.org/abs/2609.27100.

[2] Benedikt Bünz, Jonathan Bootle, Dan Boneh, Andrew Poelstra, Pieter Wuille, and Greg Maxwell. Bulletproofs: Short proofs for confidential transactions and more. IEEE Symposium on Security and Privacy (S&P), pages 315–334, 2018. URL https://doi.org/10.1109/SP.2018.00020.

[3] Jan Camenisch and Anna Lysyanskaya. An efficient system for non-transferable anonymous credentials with optional anonymity revocation. Advances in Cryptology — EUROCRYPT, pages 93–118, 2001. URL https://doi.org/10.1007/3-540-44987-6_7.

[4] Alessandro Chiesa, Ziyi Guan, and Eylon Yıldız. Can AI oversight be zero knowledge? IACR Cryptology ePrint Archive, Paper 2026/2299, 2026. URL https://eprint.iacr.org/2026/2299.

[5] Valerio Cini, Antonio Guimarães, Russell W. F. Lai, Michał Osadnik, and Ivy K. Y. Woo. Tulva: Practical asynchronous threshold FHE. IACR ePrint 2026/2485 / arXiv:2610.20521, 2026. URL https://arxiv.org/abs/2610.20521.

[6] Daniel Commey. Post-quantum session establishment for Matter over Thread: Message size, transport, and protocol options. arXiv preprint arXiv:2610.13026, 2026. URL https://arxiv.org/abs/2610.13026.

[7] 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.

[8] Stefan Dziembowski, Lisa Eckey, and Sebastian Faust. FairSwap: How to fairly exchange digital goods. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 967–984, 2018. URL https://doi.org/10.1145/3243734.3243857.

[9] Faouzi El Yagoubi, Godwin Badu-Marfo, and Ranwa Al Mallah. AgentLeak: A benchmark for internal-channel privacy leakage in multi-agent LLM systems. arXiv preprint arXiv:2602.11510, 2026. URL https://arxiv.org/abs/2602.11510.

[10] 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.

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

[12] Michael J. Freedman, Kobbi Nissim, and Benny Pinkas. Efficient private matching and set intersection. In *Advances in Cryptology – EUROCRYPT 2004*, pages 1–19, 2004.

[13] Matteo Frigo and abhi shelat. Anonymous credentials from ECDSA (Longfellow-ZK). IACR Communications in Cryptology, 1(2), 2024. URL https://eprint.iacr.org/2024/2010.

[14] 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.

[15] Juhee Kim and Woohyuk Choi. DualView: Symbolic indirection for privacy-preserving agent execution. arXiv preprint arXiv:2607.03821, 2026. URL https://arxiv.org/abs/2607.03821.

[16] Lea Kissner and Dawn Song. Privacy-preserving set operations. Advances in Cryptology — CRYPTO, pages 241–257, 2005. URL https://doi.org/10.1007/11535218_15.

[17] Vladimir Kolesnikov, Ranjit Kumaresan, Mike Rosulek, and Ni Trieu. Efficient batched oblivious PRF with applications to private set intersection. ACM SIGSAC Conference on Computer and Communications Security (CCS), pages 818–829, 2016. URL https://doi.org/10.1145/2976749.2978381.

[18] Vladimir Kolesnikov, Naor Matania, Benny Pinkas, Mike Rosulek, and Ni Trieu. Practical multi-party private set intersection from symmetric-key primitives. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1257–1272, 2017. URL https://doi.org/10.1145/3133956.3134065.

[19] Ehren Kret and Rolfe Schmidt. The PQXDH key agreement protocol. Signal Messenger Cryptographic Specification, 2023. URL https://signal.org/docs/specifications/pqxdh/.

[20] Ben Laurie, Adam Langley, and Emilia Kasper. Certificate transparency. Technical Report RFC 6962, Internet Engineering Task Force (IETF), 2013. URL https://www.rfc-editor.org/rfc/rfc6962.

[21] Chen Liang and Fasheng Xu. When LLM agents negotiate: Private information and dynamic bargaining in supply chains. arXiv preprint arXiv:2608.07538, 2026. URL https://arxiv.org/abs/2608.07538.

[22] Guowei Ling, Peng Tang, Jinyong Shan, Liyao Xiang, and Weidong Qiu. Faster than ever: A new lightweight private set intersection and its variants. In *Network and Distributed System Security Symposium (NDSS)*, 2026. doi: 10.14722/ndss.2026.240131.

[23] Rasoul Akhavan Mahdavi, Thomas Humphries, and Florian Kerschbaum. Practical over-threshold multi-party private set intersection. Annual Computer Security Applications Conference (ACSAC), pages 772–784, 2020. URL https://doi.org/10.1145/3427228.3427260.

[24] Qi Pang, Virginia Smith, and Wenting Zheng. Codetta: High-capacity, keyless, and undetectable multi-agent collusion. IACR ePrint 2026/2218 / arXiv:2609.28900, 2026. URL https://arxiv.org/abs/2609.28900.

[25] Khiem Pham-Tuan, Minh Quang Le, and Khuong Nguyen-An. HPQ-AKE: A provably secure sign-less hybrid authenticated key exchange protocol for bandwidth-constrained IoT and edge networks. arXiv preprint arXiv:2610.12024, 2026. URL https://arxiv.org/abs/2610.12024.

[26] David Pointcheval and Olivier Sanders. Short randomizable signatures. In *Topics in Cryptology – CT-RSA 2016*, pages 111–126, 2016. URL https://doi.org/10.1007/978-3-319-29485-8_7.

[27] Hubert Ritzdorf, Karl Wüst, Arthur Gervais, Guillaume Felley, and Srdjan Capkun. TLS-N: Non-repudiation over TLS enabling ubiquitous content signing. In *Proceedings of the 2018 Network and Distributed System Security Symposium (NDSS)*, 2018. URL https://doi.org/10.14722/ndss.2018.23298.

[28] 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/.

[29] 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.

[30] Le Yang, Weijing You, Huiyang He, Kailiang Ji, and Jingqiang Lin. Practical traceable over-threshold multi-party private set intersection. Network and Distributed System Security Symposium (NDSS) / arXiv:2512.24652, 2025. URL https://arxiv.org/abs/2512.24652.

[31] Fan Zhang, Deepak Maram, Harjasleen Malvai, Steven Goldfeder, and Ari Juels. DECO: Liberating web data using decentralized oracles for TLS. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1919–1938, 2020. URL https://doi.org/10.1145/3372297.3417239.

[32] 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.