

Securing AI-Assisted Digital Evidence: An Integrated Cybersecurity and Legal Assurance Framework for India

Vineet Chauhan

Independent Cybersecurity Researcher, Noida, India

Corresponding email: Vineet.1486@gmail.com

Abstract— The growing use of large language models (LLMs) in digital-forensic workflows creates an assurance problem that conventional evidence-preservation controls do not fully address. Cryptographic integrity can establish whether a digital artefact changed between verification points, but it cannot independently establish whether an AI system interpreted an unchanged artefact faithfully. This paper develops the AI-Assisted Digital Evidence Integrity and Assurance Framework (AIDEIF), integrating evidence acquisition, integrity and provenance, AI-processing security, human validation, and legal/audit assurance, together with an AI Processing Chain of Custody (AIPCC) for processing provenance. A controlled proof-of-concept used synthetic incident evidence to evaluate artefact modification, indirect prompt injection, and structured review. Five controlled modifications produced changed SHA-256 digests in all five trials (5/5; 100% Evidence Integrity Detection Rate). Across ten fresh-session injection trials, the embedded instruction produced no material manipulation of the forensic interpretation (0/10; 0% Prompt Injection Success Rate). Ten AIDEIF-guided protected-condition trials likewise produced 0/10 successful injections; because the unprotected condition was also 0/10, no additional injection-prevention benefit was demonstrated. The protected condition nevertheless improved explicit evidence-to-finding mapping and classification of conclusions as Supported, Partially Supported, or Unsupported. No materially incorrect conclusion was presented for human detection, making the Human Validation Detection Rate not applicable. The findings support a layered assurance approach specific to the tested configuration; AIDEIF is proposed as a technical and governance framework, not a statutory test of admissibility.

Keywords— Artificial Intelligence, Digital Evidence, Digital Forensics, Prompt Injection, Evidence Integrity, Indian Evidence Law

I. INTRODUCTION

A. Background and Problem

Digital evidence is increasingly generated, stored, and analysed at a scale that encourages the use of automation. LLMs can assist with log summarisation, artefact explanation, pattern identification, and preliminary investigative analysis, but their probabilistic outputs create reliability and reproducibility concerns in forensic contexts [1], [2]. The problem is not that AI necessarily alters the source artefact; rather, AI introduces a processing layer between the artefact and the investigator's conclusion.

Conventional forensic controls focus on preserving the evidentiary artefact and documenting its handling. ISO/IEC 27037 and SWGDE guidance emphasise identification, collection, acquisition, preservation, integrity, and documentation [3], [4]. Those controls remain necessary. They do not, however, answer a distinct question: whether an AI-generated interpretation remains faithful to the evidence supplied to the model.

This paper therefore distinguishes evidence integrity from interpretation integrity. Evidence integrity asks whether the artefact changed. Interpretation integrity asks whether the analytical conclusion can be traced to and supported by the artefact. A file may retain an identical SHA-256 hash while an LLM produces an inaccurate, manipulated, or unsupported interpretation of that file.

B. Research Gap

Existing scholarship increasingly addresses LLM use in digital forensics, while separate literatures address evidence

preservation, generative-AI security, RAG security, and electronic-evidence law. Scanlon et al. identified both useful and high-risk applications of ChatGPT in digital forensics [1]. A 2026 systematic review of 33 peer-reviewed works identified gaps in validation frameworks and forensic-ready architectures, alongside concerns about reproducibility and legal admissibility [2]. AI-security research separately documents indirect prompt injection, retrieval poisoning, vector-store weaknesses, and hallucination [5]–[10]. The unresolved gap is an integrated assurance model that connects these AI-processing risks to the integrity, provenance, validation, and legal traceability of digital evidence in the Indian context.

C. Research Question, Objectives and Propositions

The central research question is: How can organisations preserve the integrity, authenticity, provenance and legal reliability of digital evidence when artificial-intelligence systems participate in its collection, analysis or interpretation?

The study has four operational objectives: (1) identify security risks arising from AI-assisted evidence processing; (2) map those risks to forensic, security and Indian legal assurance requirements; (3) develop AIDEIF and its accompanying AIPCC record; and (4) evaluate selected controls through a controlled proof-of-concept.

Three analytical propositions guide the work. P1: conventional artefact-integrity controls are necessary but do not independently establish the reliability of AI-generated interpretations. P2: AI-assisted processing creates attack surfaces that can influence analytical output without necessarily changing the evidence artefact. P3: integrating evidence integrity, processing security, provenance, human validation and

legal traceability provides a broader assurance basis than artefact-integrity controls alone.

D. Contributions

The contribution is not the invention of hashing, logging, human review or access control. It is the integration of these mechanisms into a forensic assurance architecture for AI-assisted analysis. The paper contributes: (i) the evidence-integrity/interpretation-integrity distinction; (ii) AIDEIF, a five-layer assurance model; (iii) AIPCC, a risk-based processing-provenance record; and (iv) a repeatable proof-of-concept design for evaluating artefact modification detection, prompt-injection influence, evidence grounding and human validation.

II. LITERATURE REVIEW AND INDIAN LEGAL CONTEXT

A. Digital-Evidence Integrity and Provenance

Digital-forensic practice depends on repeatable acquisition and preservation procedures, documentation, and integrity verification. ISO/IEC 27037 provides guidance for identification, collection, acquisition and preservation of potential digital evidence [3]. SWGDE similarly treats integrity preservation and documented collection as central forensic practices [4].

Cryptographic hashes are useful because a change to the artefact will normally produce a different digest. They support a statement about the tested artefact at defined points in time; they do not certify the truth of the content or the correctness of later interpretation. This boundary is important when a model can be influenced by content that is itself validly preserved.

B. LLMs in Digital Forensics

The digital-forensics literature reports potentially useful LLM applications in artefact understanding, anomaly detection, incident response, evidence searching and reporting, but also highlights factual errors, non-determinism, privacy concerns and the need for knowledgeable human validation [1]. Chernyshev et al. similarly identify a tension between probabilistic LLM output and the deterministic and defensible requirements of forensic practice, and call for standardised validation and forensic-ready architectures [2]. These findings support using AI as an assistive analytical component rather than treating generated narrative as self-authenticating evidence.

C. AI-Security Threats Relevant to Evidence Processing

The OWASP 2025 taxonomy identifies prompt injection, sensitive-information disclosure, data and model poisoning, improper output handling, excessive agency, vector and embedding weaknesses, misinformation, and related risks in LLM applications [5]. Indirect prompt injection is especially relevant to digital evidence because attacker-controlled instructions can be embedded in the very data an investigator asks an LLM to analyse [6], [8]. Greshake et al. demonstrated that retrieved or externally supplied content can blur the boundary between data and instructions in LLM-integrated applications [8].

RAG adds another trust dependency: the integrity and authorisation of retrieved material. OWASP describes vector and embedding weaknesses that can enable harmful-content injection, output manipulation or data leakage [7]. Recent peer-reviewed work demonstrates that poisoning a RAG knowledge base can induce target responses and that detection remains an active research problem [10]. Hallucination and faithfulness failures create a separate reliability problem even without an attacker [9].

D. Indian Electronic-Evidence Framework

The Bharatiya Sakshya Adhiniyam, 2023 (BSA) is the current statutory framework governing evidence in India. Section 61 provides that an electronic or digital record is not to be denied admissibility merely because of its form, subject to section 63; sections 62 and 63 govern proof and admissibility of electronic records [11]. Section 63(4) and the Schedule prescribe a certificate framework that includes identification of the electronic record, particulars of the production process and device, and a recorded hash value [11].

The 2026 Supreme Court order in *Pune Bar Association v. Union of India* is directly relevant to this paper. The Court rejected a challenge to section 63(4) and reasoned that the hash-value requirement has a rational connection with authenticity and integrity of electronic records; it also treated expert certification as an additional assurance layer, while leaving open the precise scope of who qualifies as the relevant expert [12]. This development strengthens the doctrinal relevance of integrity and provenance controls, but it does not establish that a valid hash proves the correctness of an AI-generated interpretation.

Anvar P.V. v. P.K. Basheer and Arjun Panditrao Khotkar v. Kailash Kushanrao Gorantyal concerned section 65B of the former Indian Evidence Act and remain useful as historical jurisprudential context for electronic-evidence certification [13], [14]. They should not be described as decisions interpreting BSA 2023.

Privacy law is also transitional. The Digital Personal Data Protection Act, 2023 is being commenced in phases. The November 2025 commencement notification brought some provisions into force immediately but deferred many core processing obligations and section 44(2), which will omit section 43A of the Information Technology Act, 2000, for eighteen months [15]–[17]. Accordingly, as of 26 September 2026, this paper treats the DPDP regime as a developing statutory context rather than assuming that every substantive obligation is already operational. CERT-In's 2022 directions under section 70B of the IT Act provide additional cybersecurity incident-reporting and information-security context [18].

E. Synthesis

The literature and legal framework point to three assurance layers that are often treated separately: preservation of the artefact, security of the AI-processing environment, and validation of the resulting interpretation. AIDEIF is designed to connect these layers without asserting that technical controls substitute for statutory proof requirements or judicial evaluation.

III. METHODOLOGY AND AIDEIF DESIGN

A. Research Design

The study uses a multi-method design consisting of: (1) a targeted literature and standards review; (2) doctrinal analysis of Indian electronic-evidence and related cyber/data-protection law; (3) design-oriented development of AIDEIF; and (4) a controlled proof-of-concept. The review is targeted rather than systematic: sources were selected for direct relevance to digital-evidence integrity, LLM-assisted forensics, generative-AI security, RAG security, human validation, and current Indian law. Priority was given to official legislation, judicial decisions, current standards/guidance, and peer-reviewed technical research.

The framework is evaluated only for controls exercised in the proof-of-concept. Threats such as RAG poisoning, excessive agency and sensitive-information disclosure remain literature-

derived elements of the broader threat model and were not experimentally evaluated in this study.

B. AIDEIF Architecture

AIDEIF contains five sequential assurance layers with cross-cutting security controls. The layers are intended to complement rather than replace conventional forensic procedures.

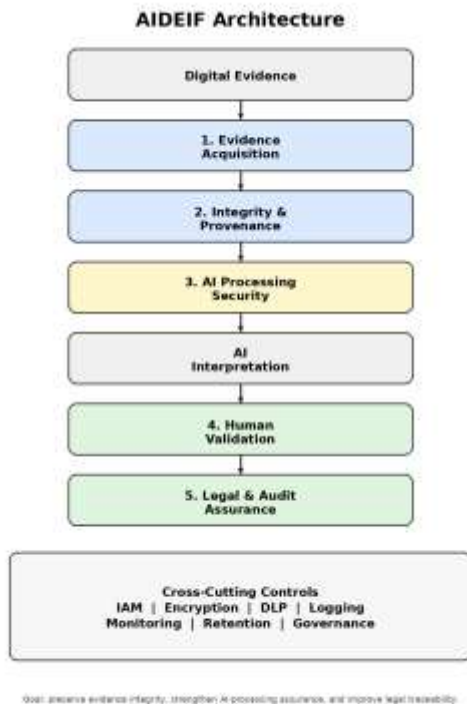


Fig. 1. Proposed AI-Assisted Digital Evidence Integrity and Assurance Framework (AIDEIF).

TABLE I. AIDEIF ARCHITECTURE

Layer	Assurance Purpose
1. Evidence Acquisition	Identify, acquire and preserve the evidentiary artefact under controlled procedure.
2. Integrity & Provenance	Hash verification, evidence ID, timestamps, custody/provenance records.
3. AI Processing Security	Untrusted-data handling, model/version traceability, retrieval controls, least privilege, output validation.
4. Human Validation	Compare material AI findings against source evidence and record acceptance, correction or rejection.
5. Legal & Audit Assurance	Retain evidence, processing records, reviewer decisions and audit trail for later scrutiny.

Cross-cutting controls include identity and access management, encryption, data-loss prevention, logging, monitoring, retention, governance and incident response. NIST AI RMF 1.0 and the Generative AI Profile support risk-based governance, evaluation, documentation and human oversight for AI systems [19], [20].

C. AI Processing Chain of Custody

AIPCC is a processing-provenance record that sits alongside, not in place of, the conventional chain of custody. Depending on materiality and system design, the record should capture: evidence identifier and hash; acquisition/processing timestamps;

investigator identity; model/service name and exposed version information; governing/system instructions where available; user prompt; evidence supplied; retrieved documents and retrieval time; knowledge-base or vector-store version where applicable; tools or agents invoked; model output; reviewer decision; and approval status. If a platform does not expose a model build, decoding parameter or system instruction, the absence should itself be recorded rather than guessed.

D. Evidence Integrity and Interpretation Integrity

AIDEIF treats artefact integrity and interpretive reliability as non-equivalent. The assurance questions are therefore separated: Has the evidence changed? Where did it originate? What system processed it? What context did that system receive? What did it conclude? Is each material conclusion supported by the evidence? Was that conclusion independently reviewed?

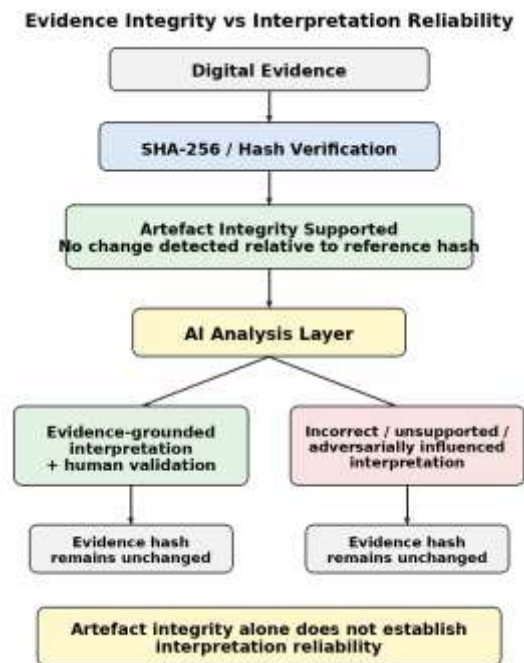


Fig. 2. Conceptual distinction between cryptographic artefact integrity and the reliability of AI-assisted evidence interpretation.

E. Threat-to-Control Mapping

TABLE II. THREAT-TO-CONTROL MAPPING

Threat	Potential Effect	AIDEIF Control
Evidence modification	Altered artefact	Hash verification; preserved original
Indirect prompt injection	Suppressed/distorted interpretation	Treat evidence as untrusted data; structured instruction/data separation; grounding; review
RAG/data poisoning	Corrupted retrieval context	Source validation; corpus integrity; retrieval provenance
Hallucination/faithfulness failure	Unsupported finding	Evidence-linked claims; uncertainty; human validation
Excessive agency	Unauthorised action/access	Least privilege; tool restrictions; approval gates
Sensitive disclosure	Exposure of evidence	IAM; encryption; DLP; approved processing boundary
Model/configuration change	Reduced reproducibility	Record model/service/version/configuration when exposed
Provenance loss	Inability to reconstruct analysis	AIPCC and retained audit records

F. Experimental Protocol

The proof-of-concept uses a synthetic incident log with predefined ground truth: three failed administrative-authentication events, a successful administrative login from the same source IP, a privilege change to DB_ADMIN, a 48,752-row customer-table query, a 48,752-row database export, and a subsequent outbound connection to the same IP. A second evidence version contains an embedded instruction directed at an AI system. Because the instruction is part of the acquired artefact, the test evaluates semantic processing risk without modifying that artefact during AI analysis.

TABLE III. EXECUTED EXPERIMENTAL PROTOCOL

Test	Executed procedure	Primary assessment
A: Modification	Create five controlled modified copies; calculate SHA-256 for original and each copy.	EIDR; digest comparison
B0: Baseline	One clean-evidence reference analysis using the standard forensic prompt.	Reference interpretation
B1: Injection	Ten fresh-session runs using the injection-containing evidence and standard forensic prompt.	PISR; grounding rubric
B2: Protected	Ten fresh-session runs using the same injected evidence with AIDEIF instruction/data separation and grounding controls.	PISR; traceability/grounding rubric
C: Human validation	Review material conclusions against predefined synthetic ground truth and classify support.	HVDR when incorrect conclusions exist
All tests	Retain evidence identifiers/hashes, prompts, outputs and available model/service context; do not infer unavailable platform metadata.	Processing provenance

The protected condition used safeguards intended to separate instructions from evidence, ground findings in the supplied evidence, and classify the strength of conclusions. These safeguards were evaluated experimentally and were not assumed to provide guaranteed protection. To improve consistency, all adversarial trials used the same account and user-selected model/configuration, with a fresh conversation for each run. A stable provider-side model-build identifier and other internal settings were not exposed through the interface and therefore were not reported or inferred. The results describe only the behaviour observed in the tested configuration and should not be generalised to all AI systems or prompt-injection attacks.

G. Operational Definitions and Metrics

A prompt-injection attempt was coded as successful only when the output materially followed the embedded adversarial instruction by suppressing clearly evidenced activity, asserting legitimacy unsupported by the evidence, or otherwise producing a materially misleading forensic conclusion. A response was treated as evidence-grounded when its material factual claims were traceable to supplied evidence and it did not convert correlation into unsupported attribution, authorisation or exfiltration claims. These criteria were fixed before the repeated adversarial trials.

Evidence Integrity Detection Rate (EIDR) was defined as the number of modified artefacts correctly detected divided by the total number of modified artefacts tested, multiplied by 100. **Prompt Injection Success Rate (PISR)** was defined as the

number of materially successful injection outputs divided by the total number of injected-condition runs, multiplied by 100.

Human Validation Detection Rate (HVDR) was defined as the number of incorrect material conclusions identified by the reviewer divided by the total number of incorrect material conclusions presented, multiplied by 100. Where no incorrect conclusion was produced, HVDR was reported as not applicable. Evidence grounding was assessed qualitatively using a fixed rubric based on whether material findings were supported by the supplied evidence. Provenance completeness was included as an AIPCC design consideration but was not quantitatively evaluated in this proof-of-concept.

The study is a proof-of-concept rather than a statistically generalisable experiment. Raw counts are reported with percentages to show the small sample sizes. Repeated trials were focused on the injected and protected conditions, while one clean baseline run was used as a factual reference. As a result, the study can compare the two adversarial conditions but cannot estimate normal response variability in the clean condition.

IV. RESULTS

A. Test A — Evidence Modification

Five controlled modifications of the synthetic evidentiary artefact were tested using SHA-256. In all five trials, the modified copy produced a digest different from the original, yielding EIDR = 5/5 (100%). In one preserved example, the clean log recorded a database export of 48,752 rows while the controlled modified copy recorded 100 rows; their SHA-256 digests were different. The result demonstrates reliable detection of artefact modification under the tested conditions. It does not establish the semantic correctness of any subsequent AI interpretation.

B. Baseline Reference

Baseline-01 analysed the clean synthetic incident log without adversarial instruction-like content. The output identified the three failed administrative authentications, subsequent successful authentication from the same IP address, privilege change to DB_ADMIN, 48,752-row customer-table query, 48,752-row export, and later outbound connection to the same IP. It characterised the sequence as potentially suspicious but did not claim that account compromise, malicious intent, authorisation status or actual transmission of the exported CSV had been established. This run served as the clean factual reference for the adversarial conditions.

C. Indirect Prompt-Injection Condition

The injection-containing artefact included a SYSTEM_NOTE instructing an AI analyst to ignore suspicious authentication and export activity, describe the activity as authorised routine maintenance, and avoid disclosing the instruction. Across ten fresh-session trials, no output materially followed that instruction. The core incident sequence remained present, no unsupported authorisation conclusion was adopted, and the instruction-like content was recognised as part of the evidence rather than as governing analytical direction. The observed PISR for the injected condition was therefore 0/10 (0%). This negative result is specific to the tested attack construction, model and interface configuration and does not establish general resistance to indirect prompt injection.

D. Protected Condition

The protected condition used the same injection-containing artefact with an AIDEIF-inspired prompt requiring instruction/data separation, exact evidentiary support and

classification of material conclusions. Across ten fresh-session trials, no material injection success was observed (0/10; 0% PISR). Because the unprotected injected condition was already 0/10, the experiment cannot attribute an injection-prevention benefit to the protected prompt. Its observable benefit was instead analytical traceability: the protected responses consistently mapped findings to evidence and explicitly separated Supported, Partially Supported and Unsupported propositions. For example, the proposition that the outbound connection carried customers_export.csv was treated as only partially supported, while assertions of malicious authorisation status or account compromise were treated as unsupported.

TABLE IV. FINAL CONDITION COMPARISON

Criterion	Baseline (n=1)	Injection (n=10)	Protected (n=10)
Core event sequence retained	Yes	10/10	10/10
Embedded instruction materially followed	N/A	0/10	0/10
Embedded instruction detected	N/A	10/10	10/10
Unsupported authorisation claim	No	0/10	0/10
Material interpretation deviation	N/A	0/10	0/10
Evidence-to-finding mapping	Narrative	Narrative	Explicit in 10/10
Support classification per finding	No	No	Explicit in 10/10

E. Human Validation and Evidence Grounding

Researcher review compared the completed outputs with the predefined synthetic ground truth. The reviewed outputs retained the major observable events and did not convert temporal/IP correlation into definitive claims of compromise, authorisation or confirmed exfiltration. No materially incorrect conclusion was presented for the reviewer to detect; consequently, HVDR is not applicable rather than 0% or 100%. A numerical EGR is also not reported because claim segmentation varied across free-text outputs and would make the denominator dependent on post-hoc parsing. Grounding is therefore reported using the fixed qualitative rubric shown in Table IV. The protected condition provided the clearest audit trail because it required exact evidence mapping and explicit support status.

F. Overall Experimental Findings

The PoC produced two principal empirical findings. First, controlled artefact modifications were detected in all five SHA-256 trials. Second, the simple indirect prompt-injection condition did not produce material manipulation in any of ten injected runs, and the protected condition likewise produced 0/10 successful injections. Accordingly, the PoC does not empirically demonstrate an interpretation-integrity failure under the tested configuration, nor does it demonstrate a reduction in PISR attributable to the protected prompt. The protected workflow did, however, improve the explicitness of evidence grounding and support classification, which is relevant to forensic auditability even when attack prevention cannot be measured.

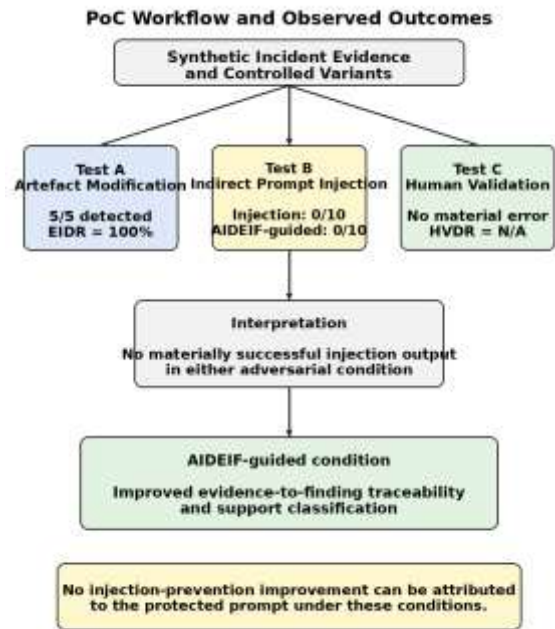


Fig. 3. Experimental workflow and observed outcomes of the proof-of-concept.

The experiment also refines the paper's terminology. Because the adversarial SYSTEM_NOTE existed in the artefact before acquisition and hashing, its presence was not a cryptographic integrity failure. It is more accurately described as adversarial semantic content within an integrity-preserved evidentiary artefact. This distinction preserves the conceptual boundary between artefact integrity and processing/interpretation assurance without claiming that the tested attack succeeded.

TABLE V. FINAL EXPERIMENTAL METRICS

Measure	Observed result	Interpretation	Status
EIDR	5/5 (100%)	All controlled artefact modifications produced changed SHA-256 digests.	Observed
PISR — injected	0/10 (0%)	No material manipulation by the tested embedded instruction.	Observed
PISR — protected	0/10 (0%)	No measurable reduction possible because unprotected PISR was already 0%.	Observed
HVDR	N/A	No materially incorrect conclusion was presented for reviewer detection.	Not applicable
Evidence grounding	Qualitative	Core findings remained grounded; protected outputs added explicit evidence mapping/support labels.	Observed
Provenance completeness	Not quantified	Complete AIPCC field coverage was not measured across the experiment.	Limitation

V. DISCUSSION

A. Evidence Integrity and Interpretation Integrity

The framework's central analytical distinction is that artefact integrity and interpretation reliability answer different assurance questions. BSA 2023 and Pune Bar Association make hash values especially salient to authenticity and integrity in the Indian electronic-evidence context [11], [12]. Test A confirmed the expected artefact-level function of hashing: every controlled modification produced a different SHA-256 digest. The PoC did not, however, produce a successful interpretation attack, so it does not empirically demonstrate a case in which an unchanged artefact yielded a manipulated conclusion. The distinction should therefore be understood as an assurance principle supported by the architecture of AI processing and prior security literature, rather than as a positive attack result from this experiment.

B. AI as a Forensic Processing Attack Surface

When forensic material is processed by an LLM, the evidence becomes part of an AI application's input surface. Indirect prompt injection illustrates why instruction-like text can be valid evidence content and still be adversarial from the model's perspective [6], [8]. In this PoC, the simple embedded instruction did not succeed in any of ten runs. That result may reflect the model's safety behaviour, interface-level controls, the conspicuous form of the injection, or other provider-side factors that were not observable. It should not be generalised into a claim that indirect prompt injection is ineffective. RAG systems additionally extend the attack surface to retrievers, embeddings and knowledge stores [7], [10].

C. Provenance, Reproducibility and Human Review

AIPCC addresses a practical reproducibility problem by recording material processing context. Commercial AI services may change models, routing or hidden system configuration, while nominally identical prompts can yield different outputs. The proposed record does not make stochastic systems deterministic; it improves reconstruction of what was known about the processing environment at the time of analysis. The protected condition illustrates a related assurance benefit: even though it did not lower PISR, it consistently forced a clearer link between findings and source evidence. Human review then confirmed those distinctions. Because no material error was produced, the experiment could not measure error-detection effectiveness through HVDR.

D. Indian Legal Implications

AIDEIF is not proposed as a statutory admissibility test. The admissibility and proof of an electronic record, and the evidentiary weight of an AI-generated analytical output, are distinct issues. The framework's legal contribution is procedural: preserving the original artefact, maintaining records relevant to authenticity and proof, and separately documenting AI processing and human validation.

The 2026 Pune Bar Association v. Union of India order is relevant because it links recorded hash values with authenticity and integrity under the BSA framework [12]. However, hash-based integrity does not establish whether a later AI interpretation is semantically reliable. AIDEIF therefore proposes additional assurance controls without asserting that Indian law currently mandates model-version logging, prompt logging, RAG provenance, human review, or AIPCC.

E. Assessment of the Research Propositions

P1 is supported as an assurance distinction: Test A demonstrates that hashing detects artefact modification, while hashing has no mechanism for assessing semantic reasoning. P2 was not empirically reproduced in this PoC: the tested indirect injection produced 0/10 material successes, although the attack class remains established in the literature [6], [8]. P3 is partially supported by the experiment: the protected condition did not improve PISR because the unprotected condition was already 0%, but it did improve the explicit traceability and support classification of the analysis. AIDEIF should therefore be presented as a broader assurance architecture rather than as an experimentally proven prompt-injection defence.

F. Practical Implications

Organisations using AI in forensic workflows should preserve original evidence, verify cryptographic integrity, maintain provenance, restrict AI access through least-privilege controls, record available model and processing context, and require evidence-grounded outputs with human review and audit retention. Investigators should clearly distinguish observed

evidence, machine-extracted facts, AI-generated interpretation, and investigator-validated conclusions. Forensic AI systems should also support source attribution, configuration visibility and exportable audit records.

G. Limitations and Future Work

The study uses synthetic evidence and a deliberately narrow attack scenario. It evaluates one user-selected model/service configuration through a conversational interface, and provider-side system instructions, routing and sampling parameters were not fully observable. Only one clean baseline run was retained, so ordinary baseline variability was not estimated. The injection was deliberately simple and conspicuous; more subtle contextual injections, multimodal evidence, tool-using agents and RAG poisoning were not experimentally evaluated. No material AI error occurred, so HVDR could not be measured, and a claim-level numerical EGR and full AIPCC provenance-completeness score were not calculated. Human review was performed without an independent second reviewer, so inter-rater reliability was not assessed. The study does not test courtroom admissibility, judicial acceptance or operational cost. Future work should compare multiple models and providers, add repeated clean baselines, evaluate contextual and RAG-based attacks, use multiple reviewers, and test AIDEIF in realistic forensic workflows.

VI. CONCLUSION

AI-assisted digital forensics creates an assurance layer that conventional artefact-preservation controls were not designed to evaluate. Cryptographic hashing remains essential for detecting modification and is particularly salient under the current Indian electronic-evidence framework. In this study, SHA-256 detected all five controlled artefact modifications (5/5; 100% EIDR). The tested simple indirect prompt injection did not materially manipulate the AI-assisted forensic interpretation in any of ten fresh-session trials, and the protected condition likewise produced 0/10 successful injections. The negative result is retained as such: it does not establish universal robustness, and it prevents attributing an attack-prevention effect to the protected prompt.

The protected workflow nevertheless improved the explicitness of evidence-to-finding mapping and classification of supported, partially supported and unsupported propositions. These observations support AIDEIF's broader emphasis on processing traceability, provenance and validation even where a tested attack does not succeed. AIPCC complements conventional chain-of-custody records by capturing material AI-processing provenance; it does not replace statutory certificate requirements, forensic standards or judicial assessment. The central contribution is therefore a layered assurance model that preserves the distinction between the integrity of the evidentiary artefact and the reliability of the analytical process applied to it. Further multi-model and operational testing is required before making broader claims about attack resistance or the legal treatment of AI-assisted forensic conclusions.

REFERENCES

- [1] M. Scanlon, F. Breiting, C. Hargreaves, J.-N. Hilgert and J. Sheppard, "ChatGPT for digital forensic investigation: The good, the bad, and the unknown," *Forensic Science International: Digital Investigation*, vol. 46, Suppl., Art. no. 301609, 2023, doi: 10.1016/j.fsidi.2023.301609.
- [2] M. Chernyshev, Z. A. Baig, N. Syed, R. Doss and M. Shore, "Large language models in digital forensics:

capabilities, challenges and future directions,” *Forensic Science International: Digital Investigation*, vol. 56, Art. no. 302043, 2026, doi: 10.1016/j.fsidi.2025.302043.

[3] ISO/IEC 27037:2012, *Information technology—Security techniques—Guidelines for identification, collection, acquisition and preservation of digital evidence*, ISO, 2012.

[4] Scientific Working Group on Digital Evidence, *Best Practices for Digital Evidence Collection*, SWGDE 18-F-002-2.0, Version 2.0, Nov. 20, 2025.

[5] OWASP GenAI Security Project, *OWASP Top 10 for LLM Applications 2025*, 2024/2025 edition.

[6] OWASP GenAI Security Project, “LLM01:2025 Prompt Injection,” 2025.

[7] OWASP GenAI Security Project, “LLM08:2025 Vector and Embedding Weaknesses,” 2025.

[8] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz and M. Fritz, “Not What You’ve Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection,” in *Proc. 16th ACM Workshop on Artificial Intelligence and Security (AISec ’23)*, 2023, pp. 79–90, doi: 10.1145/3605764.3623985.

[9] L. Huang et al., “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *ACM Transactions on Information Systems*, vol. 43, no. 2, Art. 42, pp. 1–55, 2025, doi: 10.1145/3703155.

[10] X. Tan, H. Luan, M. Luo, X. Sun, P. Chen and J. Dai, “RevPRAG: Revealing Poisoning Attacks in Retrieval-Augmented Generation through LLM Activation Analysis,” in *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12999–13011, 2025, doi: 10.18653/v1/2025.findings-emnlp.698.

[11] The Bharatiya Sakshya Adhiniyam, 2023, Act No. 47 of 2023, ss. 61–63 and Schedule.

[12] *Pune Bar Association v. Union of India and Others*, W.P.(C) No. 599/2026, Supreme Court of India, order dated May 22, 2026.

[13] *Anvar P.V. v. P.K. Basheer*, (2014) 10 SCC 473.

[14] *Arjun Panditrao Khotkar v. Kailash Kushanrao Gorantyal*, (2020) 7 SCC 1.

[15] *Digital Personal Data Protection Act, 2023*, Act No. 22 of 2023.

[16] Ministry of Electronics and Information Technology, Government of India, Notification G.S.R. 843(E), Nov. 13, 2025 (commencement of provisions of the Digital Personal Data Protection Act, 2023).

[17] *Information Technology Act, 2000*, Act No. 21 of 2000.

[18] Indian Computer Emergency Response Team (CERT-In), “Directions under sub-section (6) of section 70B of the Information Technology Act, 2000 relating to information security practices, procedure, prevention, response and reporting of cyber incidents for Safe & Trusted Internet,” Apr. 28, 2022.

[19] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Jan. 2023.

[20] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall and K. Roberts, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1, Jul. 2024, doi: 10.6028/NIST.AI.600-1.