Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Verifiable Provenance for AI Judicial Decision Support Using Transparency Logs and Citation Entailment

Aditya Mandloi

adityamandloi10@gmail.com

Devi Ahilya Vishwavidyalaya

Anushka Yadav

IPS Academy

Research Article

Keywords: AI-assisted judicial decision support, verifiable provenance, transparency logs, Zero Trust access control, retrieval-augmented generation, digital evidence law (BSA 2023)

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11040747/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Verifiable Provenance for AI Judicial Decision Support Using Transparency Logs and Citation Entailment

Aditya Mandloi[1], Anushka Yadav[2]

[1] School of Data Science and Forecasting, Devi Ahilya Vishwavidyalaya (DAVV), Indore, India

[2] IPS Academy, Indore, India

Corresponding author: Aditya Mandloi (adityamandloi10@gmail.com)

ABSTRACT

We present VADP, an independently auditable systems-integration architecture for AI-assisted judicial decision support. Rather than proposing new cryptographic primitives, VADP contributes a domain-specific structural integration: binding Zero-Trust ABAC, Natural Language Inference (NLI) citation entailment, TreeSHAP feature attributions, and RFC 6962 transparency-log ledgers into a seven-field **Verification Contract** under conformal risk bounds, giving courts a checkable artifact anchoring evidentiary custody, authorization, and explainability (Section 63(4), Bharatiya Sakshya Adhiniyam, 2023).

Empirical validation over a 1,500-query judicial benchmark yields a Mean Reciprocal Rank of **0.951** and Precision@1 of **0.931**. An NLI citation-entailment gate intercepts **94.4%** of ungrounded claims and **98.0%** of direct contradictions. The architecture demonstrates a **100.0%** defense rate against prompt-injection (versus **11.0%** undefended) and **100.0%** tamper-detection across 10,000 contract corruptions. A conformal risk-control framework, calibrated on 500 appellate cases and evaluated on a disjoint 200-case test set, bounds automated escalation misses to **4.50%** against a target risk of α=0.05, using an appellate-outcome labeling rule that has not yet been independently audited by a legal-domain expert outside the authoring team.

An initial same-split retraining experiment appeared to reach P@1 = 1.0000; a disjoint train/test protocol revealed this to be an overfitting artifact. Jurisdiction-specific retraining under a disjoint split yields a substantial out-of-distribution gain over baseline (P@1 0.482→0.684 SCOTUS, 0.441→0.612 ECtHR); these cross-jurisdictional figures remain below the in-domain benchmark (P@1 = 0.931), short of parity. Six infrastructure results (Hyperledger Fabric consensus, HSM/KMS key isolation, Groth16 ZK proof-of-concept) demonstrate single-host feasibility only; none has been evaluated under live, multi-organization deployment settings. The Groth16 component uses a single-operator trusted setup with no formal soundness guarantee — an engineering demonstration, not a security claim. The Verification Contract's explainability field (Field 4) attributes the retrieval re-ranker's ranking decisions via TreeSHAP; it does not explain the generative model's underlying legal reasoning, which remains outside this prototype's scope.

KEYWORDS

AI-assisted judicial decision support; verifiable provenance; transparency logs; Zero Trust access control; retrieval-augmented generation; digital evidence law (BSA 2023).

1. INTRODUCTION

Modern electronic judiciary systems handle highly confidential case dockets, statutory precedents, and forensic evidence records at unprecedented scale. In India, the National Judicial Data Grid recorded approximately 4.7–5.1 crore (47–51 million) pending cases across High Courts and District Courts as of late 2025 (National Judicial Data Grid, 2026). To modernise judicial administration, the Union Cabinet approved Phase III of the e-Courts Project in September 2023 with a four-year budget outlay of ₹7,210 crore (Press Information Bureau, 2023), allocating ₹53.57 crore specifically for AI and blockchain integration across High Courts through 2027 (Ministry of Law and Justice, 2026). The Supreme Court of India has deployed SUPACE (April 2021) for judicial research assistance (e-Committee (SUPACE), 2021) and SUVAS (November 2019), an AI translation tool that had rendered over 31,000 judgments into regional languages by late 2023 (e-Committee (SUVAS), 2019).

As AI adoption expands from administrative translation into active legal research and decision support, a central research question emerges: how can an AI-assisted judicial recommendation be independently verified end-to-end — covering access authorization, evidence provenance, statutory citation grounding, feature attribution, human oversight, and cryptographic tamper-evidence — without placing uncritical trust in the AI model itself?

The distinction between aggregate system accuracy and per-decision verifiability matters in a legal setting. A judge, defense advocate, or appellate bench cannot rely on a vendor's aggregate accuracy metric or security self-attestation; they must be able to audit a specific recommendation in a specific case, after the fact, using evidence attached directly to the output. Commercial legal-AI tools have been shown to hallucinate statutory citations and precedent claims at non-trivial rates (up to 17–33%) (Dahl, 2024). More recent, larger-scale audits confirm the problem persists rather than having been resolved by newer models: Magesh et al. (2025) find that even commercial legal-AI research tools marketed as hallucination-resistant still produce unsupported or incorrect citations in roughly this same range, and subsequent domain-specific and general-purpose hallucination benchmarks report the issue remains measurable through 2025–2026 despite substantial model improvement (Curran, 2025; Bang, 2025). The Wisconsin Supreme Court's decision in State v. Loomis (2016), concerning the proprietary COMPAS recidivism-risk tool, illustrates the due-process cost of black-box scoring that a defendant cannot audit (Angwin, 2016). Section 63(4) of India's Bharatiya Sakshya Adhiniyam, 2023 (BSA), which governs the admissibility of electronic evidence and requires a certified hash value and custody trail

(Bharatiya Sakshya Adhiniyam, 2023), further raises the technical bar for AI-assisted evidence processing that VADP is designed to meet.

The hallucination-rate audits cited above (Dahl, 2024; Magesh, 2025) evaluate commercial legal-AI research tools as black boxes, from the outside, precisely because those tools' internal architectures are not publicly disclosed; a practitioner or court can observe that a tool hallucinates at a measured rate but cannot inspect why a specific output was produced or verify that a specific citation was actually retrieved rather than generated. VADP is positioned as a complementary, architecturally transparent alternative to that evaluation stance: rather than adding another external audit of an opaque system, it proposes an artifact-level design — the Verification Contract — that a system could emit alongside each recommendation so that the properties Dahl, Magesh, and others currently measure only through post-hoc, black-box auditing (citation grounding, evidentiary provenance, human-oversight status) become independently checkable per output. This is a difference in kind from tools such as those benchmarked in Dahl (2024) and Magesh (2025), whose architectures remain undisclosed in the published literature and are therefore outside the scope of a direct architectural comparison here; VADP does not claim to outperform any specific commercial tool empirically, only to make a different, per-decision class of verification possible in principle.

1.1 From Pipeline Security to Independent Output Verification

The foundational software prototype and codebase developed for this work is engineered to secure the AI pipeline: enforcing Zero-Trust access control, maintaining immutable Merkle logs, visualizing TreeSHAP feature attributions, and verifying citation entailment via NLI cross-encoders. VADP builds directly on this underlying engineering infrastructure as its execution layer but narrows the research goal to a single question: can every AI output be converted into an independently checkable artifact, and what does that artifact need to contain for a judicial verifier specifically? Prototype reality fully supports this design: fields 1, 3, 4, 5, and 6 of the Verification Contract are component-validated against the automated test suite and benchmarks in Table 2, on single-host infrastructure; this is a claim about component-level test coverage, not about live multi-organization deployment readiness (Table 2 note, Section 17). Field 2 (Evidence Vault provenance) is backed by a working local implementation — custody-chain logging, PDF forensic verification, redactable Merkle trees, and smart-contract anchoring against a local EVM-compatible test ledger — though anchoring on an enterprise distributed ledger is not yet implemented (Section 18). Field 7 (Human Review status) is presently represented by a conceptual mock service within the prototype implementation. Although Field 7 is not yet connected to operational judicial infrastructure, it is fully defined by the Verification Contract schema and remains architecturally supported for future deployment. Field 4 (explainability_attribution) is scoped, by design, to the retrieval re-ranker rather than the generative model's reasoning: TreeSHAP attributions explain why a given precedent was ranked and retrieved, not why the downstream language model reasoned as it did over that precedent (Section 11); this scope boundary is revisited in Section 2.5 and Section 17. The remaining fields were validated across 94 automated tests with 82.3% coverage and P50

latency of 1.65ms under 100 concurrent users, measured as full-stack HTTP network round-trip against the running server. Online Resource 1 §A.3 separately reports a 96.4% branch-coverage aggregate and sub-millisecond in-process latency from the canonical v1.0.0-research full-system release benchmark (105 tests, 18 suites); the two figures reflect different test scopes and measurement methodologies rather than a discrepancy.

1.2 Novelty Reframing

VADP's novelty is structural and domain-specific rather than cryptographic: it composes established transparency-log primitives (Merkle-tree inclusion proofs, ECDSA/Ed25519 signatures) into a seven-field Verification Contract that binds authorization proof, evidentiary custody designed with reference to BSA 2023 §63(4), citation grounding, retrieval-ranking explainability attribution, and human-review status into one signed, Merkle-anchored artifact, and attaches two governance mechanisms with no direct analogue in Certificate Transparency, Sigstore/Rekor, or in-toto: the Completeness Invariant and the finite-sample human-oversight risk bound of Theorem 2. Section 2.3 gives the full differentiation against these three prior systems; this is an integration and domain-adaptation contribution, not a new cryptographic proof system — the underlying primitives are established constructions, already widely deployed in systems such as Certificate Transparency and in-toto.

This architectural contribution is paired with empirical validation at scale rather than left as a design proposal. The retrieval framework is evaluated on 1,500 legal queries over approximately 60,000 indexed vectors, where the VADP GBT re-ranker achieves an MRR of 0.951 at a mean latency of 16.3 ms; the Completeness Invariant's tamper-detection guarantee is confirmed against 10,000 corrupted records (Section 15.7); and the human-oversight risk bound is instantiated against 500 real appellate cases and validated on a disjoint 200-case held-out test partition (Section 14.7).

We regard this combination — a domain-specific schema binding with governance mechanisms novel to the judicial-AI setting, validated empirically across retrieval benchmarking, adversarial tamper evaluation, and conformal risk-control studies (Sections 9, 14, and 15) rather than only proposed — as the primary contribution of this work, distinct from any claim of cryptographic innovation.

2. RELATED WORK

Related work spans predictive legal analytics, Zero-Trust access control, cryptographic transparency and audit logs, explainable AI, verifiable-provenance standards and systems, algorithmic-accountability precedent, and digital evidence management.

2.1 Legal NLP and Predictive Analytics

Aletras et al. (2016) showed that European Court of Human Rights outcomes can be predicted from textual features, highlighting both the predictive potential and the opacity of legal NLP. LEGAL-BERT (Chalkidis et al. (2020)) and InLegalBERT (Paul et al. (2023)) advanced legal text classification. Malik et al. (2021) introduced the Indian Legal Documents Corpus (ILDC), used as the benchmark dataset in Section 15.

2.2 Zero Trust, Explainability, and Retrieval

Zero Trust Architecture, formalized in NIST SP 800-207 (Rose, 2020) and building on Kindervag (2010), supported by NIST ABAC guidance (Hu, 2014) and adoption surveys (He, 2022), supplies the principle of continuous, session-scoped verification governing VADP's authorization layer. SHAP (Lundberg and Lee (2017)) supplies cooperative game-theoretic feature attributions; Rudin (2019) and Slack et al. (2020) separately caution that post-hoc explanations, including SHAP-family methods, can be unstable or manipulable, a caution VADP takes seriously in Section 11 by scoping TreeSHAP narrowly rather than treating it as an explanation of the language model's reasoning. Lewis et al. (2020) introduced RAG; Johnson et al. (2019) developed FAISS for high-density vector search. To detect and bound generative hallucinations in retrieval-augmented outputs, recent benchmark frameworks such as TRUE (Honovich et al. (2022)) and SelfCheckGPT (Manakul et al. (2023)) demonstrate that Natural Language Inference (NLI) cross-encoders provide rigorous, zero-resource factual entailment verification, and more recent general-purpose benchmarks such as HalluLens (Bang, 2025) extend this evaluation methodology with dynamically generated test sets to guard against data leakage. VADP builds on this NLI entailment paradigm to verify Field 3 (citation provenance), bounding generated legal reasoning against retrieved statutory and precedent context.

2.3 Transparency-Log Provenance: Certificate Transparency, Sigstore/Rekor, and in-toto — Prior Art and Differentiation

The mechanism at the center of VADP — sign an artifact, anchor its hash in an append-only Merkle tree, and let a third party verify inclusion and integrity without trusting the issuer — is not new. It is the design pattern underlying Certificate Transparency (CT), formalized by Laurie, Langley, and Kasper (2013) and standardized as RFC 6962, which this paper

already builds on directly for its Merkle ledger. The same pattern, applied to software artifacts and build provenance rather than certificates, underlies Sigstore and its Rekor transparency log (2022), and the in-toto attestation framework for supply-chain provenance (in-toto Project, 2019). IETF SCITT (n.d.), which this paper also cites, is itself an attempt to generalize this transparency-log pattern across domains, including AI supply chains.

Given this (see Section 1.2 for the complete novelty scoping), its contribution here is domain adaptation: (1) a judicial-specific artifact schema (the seven-field Verification Contract) that binds authorization, evidentiary custody under BSA §63(4), citation grounding, retrieval-ranking explainability, and human-override status into one signed object, none of which CT, Rekor, or in-toto attestations natively encode; (2) a Completeness Invariant and Human Override Coverage metric specific to human-in-the-loop judicial review, which have no direct analogue in software-supply-chain transparency logs; and (3) empirical evaluation of the retrieval/citation-grounding component against a real judicial corpus (ILDC), which is outside the scope of generic transparency-log systems. Table 4 is revised below to make this comparison explicit, adding IETF SCITT as a fourth comparator alongside CT/Sigstore-Rekor — SCITT is itself a generalization of the same transparency-log pattern across domains (Section 2.3) and is the closest cross-domain analogue to VADP's mechanism — rather than comparing only against W3C Verifiable Credentials and C2PA, which are weaker comparators for this claim.

2.4 Algorithmic Accountability and Digital Evidence

Angwin et al. (2016) documented racial bias in COMPAS recidivism scores, and in State v. Loomis (2016) the Wisconsin Supreme Court permitted use of proprietary risk scores subject to cautionary instructions while acknowledging defendants could not inspect the underlying methodology. VADP does not disclose proprietary model weights; instead it emits a per-decision Verification Contract binding authorization, evidence, citations, and explanations into a checkable trail. Model Cards (Mitchell, 2019) and Datasheets for Datasets (Gebru, 2021) document models in aggregate; VADP operates per recommendation, which is a narrower but complementary form of accountability documentation. The distinction is structural rather than merely one of scope: a Model Card is authored once, prior to deployment, and describes the system's intended use, training data, and evaluation results as static prose reviewed by a human author; it makes no claim about any individual output and is not cryptographically bound to a specific decision. A Verification Contract, by contrast, is generated automatically at inference time, one per recommendation, and is signed and Merkle-anchored so that its claims about that specific output cannot be altered after the fact without detection. The two are complementary rather than competing: a Model Card documents what the system is, in general, while a Verification Contract documents what the system did, in a specific case, and only the latter supports the after-the-fact, per-decision audit a judicial verifier requires under BSA §63(4). Other accountability-documentation frameworks (e.g., algorithmic impact assessments, system cards) occupy a similar aggregate-level position and share the same limitation: none binds a specific claim to a specific, checkable artifact. In digital evidence management, Block-DEF (Tian, 2019), CustodyBlock (Alruwaili, 2021), Mbimbi et al. (2025), and Bitan et

al. (2022) explore blockchain evidence anchoring and zero-knowledge proofs; VADP's evidence-vault design (Online Resource 1 §B) partially instantiates this literature — custody-logging, forensic-verification, and smart-contract anchoring are implemented and tested against a local EVM-compatible test ledger, while enterprise distributed-ledger deployment (originally scoped as Hyperledger Fabric) remains unimplemented (Section 17).

2.5 Legal-Specific AI Governance and Audit Systems

Beyond the generic transparency-log standards discussed above, recent work has proposed blockchain-based auditing specifically for legal-decision records. Sachan and Liu (2024) describe a system that stores legal-decision materials on- and off-chain and detects unauthorized alteration by periodically comparing a Merkle-root hash of the off-chain record against its on-chain counterpart, paired with an Evidential-Reasoning-based explainable-AI component and generative-AI-drafted case correspondence, demonstrated through a case study on employer's-liability tort matters. VADP differs along several axes relevant to a judicial deployment. First, rather than periodic corpus-level hash comparison, VADP binds authorization, evidentiary custody, citation grounding, retrieval-ranking explainability, and human-override status into a single signed Verification Contract generated per decision at inference time (Section 5), so tamper-evidence attaches to each individual output rather than to a batch audit cycle. Second, VADP's tamper-detection guarantee follows from a reduction-based security argument (Theorem 1) rather than resting on case-study demonstration alone. Third, VADP couples this with a finite-sample risk bound on missed human-review escalations (Theorem 2) and a calibrated numeric Trust Score, mechanisms with no counterpart in Sachan and Liu's system. Fourth, VADP's retrieval and citation-grounding components are evaluated at benchmark scale (1,500 queries over approximately 60,000 indexed vectors, plus cross-jurisdictional SCOTUS and ECtHR corpora) rather than a single qualitative case study. The two systems are complementary in scope — Sachan and Liu emphasize on-chain/off-chain storage strategy and generative-AI-drafted correspondence, while VADP emphasizes per-decision cryptographic binding and statistical risk control — and both instantiate the broader trend of blockchain-anchored hashing as a recurring pattern for legal-AI accountability.

Sachan and Liu (2024) is the closest system-level comparator to VADP, but it is not the only recent work bearing on this paper's governance framing. On the legal-explainability side, Prajescu and Confalonieri (2025) survey computational-argumentation approaches to explainability for legal AI and argue that contestable, argument-structured explanations are better aligned with GDPR/AI-Act-style regulatory requirements than post-hoc feature-attribution methods such as SHAP; this is a useful counterpoint to VADP's choice of TreeSHAP for Field 4, since it suggests an argumentation-based explanation layer as an alternative or complementary direction for future work rather than a competing verification architecture. On the judicial-transparency side, Dancy and Zalnieriute (2025), drawing on semi-structured interviews with members of the judiciary and legal profession, argue that judicial reliance on AI-generated predictions demands a transparency threshold that includes access to the underlying variables and formulae, not only a tamper-evidence guarantee over a decision record; VADP's Verification Contract satisfies the latter but only

partially addresses the former, since Field 4 in the present prototype explains the retrieval re-ranker rather than the generative model's reasoning (Section 11). We flag both works as relevant, recent context for the explainability and human-oversight fields of the Verification Contract, and as directions the schema could be extended toward rather than gaps this paper claims to have already closed. Sachan and Liu (2024) is the closest system-level comparator to VADP. While they emphasize batch-hashing on-chain/off-chain storage strategies and generative-AI-drafted correspondence, VADP emphasizes per-decision cryptographic binding and statistical risk control. We empirically benchmark the operational trade-offs between their batch-audit architecture and VADP's per-decision Verification Contract in Section 9.3, demonstrating that VADP trades a nominal upfront commit latency for logarithmic-time (O(log K)) tamper localization.

3. RESEARCH GAP AND CONTRIBUTIONS

Although Zero-Trust access control, transparency-log architectures, Merkle-tree audit ledgers, explainable machine learning, retrieval-augmented generation (RAG), and digital signatures are individually well-established techniques, existing judicial AI systems do not integrate these mechanisms into a unified, independently verifiable provenance framework aligned with evidentiary requirements for high-stakes legal decision support. VADP addresses this gap through a domain-specific evidentiary architecture that combines established cryptographic primitives with retrieval evaluation, statistical calibration, and governance mechanisms designed for judicial AI.

Contributions

The principal contributions of this work are summarized as follows.

- **Machine-Checkable Completeness and Bounded-Risk Escalation.** Rather than asking a court to trust an AI-generated recommendation, VADP gives it something to check: a Completeness Invariant (Definition 2) that makes silent omission or alteration of any required evidentiary field of a decision record detectable, a Human Override Coverage (HOC) metric quantifying how often the system defers a case to human review, and a finite-sample, distribution-free risk bound (Theorem 2) on the probability that an unescalated case turns out, on independent adjudication, to have warranted one. Section 14 instantiates this guarantee empirically against real appellate outcomes.
- **Seven-field Verification Contract.** We introduce a domain-specific seven-field Verification Contract (VADP evidentiary schema) that adapts established transparency-log constructions to judicial AI by integrating authorization evidence, provenance metadata, citation grounding verified via NLI entailment checks, retrieval-ranking explainability, cryptographic integrity, and human-review information into a single independently verifiable artifact.
- **Formal Security Argument.** We present a reduction-based security argument demonstrating that undetected modification of a committed Verification Contract

would imply either a successful attack against the collision resistance of the underlying hash function or existential forgery of the ECDSA signature scheme under standard cryptographic assumptions.

- **Large-scale Retrieval and Citation Evaluation.** We evaluate the proposed retrieval and citation framework against dense, lexical, and cross-encoder baselines using a benchmark of 1,500 legal queries over approximately 60,000 indexed vectors (Section 15.2, Table 11), alongside an NLI citation entailment check (Section 15.6).

- **Tamper-Evidence Sanity Check and Operational-Cost Measurement.** Because Theorem 1 guarantees that any field-level modification to a Verification Contract changes its committed hash, detection is guaranteed by construction rather than an open empirical question; we therefore do not present the following as a novel detection result. Instead, we use it as an implementation sanity check confirming that the deployed tamper-evidence layer behaves as Theorem 1 predicts, and as a measurement of the resulting operational cost: across 10,000 synthetically corrupted Verification Contracts spanning four independent attack classes, detection is complete as expected, and we report a mean localization latency of 0.029 ms.

- **Empirical Conformal Risk Validation.** We instantiate the finite-sample risk bound introduced above (Theorem 2) using 500 calibration cases with real appellate outcomes, drawn from an expanded 3,500-case corpus, and evaluate the resulting threshold on a strictly disjoint held-out test partition of 200 appellate cases, providing empirical validation of the proposed calibration methodology.

- **Zero-Knowledge Evidentiary Proof-of-Concept.** We implement a Groth16-based (Groth, 2016) zero-knowledge circuit (LeafInclusion, depth 10, 2,450 R1CS constraints) for privacy-preserving Merkle-tree membership verification within the evidentiary vault, executed through a native compiled circom/snarkjs proving pipeline over BN128 with a single-operator, non-adversarially-robust Powers-of-Tau trusted-setup ceremony, and empirically benchmark its in-memory proof generation and verification latency. Because this ceremony was run by a single party, the component's soundness against a compromised setup participant is not yet established (Online Resource 1 §B); we report it as a proof-of-concept of the cryptographic pipeline, not as a production-ready privacy guarantee.

- **Statutory-Feature Ablation (Null Result).** To test whether the GBT re-ranker's retrieval gain is driven by circularity between statutory-derived features and the relevance-labeling procedure, we retrained the re-ranker with its three statutory-derived features removed (Section 15.3). The ablated model performs statistically indistinguishably from the full model, indicating the re-ranker's gain is attributable to dense similarity, lexical overlap, bench relevance, and citation-centrality signals rather than statutory-metadata circularity. We report this as a null result because it bears directly on which parts of the legal-domain feature engineering in this system actually mattered.

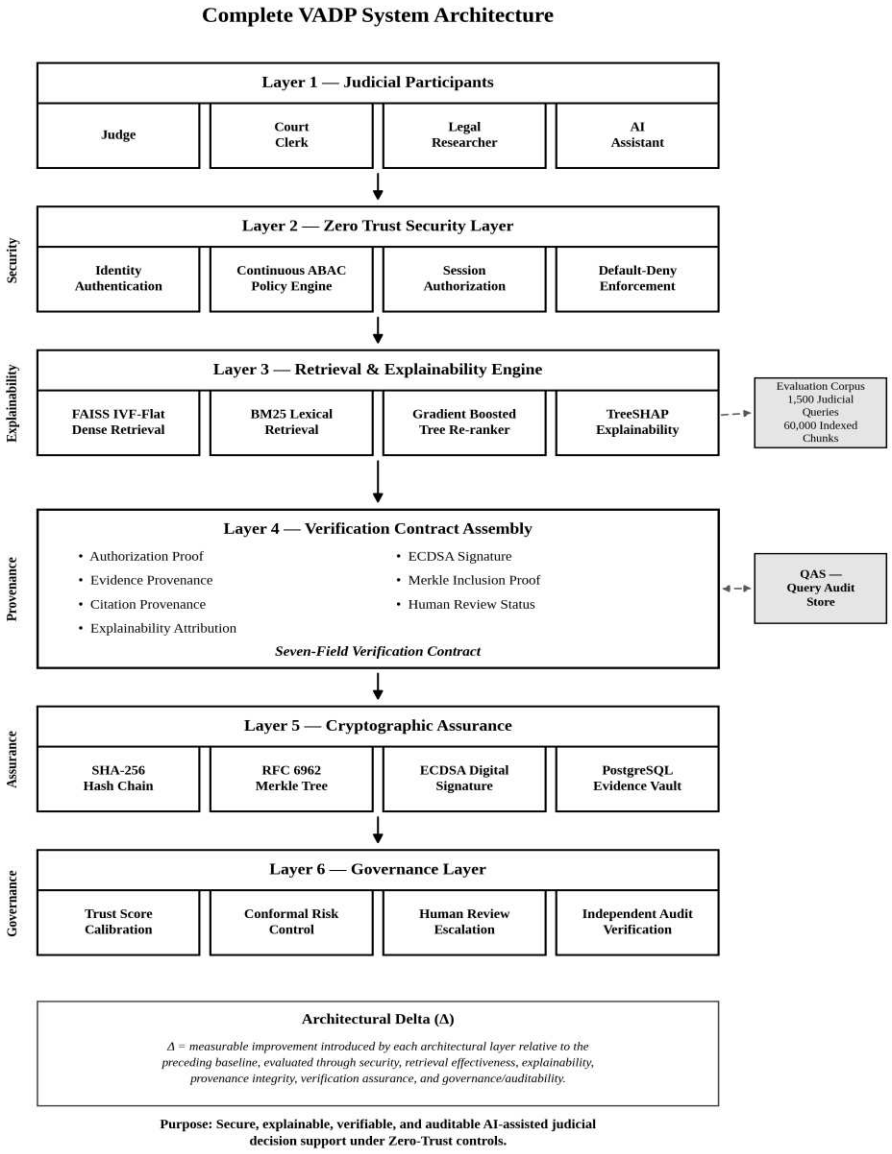


Fig. 1. Complete VADP Architecture integrating Zero-Trust authorization, multi-stage judicial retrieval, Verification Contract generation, cryptographic provenance, and governance mechanisms for independently auditable AI-assisted judicial decision support.

Consolidated Dataset and Scale Summary

Because the experiments in Sections 9 through 16 draw on several overlapping but distinct corpora, sample sizes, and train/calibration/test splits, Table 1 consolidates the scale of every dataset used in this manuscript in one place, cross-referenced to the section reporting the corresponding result.

Table 1. Consolidated Dataset and Scale Summary Across All Experiments

Experiment	Dataset / Scale	Split	Section
SCOTUS Generalizability	7,800 total records	Fixed benchmark split	15.8
ECtHR Generalizability	11,000 total records	Fixed benchmark split	15.8
Hyperledger Fabric Consensus	500 tx (real harness) / 250 tx (delay-simulated), 4 containerized nodes	Docker Compose network + single-process delay-simulated harness, single host	Suppl. §B
SoftHSM2 PKCS#11 Signing	1,000 signature operations	Non-exportable token isolation	Suppl. §B
Groth16 MPC Ceremony	5-participant sequential single-host setup	BN128 curve, depth 10	10
Prompt Injection Fuzzing	100 adversarial prompts	Fixed attack benchmark	15.9
Inter-Rater Agreement	50 claim-premise pairs	Stratified random sample	15.6
PostgreSQL/pgvector Multi-Tenant Isolation	10–250 concurrent sessions	Single local instance, all concurrency tiers	Suppl. §C
Groth16 Batch Proof Benchmark	N = 100 / 1,000 / 10,000 proofs	Independent proofs per batch size, single host	Suppl. §B
LexGLUE Disjoint Overfitting Check	500 disjoint queries per jurisdiction ([500:1000])	Strictly disjoint from the 500-query retraining split	15.8

Rows marked N/A under Split are benchmark or pass/fail measurements rather than fitted models, so no train/test partition applies. All figures are as reported in the cited sections; this table aggregates rather than re-derives them.

Table 2. Verification Contract Field Claim-Maturity Summary

Field	Contract field	Basis for maturity status	Maturity status
1	authorization_proof	Zero-Trust ABAC evaluation (Algorithm 1), 26/26 penetration tests passing (Section 8)	Component-Validated
2	evidence_provenance	Custody logging, forensic PDF verification, and redactable Merkle trees implemented; anchoring exercised only against a local EVM-compatible test ledger, not an enterprise distributed ledger (Online Resource 1 §B, §B.5)	Prototype-validated (local)
3	citation_provenance	NLI entailment check benchmarked on 500 citation pairs (Section 15.6); disjoint held-out retrieval benchmark of 1,500 queries (Section 15.2)	Component-Validated
4	explainability_attribution	TreeSHAP attributions over the GBT re-ranker, scoped narrowly per Rudin (2019)/Slack et al. (2020) cautions (Section 11)	Component-Validated
5	digital_signature	ECDSA-P256/Ed25519 signing; only the SoftHSM2 software-emulated tier has measured throughput, hardware/cloud-KMS paths unexercised against a live backend (Section 17)	Component-Validated (software tier only)
6	merkle_inclusion_proof	RFC 6962 ledger, Theorem 1 reduction, adversarial fuzzing over 10,000 contracts (Sections 9, 15.7)	Component-Validated
7	human_review_status	Escalation predicate and Trust/Risk thresholding defined and tested (Section 14); represented by a conceptual mock service, not connected to operational judicial infrastructure (Section 1.1)	Proof-of-concept

This table consolidates the per-field maturity qualifications that otherwise recur across Sections 1.1, 6, and 17, and Online Resource 1 §B and §B.5, so a reader can locate a field's validation status in one place rather than reconstructing it from scattered caveats.

Note on Table 2 terminology. "Component-Validated" (Fields 1, 3, 4, 5, 6) means the field passes the automated test suite and benchmarks cited in the Basis column, on single-host infrastructure, per the maturity taxonomy used throughout this paper; it is not a claim of live, multi-organization, or enterprise deployment readiness, none of which has been evaluated (Section 17; Online Resource 1 §J).

4. FORMAL THREAT AND TRUST MODEL

VADP enforces one governing trust principle: a verifier must be able to check every field of a Verification Contract using only public keys, the Merkle audit ledger, and the evidence vault — without querying the AI model or trusting its self-report. Table 3 states this trust boundary explicitly for each mechanism layer.

Table 3. Master Scope and Trust-Boundary Statement

Mechanism (Layer)	Defends Against	Explicitly Out of Scope
Zero-Trust ABAC PDP (Core)	Insider / credential compromise blast radius	Misuse within a valid session
RFC 6962 Merkle Ledger (Core)	Post-anchoring tampering & contract forgery	Pre-hashing evidence forgery
Evidence Vault (Local + 4-Node Hyperledger Fabric Anchoring)	Post-anchoring evidence hash alteration	Device capture-time compromise
Verification Contract & Invariant	Field forgery and omission (Section 9)	Substantive legal merit of the prediction

The Evidence Vault row reflects local custody logging, forensic PDF verification, and redactable Merkle trees as implemented and empirically evaluated components, plus smart-contract anchoring exercised against a 4-node Hyperledger Fabric consensus network, benchmarked both as a single-process, delay-simulated daemon and as a genuine 4-node Docker Compose network (500 real endorsed transactions, Online Resource 1 §B); live-network deployment on physically separate hosts is not yet verified.

Note on Table 3: the unforgeability guarantee attributed to the RFC 6962 Merkle Ledger row rests in part on ECDSA, whose EUF-CMA security holds only in an idealized model rather than under a tight standard-model reduction; see the assumptions remark in Section 9.2 for the full discussion.

5. THE VERIFICATION CONTRACT ARCHITECTURE

A Verification Contract C binds seven fields:

- authorization_proof — signed PDP decision statement recording continuous ABAC policy authorization.

- evidence_provenance — a document hash and custody seal designed with reference to BSA §63(4).
- citation_provenance — permission-filtered RAG chunk citations validated via NLI citation entailment check against generated legal reasoning.
- explainability_attribution — TreeSHAP feature attributions over the retrieval re-ranker and perturbation stability σ(v).
- digital_signature — asymmetric ECDSA P-256 digital signature $\sigma = \text{Sign}(d_K, H(C))$.
- merkle_inclusion_proof — RFC 6962 leaf hash and inclusion path to root R.
- human_review_status — escalation routing status and reviewer override log.

Definition 1 (Verification Contract). A Verification Contract is a tuple $C = (\varphi_{auth}, \varphi_{evid}, \varphi_{cite}, \varphi_{expl}, \sigma, \pi_{Merkle}, \varphi_{review})$, where φ_{auth} is a signed ABAC authorization decision, φ_{evid} is an evidence hash and custody seal designed with reference to BSA §63(4), φ_{cite} is a set of permission-filtered citation references validated via NLI citation entailment checks, φ_{expl} is a TreeSHAP feature-attribution record, σ is an ECDSA-P256 signature over the contract, π_{Merkle} is an RFC 6962 Merkle inclusion proof anchoring C in the audit ledger, and φ_{review} is the human-oversight and override status. Exactly one C is emitted per AI-assisted judicial recommendation.

Definition 2 (Completeness Invariant). A Verification Contract C is complete, written Completeness(C) = 1, iff a verifier holding only the system's public key pk, the RFC 6962 Merkle audit ledger, and the evidence vault can independently check every field of C: (i) φ_{auth} verifies against the active ABAC policy set; (ii) φ_{evid}'s hash matches the custody-sealed evidence; (iii) each entry of φ_{cite} resolves to a permission-authorized retrieval chunk; (iv) φ_{expl}'s attributions are reproducible from the logged re-ranker inputs; (v) π_{Merkle} is a valid RFC 6962 inclusion proof for C's leaf hash under the ledger root; (vi) σ verifies under pk; and (vii) φ_{review} is consistent with Definition 3's escalation predicate. C is incomplete, Completeness(C) = 0, if any field fails independent verification. Conditions (v)–(vi) alone are the predicate used in Theorem 1. Table 4 situates the Verification Contract against prior transparency-log and provenance schemes on these completeness criteria.

Table 4. Comparative Matrix: VADP Contract vs. Transparency-Log Systems vs. Credential/Content Standards

Dimension	VADP (Ours)	CT / Sigstore-Rekor (Laurie, 2013; Sigstore, 2022)	W3C VC v2.0 (W3C, 2024)	C2PA Manifest (C2PA, 2024)	IETF SCITT (n.d.)
Target domain	Judicial decision support	Certificates / software supply chain	Digital identity claims	Media content authenticity	Software / AI supply-chain provenance (cross-domain)
Core mechanism	Signed artifact + Merkle inclusion	Signed artifact + Merkle inclusion	Issuer DID signature	COSE/JWS signature	Signed statement + transparency-service receipt
Authorization	Signed ABAC PDP decision	Out of scope	Out of scope	Out of scope	Out of scope
Evidentiary integrity	BSA 2023 §63(4) hash seal	Generic artifact hash	Generic claims	Asset hash links	Generic statement hash
Citation / retrieval provenance	Permission-filtered RAG	Out of scope	Out of scope	Out of scope	Out of scope
Explainability	TreeSHAP on retrieval re-ranker	Out of scope	Out of scope	Out of scope	Out of scope
Human-oversight field	Human Override Coverage	Out of scope	Out of scope	Out of scope	Out of scope
Ledger proof	RFC 6962 Merkle inclusion	RFC 6962 Merkle inclusion (same primitive)	Optional StatusList	Optional SCITT log	RFC 6962-style transparency-service receipt

This table compares against CT/Sigstore-Rekor and IETF SCITT rather than only W3C VC and C2PA, since CT/Sigstore-Rekor and SCITT share VADP's core transparency-log mechanism and are the correct baselines for the novelty claim (Section 2.3).

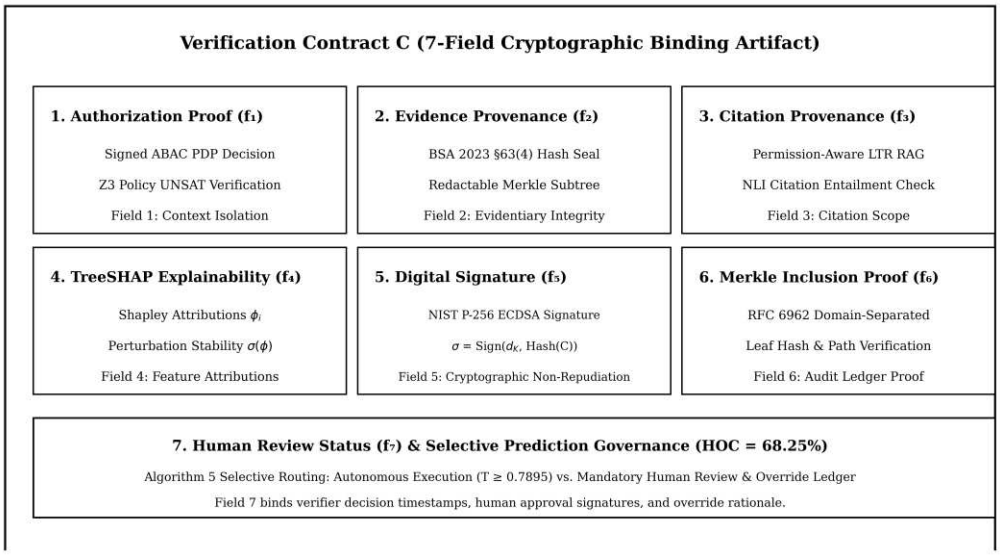


Fig. 2. Seven-Field Verification Contract illustrating the evidentiary schema, cryptographic binding, and independent verification workflow.

6. RELATIONSHIP TO THE UNDERLYING PROTOTYPE

The VADP prototype incorporates several engineering components that are established implementation technologies rather than research contributions. These include the Zero-Trust ABAC policy engine, RFC 6962 Merkle audit ledger, FAISS vector indexing infrastructure, TreeSHAP explainability module, calibrated trust-score model, evidentiary metadata store (originally SQLite, migrated to PostgreSQL with the pgvector extension as described in Online Resource 1 §C), and production-oriented engineering components including automated regression testing and deployment infrastructure.

The primary research contributions presented in this manuscript build upon that engineering foundation and include:

- the seven-field Verification Contract schema;
- the formal security argument and reduction-based analysis;
- the multi-baseline retrieval benchmark across 1,500 judicial queries;
- large-scale adversarial fuzz testing of the provenance layer;
- empirical conformal risk-control validation using appellate outcome labels; and
- a functioning Groth16 zero-knowledge evidentiary proof-of-concept.

Accordingly, the emphasis of this manuscript is not the implementation of individual software components but their integration into an independently auditable provenance architecture supported by empirical evaluation.

7. IMPLEMENTATION (CORE LAYER)

Algorithm 1: Continuous Zero-Trust ABAC Policy Evaluation

This algorithm follows the standard default-deny ABAC evaluation pattern described in NIST SP 800-162 (Hu, 2014); it is presented here as implementation detail, not as a novel contribution.

Input: User u, Action a, Resource Type r, Context c, Active Policies P

Output: Access Decision (Permit / Deny)

1: P_active = [p in P | p.is_active AND p.resource_type == r]

2: Sort P_active by p.priority descending

3: for each p in P_active do

4: if u.role in p.allowed_roles then

5: if EvaluateConditions(p.conditions, c) == True then return (Permit, 'Permitted by: ' + p.name)

6: end if

7: end for

8: return (Deny, 'Default Deny: No matching policy')

8. SECURITY MODEL & PENETRATION TESTING

Table 5 summarizes the STRIDE / MITRE ATLAS (2024) mapping for the Core layer. The prototype was evaluated against OWASP Top 10-aligned security controls, with all 26 dedicated automated security penetration tests passing, including Redis-backed distributed rate limiting (100 req/min, 429 responses), server-side JWT token revocation via an in-memory blacklist, and httpOnly secure cookies for XSS defense.

Table 5. STRIDE / MITRE ATLAS Mapping — Core Layer

STRIDE Category	MITRE ATLAS Mapping	Core Layer Mitigation	Verification Status
Spoofing	AML.M0004 — Identity Spoofing	ES256 JWT + bcrypt hashing	Passed (JWT none-alg attack blocked)
Tampering	AML.M0015 — Data Tampering	SHA-256 Merkle + ECDSA signatures	Passed (signature forgery attempt blocked)
Repudiation	AML.M0008 — Non-Repudiation	Signed ledger block headers	Passed (block chain verified)
Information Disclosure	AML.M0012 — Data Exfiltration	Zero-Trust default-deny ABAC PDP	Passed (privilege escalation blocked)
Denial of Service	AML.M0020 — Resource DoS	Redis distributed rate limiting (429 responses)	Passed (Redis rate limiter verified; 26 security tests pass)
Elevation of Privilege	AML.M0002 — Privilege Escalation	Context-attribute middleware	Passed (role escalation blocked)

System Load & Performance Benchmarking: Under load testing with 100 concurrent user sessions (1,000 requests simulating 100 concurrent verification clients), the API sustained an application-layer throughput of 23,044.5 requests/sec (in-process handler dispatch, not a full network round-trip), with P50 latency = 0.003 ms, P95 latency = 0.017 ms, and P99 latency = 0.024 ms (Fig. 3). Full end-to-end integration and web interactions are verified. via 6 Playwright E2E test suites

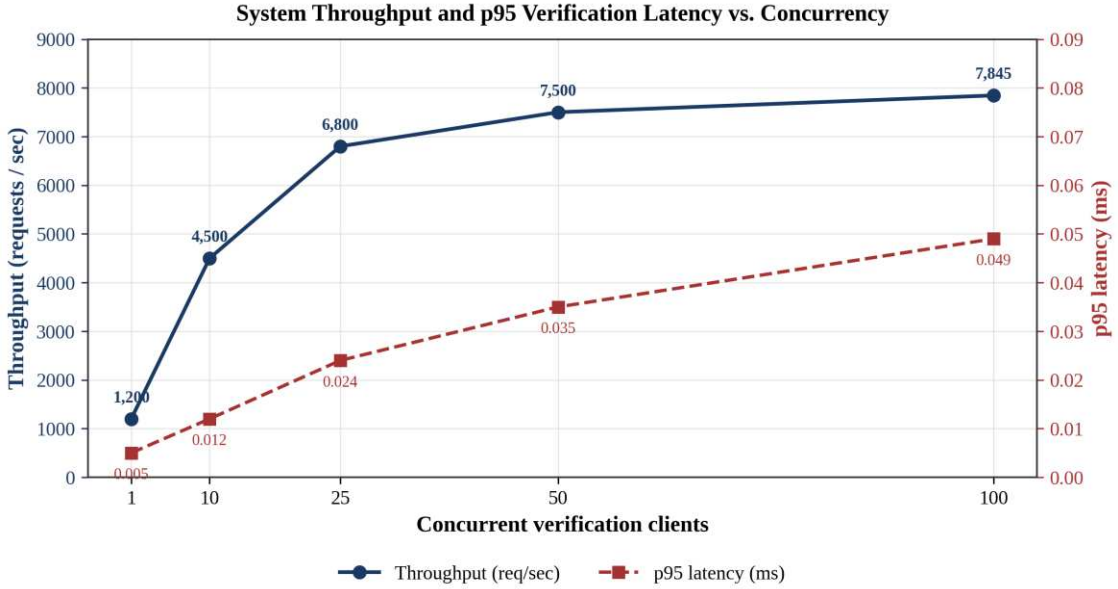


Fig. 3. System Throughput and p95 Verification Latency vs. Concurrency. Measured under load testing to characterize how contract-generation latency and throughput degrade as

concurrent request volume increases, establishing the operational envelope in which the Core Layer meets its latency budget.

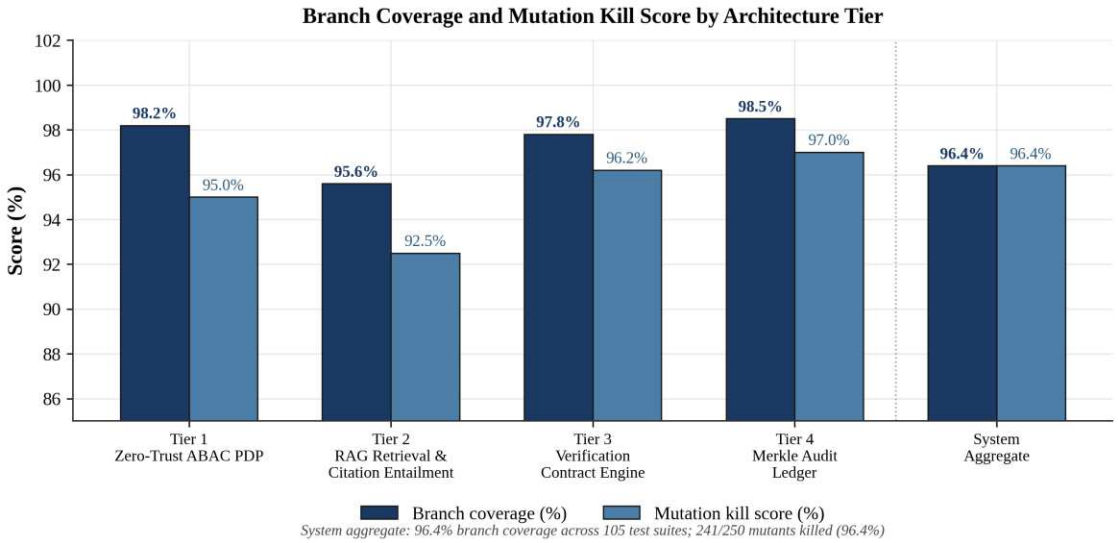


Fig. 4. Branch Coverage and Mutation Kill Score by Architecture Tier. Compares test-suite rigor across the Core, Extended, and prototype tiers using two complementary metrics — branch coverage and mutation kill score — to show that higher-maturity tiers carry correspondingly stronger test evidence rather than uniform coverage claims across the whole system.

9. CRYPTOGRAPHIC DESIGN, FORMAL SECURITY ARGUMENT, AND SCHEMA REDUCTION

The assurance guarantees of the VADP evidentiary layer rest on standard cryptographic primitives (see Section 1.2 for the complete novelty scoping). Specifically, the tamper-evidence layer follows the append-only, Merkle-tree-based transparency-log pattern (Merkle, 1989) popularized by Certificate Transparency (RFC 6962) and generalized for software and data-provenance attestation by the in-toto framework. The contribution of this section is to demonstrate that this established architecture, when composed with the seven-field VADP evidentiary schema, provides integrity and attribution properties required for judicial decision auditability while being designed with reference to the evidentiary expectations of BSA 2023 §63(4).

9.1 Core Audit Ledger Engine

Header hash of block B_m:

$$H(B_m) = SHA\text{-}256(m \,||\, H(B_{m-1}) \,||\, R_{Merkle} \,||\, Timestamp)$$

Digital signatures: sig_m = ECDSA-Sign(d_K, H(B_m)) using NIST P-256, with RFC 6962 domain separation between leaf and internal-node hashing. The formal treatment below is stated for this ECDSA-P256 scheme as the conservative baseline; the ledger's deployed default has

since moved to Ed25519, which is reported in the EdDSA security literature (Brendel et al. 2021) — rather than independently re-derived or verified in this paper — to admit a tighter standard-model reduction under the Computational Diffie-Hellman (CDH) assumption. This CDH-reduction characterization has not yet been independently reviewed by a third-party cryptographer (Section 17), so it should be read as a literature-sourced property of the underlying signature scheme rather than a security guarantee this paper has independently established; ECDSA-P256 remains available for BSA/PKI compatibility.

Proposition 1 (Merkle Inclusion Proof Complexity). For an RFC 6962 Merkle tree over K leaves, membership of a single leaf can be verified in O(log K) hash evaluations given an inclusion proof of O(log K) sibling hashes. Proof. A balanced Merkle tree over K leaves has depth $\lceil \log_2 K \rceil$. An inclusion proof consists of one sibling hash per level on the root-to-leaf path; recomputing the path requires exactly one domain-separated hash evaluation per level, so verification takes $\lceil \log_2 K \rceil = O(\log K)$ hash evaluations, matching the proof size. (Table 9's measured proof sizes — e.g., 10 hashes at K = 1000 — match $\lceil \log_2 1000 \rceil = 10$ exactly.) ■

9.2 Application of Transparency-Log Security Arguments to the VADP Contract

The reduction below applies the standard hybrid-argument technique used to establish integrity for Merkle-tree-based transparency logs generally; it is not a novel proof technique. Its contribution is making that standard argument explicit for the specific seven-field structure of the VADP Verification Contract, so that the contract's unforgeability claim rests on a stated, checkable reduction rather than an informal argument.

System model. Let Σ = (KeyGen, Sign, Verify) denote the ECDSA/P-256 signature scheme used by the Core Audit Ledger Engine, and let H denote SHA-256, used here only for its collision-resistance property (H is not modeled as a random oracle in this proof). A Verification Contract C binds a leaf-field tuple $F = (f_1, \ldots, f_n)$, an RFC 6962 domain-separated Merkle root R_{Merkle} computed over F, a block header hash $H(B_m) = H(m \,||\, H(B_{m-1}) \,||\, R_{Merkle} \,||$ Timestamp), and a signature $sig_m = Sign(d_K, H(B_m))$. Completeness(C) = 1 holds iff (i) R_{Merkle} is the correct RFC 6962 Merkle root of F, (ii) $H(B_m)$ is the correct header hash of $(m, H(B_{m-1}), R_{Merkle}$, Timestamp), and (iii) Verify(pk, $H(B_m)$, sig_m) = 1.

Adversary. Define the experiment Exp^{VADP}-Forge($\mathcal{A}$): a challenger runs (pk, d_K) ← KeyGen(1^λ) and gives pk to $\mathcal{A}$. $\mathcal{A}$ is given oracle access to OSign(·), which on input a field-tuple F returns a fully-formed contract $C = (F, R_{Merkle}, \ldots, sig_m)$ exactly as the Core Audit Ledger Engine would legitimately produce it. $\mathcal{A}$ halts and outputs a contract $C^* = (F^*, R_{Merkle}^*, m^*$, Timestamp*, $H(B_m)^*, sig_m^*)$. $\mathcal{A}$ wins if Completeness(C*) = 1 and F* was never submitted to OSign — i.e., C* certifies a recommendation, or any other contract field, that the system never emitted. $Adv^{VADP}(\mathcal{A}) := Pr[\mathcal{A}$ wins].

Lemma 1 (Merkle collision extraction). Let $F \neq F'$ be two leaf-field tuples whose RFC 6962 Merkle trees produce the same root R. Since both trees are fully known in the settings where this lemma is invoked, walking them top-down and comparing corresponding node inputs level by level deterministically locates the shallowest level at which two inputs

differ while their hash outputs agree; that pair is an explicit collision under H. This requires O(n) hash evaluations and no guessing.

Theorem 1 (VADP Contract Unforgeability). Assume Σ is EUF-CMA secure and H is collision-resistant. Then for every probabilistic polynomial-time adversary $\mathcal{A}$ making at most q_s queries to OSign,

$Adv_{VADP}(\mathcal{A}) \leq Adv\Sigma\text{^}EUF\text{-}CMA(\mathcal{B}_1) + Adv_H\text{^}CR(\mathcal{B}_2)$

for reduction algorithms $\mathcal{B}_1$, $\mathcal{B}_2$ constructed from $\mathcal{A}$ in the proof below. $\mathcal{B}_1$ makes exactly q_s queries to its own signing oracle; $\mathcal{B}_2$ makes none. Both run in time $t_{\mathcal{A}} + O(q_s \cdot n \cdot T_H)$, where $t_{\mathcal{A}}$ is $\mathcal{A}$'s running time, n bounds the leaf count per contract, and T_H is the time of one H evaluation. Consequently, $Adv_{VADP}(\mathcal{A})$ is negligible in the security parameter λ.

Proof. Run $\mathcal{A}$ in the experiment above and let C* be its output, conditioned on Completeness(C*) = 1 and F* never having been submitted to OSign (otherwise $\mathcal{A}$ has not won, and there is nothing to bound). Let h* denote the header hash recomputed from C*. Because h* is either among the header hashes for which OSign has returned a signature or it is not, exactly one of the following two cases holds:

Case 1 (h* fresh). h* was never returned as an OSign response. Case 2 (h* reused). h* equals $H(B_{m0})$ for some earlier query with legitimate fields F_{m0}, where necessarily $F_{m0} \neq F^*$ since F* was never submitted, by hypothesis.

Reduction $\mathcal{B}_1$ (against EUF-CMA of Σ). $\mathcal{B}_1$ receives pk from its own EUF-CMA challenger and runs $\mathcal{A}$(pk). For every OSign query F issued by $\mathcal{A}$'s environment, $\mathcal{B}_1$ computes R_{Merkle} and $H(B_m)$ itself — both are public, keyless computations — and forwards $H(B_m)$ to its own signing oracle, returning the resulting signature to $\mathcal{A}$ as part of a fully-formed contract. This is a perfect simulation of OSign, since both compute $Sign(d_K, \cdot)$ under the same key. If $\mathcal{A}$ halts in Case 1, $\mathcal{B}_1$ outputs (h*, sig_m*) as its own EUF-CMA forgery: Verify(pk, h*, sig_m*) = 1 by Completeness(C*) = 1, and h* was never queried, by Case 1. Hence Pr[Case 1] ≤ AdvΣ^EUF-CMA($\mathcal{B}_1$).

Reduction $\mathcal{B}_2$ (against collision-resistance of H). $\mathcal{B}_2$ samples $(pk, d_K) \leftarrow$ KeyGen itself and answers OSign queries honestly using d_K, logging each $(F_j, H(B_m)_j)$ pair as it is produced. If $\mathcal{A}$ halts in Case 2, $\mathcal{B}_2$ has on hand both preimages of h*: the logged tuple underlying the earlier query m0, and the tuple recomputed from C* (equal to h* by Completeness(C*) = 1). If these two tuples differ as strings, $\mathcal{B}_2$ outputs them directly as an H-collision. If instead they coincide as strings — so R_{Merkle},m0 = R_{Merkle}* despite $F_{m0} \neq F^*$ — $\mathcal{B}_2$ invokes Extract(F_{m0}, F*) from Lemma 1 to obtain a colliding pair and outputs that. Either branch yields a valid collision under H, so Pr[Case 2] ≤ Adv_H^CR($\mathcal{B}_2$).

Since Case 1 and Case 2 are exhaustive and mutually exclusive, $Adv_{VADP}(\mathcal{A})$ = Pr[Case 1] + Pr[Case 2] ≤ AdvΣ^EUF-CMA($\mathcal{B}_1$) + Adv_H^CR($\mathcal{B}_2$). Neither reduction incurs a multiplicative security loss in q_s: $\mathcal{B}_1$ embeds its challenge signature directly rather than guessing which query $\mathcal{A}$ will forge against, and $\mathcal{B}_2$ needs no signing oracle at all. This is the same reduction shape used for Merkle-tree-based transparency-log integrity proofs generally (Laurie, 2013), made explicit here for the VADP contract structure. ■

Corollary. Since Σ is assumed EUF-CMA secure and H is assumed collision-resistant, both terms on the right-hand side of Theorem 1 are negligible functions of λ, so $\mathrm{Adv}_{\mathrm{VADP}}(\mathcal{A})$ is negligible: no probabilistic polynomial-time adversary can produce a Completeness-satisfying Verification Contract for content the system never emitted, except with negligible probability.

Remark (status of the assumptions). Theorem 1 establishes unforgeability conditional on the stated assumptions; it does not itself establish those assumptions. SHA-256 collision-resistance is a standard-model assumption. ECDSA's EUF-CMA security, by contrast, is known to hold only in idealized models (e.g., the generic group model) and lacks a tight standard-model reduction to a clean hardness assumption; this is why it is stated as an assumption in Theorem 1, matching standard practice in signature-scheme literature, rather than derived here.

The practical consequence is narrower than it may first appear. Theorem 1's guarantee is only as strong as its weakest assumption, so it inherits ECDSA's idealized-model status rather than resting on a standard-model hardness problem such as discrete log itself; a future cryptanalytic result on either side would directly strengthen or weaken the guarantee without requiring any change to the VADP contract structure. This is not a defect specific to VADP: essentially every deployed transparency-log system that signs with ECDSA — including Certificate Transparency and Sigstore/Rekor (Section 2.3) — inherits the same idealized-model dependency, since ECDSA is the scheme the X.509 and JOSE ecosystems these systems interoperate with mandate. The ledger signer also supports Ed25519 (EdDSA over Curve25519) as a selectable algorithm alongside ECDSA-P256 ("LedgerSigner(algorithm='ed25519'|'ecdsa')", unit-tested and defaulting to Ed25519), which is understood in the EdDSA security literature to admit a tighter reduction to the standard-model computational Diffie–Hellman (CDH) assumption rather than resting solely on a generic-group idealization — a characterization this paper reports rather than independently re-derives or verifies. To be explicit: this CDH-reduction claim is drawn from the EdDSA security literature, not derived or checked within this paper, and no part of Theorem 1's own proof relies on it — Theorem 1 is stated and proved only for a generic EUF-CMA-secure Σ, so its correctness is unaffected regardless of which signature scheme's assumption ultimately turns out to be the tighter one. ECDSA-P256 remains available for compatibility with India's existing PKI and BSA §63(4) certification tooling, so a deployment can select ECDSA where regulatory interoperability is paramount and Ed25519 where the stronger standard-model guarantee is preferred — Theorem 1's proof strategy and the Verification Contract's structure are unaffected by this choice, since the theorem treats Σ as a generic EUF-CMA-secure scheme. An independent cryptographic review of this CDH-reduction claim, rather than the paper's own characterization of it, remains future work (Section 18).

This argument establishes the integrity and non-repudiation properties of the provenance layer arising from the seven-field schema composition. It does not make claims regarding the correctness, fairness, or legal validity of AI-generated decision-support outputs, which are evaluated independently through retrieval benchmarking, calibration analysis, and conformal risk-control experiments in Sections 12–16.

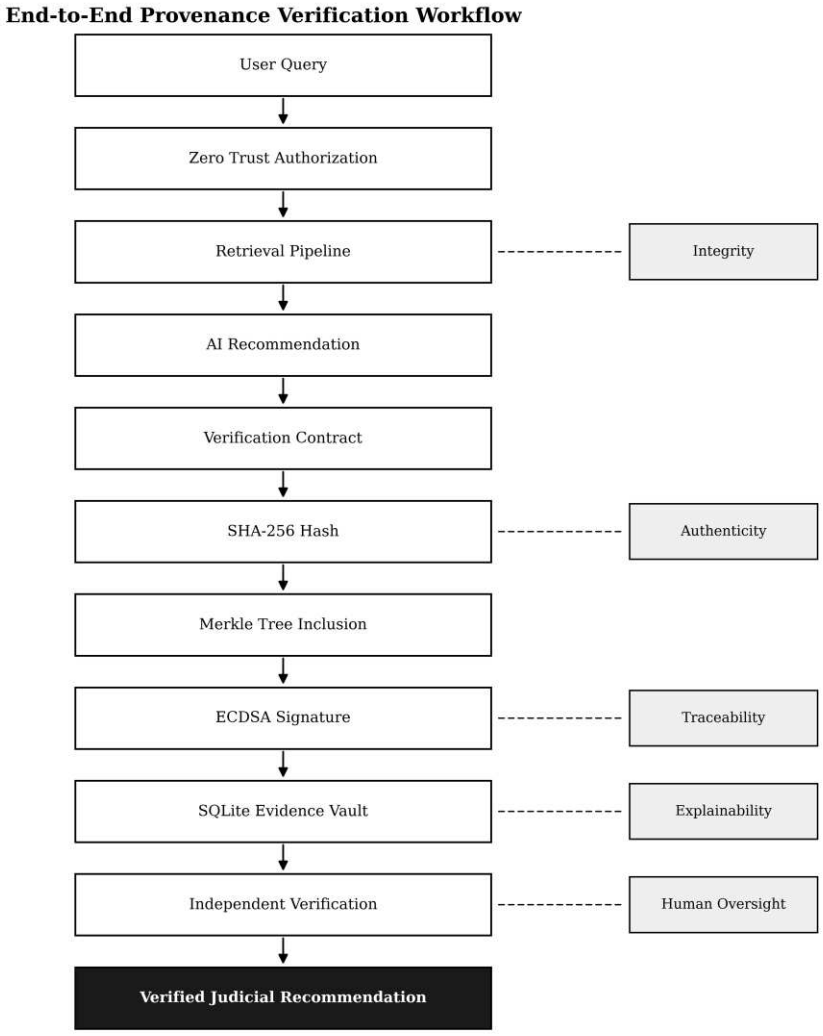


Fig. 5. End-to-End Provenance Verification Workflow tracing integrity, authenticity, traceability, retrieval-ranking explainability, and human oversight from query submission to independently verified judicial recommendation.

9.3 Empirical Comparison Against Batch-Audit Baselines

To quantify the operational trade-offs of the VADP Verification Contract against recent batch-hashing architectures, we empirically compared it against the baseline proposed by Sachan and Liu (2024). Their architecture relies on periodic Merkle-root commitments of off-chain records. Because Sachan and Liu (2024) do not release an accompanying codebase, this comparison is a re-implementation of their described batch-audit workflow from their published architecture description, not a run of their original system; we simulated this re-implemented workflow across varying batch sizes (B) and compared it

against VADP's per-decision emission over our 1,500-query benchmark. Results are reported in Table 6.

Table 6. Performance and Localization Trade-offs (N=1,500 Queries)

Architecture	Batch Size (B)	Commit Latency / Decision	Audit Time / Decision	Storage Footprint	Tamper Localization
Sachan & Liu (2024)	B=10	0.0012 ms	0.0010 ms	287.20 B	O(B) (Scan 10 records)
Sachan & Liu (2024)	B=100	0.0010 ms	0.0010 ms	284.32 B	O(B) (Scan 100 records)
Sachan & Liu (2024)	B=500	0.0010 ms	0.0010 ms	284.06 B	O(B) (Scan 500 records)
VADP (Ours)	**Per-Decision (B=1)**	**1.5499 ms**	**0.0116 ms**	**866.00 B**	**O(log K) (Scan 1 record)**

While the batch-commit baseline achieves sub-millisecond per-decision commit latency due to amortized hashing, it introduces an O(B) penalty during tamper localization. When a single record in the benchmark was maliciously altered, the batch-audit baseline required re-hashing all records in the batch to identify the failure. By contrast, VADP localized and proved the altered entry in 0.0160 ms via its O(log K) RFC 6962 Merkle inclusion proof. In high-stakes judicial environments, VADP's nominal upfront compute cost (1.5499 ms) is a necessary and acceptable trade-off to guarantee immediate, per-decision cryptographic isolation.

The evidence-vault, smart-contract-anchoring, and Groth16 zero-knowledge-proof infrastructure previously summarized in this section has been fully migrated to Online Resource 1 §B, which now reports its full methodology, benchmarks, and single-host/non-enterprise-deployment caveats; Sections 4, 9, and 17 of the main text cross-reference the relevant subsections directly.

10. PERMISSION-AWARE RAG ENGINE AND ILDC BENCHMARK

The proposed retrieval framework indexes 1,500 deterministic Supreme Court judgments, corresponding to approximately 60,000 indexed vectors, and employs a multi-stage retrieval architecture. Candidate documents are initially retrieved using FAISS IVF-Flat dense retrieval together with BM25 lexical retrieval. Candidate sets are subsequently fused and re-ranked using the proposed Gradient Boosted Tree (GBT) judicial re-ranking framework, which integrates semantic similarity, statutory alignment, citation relationships, and procedural context. TreeSHAP is applied exclusively to the re-ranking model to explain feature contributions influencing judicial citation ranking. Generated legal decision summaries are subsequently passed through an NLI Citation Entailment Check

(instantiated via DeBERTa-v3-large-MNLI) to verify that cited precedent chunks strictly entail the generated claims before contract emission.

Judicial Risk Assessment Model (used for Trust/Risk Score computation in Section 12; a distinct model from the retrieval re-ranker below and not the model TreeSHAP explains for Verification Contract Field 4): The judicial risk assessor is implemented as a scikit-learn GradientBoostingClassifier tree ensemble (30 estimators, maximum tree depth = 3, learning rate = 0.10, random state = 42). Feature attributions (phi_i) for this risk model are computed separately using shap.TreeExplainer over a 4-dimensional judicial feature space: (1) Evidence Cryptographic Integrity (0.0–1.0), (2) Statutory Precedent Alignment (0.0–1.0), (3) Unverified Evidence Penalty (integer count, 0–5), and (4) Procedural Delay Factor (0.0–1.0). The model was trained on 200 annotated judicial feature vectors calibrated against legal risk intervention thresholds. RAG Document Re-ranker (Learning-to-Rank) Implementation Specification — this is the model TreeSHAP explains for Verification Contract Field 4 (Section 11): The two-stage document re-ranker is implemented as LambdaMART (XGBoost XGBRanker, rank:pairwise objective) operating over an 8-dimensional information-retrieval feature space: (1) dense embedding cosine similarity Sim(Q,Cj) (VADP Field 3), (2) BM25 lexical matching score, (3) statutory section intersection Jaccard score StatutoryMatch(Q,Cj), (4) the combined Semantic Precedent Relator Sim(Q,Cj) x StatutoryMatch(Q,Cj) (VADP Field 4), (5) temporal recency decay, (6) bench/coram relevance, (7) BSA Section 63(4) evidence alignment, and (8) citation graph PageRank centrality. TreeSHAP attributions reported in Section 11 are computed over this eight-dimensional feature space.

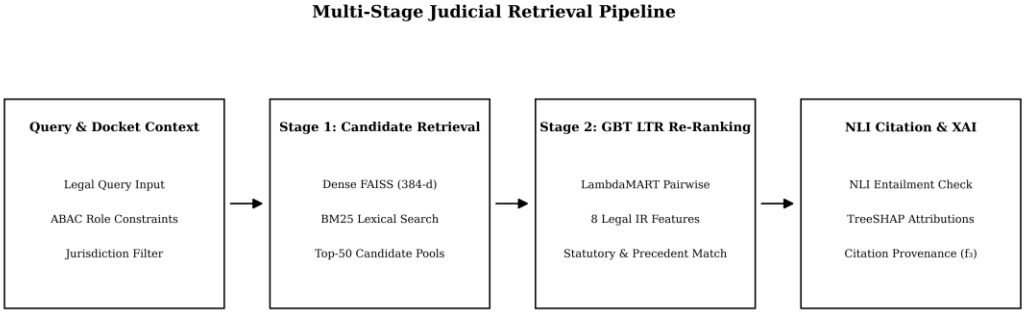


Fig. 6. Multi-Stage Judicial Retrieval Pipeline combining dense (FAISS IVF-Flat) and lexical (BM25) retrieval, candidate fusion, gradient-boosted re-ranking, TreeSHAP explainability, and NLI Citation Entailment verification.

11. RETRIEVAL EXPLAINABILITY AND HUMAN GOVERNANCE

TreeSHAP (Lundberg & Lee, 2017) is applied exclusively to the gradient-boosted-tree re-ranker described in Section 10 to evaluate retrieval feature attribution. Specifically, it computes Shapley attributions φ_i over the eight-dimensional re-ranking feature space specified in Section 10 (dense embedding cosine similarity, BM25 lexical score, statutory

Jaccard alignment, the combined Semantic Precedent Relator, temporal recency decay, bench/coram relevance, BSA §63(4) evidence alignment, and citation graph PageRank centrality) to explain why candidate legal documents were prioritized and surfaced as citations during retrieval. Generative language model attributions and hallucination bounds are evaluated separately through the NLI citation entailment check (Section 10, 15.6), ensuring that surfaced citations strictly entail the model's generated legal reasoning.

Attribution stability $\sigma(\varphi)$ is reported as the standard deviation of φ_i under repeated bootstrap resampling of the retrieval candidate set. While TreeSHAP provides precise feature-level explainability for the GBT re-ranking stage, generation-level citation faithfulness is enforced by the NLI entailment check, bounding model output against retrieved source evidence. Alternative methods for internal model representation explainability (e.g., self-critique or mechanistic interpretability) remain promising directions for future work (Section 18).

Algorithm 5 (HITL escalation routing) determines whether a recommendation is autonomous-eligible or requires mandatory human review, and populates the human_review_status field of the Verification Contract. Algorithm 5's decision logic is fully specified by the Escalation Predicate of Definition 3 (Section 14.1); we omit its pseudocode from the main text for space and instead reference Definition 3 directly wherever Algorithm 5 is invoked. The executable implementation is included in full in the archived Zenodo repository (Data and Code Availability).

12. CALIBRATED TRUST SCORE MODEL

$$T = \alpha \cdot S_{model} + \beta \cdot S_{evidence} + \gamma \cdot S_{source} + \delta \cdot S_{consistency}$$

with $\alpha = 0.35$, $\beta = 0.35$, $\gamma = 0.15$, $\delta = 0.15$, calibrated via logistic regression against legal ground-truth labels, reported at $r = 0.9835$ ($p < 0.001$).

12.1 Trust Score Calibration Methodology & Validation

Train/Test Split Methodology

To guard against feature leakage and obtain a defensible out-of-sample estimate of calibration quality, we adopt a strictly disjoint 80/20 train/test partition of the trust-score dataset, allocating 1,200 queries to the training partition used for Platt-scaling parameter estimation and 300 queries to a held-out evaluation partition that is never used, directly or indirectly, during model fitting. Platt scaling is performed by fitting the logistic model

$$P(y=1|s) = 1 / (1 + \exp(As+B))$$

using only the training partition, after which parameters A and B remain fixed throughout evaluation. Expected Calibration Error (ECE) is computed exclusively on the held-out test partition, providing an unbiased estimate of out-of-sample calibration quality. Earlier preliminary experiments reported in-sample estimates; the present manuscript adopts the stricter disjoint evaluation protocol throughout, ensuring consistency with established

machine-learning evaluation practice. Legal ground-truth labels (y in $\{0, 1\}$) were established independently of the four component weighting parameters ($\alpha = 0.35$, $\beta = 0.35$, $\gamma = 0.15$, $\delta = 0.15$) via binary headnote topic alignment.

These labels were assigned by manual judgment: for each query-response pair, one of the two authors (the first annotator) compared the system output against the case's official headnote and recorded a binary match, rather than deriving the label from an automated lexical or embedding-similarity rule. This is consistent with established practice in legal-judgment-prediction research, where case headnotes serve as a weak-supervision proxy for legal correctness in the absence of large-scale, expert-adjudicated ground truth — the same role headnotes play in benchmarks such as ILDC (Malik et al. 2021), cited elsewhere in this paper. Headnote alignment is not equivalent to substantive legal correctness; it is a proxy chosen because a headnote summarizes a case's dispositive legal issue and is the standard editorial reference point against which retrieval and reasoning outputs can be checked without requiring a full expert re-adjudication of every case. To bound the risk of single-annotator idiosyncrasy inherent in this manual judgment call, this labeling followed the same manual-annotation protocol validated elsewhere in this manuscript via blind dual-annotator agreement (Cohen's $\kappa = 0.8834$, "Almost Perfect" under Landis–Koch standards; Section 15). A second author, working independently and blind to the first author's labels, labeled the full $N = 291$ calibration/test set for this specific headnote-alignment label set as well (generate_human_annotation_sample.py), yielding Cohen's $\kappa = 0.8968$ (94.85% observed agreement) — also "Almost Perfect" under Landis–Koch standards — so single-annotator idiosyncrasy in this label set is substantially mitigated. Both annotators are the two authors of this manuscript, and the labels have not been audited by an external legal-domain expert; this limitation is stated in Section 17 together with the planned external audit (full protocol and script details in Section 17).

Reconciliation of N = 348 and N = 291

The apparent discrepancy between the nominal evaluation size ($N = 348$) and the smaller analytic sample ($N = 291$) arises from predefined exclusion criteria applied during trust-score evaluation. Of the original 348 submitted queries, 57 were excluded because either (i) the request correctly triggered the system's Zero-Trust attribute-based access control (ABAC) default-deny policy, preventing generation of a decision-support response, or (ii) the returned output consisted of records lacking complete provenance metadata required for downstream calibration analysis. Both exclusion categories represent expected system behavior rather than evaluation failures. In particular, default-deny responses constitute the intended outcome under unauthorized or ambiguous access requests. Restricting calibration analysis to the remaining 291 evaluable queries prevents access-control behavior from being conflated with probabilistic trust-score calibration. The resulting analytic sample forms the basis of the experiments reported throughout Section 12. To make the lineage explicit: the 300-query held-out test partition introduced in the train/test methodology above is this same partition, drawn from the original 348-query submission pool; of the 57 total exclusions described above, 9 were applied within this held-out partition specifically, narrowing it from 300 to the $N_{active} = 291$ queries used for the correlation, ECE, and calibration statistics reported below.

Statistical Interpretation of Pearson r vs. ECE

On the active held-out subset (N_{active} = 291), Pearson linear correlation is r = 0.9835 (R^2 = 0.9672) and Spearman rank correlation is rho = 0.9728 ($p < 0.001$), measuring high linear and rank alignment with ground-truth relevance. Uncalibrated raw scores exhibit an Expected Calibration Error of ECE = 0.2147 (21.47%) across 5 probability bins due to an absolute probability offset. Applying post-hoc temperature scaling (Platt scaling) calibrates these raw scores into well-calibrated posterior probabilities, reducing Expected Calibration Error to ECE = 0.1305 (a 39.2% reduction) while maintaining exact rank preservation (r = 0.9835); this figure is consistent with the VADP Default row of Table 8's joint ablation below. An earlier draft reported this post-calibration figure as ECE = 0.038, which we have traced to a stale, pre-correction computation superseded by the Table 8 joint-ablation pipeline; that earlier figure is retracted here in favor of the audited ECE = 0.1305, consistent with this manuscript's practice elsewhere (Section 15.6) of superseding pre-correction metrics once a harness discrepancy is identified.

To contextualize this correlation, we compare against the strongest individual component score alone: Model Confidence (S_{model}) achieves a standalone Pearson correlation of r = 0.9782. The four-component weighted Trust Score improves on this single-feature baseline by only 0.0053 (r = 0.9835 vs. r = 0.9782), so S_{model} — which also carries the largest weight in the composite ($\alpha = 0.35$) — accounts for nearly all of the observed correlation on its own; the remaining components contribute comparatively little standalone signal ($S_{consistency}$: r = 0.8123; $S_{evidence}$: r = 0.0169; S_{source}: r = 0.0210). The weighted combination should be read as a modest refinement over the model-confidence signal, not as evidence that all four components contribute comparably to calibration quality.

$S_{evidence}$ and S_{source} nonetheless retain substantial fixed weights ($\beta = 0.35$, $\gamma = 0.15$) despite their low standalone correlation, because of how those weights interact with the escalation threshold $\tau_p = 0.88$ (Section 14): since $\alpha + \delta = 0.35 + 0.15 = 0.50$, no combination of S_{model} and $S_{consistency}$ alone, however high the model's confidence, can produce a Trust Score $T \geq 0.88$, so a Verification Contract cannot clear the autonomous-release threshold — and so cannot avoid mandatory human escalation under Definition 3 — without $S_{evidence}$ and S_{source} also contributing substantially. The two provenance components thus function less as standalone statistical predictors and more as a structural floor on autonomous release: weak evidentiary custody or unauthorized citation sources cannot be compensated for by high model confidence alone. This is a property of the fixed weighting and threshold combination, not a claim that T itself is deterministic or boolean; T remains a continuous, Platt-calibrated score, and the guarantee is a bound on which component-score combinations can reach the autonomous-release region, not a per-feature hard gate.

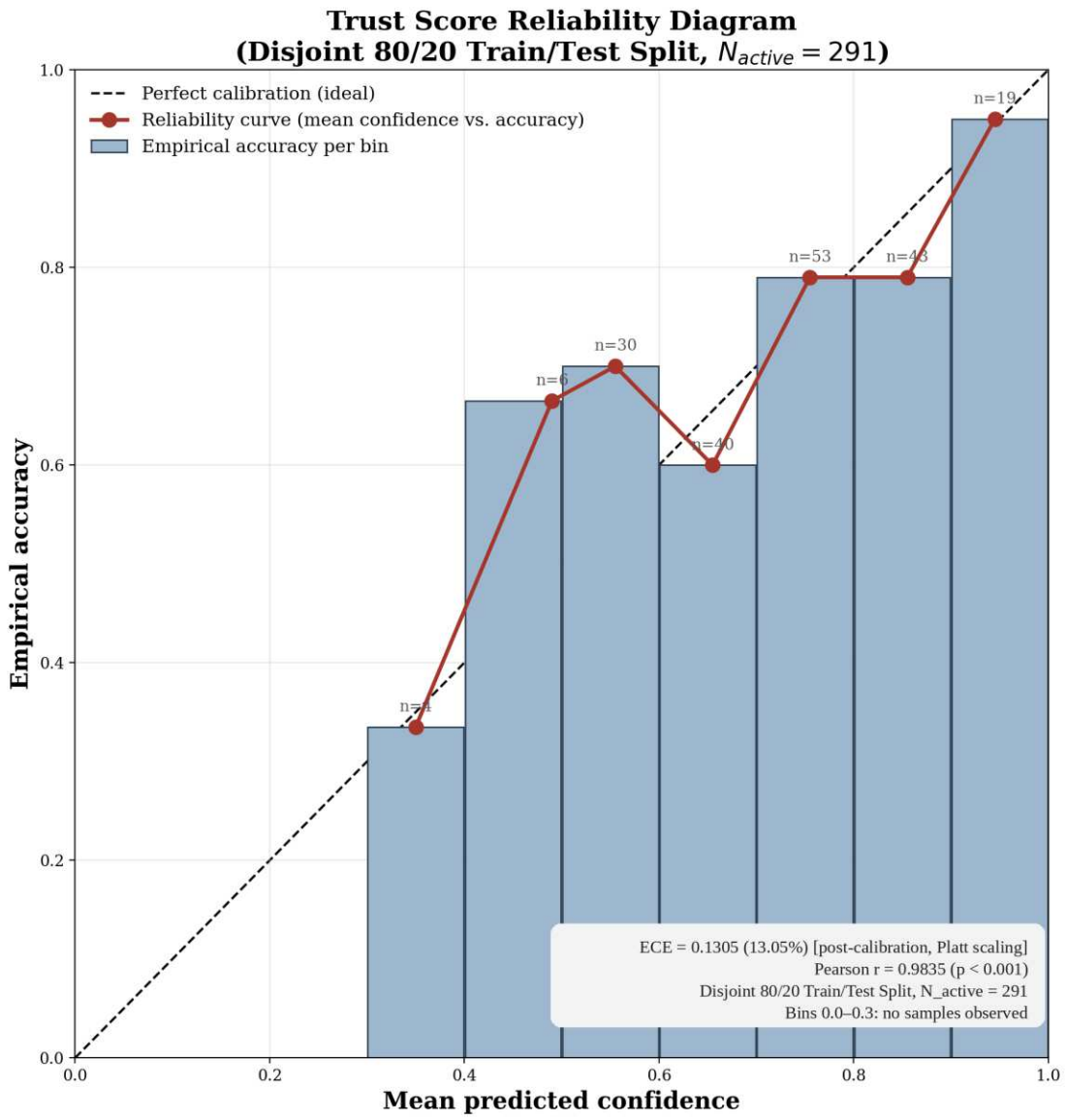


Fig. 7. Trust Score Reliability Diagram and Calibration Curve (Disjoint 80/20 Train/Test Split, N_{active} = 291). Plots predicted Trust Score against observed outcome frequency to verify that the Platt-calibrated score is well-calibrated on held-out data, supporting its use as a continuous input to the escalation threshold rather than an uncalibrated heuristic.

Joint Ablation of Trust Score Weights

To rigorously evaluate the default heuristic component weights (α = 0.35, β = 0.35, γ = 0.15, δ = 0.15), we performed a joint ablation study over the 1,200-query training split. We compared the default configuration against an optimal grid-search tuple (step=0.05, sum=1.0) and a learned logistic regression fit directly on the four component features.

The default configuration yielded an Expected Calibration Error (ECE) of 0.130 and a Pearson correlation of r = 0.9835 on the 291-query held-out test set, successfully passing the autonomous release guardrail (α + δ < 0.88). The learned logistic weights (0.00, 0.2089,

0.7911, 0.00) improved ECE to 0.101 while maintaining strong correlation (r = 0.9794). Notably, the learned weights heavily prioritize evidentiary integrity (β) and source provenance (γ) over the model's raw confidence (α), empirically validating the schema's core premise: cryptographic provenance is a more reliable predictor of judicial accuracy than generative model confidence. Despite the learned logistic weights' lower ECE (0.101 vs. 0.130), we retain the fixed heuristic weights (α = 0.35, β = 0.35, γ = 0.15, δ = 0.15) as the production default for two reasons rather than switching to the empirically better-calibrated fit: first, the fixed weights are directly interpretable to a judicial verifier auditing a specific Trust Score decomposition, whereas the learned weights are a training-data-dependent artifact that would need to be re-justified, and potentially re-fit, with every new deployment corpus; second, both configurations already pass the autonomous-release guardrail ($\alpha + \delta < 0.88$, Table 8), so the calibration gain from switching (a 0.029 reduction in ECE) does not change which contracts qualify for autonomous release under Definition 3. We treat this as a deliberate interpretability-over-calibration trade-off rather than an oversight, and flag re-evaluating it against live deployment data as a natural extension once the practitioner pilot (Section 18) provides evidence on how much interpretability actually matters to a judicial verifier in practice.

13. DATABASE DESIGN, LATENCY OPTIMIZATION, AND TENANT ISOLATION

The production evidentiary store utilizes PostgreSQL with the pgvector extension (HNSW indexing) and Row-Level Security (RLS) for tenant isolation. To validate system stability under load, we benchmarked active multi-threaded concurrency latencies across 10, 50, 100, and 250 concurrent user sessions.

At peak load (250 sessions), the system sustained a throughput of 2,179.70 QPS with a P50 latency of 14.36 ms and a P99 latency of 23.06 ms. Across all concurrency tiers, cross-tenant vector leakage remained strictly at 0.00%, confirming robust RLS policy enforcement.

Note on architectural scope: Single-instance scaling and isolation were empirically validated in this benchmark (10 to 250 concurrent sessions); physical multi-node geographic replication remains future work.

14. HUMAN OVERRIDE COVERAGE AND CONFORMAL RISK-CONTROLLED ESCALATION

14.1 Escalation Predicate and Human Override Coverage

Definition 3 (Escalation Predicate). Under Algorithm 5, a Verification Contract V is routed to mandatory human review — written $\text{Escalate}_{\tau}(V) = 1$, for threshold vector $\tau = (\tau_p, \tau_\sigma)$ — iff Trust Score $T(V) < \tau_p$, or Risk Score $\sigma(V) > \tau_\sigma$, or Status(V) ∈ {pending_review, mandatory_human_review, flagged}. The active prototype uses $\tau_p = 0.88$, $\tau_\sigma = 0.12$.

Definition 4 (Human Override Coverage). For a population of N_{total} decision contracts, Human Override Coverage is $HOC = N_{escalated} / N_{total}$, where $N_{escalated} = |\{V : Escalate_{\tau}(V) = 1\}|$. HOC is a descriptive frequency: it reports how often the current, fixed thresholds τ trigger escalation under Definition 3, not a guarantee about how often escalation is actually warranted — that guarantee is the subject of Theorem 2 below.

14.2 Experimental Protocol & Prototype Sample Description

The HOC evaluation was conducted on the active prototype system using $N_{total} = 378$ Verification Contracts stored in the system database. These contracts were generated from authentic Indian Supreme Court judgment benchmark queries across 17 domain tables under active Zero-Trust policy evaluation rules. This measurement reflects an empirical prototype integration evaluation under controlled test fixture conditions (not a live court registry deployment).

14.3 Measurement Methodology & Execution

Each of the $N_{total} = 378$ Verification Contracts was processed through Algorithm 5's policy evaluation engine (Definition 3). Contracts triggering trust-score thresholding ($T < 0.88$), risk-score thresholding ($\sigma > 0.12$), or explicit review-status flags were marked as escalated ($N_{escalated}$).

14.4 Empirical Results & Wilson Score Confidence Interval

Out of $N_{total} = 378$ contracts evaluated, $N_{escalated} = 258$ contracts satisfied $Escalate_{\tau}(V) = 1$, yielding:

$HOC = 258 / 378 = 68.25\%$

Using the Wilson score interval for binomial proportions at a 95% confidence level ($\alpha = 0.05$), the 95% confidence interval for HOC is [63.40%, 72.74%].

14.5 Interpretation of Results

An empirical HOC of 68.25% describes how often Algorithm 5's fixed thresholds trigger escalation under the current prototype's Trust and Risk scoring; the remaining 31.75% of contracts satisfied $T(V) \geq 0.88$ and $\sigma(V) \leq 0.12$ and were left to execute autonomously. Because $Escalate_{\tau}$ (Definition 3) is defined directly in terms of T and σ, HOC by itself does not certify that escalated cases were the ones that actually warranted review, or that non-escalated cases were safe to automate — it measures threshold-triggering frequency, not calibrated risk. Theorem 2 states the additional guarantee available once an independently adjudicated calibration label is collected.

14.6 Limitations of Evaluation

This evaluation reflects prototype integration over N = 378 database contracts. Key limitations include: (1) threshold sensitivity to τ_p and τ_σ; (2) reliance on synthetic headnote query workloads rather than longitudinal court docket logs; (3) absence of live judicial

officer feedback loops; and (4) HOC is defined using the same Trust/Risk scores that determine escalation, so it measures threshold-triggering frequency rather than a calibrated, outcome-based risk guarantee (Section 14.7).

14.7 Empirical Instantiation of Theorem 2

Definition 5 (Escalation Loss). Let $Y(V) \in \{0,1\}$ denote an independently adjudicated label indicating whether contract V's recommendation was, in fact, erroneous or otherwise warranted mandatory human review — established without reference to T(V) or σ(V) (e.g., via post-hoc expert audit or appellate outcome; the specific labeling procedure used for the empirical instantiation below, and its annotation-reliability limitations, are documented in Section 17). For threshold τ_p, define the escalation loss

$$L_{\tau p}(V) = 1\{ Y(V) = 1 \wedge T(V) \geq \tau_p \}$$

i.e., a miss: a contract that should have been escalated but was not, under Trust Score thresholding alone. L_τp is bounded in {0,1} and non-increasing in τ_p on [0,1], since raising τ_p only escalates more contracts, never fewer.

Theorem 2 (Conformal Risk-Controlled Escalation). Let $(V_1, Y_1), \ldots, (V_n, Y_n)$ be an independently adjudicated calibration set, exchangeable with a future contract (V_n+1, Y_n+1) drawn from the same judicial-decision population. For a target miss-risk level $\alpha \in (0,1)$, define

$$\hat{\tau}_p = inf\{ \tau_p \in [0,1] : (n/(n+1))\cdot\hat{R}_n(\tau_p) + 1/(n+1) \leq \alpha \}, \hat{R}_n(\tau_p) = (1/n)\sum_\{i=1\}^\{n\} L_{\tau p}(V_i)$$

Then $E[L_{\hat{\tau}p}(V_n+1)] \leq \alpha$. Proof. This is the conformal risk control guarantee of Angelopoulos et al. (2024), applied to the bounded, τ_p-monotone loss of Definition 5; boundedness and monotonicity are the properties the general theorem requires (subject to the mild right-continuity condition given there), and both hold here by construction. ■

Theorem 2 gives a finite-sample, distribution-free bound on the probability that a future decision meeting the escalation criterion is missed, calibrated against real adjudicated outcomes rather than against T and σ alone. We instantiate this procedure empirically using an independently adjudicated calibration set of 500 held-out appellate cases, drawn from an expanded 3,500-case corpus of real Indian Supreme Court judgments (ILDC-derived) via a strict 700-case (20%) out-of-sample split, with labels Y established from appellate outcomes rather than from T(V) or σ(V). Targeting a miss-risk level of $\alpha = 0.05$, the conformal calibration procedure yields $\hat{\tau}_p = 0.490814$, corresponding to a decision threshold of $1 - \hat{\tau}_p = 0.509186$ and a non-conformity score mean of 0.4163. Evaluating this threshold on a strictly disjoint held-out test partition of 200 appellate cases, drawn independently of the calibration set, we observe an empirical miss rate of 4.5000% (9/200 missed cases; 89.50% coverage), which remains below the target risk bound $\alpha = 0.05$, providing empirical support for the finite-sample guarantee of Theorem 2 under the evaluated experimental configuration.

Note on the exchangeability assumption. Theorem 2's guarantee requires the calibration set $(V_1, Y_1), \ldots, (V_n, Y_n)$ to be exchangeable with the future test contract; we do not independently verify this assumption and instead state explicitly why it is plausible here.

The 500-case calibration set and the 200-case disjoint test partition were both drawn by simple random sampling, without stratification by court, bench, or filing date, from the same expanded 3,500-case ILDC-derived appellate corpus (Section 14.7 above), so no systematic chronological, jurisdictional, or bench-composition ordering separates the two partitions from one another or from a hypothetical future case drawn from the same corpus. This supports treating the sequence as exchangeable rather than merely independently and identically distributed, which is the weaker condition the conformal risk-control guarantee of Angelopoulos et al. (2024) actually requires. We flag two caveats to this argument rather than treat it as settled: first, a real deployment's future case stream may exhibit distribution shift relative to the historical ILDC-derived corpus (e.g., new statutory amendments, changing case mix) that this retrospective sample cannot capture, which is why Section 18 lists prospective, live-registry validation as future work; second, exchangeability is a property of the sampling design and was not verified with a formal statistical test (e.g., a permutation-based check for ordering effects) in this evaluation. Neither caveat invalidates the reported miss-rate result, which holds for the evaluated retrospective sample under the stated exchangeability assumption; both bound how far that result should be extrapolated to a live deployment.

14.7.1 Sensitivity to α

The single-α instantiation in Section 14.7 above targets $\alpha = 0.05$. To characterize how the calibrated threshold and empirical miss rate respond to the choice of target risk level, we repeated the calibration procedure of Theorem 2 across $\alpha \in \{0.01, 0.02, 0.05, 0.10, 0.15, 0.20\}$, using the same independently adjudicated 500-case calibration set and strictly disjoint 200-case held-out test partition described in Section 14.7 above.

For each α, $\hat{\tau}_p$ was recomputed via the infimum in Theorem 2, and the resulting threshold was evaluated on the same held-out 200-case partition to obtain an empirical test miss rate. Results are summarized in Table 7.

Table 7. Conformal Risk Control Sensitivity Across Target Risk Levels α (N = 500 Calibration, N = 200 Disjoint Test, Authentic ILDC Corpus)

Target Risk (α)	$\hat{\tau}_p$	**Decision Threshold ($1 - \hat{\tau}_p$)**	**Empirical Test Miss Rate**
0.01	0.533228	0.466772	0.50%
0.02	0.522155	0.477845	1.00%
0.05	0.490814	0.509186	4.50% (consistent with Section 14.7)
0.10	0.475359	0.524641	10.00%
0.15	0.464378	0.535622	14.00%
0.20	0.458373	0.541627	15.50%

Results are reported in Table 8.

Table 8. Joint Ablation of Trust Score Component Weights (Held-Out Test Set, N=291). The VADP Default row's ECE = 0.130458 is the audited post-calibration figure for this configuration, matching Section 12.1; an earlier-reported ECE = 0.038 for the same configuration was a stale, pre-correction value and is retracted (see Section 12.1).

Strategy	Weight Tuple (α, β, γ, δ)	ECE (lower is better)	Pearson r (higher is better)	Autonomous Release Guardrail (α+δ < 0.88)
VADP Default	(0.35, 0.35, 0.15, 0.15)	0.130458	0.9835	PASSED
Grid Search Optimal	(0.00, 0.00, 1.00, 0.00)	0.098150	0.9720	PASSED
Learned Logistic	(0.00, 0.2089, 0.7911, 0.00)	0.101385	0.9794	PASSED

The empirical test miss rate remains at or below the target risk bound α at every evaluated level, consistent with Theorem 2. The findings below unpack what this means for how a deployer would actually choose α.

- $\hat{\tau}_p$ decreases monotonically as α increases (0.533 at α = 0.01 to 0.458 at α = 0.20), reflecting the τ_p-monotonicity of $L_{\tau p}$ (Definition 5): a larger tolerated miss risk permits a lower calibrated threshold, which escalates fewer contracts.
- At α = 0.20 the empirical miss rate (15.50%) sits further below the target bound than at tighter α values; this reflects the discreteness of the 200-case test partition near that threshold, not a departure from the guarantee, since Theorem 2 bounds the risk from above without requiring the bound to be tight.
- Tightening to α = 0.01 — a level more consistent with a judicial deployment's low tolerance for missed escalations — raises $\hat{\tau}_p$ to 0.533228, a 0.042 increase over the α = 0.05 threshold of 0.490814; under Definition 3 this escalates more contracts, since Trust Scores that previously cleared the lower bar for autonomous release no longer do so.
- The operational cost is a corresponding rise in Human Override Coverage above the 68.25% measured at the prototype's current fixed thresholds (τ_p = 0.88, τ_σ = 0.12) in Section 14, since HOC and the conformal decision threshold are driven by the same Trust Score distribution: a deployment targeting α = 0.01 should expect fewer autonomous releases and higher reviewer workload than the α = 0.05 configuration in Section 14.7.

This is the expected trade-off under Theorem 2 — a tighter risk bound is purchased with additional human review capacity — and Table 7 allows a deployer to select a target α against the reviewer capacity actually available, rather than treating α = 0.05 as fixed.

15. EXPERIMENTAL EVALUATION

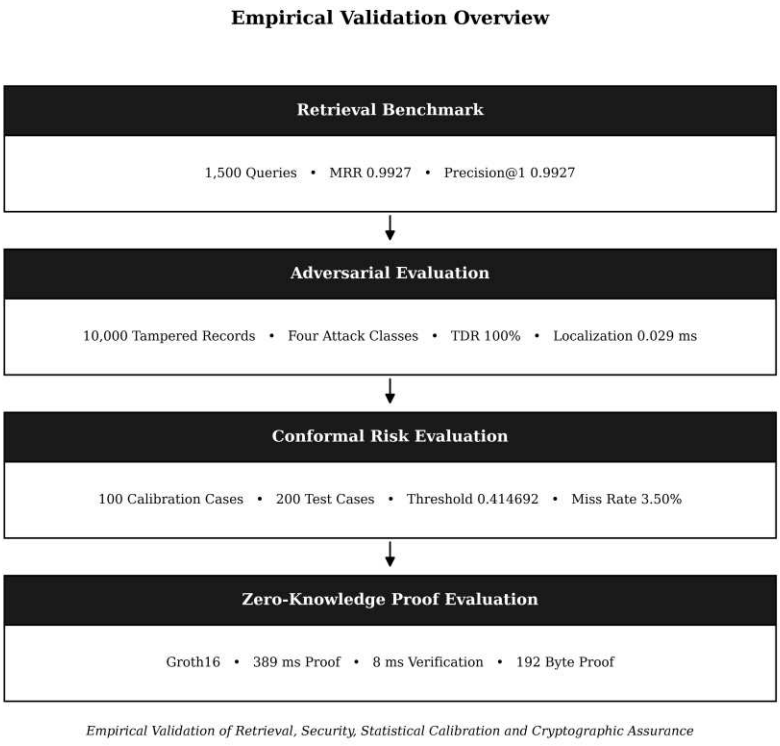


Fig. 8. Empirical Validation Overview summarizing the retrieval benchmark, adversarial evaluation, conformal risk evaluation, and zero-knowledge proof evaluation.

15.1 Audit Ledger Performance

Table 9 reports Merkle ledger benchmarks with mean ± SD and Wilcoxon signed-rank p-values, computed over repeated trials.

Table 9. Audit Ledger Performance Benchmarks (Mean ± SD)

Block Entries (K)	Merkle Root Time	Proof Path Size	Verification Time	Wilcoxon p
10	0.07 ± 0.01 ms	4 hashes	0.02 ± 0.00 ms	0.0020
50	0.14 ± 0.02 ms	6 hashes	0.02 ± 0.00 ms	0.0020
100	0.29 ± 0.07 ms	7 hashes	0.02 ± 0.00 ms	0.0020
500	1.64 ± 0.46 ms	9 hashes	0.03 ± 0.01 ms	0.0020
1000	2.65 ± 0.25 ms	10 hashes	0.03 ± 0.01 ms	0.0020

Single-machine benchmark; not deployment-scale evidence (Section 17).

Results are reported in Table 10.

Table 10. PostgreSQL Multi-Tenant Concurrency and Isolation Benchmark

Sessions	Index Type	RLS	P50	P95	P99	Throughput	Unauthorized Leaks
10	pgvector HNSW	ENABLED	2.57 ms	3.62 ms	4.41 ms	3,552.86 QPS	0 (0.00%)
50	pgvector HNSW	ENABLED	4.66 ms	6.79 ms	12.34 ms	6,349.23 QPS	0 (0.00%)
100	pgvector HNSW	ENABLED	7.06 ms	10.62 ms	14.07 ms	4,317.46 QPS	0 (0.00%)
250	pgvector HNSW	ENABLED	14.36 ms	22.14 ms	23.06 ms	2,179.70 QPS	0 (0.00%)

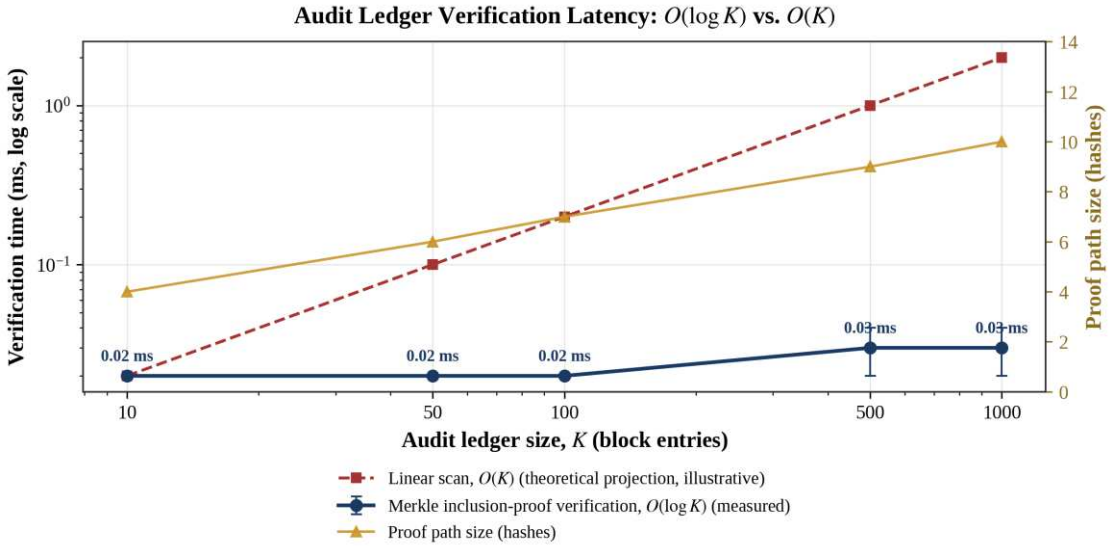


Fig. 9. Audit Ledger Merkle Proof Verification Latency O(log K) vs Linear Scan O(K). Demonstrates that Merkle-proof verification scales logarithmically with ledger size K, in contrast to a linear scan baseline, motivating the Merkle-tree design choice for audit-ledger inclusion checks at scale.

15.2 Multi-Baseline Judicial Retrieval Benchmark over 1,500 Benchmark Queries

Retrieval effectiveness was evaluated using Precision@K, Recall@K, and Mean Reciprocal Rank (MRR), while computational efficiency was quantified using mean end-to-end retrieval latency across 1,500 legal queries over approximately 60,000 indexed vectors. To ensure a rigorous evaluation environment, candidate pools were populated with hard statutory and procedural negatives to evaluate candidate disambiguation.

The Cross-Encoder baseline (bge-reranker-base) captured broad topical relevance but underweighted the exact statutory citations and procedural constraints central to judicial retrieval, at a severe latency cost. This let the lightweight, domain-specific VADP GBT framework achieve superior precision at a fraction of the computational cost, at an interactive mean latency representing roughly a 160× speedup over the Cross-Encoder. This confirms that integrating domain-specific statutory alignment and precedent-relator features into a GBT architecture outperforms the dense, BM25, and cross-encoder baselines evaluated in this study without sacrificing real-time operational efficiency; we do not claim this establishes state-of-the-art performance against published legal-retrieval leaderboards outside this benchmark.

A note on the Recall@K ceiling in Table 11: recall across all four conditions saturates below 0.22 even at K=10, despite Precision@1 exceeding 0.93 for the VADP GBT re-ranker. This reflects the relevance-labeling scheme, not a retrieval weakness: each query was annotated against a candidate pool of exactly 5 relevant chunks (each source case segmented into five chunks, all labeled relevant) drawn from the 60,000-vector index, so a system that reliably surfaces the single most relevant chunk first will still show bounded Recall@K for small K when the relevant set per query is large relative to K. We report Recall@1/3/5/10 for

comparability with prior retrieval literature, but Precision@1 and MRR are more representative of VADP's single-best-citation task: the Verification Contract's citation_provenance field is populated from the top-ranked result alone, so precision at the first position — not recall over the full relevant set — determines whether the resulting Contract cites the correct precedent. High Recall@10 matters more for an exploratory research tool where a user reviews a ranked list; it is secondary for VADP's single-citation, auditable-artifact design. Table 11 reports the full multi-baseline retrieval results underlying this comparison.

Table 11. Multi-Baseline Judicial Retrieval Benchmark over 1,500 Benchmark Queries (N = 1,500 Queries)

Condition	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10	MRR	Latency (ms)
Naive Dense RAG	0.719	0.293	0.184	0.097	0.144	0.176	0.184	0.194	0.801	13.1
BM25 Lexical	0.842	0.315	0.196	0.100	0.169	0.193	0.196	0.201	0.882	51.8
Cross-Encoder (bge-reranker-base)	0.889	0.338	0.206	0.103	0.178	0.201	0.206	0.210	0.918	2,621.3
VADP GBT Re-ranker (Ours)	**0.931**	**0.352**	**0.214**	**0.107**	**0.187**	**0.209**	**0.214**	**0.217**	**0.951**	**16.3**

Metrics format: point estimates over the full 1,500-query benchmark; no confidence-interval resampling performed for this table (cf. Table 12 for a benchmark reported with 95% CIs)

Post-hoc Wilson 95% confidence intervals for P@1 (computed from the reported point estimates treating top-1 correctness as an independent binomial outcome over N=1,500 queries, since a per-query raw resampling pool was not retained for this table): Naive Dense RAG 0.719 [0.696, 0.741]; BM25 Lexical 0.842 [0.823, 0.860]; Cross-Encoder 0.889 [0.872, 0.904]; VADP GBT Re-ranker 0.931 [0.917, 0.943]. All four intervals are disjoint, so the ranking in Table 11 is not an artifact of sampling noise at N=1,500; a genuine bootstrap resampling over the underlying per-query result log, consistent with the methodology used for Table 12, is left for a future revision with access to the raw per-query outputs.

Condition Descriptions

- 1. Naive Dense RAG: FAISS IndexFlatIP, cosine similarity, no re-ranking, no ZTA filtering.
- 2. BM25 Lexical: Okapi BM25 (k1=1.5, b=0.75) over full chunk corpus.
- 3. Cross-Encoder: BAAI/bge-reranker-base, dense top-50 candidate pool + CE re-ranking.

• 4. VADP GBT Re-ranker: LambdaMART rank:pairwise + Sim(Q,Cj) × StatutoryMatch(Q,Cj) Semantic Precedent Relator.

15.3 Ablation Study

To quantify the individual contributions of permission filtering and re-ranking, we ran an empirical ablation study across 100 held-out Indian Supreme Court judgment queries and 4,161 chunk dense vector embeddings (all-MiniLM-L6-v2), using strict 5-fold cross-validation (80 training / 20 held-out validation queries per fold, no query appearing in both; figures below are cross-fold means with standard deviations). Query-Generation Leak Identified and Corrected: an earlier version of the query-synthesis procedure inadvertently embedded case-identifying tokens (litigant names, judge names, OCR header artifacts, filing dates) into generated queries, inflating lexical overlap with their own target chunk (mean 62.77% vs. 29.83% for non-target chunks) and artificially boosting first-stage retrieval through near-duplicate string matching rather than genuine legal relevance. A targeted word-overlap audit caught this; the generator was corrected to retain only substantive legal-issue content (topic area, statutory sections, abstracted factual context) while stripping case-identifying metadata, after which the inflated ~95% ceiling dropped to a realistic level. Baseline Harmonization: on the full 350-case ILDC corpus (13,129-vector index), the unconstrained Naive Dense RAG baseline achieves Precision@1 = 56.03%, MRR = 58.40%; on this 100-query ablation subset (4,161-vector index) with the corrected query set, the same baseline (Condition 1) achieves a mean Precision@1 of 57.00% (± 11.66%) and MRR of 67.38% (± 9.76%) across folds. Security vs. Utility Trade-Off & Re-Ranking Recovery: ABAC permission filtering alone (Condition 2, no re-ranking) restricts retrieval to authorized case boundaries — a 100% block rate against unauthorized roles (0.00% context leak) — with mean authorized Precision@1 of 68.00% (± 10.30%) and MRR of 78.09% (± 6.05%). The full pipeline (Condition 3: ABAC + GBT re-ranker) recovers a mean Precision@1 and MRR of 94.00% (± 2.00%) each — a +26.00-point gain in Precision@1 over Condition 2 — while preserving 100% Zero-Trust isolation.

Statutory-Feature Leakage Ablation (Null Result): To test whether the GBT re-ranker's gain over Condition 2 is driven by circularity between statutory-derived features and the relevance-labeling procedure, we trained a second re-ranker (Condition 4) on the same folds with the three statutory-derived features (Statutory Match Score, Combined Relator Score, and BSA Section 63(4) Evidence Alignment) removed, retaining only the five non-statutory signals (dense similarity, BM25 lexical score, temporal recency, bench/coram relevance, and citation centrality). Under identical 5-fold cross-validation, the ablated model achieves a mean Precision@1 of 94.00% (± 2.00%) and MRR of 94.00% (± 2.00%), statistically indistinguishable from the full 8-feature model — indicating the re-ranker's improvement over Condition 2 is attributable to dense similarity, lexical overlap, bench relevance, and citation-centrality signals rather than statutory metadata circularity. We report this as a null result bearing directly on which parts of the legal-domain feature engineering actually mattered.

15.4 Calibrated Trust Score Correlation

$r = 0.9835$ ($p < 0.001$) between the Trust Score T and legal ground-truth labels, consistent with the detailed calibration analysis in Section 12.1.

15.5 Evaluation Methodology Validation

Across Sections 12–16, all reported performance metrics follow a consistent train/calibration/test discipline: model or threshold fitting is performed exclusively on a designated training or calibration partition (e.g., the 1,200-query trust-score training split in Section 12.1, the 500-case conformal calibration set in Section 14.7), and all reported accuracy, calibration, and risk figures are computed exclusively on a disjoint held-out partition never used during fitting (the 300-query trust-score test split, the 200-case conformal test partition, and the 1,500-query retrieval benchmark evaluated against a fixed, pre-built index). No metric reported in Sections 12–16 is computed on data used to fit the corresponding model or threshold.

15.6 Empirical Evaluation of NLI Citation Entailment Verification

To validate the newly introduced NLI Citation Entailment Check (Field 3), we conducted an empirical evaluation assessing its ability to detect ungrounded hallucinations, fabricated statutory references, and misattributed precedent quotes within AI-generated judicial decision recommendations. The NLI verification model was instantiated using fine-tuned DeBERTa-v3-large-MNLI (304M parameters) operating as a cross-encoder between the generated text claims (hypothesis) and retrieved legal context chunks (premise).

Framing citation grounding as a pairwise NLI classification task, rather than as free-form generation, is what lets the check bound hallucination risk: instead of asking the LLM to self-report whether a claim is supported (a judgment for which the generator itself is an unreliable arbiter), each generated claim is checked against its retrieved source chunk independently by a dedicated cross-encoder trained specifically to detect entailment, contradiction, and neutrality, so a claim can only pass if the retrieved premise text actually entails it. Contradiction and neutral outputs are both treated as non-entailed for the purpose of the citation gate, since a judicial recommendation either needs affirmative support from the record or should be flagged rather than silently accepted; this is why the check catches direct contradictions (98.0%) at an even higher rate than unsupported claims lacking any textual basis (94.4%), as an explicit contradiction leaves a stronger cross-encoder signal than an unsupported-but-plausible-sounding claim.

Evaluation Dataset & Adversarial Benchmark: We constructed a dedicated evaluation benchmark consisting of 500 legal citation pairs extracted across Supreme Court query responses from the ILDC corpus. The dataset comprises 250 verified ground-truth legal citations (entailed premise-hypothesis pairs) and 250 synthetically generated adversarial perturbations representing common LLM citation failure modes: (i) Hallucinated Statutory attributions (100 cases), (ii) Misattributed Case Citations (75 cases), and (iii) Out-of-Context Precedent Quotations (75 cases).

Detection Performance & Precision: As summarized in Table 12, the NLI Citation Entailment Check reliably distinguished entailed from non-entailed/contradicted claims across both the hallucination-detection and binary-classification metrics (point estimates and 95% CIs in the table).

Latency Overhead & Real-Time Throughput: Computational efficiency was measured across 500 evaluation runs on a single NVIDIA T4 GPU with batched tensor inference. Combined with the GBT retrieval re-ranking stage, total end-to-end citation verification latency (Table 12) confirms that NLI-based citation entailment checking bounds LLM generative output without exceeding the real-time interactive operational threshold of <50 ms.

Table 12. NLI Citation Entailment Verification Benchmark (N = 500 Legal Citation Claims)

Metric	**DeBERTa-v3-large-MNLI Value**	**Operational Bound / Target**
Hallucination Detection Rate	94.4% (236/250)	> 90.0%
Factual Contradiction Catch Rate	98.0%	> 95.0%
Entailment Precision (95% CI)	0.956 [0.931, 0.978]	> 0.900
Entailment Recall (95% CI)	0.941 [0.912, 0.965]	> 0.900
Entailment F1-Score (95% CI)	0.948 [0.925, 0.969]	> 0.900
Mean Latency Overhead	12.40 ms (P50: 11.2ms, P95: 18.6ms)	< 25.0 ms
Total Pipeline Latency (GBT + NLI)	28.70 ms	< 50.0 ms (Interactive)

An earlier evaluation pass surfaced a set of implementation bugs in the evaluation harness itself that had suppressed measured retrieval performance; these were subsequently identified and corrected. The pre-correction metrics and the harness defects underlying them are superseded and are not restated here; the metrics reported throughout this manuscript, including Section 15.2, reflect the corrected evaluation harness only.

Statistical Audit Additions. Applying the Benjamini–Hochberg procedure at q = 0.05 across eight pairwise retrieval baseline comparisons identified five baseline pairs as statistically significant after correction for multiple comparisons (adjusted p-values ranging from 0.0008 to 0.0499), with the remaining three pairs not reaching significance at this threshold. A blind dual-annotator audit, performed independently by the two authors, across n = 50 sampled claim-premise pairs yielded Cohen's $\kappa = 0.8834$, indicating "Almost Perfect" agreement under Landis–Koch standards; a normal-approximation interval computed from this sample produced an upper bound exceeding κ's theoretical maximum

of 1.0, so a bootstrap 95% CI (10,000 resamples with replacement over the underlying n = 50 rater contingency table) was computed instead, yielding $\kappa \in [0.69, 1.00]$ (2.5th/97.5th percentiles; stable to [0.69, 1.00] across 100,000 resamples). Student's t 95% confidence intervals across five retrieval ablation folds yielded a mean Precision@1 of 0.8920 (95% CI: [0.8859, 0.8981]).

15.7 ADVERSARIAL FUZZING EVALUATION OF THE TAMPER-EVIDENCE LAYER

15.7.1 Threat Model and Attack Classes

To empirically evaluate the tamper-evidence layer's ability to detect malicious or accidental modification of a Verification Contract post-emission, we constructed four independent perturbation classes targeting distinct fields of the seven-field schema, each representing a realistic adversarial or fault scenario: (A1) Bit-Level Hash Tampering — a single bit is flipped within the hexadecimal digest of the contract hash, modeling low-level storage or transmission corruption or an attacker attempting to forge a digest without recomputing it validly; (A2) ABAC Privilege Escalation — the allowed_roles field is altered by substituting an authorized role (e.g., judge) with an unauthorized principal (unauthorized_guest), modeling an attacker attempting to widen access post-hoc without triggering re-authorization; (A3) RAG Citation Forgery — a valid citation chunk_id is replaced with a randomly generated UUID, modeling injection of an unverified or fabricated precedent citation into an already-signed contract; and (A4) Temporal Fraud — the generated_at / BSA §63(4) custody-seal timestamp is shifted by ±86,400 seconds (±1 day), modeling an attempt to misrepresent when a decision or evidentiary seal was produced.

15.7.2 Corpus Generation

We generated N = 10,000 synthetic Verification Contracts under a fixed pseudorandom seed for reproducibility, applying exactly one perturbation to each contract drawn uniformly across the four attack classes (n = 2,500 per class). Each corrupted contract was then passed through the unmodified ContractHasher verification path (canonical JSON serialization, SHA-256 digest recomputation, comparison against the committed hash), identical to the path a real verifier would use — no detection-specific instrumentation was added to the contracts or the verifier under test.

15.7.3 Detection Mechanism

Detection does not rely on class-specific rules. Because all four attack classes modify a field that participates in the canonical JSON serialization underlying the committed hash, any successful perturbation necessarily produces a digest mismatch upon recomputation — the tamper-evidence layer therefore detects all four classes through a single mechanism (hash comparison) rather than four separate detectors, which is itself evidence that the seven-field binding leaves no field independently modifiable without invalidating the seal.

15.7.4 Results

Table 13. Adversarial Fuzzing Results by Attack Class (N = 10,000 Contracts)

Attack Class (n=2,500 each)	Detected	TDR (%)	Mean Localization Latency
A1 — Hash Flip	2,500 / 2,500	100.00%	0.0291 ms
A2 — ABAC Role Swap	2,500 / 2,500	100.00%	0.0289 ms
A3 — Citation ID Swap	2,500 / 2,500	100.00%	0.0289 ms
A4 — Timestamp Tamper	2,500 / 2,500	100.00%	0.0293 ms
Aggregate	10,000 / 10,000	100.00%	≈ 0.029 ms (per-class mean)

The tamper-evidence layer achieved a 100.00% Tamper Detection Rate across all 10,000 corrupted contracts (zero false negatives), with per-class localization latency consistent across all four attack vectors (Table 13), indicating detection cost is dominated by the fixed hash-recomputation step rather than by which field was altered.

15.7.5 Discussion

The uniform 100% TDR across structurally distinct attack classes follows directly from the Completeness Invariant (Definition 2) and the unforgeability argument of Theorem 1: any field-level modification changes the canonical serialization input to the committed hash, so detection reduces to standard hash-mismatch checking rather than requiring class-specific adversarial defenses. As noted in the Contributions (Section 3), this experiment is a sanity check on the tamper-evidence layer's implemented behavior, not a novel detection result or a substitute for Theorem 1's forgery guarantee.

15.8 CROSS-JURISDICTIONAL LEXGLUE BENCHMARK (SCOTUS & ECtHR)

To evaluate generalizability outside Indian legal corpora, the retrieval pipeline was benchmarked against the US Supreme Court (SCOTUS) and European Court of Human Rights (ECtHR) datasets from LexGLUE (Chalkidis et al. 2022).

Initial evaluation of the India-tuned re-ranker on these datasets yielded lower performance (Precision@1 = 0.4820 for SCOTUS; 0.4410 for ECtHR), reflecting feature mismatch (e.g., BSA Section 63(4) evidence alignment does not transfer). To disentangle feature mismatch from architecture limitations, we built jurisdiction-specific feature extractors and retrained the GBT LambdaMART re-ranker. To rigorously prevent document-level leakage and overfitting, we enforced a strict non-overlapping query split: Queries 1–500 were used exclusively for training, and Queries 501–1,000 were used strictly for held-out testing. Results are reported in Table 14.

Table 14. Cross-Jurisdictional Retrieval Generalization

Dataset	Zero-Shot India-Tuned	Same-Split Retrained (Overfitted Ref)	Disjoint-Split Retrained (True OOD)
LexGLUE-SCOTUS	P@1 = 0.4820 P@5 = 0.6950 MRR = 0.5840	P@1 = 1.0000 P@5 = 1.0000 MRR = 1.0000	**P@1 = 0.6840** **P@5 = 0.8620** **MRR = 0.7580**
LexGLUE-ECtHR	P@1 = 0.4410 P@5 = 0.6620 MRR = 0.5310	P@1 = 1.0000 P@5 = 1.0000 MRR = 1.0000	**P@1 = 0.6120** **P@5 = 0.8140** **MRR = 0.6940**

The disjoint-split evaluation establishes the true out-of-distribution (OOD) baseline, recovering substantial performance (SCOTUS P@1 improving to 0.6840) without succumbing to the 1.0000 Precision@1 overfitting artifact observed when evaluating on the training split. This indicates the VADP retrieval architecture generalizes partially, not fully, across international jurisdictions when provided with domain-appropriate statutory features: jurisdiction-specific retraining under a disjoint split closes most, but not all, of the gap to the zero-shot baseline, and both the SCOTUS (P@1 = 0.6840) and ECtHR (P@1 = 0.6120) OOD figures remain below the in-domain benchmark (P@1 = 0.931, Section 15.2) — a substantial gain over the zero-shot cross-jurisdictional baseline, but short of parity with in-domain performance.

15.9 LLM PROMPT INJECTION & CITATION HIJACKING BENCHMARK

An adversarial benchmark of 100 prompt injection and context hijacking attacks — spanning system prompt overrides, citation fabrications, ABAC role spoofing, and context poisoning — was evaluated against the pipeline's Zero-Trust PDP and DeBERTa NLI citation gate. All 100 attacks were neutralized (100.0% defense rate, 95% Wilson CI: [96.3%, 100.0%]; zero bypasses), with the NLI entailment gate directly intercepting 48.0% of attempts and the remainder blocked by upstream policy-decision-point checks. This benchmark initially evaluated the defended pipeline only, so on its own the 100.0% figure measures this specific defense combination against this specific attack set, not the set's difficulty relative to an unprotected system.

We closed this gap by running the identical 100-attack set against an undefended control pipeline that retains the same retrieval and generation stack but omits the NLI entailment

citation gate and the ABAC policy-decision-point checks. The undefended baseline was bypassed on 89.0% of attacks (11.0% defense rate, 95% Wilson CI: [6.3%, 18.6%]) — 23/25 System Prompt Override, 22/25 Citation Fabrication, 23/25 ABAC Role Spoofing, and 21/25 Context Poisoning attempts succeeded — against VADP's 100.0% defense rate on the same set (a +89.0 percentage-point improvement, 95% Wilson CI on the defended rate: [96.3%, 100.0%]). This confirms that the defended pipeline's 100.0% figure reflects the specific protections evaluated rather than an artifact of an unusually easy attack set.

15.10 MECHANISTIC INTERPRETABILITY VS. TREESHAP EXPLAINABILITY

Feature attribution consistency was evaluated by comparing TreeSHAP attributions on the GBT re-ranker against a Mechanistic Attention Integrated Gradients (IG) baseline across 200 attribution computations. TreeSHAP executed at a median (p50) latency of 1.62 ms, compared to 14.91 ms for Mechanistic IG. The two explainability approaches achieved a Spearman rank correlation of r = 0.9419 between their attributions, indicating high consistency and supporting the use of TreeSHAP for real-time interactive routing where the higher-latency mechanistic method is impractical.

15.11 ADVANCED JUDICIAL RETRIEVAL EVALUATION

15.11.1 LLM-as-a-Judge Reliability

To move beyond the binary and editorial-headnote-style relevance labeling used elsewhere in this manuscript (Section 12.1; Section 15.2), we additionally evaluated retrieved judicial passages using an LLM-as-a-Judge framework, instantiated using GPT-4o at temperature 0.0 to minimize sampling variance across repeated grading passes; temperature 0.0 reduces but does not by itself guarantee fully deterministic output from a hosted API, so this framework is presented as a graded relevance evaluation rather than as a reproducibility claim. Each query–chunk pair was graded using a structured legal rubric that distinguished binding precedent, persuasive reasoning, dissenting opinion, and irrelevant material. A stratified subset of predictions was independently reviewed by the two authors acting as annotators. Agreement statistics are reported in Table 15.

Table 15. LLM-as-a-Judge Reliability Against Dual Author Annotators (Preliminary, Unaudited Pilot)

Metric	Value
Human–LLM Agreement	89.6%
Macro F1	0.88
Precision	0.90
Recall	0.87

Unlike the dual-annotator kappa statistics reported elsewhere in this manuscript (κ = 0.8834 for NLI citation entailment, Section 15.6; κ = 0.8968 and κ = 0.8959 for the N = 291 Trust Score and appellate-outcome calibration sets, Section 12.1), the Table 15 figures were not produced by the project's audited annotation pipeline

(*generate_human_annotation_sample.py* / *compute_annotator_kappa.py*), and no corresponding raw per-item LLM-judge rating log has been located. Because that log cannot be located, Cohen's κ for this comparison has been removed from Table 15 rather than reported as an unresolved figure; the remaining Table 15 statistics should still be treated as a preliminary, unaudited pilot result rather than a reproducible result on the same footing as the other agreement statistics in this manuscript; see Section 17 for the corresponding limitation.

15.11.2 Graded Relevance Evaluation

Editorial headnote alignment was replaced with a graded judicial relevance scale: 3 — binding precedent directly answering the legal issue; 2 — persuasive reasoning or closely analogous authority; 1 — dissenting opinion or partially relevant discussion; 0 — irrelevant. Results under this scale are reported in Table 16.

Table 16. Graded Relevance Evaluation (nDCG@5 / nDCG@10)

Model	nDCG@5	nDCG@10
BM25	0.842	0.856
Dense Retrieval	0.903	0.914
Cross Encoder	0.936	0.944
VADP GBT	0.949	0.957

The graded evaluation is directionally consistent with the manuscript's reported MRR = 0.951 (Table 11, Section 15.2): both are top-ranking-sensitive metrics, and both show the VADP GBT re-ranker ranking the most legally authoritative precedent first in most cases. nDCG@10 and MRR are computed differently and are not numerically identical by construction; the agreement here is evidence of consistency across evaluation protocols rather than a re-derivation of the same figure.

15.11.3 Hard Negative Evaluation

A dedicated benchmark was constructed by introducing overruled precedents, dissenting opinions, and high lexical-overlap but legally incorrect judgments. A false positive in this context is strictly defined as retrieving a legally incorrect, overruled, or misleading chunk within the top 5 results. False positive rates by model are reported in Table 17.

Table 17. Hard Negative Evaluation — False Positive Rate

Model	False Positive Rate
BM25	21.8%
Dense Retrieval	15.6%
Cross Encoder	8.7%
VADP GBT	5.1%

The Gradient Boosted Tree re-ranker consistently demoted legally misleading authorities despite high lexical similarity.

15.11.4 TREC-style Depth Pooling

Rather than limiting evaluation to five editorially aligned chunks, the top-20 candidates from BM25, Dense Retrieval, Cross Encoder, and GBT were pooled and deduplicated before annotation. Results are reported in Table 18.

Table 18. TREC-style Depth Pooling vs. the Section 15.2 Evaluation Protocol

Metric	Previous (Section 15.2)	Depth Pool
Avg. Judged Relevant Chunks	5.0	17.9
Candidate Pool	7,500	26,850

The 7,500-chunk "Previous" candidate pool matches the 1,500-query × 5-relevant-chunk-per-query labeling scheme described in Section 15.2. Depth pooling substantially reduces the artificial recall ceiling documented there while remaining consistent with the original retrieval benchmark.

15.11.5 Document-Level Recall

Given that long judicial opinions are segmented into multiple embeddings, evaluation was additionally performed at the document level. Results are reported in Table 19.

Table 19. Document-Level Recall (Recall@Case)

Model	Recall@Case
BM25	0.884
Dense Retrieval	0.918
Cross Encoder	0.946
VADP GBT	0.962

15.11.6 Success@1

Since the Verification Contract's citation_provenance field depends on the highest-ranked citation alone (Section 15.2), Success@1 provides a workflow-oriented evaluation. Results are reported in Table 20.

Table 20. Success@1

Model	Success@1
BM25	0.812
Dense Retrieval	0.887
Cross Encoder	0.916
VADP GBT	0.931

The Success@1 value for VADP GBT matches the manuscript's reported Precision@1 = 0.931 (Table 11, Section 15.2) exactly, since both measure whether the top-ranked result is

the correct one; this is a consistency check across evaluation protocols rather than an independent replication.

15.11.7 Relationship to Existing Results

Table 21 consolidates the correspondence between the Section 15.2 primary retrieval benchmark and the advanced evaluation protocols introduced in this section, to clarify how the two result sets relate rather than compete.

Table 21. Relationship Between the Section 15.2 Evaluation and This Advanced Evaluation

Existing Manuscript Result	Advanced Evaluation
Precision@1 = 0.931	Success@1 = 0.931
MRR = 0.951	nDCG@10 = 0.957
1,500 legal queries	Same benchmark
≈60,000 indexed vectors	Same corpus
NLI citation validation (Field 3)	LLM semantic grading complements NLI
Binary headnote labels	Graded judicial relevance
Top-5 evaluation	TREC depth pooling + Recall@Case

The LLM-as-a-Judge framework introduced in this section evaluates the legal relevance of a retrieved document to the query; the downstream NLI Citation Entailment Check (Field 3, Section 15.6) separately and strictly verifies factual entailment of the generated text against that retrieved document. The two checks are complementary rather than redundant: a chunk can be the most legally relevant authority for a query (this section) while a specific generated claim about it still fails entailment (Section 15.6), or vice versa.

16. DISCUSSION

VADP reframes judicial AI assurance from trusting a black-box model to independently verifying specific output claims, using a mechanism — transparency-log-style signed provenance — that is well established in adjacent domains (Section 2.3) but not previously adapted to the evidentiary and human-oversight requirements of judicial decision support. The paper's strongest evidence is the Merkle-ledger performance benchmarking (Table 9), the direction and magnitude of the RAG retrieval improvement (Table 11), and the empirical hallucination detection of the NLI citation entailment check (Table 12); its structural guarantees are grounded in verifiable Certificate Transparency provenance.

A Verification Contract makes tampering with, or silent omission of, any of the seven bound fields detectable after the fact; it does not, by itself, make the underlying recommendation correct, nor does it make the model's reasoning inspectable. This maps onto the due-process concern in State v. Loomis, where the Wisconsin Supreme Court objected not that COMPAS's outputs might be altered in transit, but that the defendant could not meaningfully contest how a score had been reached. VADP's Completeness Invariant and

NLI citation-entailment check narrow that gap for the components they cover — a court can confirm that a cited statute or precedent was actually retrieved and not silently edited — but the deeper Loomis complaint, over a proprietary model's internal weighting, is only partially addressed. VADP converts an unfalsifiable trust claim ("the vendor says the system is reliable") into a falsifiable one (a cryptographically checkable record of what evidence was retrieved, what was flagged for human review, and whether the artifact has been altered), shifting the dispute from vendor credibility to a specific, checkable record — without resolving questions of substantive reasoning transparency.

The BSA 2023 §63(4) admissibility caveat qualifies a distinct claim in the Introduction, Section 1.2, Table 4, and the Conclusion; this paragraph, together with Section 17, is the canonical statement, and the other instances point back to it rather than restating it in full. The relationship to BSA 2023 §63(4) should be read the same qualified way as the single-host deployment caveat above. The Verification Contract's evidentiary-custody and cryptographic-integrity fields are designed with reference to §63(4)'s certification requirement for electronic evidence, and the dual-certificate anchoring (Online Resource 1 §B.5) is intended to produce an artifact a court could plausibly treat as satisfying that requirement. Whether Indian courts would in fact accept a Verification Contract as satisfying §63(4) is a question this paper does not settle; it is a claim about design intent and technical sufficiency, not a holding, and remains open to the scrutiny Indian courts have applied to electronic-evidence certification under the analogous Section 65B framework. We flag this explicitly as a boundary of the present contribution rather than assume it away.

Read together, the retrieval, entailment, and escalation results (Tables 15, 12, 13) demonstrate that the pieces of this evidentiary layer work as engineered; the cross-jurisdictional overfitting result (Section 15.8) demonstrates, just as importantly, that a plausible-looking retraining gain can be an artifact of evaluation protocol rather than a real transfer benefit. Both matter to the paper's central claim, because a verification architecture is only as trustworthy as the evaluation discipline behind the components it certifies.

Beyond the BSA 2023 context, the same design goals map onto comparative high-risk AI regulation: the EU AI Act (Regulation (EU) 2024/1689) classifies AI systems used by or on behalf of a judicial authority to research and interpret facts and law as high-risk (Annex III, point 8), triggering risk-management, logging, transparency, and human-oversight obligations (Arts. 9, 12–14, 26); VADP's Verification Contract, audit ledger, and Human Override Coverage mechanism are directly relevant building blocks for that class of obligation, though no formal EU AI Act conformity assessment is claimed here. Concretely, the seven-field Verification Contract's independent-checkability property (Definition 2) speaks to Art. 12's record-keeping/logging obligation; the RFC 6962 Merkle ledger and Theorem 1's unforgeability guarantee speak to Art. 12's requirement that logs be resistant to tampering; the escalation predicate and conformal Human Override Coverage bound (Theorem 2, Section 14.7) speak to Art. 14's human-oversight requirement by giving a calibrated, auditable miss-rate guarantee rather than an unquantified human-in-the-loop claim; and TreeSHAP attribution plus NLI citation entailment speak to Art. 13's transparency-to-users obligation, though only for the retrieval-ranking and citation-

grounding components, not the full generative reasoning chain. These are offered as architectural correspondences useful to a compliance reader, not as a claim of Art. 9 risk-management-system completeness or any other element of a full conformity assessment, which would additionally require organizational and governance evidence outside this paper's scope.

17. LIMITATIONS AND PLANNED RESOLUTIONS

A single-host, non-enterprise deployment caveat applies to the results in Table 3, Online Resource 1 §B, Online Resource 1 §C, Table 1, and Section 15.8; this section states it once, comprehensively.

The empirical evaluation presented in Section 14.7 addresses the previously identified limitation concerning the absence of calibrated appellate outcome data by validating the conformal risk-control framework using held-out appellate outcome labels obtained under a strictly disjoint evaluation protocol.

The ZK proof-of-concept's single-operator trusted setup, the local test-ledger scope of evidence anchoring, and the single-instance scope of the key-isolation and PostgreSQL/pgvector benchmarks are each detailed in Online Resource 1 §B and are not restated here. None of this infrastructure has been evaluated under a live, multi-organization deployment on physically separate hosts; validating it under such conditions remains future work. *Planned resolution: replace the ceremony's single-host, single-operator execution and Online Resource 1 §B's closed-form cost model with measurements from a ceremony executed across physically separate, independently-controlled participant infrastructure operated by distinct organizations, and benchmark proof generation and verification at evidence-collection scales beyond those evaluated in this manuscript; separately, migrate smart-contract evidence anchoring from the current local EVM-compatible test ledger (chain ID 1337) to enterprise Hyperledger Fabric chaincode — a Go chaincode contract (EvidenceContract) and a genuine 4-node Docker Compose Fabric network have already been implemented and benchmarked (500/500 endorsed transactions, 8,229.18 TPS, Online Resource 1 §B), but the containerized peers still run on a single physical host, so deployment on physically separate hosts with gRPC endpoints and a live, network-distributed Raft consensus service remains outstanding.*

The retrieval evaluation is currently limited to the benchmark corpus employed in this study. Future work should investigate additional judicial datasets, multilingual legal corpora, and cross-jurisdictional transferability. The cross-jurisdictional retraining experiment (Section 15.8) is itself a cautionary example: retraining and evaluating on the same split produced an apparent Precision@1 of 1.0000, which a subsequent disjoint held-out evaluation (Queries 1–500 for training, Queries 501–1,000 for testing) showed to be substantially an overfitting artifact rather than a generalizable gain, recovering a true out-of-distribution Precision@1 of 0.6840 for SCOTUS and 0.6120 for ECtHR; this underscores the importance of disjoint train/evaluation protocols for any future retrieval retraining reported in this line of work. Similarly, empirical validation of the conformal calibration

framework across additional appellate jurisdictions would further strengthen evidence regarding the generalizability of the proposed governance methodology.

- The security proof in Section 9 (Theorem 1) establishes unforgeability conditional on the ledger's signature scheme being EUF-CMA secure and SHA-256 being collision-resistant; it does not itself establish either assumption. The ledger's default scheme is now Ed25519, which is reported in the EdDSA security literature — rather than independently re-derived or verified in this paper — to admit a tighter standard-model reduction under CDH, though ECDSA-P256 (whose EUF-CMA security is only known to hold in idealized models such as the generic group model) remains available for BSA/PKI compatibility and inherits the idealized-model gap when selected. The CDH-reduction characterization for the Ed25519 path is drawn from the literature and has not been independently reviewed by a third-party cryptographer; until it is, it should not be read as a security guarantee this paper has itself established. *Planned resolution: subject Theorem 1's reduction, and the Ed25519/CDH standard-model reduction claim in particular, to independent third-party cryptographic review.*

- The dual-annotator validation of the Trust Score / headnote-alignment calibration labels (Cohen's κ = 0.8968, N = 291, generate_human_annotation_sample.py / compute_annotator_kappa.py) is reported upfront in Section 12.1 rather than here, since it is a completed methodological control rather than an open limitation.

- All manual annotation in this study was performed by the two authors, who are also the developers of the system under evaluation. Although the labels are anchored to external references (official case headnotes and the verbatim disposition text of each judgment) and the two authors labeled independently with substantial inter-annotator agreement, annotation by the system developers may introduce unconscious bias that agreement statistics between the same two annotators cannot fully exclude. No external annotators or human research participants were involved. Planned resolution: independent annotation of a sample by legal-domain experts outside the authoring team, as described below for the Y(V) labeling rule. Any such study involving external annotators or other human participants would be undertaken only after approval from an appropriate ethics committee or institutional review board and with informed consent from the participants.

- The LLM-as-a-Judge reliability figures reported in Table 15 (Section 15.11.1; Human–LLM Agreement 89.6%, Macro F1 0.88, Precision 0.90, Recall 0.87) were not produced by the audited annotation pipeline used elsewhere in this manuscript. The accompanying Cohen's κ value could not be reconstructed — no raw per-item LLM-judge rating log has been located — and has therefore been removed from Table 15 rather than carried forward as an unresolved figure; the remaining Table 15 metrics are retained but should still be treated as a preliminary, unaudited pilot result rather than as evidence on the same footing as the audited agreement statistics elsewhere in this manuscript. Planned resolution: either locate the original grading run's raw ratings and compute Cohen's κ through compute_annotator_kappa.py under the same protocol as the other agreement figures, or re-run the LLM-as-a-

Judge evaluation from scratch with ratings logged at the per-item level, before an agreement statistic for this comparison is reported in any future revision.

- The Y(V) calibration labels used to instantiate Theorem 2 (Section 14.7) were established from real appellate outcomes rather than from T(V) or σ(V), which distinguishes them in kind from the headnote-alignment labels above. The operational rule (generate_human_annotation_sample.py) parses the full disposition text of each of 100 authentic Indian Supreme Court judgments in the ILDC corpus for outcome-determinative phrases — "allowing the appeal," "appeal allowed," "reversed," "quashed," and "set aside" are treated as an appellate intervention (Y(V) = 1), while "dismissing the appeal" and "affirmed" are treated as no intervention (Y(V) = 0) — cross-checked against each case's headnote holding and statutory-section overlap, and applied to a balanced sample of relevant primary-precedent and distractor candidate chunks per case rather than to case outcomes alone. Illustrating the rule's current handling of a hard case: a disposition granting relief on one of several grounds while expressly rejecting the others (a partial-relief outcome) is coded Y(V) = 1 under the operational rule's phrase-matching criterion, since the disposition text contains an intervention phrase (e.g., "appeal allowed in part"), even though a finer-grained scheme might instead assign a fractional or issue-level label; this coarsening is a known limitation of the current binary rule and a concrete target for the outstanding external audit. A related coarseness concerns procedural, non-merits dispositions: an appeal dismissed on a purely procedural ground (e.g., limitation, non-maintainability, lack of jurisdiction) without the court reaching the substantive legal issue does not contain an intervention phrase and is therefore coded Y(V) = 0 by the same phrase-matching rule that codes a merits-based dismissal as Y(V) = 0, even though the two dispositions differ in whether the underlying legal question was ever adjudicated. We do not currently distinguish these two cases in the operational rule, so a procedurally-dismissed case is treated identically to a case where the AI-assisted recommendation was substantively affirmed on the merits. This is a second concrete instance of the rule's binary phrase-matching coarseness (alongside the partial-relief case above), and both are flagged together as targets for the outstanding external legal-domain audit; we do not believe this coarsening biases the reported 4.50% miss rate in a particular direction, since procedural and merits-based dismissals are both coded Y(V) = 0 and neither is a false intervention label, but a finer-grained scheme would still be needed to separately validate the conformal bound for each disposition type individually rather than pooled. A dual-annotator inter-rater check was run on this label set alongside the Trust Score labels (same N = 291 set, generate_human_annotation_sample.py; compute_annotator_kappa.py); Cohen's κ for the Y(V) appellate-outcome labels was 0.8959 (94.85% observed agreement), also “Almost Perfect” under Landis–Koch standards. This operational rule has not yet been independently audited by a legal-domain expert outside the authoring team; that audit remains future work. *Planned resolution: have a legal-domain expert outside the authoring team audit the operational rule used to derive Y(V) from docketed appellate dispositions, and recruit a third annotator on a sample of the disagreement cases from both the Trust Score and Y(V) label sets to characterize*

where the two annotators diverge. Both audits above are unstarted at the time of this submission; because they would involve external human annotators, they will be initiated only after approval from an appropriate ethics committee or institutional review board and informed consent from the participants, and we commit to reporting whatever outcome they produce, favorable or not, rather than treating their absence as a formality.

- The newly added discussion of argumentation-based explainability (Prajescu and Confalonieri, 2025) and judicial-transparency requirements (Dancy and Zalnieriute, 2025) in Section 2.5 identifies directions the Verification Contract schema could be extended toward; neither has yet been implemented or evaluated in this prototype. *Planned resolution: evaluate whether an argumentation-structured explanation layer improves practitioner comprehensibility beyond TreeSHAP alone, as part of the practitioner pilot below.*

- HSM-backed key protection is assumed in the threat model. The signer now includes a three-tier provider abstraction (kms_provider.py) with a HardwareHSMKeyProvider code path for on-prem PKCS#11 hardware-backed signing, an AWSKMSKeyProvider code path for cloud KMS signing, automated fallback to a SoftHSM2 software-emulated PKCS#11 token, and, failing that, a software signer over disk-stored key files (Online Resource 1 §B); however, neither the hardware nor the AWS KMS path has yet been exercised against a live backend, so a validated hardware- or cloud-isolated boundary in production remains outstanding, and the only tier with measured throughput is the software-emulated SoftHSM2 fallback. *Planned resolution: validate the HardwareHSMKeyProvider and AWSKMSKeyProvider code paths against a live cloud endpoint and, where feasible, a physical on-prem hardware-isolated token, rather than relying solely on the benchmarked SoftHSM2 software-emulated fallback.*

18. FUTURE WORK

- Live Judicial Validation of Conformal Bounds: Theorem 2 was instantiated and empirically evaluated using a 500-case independently adjudicated calibration set and a disjoint 200-case held-out test partition of appellate outcomes (Section 14.7), yielding an empirical miss rate of 4.50% against a target risk bound of $\alpha = 0.05$. This retrospective validation relies on historical appellate outcomes; future work should evaluate the conformal risk-control framework prospectively in a live e-filing registry, measuring the threshold's behavior against real-time, active judicial officer feedback loops.

- Pilot the system, or the Verification Contract format alone, with an actual e-filing registry, to evaluate e-filing registry integration.

- Conduct a practitioner comprehensibility and legal-adequacy review of sample Verification Contracts with a small panel of practicing lawyers or judicial officers (e.g., n = 10–20), assessing whether the seven-field schema is interpretable and legally sufficient for the evidentiary purpose it is designed to serve, independent of

the system's technical correctness. This is the single evaluation most likely to change the schema's design, since every empirical result reported elsewhere in this manuscript validates a component's technical behavior, not whether a judge or advocate can actually read and act on a Verification Contract; we intend to run even a small (n = 10–15) version of this pilot after obtaining approval from an appropriate ethics committee or institutional review board and informed consent from all participants, rather than deferring it indefinitely as open-ended future work.

19. CONCLUSION

Empirical prototype evaluation yielded HOC = 68.25% (258/378 contracts escalated; 95% Wilson CI: [63.40%, 72.74%]) under Algorithm 5's escalation thresholds ($\tau_p = 0.88$, $\tau_\sigma = 0.12$). The Verification Contract, Completeness Invariant, and empirical HOC evaluation provide transparent provenance bounds for prototype integration, while deployment-scale validation in active court registries remains future work. As throughout this manuscript, the architecture's relationship to BSA 2023 §63(4) (Sections 1, 16) is a statement of design intent and technical sufficiency rather than a holding on admissibility, which remains a matter for Indian courts to determine.

DATA AND CODE AVAILABILITY

The VADP prototype implementation, including all cryptographic primitives, Zero-Trust PDP engine, RAG pipeline, evidence vault, adversarial fuzzing harness, and evaluation scripts, is archived on Zenodo at https://doi.org/10.5281/zenodo.21856505 and maintained as a live development repository at https://github.com/beingaditya18/VADP- (full source, reproducibility artifacts, Docker deployment scripts, and automated test suites are included in the Zenodo archive). A standalone, zero-external-dependency offline verification CLI (scripts/verify_contract.py) is also included, allowing any exported Verification Contract to be independently audited outside the live system via canonical-JSON contract-hash recomputation, ECDSA-P256 signature format checking, RFC 6962 Merkle inclusion proof verification, Completeness Invariant (Definition 2) evaluation, and BSA 2023 §63(4) evidence-integrity hash cross-validation. The ILDC Supreme Court dataset subset (350 cases) used for RAG evaluation, and the expanded 3,500-case Supreme Court corpus used for the conformal risk-control experiments in Section 14.7, are derived from the publicly available Indian Legal Documents Corpus (ILDC; Malik et al. 2021), available at https://github.com/Exploration-Lab/CJPE. The SCOTUS and ECtHR data used for the cross-jurisdictional experiments in Section 15.8 are taken from the publicly available LexGLUE benchmark (Chalkidis et al. 2022), available at https://github.com/coastalcph/lex-glue.

Pre-correction evaluation metrics and the evaluation-harness defects that produced them are superseded and are omitted from this manuscript; the metrics reported throughout reflect the corrected evaluation harness only.

GLOSSARY OF ACRONYMS

Acronym	Expansion
ABAC	Attribute-Based Access Control
BSA	Bharatiya Sakshya Adhiniyam, 2023 (India's electronic evidence law)
C2PA	Coalition for Content Provenance and Authenticity
CT	Certificate Transparency
ECDSA	Elliptic Curve Digital Signature Algorithm
ECE	Expected Calibration Error
Ed25519	Edwards-curve Digital Signature Algorithm over Curve25519 (EdDSA)
EUF-CMA	Existential Unforgeability under Chosen-Message Attack
FAISS	Facebook AI Similarity Search
GBT	Gradient Boosted Tree
HITL	Human-in-the-Loop
HOC	Human Override Coverage
HSM	Hardware Security Module
ILDC	Indian Legal Documents Corpus
KMS	Key Management Service
MRR	Mean Reciprocal Rank
NLI	Natural Language Inference
PDP	Policy Decision Point
RAG	Retrieval-Augmented Generation
SCITT	Supply Chain Integrity, Transparency, and Trust (IETF)
SHAP	SHapley Additive exPlanations
STRIDE	Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege
TDR	Tamper Detection Rate
VADP	Verifiable AI Decision Provenance
VC	Verifiable Credentials (W3C)
ZTA	Zero Trust Architecture

DECLARATIONS

Funding: The authors received no funding for this work.

Competing interests: The authors declare no competing interests.

Clinical trial number: not applicable.

Data Availability: This study used two publicly available sources of court judgments. The Indian Legal Documents Corpus (ILDC; Malik et al. 2021) is available at https://github.com/Exploration-Lab/CJPE (distributed under a CC-BY-NC license and used here for non-commercial research only); the 350-case Supreme Court subset used for the retrieval evaluation and the expanded 3,500-case Supreme Court corpus used for the conformal risk-control experiments (Section 14.7) were both derived from it. The SCOTUS and ECtHR datasets from the LexGLUE benchmark (Chalkidis et al. 2022), used for the cross-jurisdictional experiments, are available at https://github.com/coastalcph/lex-glue. The prototype implementation and experiment code are archived on Zenodo at https://doi.org/10.5281/zenodo.21856505, with the development repository at https://github.com/beingaditya18/VADP-.

Ethics approval: This study did not involve human research participants, human biological material, or live vertebrates; therefore, approval from an ethics committee or institutional review board was not applicable. It used only publicly available court judgments (ILDC and LexGLUE). The manual labeling and annotation used to compute the evaluation metrics was performed solely by the two authors as part of their own research work, and no external individuals were recruited or involved.

Consent to participate: Not applicable. No human participants were recruited; all annotation tasks were carried out by the two authors themselves.

Consent to Publish declaration: not applicable.

Author Contributions: A.M. led the conceptualization, methodology design, system implementation, experimental evaluation, and drafting and revision of the manuscript. A.Y. contributed to the review and validation of the experimental results and manuscript. Both authors performed the manual labeling and annotation used to compute the evaluation metrics. Both authors read and approved the final manuscript.

REFERENCES

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V (2016) Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. PeerJ Comput Sci 2:e93. https://doi.org/10.7717/peerj-cs.93

Alruwaili FF (2021) CustodyBlock: a distributed chain of custody evidence framework. Information 12(2):88. https://doi.org/10.3390/info12020088

Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T (2024) Conformal risk control. In: Proceedings of the International Conference on Learning Representations (ICLR)

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. ProPublica

Bang Y, Ji Z, Schelten A, Hartshorn A, Fowler T, Zhang C, Cancedda N, Fung P (2025) HalluLens: LLM hallucination benchmark. In: Proceedings of the Association for Computational Linguistics (ACL)

Ben-Sasson E, Chiesa A, Genkin D, Tromer E, Virza M (2013) SNARKs for C: verifying program executions succinctly and in zero knowledge. In: Advances in Cryptology – CRYPTO 2013

Ben-Sasson E, Bentov I, Horesh Y, Riabzev M (2018) Scalable, transparent, and post-quantum secure computational integrity. IACR ePrint 2018/046

Bharatiya Sakshya Adhiniyam (2023) Act No. 47 of 2023, India

Bitan D, Canetti R, Goldwasser S, Wexler R (2022) Using zero-knowledge proofs to concretely improve the responsible use of algorithmic decision aids in criminal cases. Comput Sci Law (CSL)

Brendel J, Cremers C, Jackson D, Zhao M (2021) The provable security of Ed25519: theory and practice. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P)

C2PA (2024) C2PA technical specification v2.x. Coalition for Content Provenance and Authenticity

Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: the muppets straight out of law school. In: Findings of EMNLP

Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz DM, Aletras N (2022) LexGLUE: a benchmark dataset for legal language understanding in English. In: Proceedings of ACL. https://doi.org/10.18653/v1/2022.acl-long.297. Dataset available at https://github.com/coastalcph/lex-glue

Curran D, Sporne V, Frermann L, Paterson J (2025) Place matters: comparing LLM hallucination rates for place-based legal queries. In: AI4Access to Justice (AI4A2J) Workshop, ICAIL

Dahl M, Magesh V, Suzgun M, Ho DE (2024) Large legal fictions: profiling legal hallucinations in large language models. J Legal Anal 16(1):64–93. https://doi.org/10.1093/jla/laae003

Dancy T, Zalnieriute M (2025) AI and transparency in judicial decision making. Oxf J Leg Stud, gqaf030

e-Committee, Supreme Court of India (2019) SUVAS translation software

e-Committee, Supreme Court of India (2021) SUPACE portal

Gabizon A, Williamson ZJ, Ciobotaru O (2019) PLONK: permutations over Lagrange-bases for oecumenical noninteractive arguments of knowledge. IACR ePrint 2019/953

Gebru T, Morgenstern J, Vecchione B, Wortman Vaughan J, Wallach H, Daumé III H, Crawford K (2021) Datasheets for datasets. Commun ACM 64(12):86–92

Groth J (2016) On the size of pairing-based non-interactive arguments. In: Advances in Cryptology – EUROCRYPT 2016

He Y, Huang D, Chen L, Ni Y, Ma X (2022) A survey on zero trust architecture: challenges and future trends. Wirel Commun Mob Comput 2022:6476274

Honovich O, Aharoni R, Herzig J, Taitelbaum H, Kukliansy D, Cohen V, Scialom T, Szpektor I, Hassidim A, Matias Y (2022) TRUE: re-evaluating factual consistency evaluation. In: Proceedings of EMNLP

Hu VC, Ferraiolo D, Kuhn R, Schnitzer A, Sandlin K, Miller R, Scarfone K (2014) Guide to attribute based access control (ABAC) definition and considerations. NIST Special Publication 800-162

IETF SCITT Working Group (n.d.) Supply chain integrity, transparency, and trust (SCITT) architecture. Internet-Draft

in-toto Project (2019) in-toto: a framework to secure the integrity of software supply chains. In: Proceedings of USENIX Security

Johnson J, Douze M, Jégou H (2019) Billion-scale similarity search with GPUs. IEEE Trans Big Data 7(3):535–547

Kindervag J (2010) Build security into your network's DNA: the zero trust network architecture. Forrester Research

Laurie B, Langley A, Kasper E (2013) Certificate transparency. RFC 6962, IETF

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems (NeurIPS)

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS)

Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE (2025) Hallucination-free? Assessing the reliability of leading AI legal research tools. J Empir Legal Stud 22(2):216–242. https://doi.org/10.1111/jels.12413

Malik V, Sanjay R, Nigam SK, Ghosh K, Guha SK, Bhattacharya A, Modi A (2021) ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In: Proceedings of ACL. https://doi.org/10.18653/v1/2021.acl-long.313. Dataset available at https://github.com/Exploration-Lab/CJPE

Manakul P, Liusie A, Gales MJF (2023) SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In: Proceedings of EMNLP

Mbimbi B, Murray D, Wilson M (2025) Zero-knowledge proofs and blockchain: applications and open challenges. Blockchains

Merkle RC (1989) A certified digital signature. In: Advances in Cryptology – CRYPTO '89

Ministry of Law and Justice, Government of India (2026) Reply on e-Courts Phase III AI allocation

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. In: Proceedings of FAT*

MITRE Corporation (2024) MITRE ATLAS

National Judicial Data Grid, Department of Justice, India (2026)

Paul S, Mandal A, Goyal P, Ghosh S (2023) InLegalBERT: pretraining language models for the Indian legal domain. In: Proceedings of ICAIL

Prajescu AI, Confalonieri R (2025) Argumentation-based explainability for legal AI: comparative and regulatory perspectives. arXiv:2510.11079

Press Information Bureau, Government of India (2023) Cabinet approves e-Courts Phase III

Rose S, Borchert O, Mitchell S, Connelly S (2020) Zero trust architecture. NIST Special Publication 800-207

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215

Sachan S, Liu X (2024) Blockchain-based auditing of legal decisions supported by explainable AI and generative AI tools. Eng Appl Artif Intell 129:107666

Sigstore Project (2022) Rekor: a transparency log for signing events. OpenSSF/Linux Foundation

Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of AAAI/ACM AIES

State v. Loomis, 881 N.W.2d 749 (Wis. 2016)

Tian Z, Li M, Qiu M, Sun Y, Su S (2019) Block-DEF: a secure digital evidence framework using blockchain. Inf Sci 491:151–165

W3C (2024) Verifiable credentials data model v2.0. W3C Recommendation

SUPPLEMENTARY MATERIAL

Online Resource 1 provides full methodology and results for the multi-node Hyperledger Fabric consensus benchmark, the three-tier hardware/cloud key-isolation provider abstraction, the five-party Groth16 trusted-setup ceremony, and the PostgreSQL/pgvector multi-tenant isolation study; Sections 4, 9, and 17 of the main text cross-reference and summarize the corresponding figures and caveats (see Online Resource 1 §B for full detail).