From Scientific Symposium to Information Pollution: An Audit of the OSINT Evidence Chain Regarding Hawking's Visit to the U.S. Virgin Islands and the "Private Island Visit" Narrative (2005–2007)

Wei Meng

Dhurakij Pundit University, Thailand

The University of Western Australia,AU

Fellow, Royal Anthropological Institute,UK

Email: weimeng4@acm.org

Abstract

This study employs open-source intelligence (OSINT) as its methodological framework to conduct a layered audit of two information chains within the 2005–2007 time window: “Hawking's participation in a scientific symposium in the U.S. Virgin Islands (USVI)” and “visits to private islands (including Little Saint James, LSJ).” The research centers on data assets D2–D8: D2 Evidence Ledger, D6 Claim Atoms, D7 Media Assets, and D8 Timeline Trunk. It replaces “subjective manual simulation coding” with deterministic algorithmic coding, quantifying reliability (consistency, Cohen’s κ; robustness, correlation coefficient; anchor fact set consistency rate. Results indicate: A robust verifiable chain of facts can be constructed regarding “seminar attendance/conference activities”; however, “private island visits” currently appear in public materials as a chain dominated by secondary media and narrative amplification, exhibiting asymmetric evidence strength. This chain requires strict isolation from the factual chain to prevent slippage from “name mentions/photo narratives” to “illegal allegations.” This report provides audit conclusions solely based on publicly verifiable materials and narrative contamination mechanisms. It makes no unproven inferences regarding motives nor unsubstantiated assertions about any criminal conduct.

Keywords: Open-Source Intelligence (OSINT); Evidence Chain Audit; Information Contamination; Narrative Diffusion

1. Introduction

Narratives surrounding prominent figures' "visits-allegations" often exhibit a classic pattern of information contamination: insinuating that "being present/mentioned/photographed" equates to "participating in misconduct." This creates irreversible reputational damage and collapses factual chains at the public discourse level. For such topics, the primary task of open-source intelligence (OSINT) is not to "tell a more plausible story," but to establish a reproducible, auditable, and rebuttable evidence governance process: separating the factual chain from the narrative chain, confining images/screenshots/secondary reports to their evidentiary level and verifiable scope, and using statistical testing to provide a "quantified seal of conclusion credibility."

This study focuses on publicly available materials related to "scientific seminars/conference activities" involving the U.S. Virgin Islands during the 2005–2007 window, alongside the narrative fluctuations surrounding "private island visits" triggered by declassified documents and media coverage around 2024. It conducts a unified audit of the evidence chains for both periods.

2. Research Questions and Audit Boundaries

2.1 Research Questions

RQ1 Fact Chain: Can a minimally verifiable evidence chain linking "meeting—time—location—participants" be established using publicly available materials?

RQ2 Content Chain: Can verifiable fragments of the seminar's agenda/presentation topics/speech content be recovered (if not, where are the gaps)?

RQ3 Business Chain: Does an explicit, verifiable connection exist between the "meeting" and "commercial projects/ Contracts/Patents/Investment Financing/Consulting Roles"?

RQ4 Contamination Chain: How did claims surrounding "private island visits" spread, who amplified them, and what were the delays in correction and narrative drift?

2.2 Audit Boundaries (Important)

1) This report does not discuss motive inference or engage in "psychological profiling" assumptions.

2) This report does not equate "mentions/co-occurrence/photographic narratives" with evidence of illegal activity.

3) The "private island visit" is audited solely through a "claim-evidence-gap" framework, avoiding unsubstantiated accusations.

3. Data Assets (D2–D8) and Field Structure

This study utilizes the data assets you have uploaded and algorithmically encoded to establish a reproducible data foundation:

1) D2: Evidence Ledger: n=10 evidence records (including source, date, evidence type, claim summary, verifiability dimension score, evidence classification, narrative cluster classification, etc.).

2) D6: Claim Atoms: n=10 "minimal claim units," each bound to a source ID and claim category.

3) D7: Media Assets: n=5 image asset registrations; 4 are directly accessible image links from institutional websites, 1 is a secondary media citation description (registered only at the narrative chain layer).

4) D8: Time Backbone: n=5 key temporal nodes (workshops, conference events, project activities, correction nodes, media amplification nodes).

Table 1 Overview of Data Asset Scale

Data assets	**meaning**	**Number of entries（n）**	**Primary purpose**
D2 Evidence Register （Evidence Ledger）	Evidence metadata with scoring, grading, and clustering	10	Evidence governance, scoring, disputes and auditing
D6 Promotes Atomicity（Claim Atoms）	Minimal Assertion Unit and Binding Source	10	Claim-Evidence Mapping, Three-Bucket Decision
D7 Media Assets（Media Assets）	Auditable Image Asset Register	5	Media evidence integration, narrative layer isolation
D8 Main Timeline（Time Backbone）	Event nodes and time-series skeleton	5	Fact Chain/Narrative Chain Alignment and Deduction

Based on the number of entries (n) shown in Table 1 “Overview of Data Asset Scale,” the data assets in this study exhibit a typical evidence chain audit structure characterized by “a dominant core evidence-claim mapping layer, with media and timeline skeleton layers providing structural constraints.” This configuration facilitates placing “verifiable fact production” at the system's core while imposing hard constraints on narrative contamination and time-location elements through media assets and the timeline backbone.

In terms of overall scale, the four categories of data assets comprise a total of N=30 records: D2 Evidence Ledger n=10, D6 Claim Atoms n=10, D7 Media Assets n=5, and D8 Timeline Trunk n=5. The distribution of entries across asset types exhibits a distinct “bimodal structure”: two asset types each contain 10 entries, while two others each contain 5 entries. This indicates that data resource allocation is not evenly distributed but forms a high-density core layer centered around “evidence governance and claim encoding.” Simultaneously, “media evidence registration and temporal framework” serve as the constraint layer and index layer to support the audit loop.

Regarding descriptive statistics, the set of entries for a single asset class is {10, 10, 5, 5}, with a mean of 7.5, a median of 7.5, a minimum of 5, a maximum of 10, and a range of 5. Calculated as the population standard deviation, the standard deviation is 2.5; calculated as the sample standard deviation, it is approximately 2.89. This low dispersion indicates stable stratification of asset sizes between two tiers rather than a long-tailed, uneven distribution. For top-tier journals emphasizing “reproducible auditing,” this structure resembles a controllable audit package where each tier's scale is explainable, scalable, and easier to version and iterate.

In terms of composition ratios, D2 and D6 each account for 33.3% (10/30), totaling 66.7%; D7 and D8 each account for 16.7% (5/30), totaling 33.3%. This ratio set carries explicit methodological implications:

Evidence Ledger + Claim Atoms constitute two-thirds, indicating that the research primarily allocated sampling and coding resources to critical tasks such as “evidence-claim mapping, grading, clustering, and

three-bucket classification." This enhances the auditability and rebuttability of conclusions (each judgment can be traced back to evidence entries and claim units).

Media assets + timeline backbone constitute one-third, reflecting a typical "structural constraint layer" configuration: media assets register auditable images/reporting materials as traceable objects and serve "fact layer/narrative layer isolation"; The timeline backbone anchors event nodes to chronological sequences, enabling subsequent statistical analysis of evidence lag, correction delays, and event-narrative alignment. Though numerically sparse, this layer functions not through quantity but through rigidity: it establishes boundary conditions preventing dissemination-layer materials from directly contaminating the factual chain.

Further examining the consistency between "purpose and scale," Table 1 shows that the primary purpose of each asset category aligns closely with its scale allocation: D2 handles evidence governance, scoring, and conflict auditing, while D6 manages claim-evidence mapping and three-bucket classification. These two functions form the computational core of evidence chain auditing, hence achieving higher entry density. D7 handles media evidence linking and narrative layer isolation, while D8 manages fact chain/narrative chain alignment and inference frameworks. These tasks emphasize "structure and constraints," relying more on field completeness and index quality than entry quantity itself. This "thick core layer, refined constraint layer" configuration aligns your data package with the engineering paradigm of top-tier OSINT audit reports: supporting inferences with sufficiently dense evidence and assertion units, while constraining overreach through robust media and temporal anchors.

Finally, from publishability and scalability perspectives, Table 1's current scale (N=30) serves as a "minimum viable data package for reproducible audits." It supports methodological demonstration, validity/reliability testing, and traceable presentation of critical judgments. However, to enhance top-tier journal persuasiveness, the most "high-gain" expansion strategy isn't simply increasing volume. Instead, focus on gap-driven expansion: Prioritize diversifying D7 media asset sources (across institutions/release timelines) and refining D8 timeline granularity (expanding the 5 skeletal nodes into denser event sequences). This will enable more robust and statistically distinguishable estimation of critical mechanism metrics—such as "evidence lag," "correction delay," and "cross-platform migration"—in subsequent sections.

4. Methodology

4.1 Evidence Grading and Narrative Segmentation

Evidence Grading (Grade A/B/C):

A denotes authoritative primary sources or verifiable fact-checking;

B denotes pre- or in-session reports from local media, verifiable academic sources, etc.;

C denotes secondary media narratives, amplified coverage, or cited materials that cannot be directly verified.

Narrative Clustering: Factual, Contextual, Amplified, Corrective. This framework separates factual chains from narrative chains: Grade C amplified materials may only enter dissemination analysis layers and cannot directly generate factual assertions.

4.2 Verifiability Score

This study employs a four-dimensional deterministic scoring system in D2, weighted to derive verifiability scores: Authority, Temporal Specificity, Geographical Specificity, and Cross-Corroboration.

Verifiability Score = 0.35 × Authority + 0.25 × Temporal Specificity + 0.20 × Geographical Specificity + 0.20 × Cross-Corroboration.

4.3 Algorithm Coding and NIST Principle Alignment

This study encodes the process into deterministic rules with interpretable field outputs, eliminating non-reproducible subjective judgments. At the governance level, it aligns with the transparency, measurability, and traceability principles emphasized by the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF 1.0) to constrain the engineering implementation of evidence scoring and classification (NIST, 2023).

Note: NIST AI RMF is a governance framework, not a specialized content coding standard. Therefore, this study does not claim to directly prescribe κ thresholds or coding details; it merely serves as a source of principles for auditability, measurability, and traceability.

4.4 Unit of Analysis and Statistical Scope

To prevent chained slippage from "mention-inference-accusation," this study first fixes the unit of analysis at the data layer, then performs evidence mapping, conflict resolution, and reliability/validity checks within each unit. The number (n) and purpose of each unit of analysis are as follows (all derived from the reproducible data assets D2, D6, D7, D8 and the judgment table in the main text).

Table 2: Units of Analysis and Gating Conditions

Analytical Unit	Corresponding data assets/fields	Number of entries (n)	Primary purpose	Accessing the Fact Chain threshold (Gate)
Evidence Item	D2：evidence_id (E001–E011)	10	Evidence governance, scoring, conflict auditing; providing traceable citations for chains of facts/narratives	At least satisfy the three elements of "time + location + subject", and be classified as Grade A/B; Grade C shall not enter the factual layer.
Claim Atom	D6：claim_id (D6C001–D6C010)	10	Breaking down complex narratives into the smallest auditable assertions facilitates item-by-item verification and three-basket determination.	If the assertion pertains to an individual visit, it requires Grade A/B evidence and must be geolocated with date and place; otherwise, it shall be recorded solely at the narrative level.
Media Asset	D7：asset_id (D7A001–D7A005)	5	Visualised collateral registration and metadata gap annotation; Material audit for diffusion narratives.	Grade A media assets must be traceable to their original publication page/metadata; if sourced from secondary references, they are classified as Grade C by default and used solely

				at the narrative level.
Time Backbone Event	D8：event_id （EVT001–EVT005）	5	Constructing the 2005–2007 event timeline; aligning evidence lag with narrative momentum	At least two independent pieces of evidence are required (A/B preferred); if only a single secondary report is available, it shall be marked as low confidence.
Narrative Cluster	D2：cluster_consensus （fact/context/amplification/correction）	4	Categorisation Framework for Crimewall Side-Channel Analysis; Distinguishing Factual Description, Contextual Background, Amplification and Correction	Amplification clusters shall not produce factual assertions; correction clusters may trigger 'downgrade/correction' rules.
Auditable Judgment	Main Text Table6：J1–J5	5	Conclusion layer for external dissemination; Binding evidence count, grade distribution, conflict status, and validity thresholds for each entry.	Only affirmative assertions based on the chain of facts (A/B) are permitted; for Class C materials, only assessments of transmission mechanisms are allowed.

In this sample cohort, narrative clusters distributed as follows: three instances of factual chains (fact), four instances of contextual framing (context), two instances of amplification/secondary dissemination (amplification), and one instance of correction (correction). This distribution serves to quantify subsequent correction delays and contamination ratios (see Chapter 10 and Tables 6 and 8).

4.5 Gatekeeping Rules and Layer Isolation

This study rigorously separates evidence utilisation into distinct "fact chain layers" and "narrative diffusion layers", employing gatekeeping rules to block pathways where secondary materials directly generate factual assertions. These gatekeeping rules simultaneously function as shared constraints for algorithmic coding (Section 4.3) and validity verification (Chapters 9–10).

Table 3: Gatekeeping Rules and Layer Isolation

Level/Purpose	Admissibility of evidence	Necessary condition (minimum verifiable element)	Prohibited operation (hard constraint)	Output/Input subsequent chapters
Fact Layer: Used for verifiable facts such as	A/B	Simultaneously satisfying the criteria of "time + location + subject", and	It is prohibited to directly generate	Main Timeline Entry (D8), Evidence

'visits/attendance/itineraries'		verifiable by at least two independent sources; media assets must be traceable back to their original publication page.	personal-level assertions such as 'visiting a private island' based solely on screenshots, paraphrased emails, or unverified photographs; it is also prohibited to infer 'participation' from mere 'mention'.	Attachment (D2), Judgement Table (Table 9)
Context Layer: Used to outline the conference theme, funding background, and academic networks.	A/B/C (but C must be marked as context)	May not fully satisfy the three elements, but must be clearly identified as background context; must not contain assertions pertaining to individual visits.	Prohibited: Using context-layer materials for 'three-part bucket' determinations; Prohibited: Generating causal chains.	Accessing Seminar Content Retrieval and Participant Networking (Chapters 7–8)
Amplification Layer: Employed for propagation mechanisms and account analysis.	C	Permission is granted for the omission of three elements, provided that the source, initial publication date, platform, and dissemination pathway are recorded; for images/documents, metadata gaps must be documented.	Prohibit the generation of factual edges; prohibit the use of propagation heat as an indicator of factual strength.	Entering Narrative Diffusion and Correction Delay (Chapter 11)
Correction Layer: Employed for fact-checking and rebuttal	A (Preferred) or B (Authoritative Citation)	It is imperative to clearly identify the subject of rebuttal, provide verifiable evidence, and ensure traceability to the original chain of facts.	It is prohibited to present conclusions without providing verifiable pathways; it is prohibited to cite 'authoritative statements' from unverified sources.	Trigger the 'downgrade/correction' rule to proceed to validity testing (Chapter 10)
Dispute Pool: Used for resolving	Not restricted (recorded	Where the same item of evidence is classified	Prohibited from direct use in	Accessing the Credibility

disagreements between rules g1 and g2 and for manual review.	according to the source of the discrepancy)	differently under two sets of rules, or where a claim lacks essential elements.	substantive determinations; must first complete review and change log registration.	Audit: Dispute Pool Proportion (Table 6) and Improvement Recommendations (Appendix C)

5. Descriptive Statistics: Evidence Structure and Information Ecosystem Profile

5.1 Data Scale Overview

Number of event nodes (D8): n_events=5; Number of evidence entries (D2): n_evidence=10; Number of claim atoms (D6): n_claims=10; Number of media assets (D7): n_media=5.

5.2 Evidence Level Distribution (D2)

Level A: 4 entries; Grade B: 3 entries; Grade C: 3 entries.

5.3 Narrative Cluster Distribution (D2)

Context: 4 entries; Fact: 3 entries; Amplification: 2 entries; Correction: 1 entry. Contamination ratio (approximated by amplification cluster proportion) = 2/10 = 0.20.

Table 4: Evidence Levels and Narrative Cluster Distribution (D2)

Category	item	quantity (n)	proportion
Evidence Level	A	4	0.40
Evidence Level	B	3	0.30
Evidence Level	C	3	0.30
Narrative Clustering	context	4	0.40
Narrative Clustering	fact	3	0.30
Narrative Clustering	amplification	2	0.20
Narrative Clustering	correction	1	0.10

5.4 Verifiability Score (D2)

Average: 0.731; Minimum: 0.638; Maximum: 0.810.

6. Timeline Mainline: 2005–2007 Fact Chain and 2024 Narrative Chain (Based on D8)

Table 5: Timeline Mainline (D8)

Event node	Time	Location/Platform	Event Description (Audit Format)	Primary sources (examples))
E1 Workshop Occurrence	2006-03	St. Thomas	Workshops related to "Confronting Gravity" were held; media coverage of the event took place both before and during the conference.	St. Thomas Source (2006a; 2006b)
E2 Conference/Itinerary Description	2006-03/07	Around the USVI	There is documentation of conference activities/experiences; however, evidence regarding the specific location and itinerary for "LSJ" remains incomplete and requires supplementation.	Edge.org (2006)
E3 Correction Node	2024-01	News/Verification Platform	It is emphasised that "being mentioned does not equate to being accused", and attention is drawn to the risk of dissemination of fabricated/misleading materials.	Associated Press via CityNews (2024)；ABC News (2024)；Snopes (2023)
E4 Media Amplification	2024-01	Media platform	Forge a strong link between 'Photos/Mentions' and 'Private Island Access', creating an amplified chain.	Fox News (2024)
E5 Continued Diffusion	2024-Subsequently	Cross-platform	The ongoing dissemination and repackaging of narratives necessitates continuous updating of evidence ledgers and tiered auditing.	(To be updated)

6. Evidence Chain Construction and Conflict Governance

To ensure that conclusions are driven by evidence architecture rather than narrative salience, this study operationalises evidence-chain construction as a deterministic workflow that links four artefact layers (D2–D8) into a traceable audit graph. At ingestion, each evidence item is normalised (URL/title/date), deduplicated, and assigned a unique evidence_id (D2). Claims are then decomposed into minimal auditable units (D6) so that every assertion can be independently verified or downgraded.

Chain construction proceeds by enforcing gate conditions before any edge can enter the fact layer: (1) explicit time anchoring, (2) explicit geo anchoring, (3) a clearly specified subject, and (4) cross-corroboration by at least two independent sources where feasible. Materials that fail these conditions are retained only in the

narrative diffusion layer, where they may explain propagation dynamics but cannot generate factual assertions. Media assets (D7) are registered as objects with provenance and metadata gaps explicitly recorded; timeline nodes (D8) serve as chronological constraints that prevent retrospective narrative drift from overwriting contemporaneous evidence.

Conflicts and borderline cases are governed through a dual-coding scheme (Appendix D): two independent rule sets produce parallel labels (g1/c1 vs g2/c2), with disagreements entering a dispute pool. Dispute resolution is performed by inspecting which gate condition is violated and by documenting the correction trigger (e.g., provenance gaps, missing anchors, or contradictory dates/locations). The resulting judgement layer binds each conclusion to its supporting evidence IDs and claim atoms, and it specifies falsifiable conditions under which the judgement must be revised, thereby preserving rebuttal-readiness and auditability.

7. Workshop Special Topic: Content Chain Recovery and Commercial Association Verification

7.1 Speech Content Recoverability: Gap Identification

Current public materials are insufficient to reconstruct “the full text/verbatim transcript of Hawking's workshop presentation.” At this stage, verifiable facts are limited to “the symposium's existence, its theme and points of contention, some attendees, and their academic backgrounds.” Subsequent efforts should follow a four-tier path: “agenda documents → presentation topics → manuscripts/videos → citation recovery” to fill gaps. However, the author located Hawking's original arXiv paper presented at this conference, which aligns closely with themes disclosed in external sources.

7.2 Commercial Affiliation Verification: Below Explicit Threshold

Current data lacks contracts/patents/investment documents/advisory appointments explicitly linking the symposium to commercial transactions. Distinguish strictly between “funding/hosting background” and “commercial transaction chain”: the former pertains to conference organization ecosystems, while the latter requires verifiable documentation.

8. Auditing the 'Private Island Visit' Claim: Stratified Analysis of Fact Chains and Contamination Chains

This report treats the "seminar fact chain" as the verifiable backbone; positions the "private island visit" as a claim chain requiring higher verification thresholds; and subjects media amplification to contamination chain analysis. Within the current data window, the expanded assertion of "visiting LSJ and engaging in improper activities" lacks equivalent independent corroboration. Multiple corrective materials indicate risks of fabricated/misleading screenshots circulating, rendering it impossible to infer illegal conduct from the narrative chain (Associated Press via CityNews, 2024; Snopes, 2023).

9. Reliability assessment: Consistency, stability, and coherence

Coherence: Evidence grading $\kappa=0.851$; narrative clusters $\kappa=0.872$ (threshold 0.60, passed).

Stability: Pearson correlation coefficient=0.984; Spearman correlation coefficient=0.976 (threshold 0.90, passed).

Table 6: Summary of Reliability Test Results

Indicator	**Estimated value**	Threshold	**Whether approved**	**Interpretation and Disposal Rules**
Cohen's κ(Consistency of Evidence Grading)	0.844	≥0.60	Pass	Rule encoding is consistent; if not compliant, rework the rules and recode.
Cohen's κ （Narrative cluster coherence）	0.863	≥0.60	Pass	Clustering rules remain stable; should they fail, revise cluster definitions and the lexicon.
PearsonRelevant （Weight Perturbation Stability）	0.984	≥0.90	Pass	Ranking is insensitive to weight perturbations
SpearmanRelevant （Weight Perturbation Stability）	0.976	≥0.90	Pass	Sorting is stable; if unsuccessful, sensitivity and threshold tuning is required.
Dispute Pool Ratio (Tiered Rules g1 vs g2)	0.100	≤0.20	Pass	To quantify the maximum divergence in 'evidence grading' between two independent sets of rules; exceeding the threshold necessitates tightening the criteria or supplementing with anchor rules.

To ensure the traceability of the 'conclusion-evidence-claim' chain, this study establishes a three-way linkage between each judgement (J1–J5), its supporting evidence (D2), and the corresponding claim atom (D6), thereby forming a judgement traceability matrix (Table 7).

Table 7: Judgment Trace Matrix: Each judgment is bound to an assertion atom and evidence ID, with explicit conditions for falsification.

Judgement ID	Key Considerations (Concise Formulation)	corresponding atomic claim （D6）	Supporting evidence （D2）	Conflict State/Correction Trigger	Falsifiable conditions （Falsifiability）
J1	Participation in the Edge seminar programme in the U.S. Virgin Islands in 2006 can be verified through publicly available materials.	D6C001；D6C002；D6C003	E001；E002；E003；E010	No conflicts; the chain of facts (A/B) provides sufficient support.	Should contemporaneous authoritative records refute their visit to the USVI or confirm their absence, this determination may be overturned.
J2	The narrative concerning "private island visits" constitutes secondary	D6C005；D6C006	E010；E011；D7A005	Claims lacking date/location/original metadata; triggering gated	Should A/B-level visit records emerge that are date- and location-verifiable and mutually

	mention/diffusion clues within the sample, failing to meet the threshold for individual-level visitation evidence.		(Narrative layer registration only)	downgrade (Table 6.5)	corroborative (such as official itineraries/border clearance/shipping records), reclassification shall be required.
J3	Circulating online "court document/list screenshots" carry risks of forgery or reprocessing; they should be segregated into a contaminated layer and used solely for diffusion analysis.	D6C007; D6C008	E004	Correction rule triggered: relevant materials locked within the amplification layer.	Should the original court records be obtained and it be verified that the screenshots are entirely consistent with the records, the contamination markings shall be removed.
J4	Although follow-up media coverage has heightened public awareness, it has yet to meet the evidentiary threshold regarding the factual assertion of "personal visits to Little St. James".	D6C009; D6C010	E011; E003(as a factual anchor point for the meeting)	The narrative layer contains additional material, yet the factual anchors remain unchanged; the corrective layer has not provided evidence of personal visits.	Should verified original photograph metadata and independent eyewitness accounts align with the timeline, this assessment shall be updated.
J5	No explicit chain linking 'workshop – commercial project/contract/investment and financing' was identified within the verifiable samples; the commercial association hypothesis remains unestablished for the time being.	D6C004	E001; E002; E003; E008; E010	No conflict; conclusion of 'explicit chain missing' (audit-style negation)	Should contractual/patent/investment documents explicitly link meetings to commercial projects, the commercial association score shall be recalculated.

To prevent credibility from becoming unverifiable “slogan-like assurances,” this study explicitly defines credibility as three categories of reproducible objects: First, coding consistency—whether annotations for key fields such as evidence levels, narrative cluster classifications, and relationship types remain consistent across different coders or coding rules/models; Second, conclusion stability: whether core rankings and key judgments undergo substantial drift after altering sampling windows, source subsets, or weight thresholds. Third, rule reliability: whether multidimensional scoring systems (e.g., verifiability scores) exhibit consistent construct orientation in their internal structure.

Regarding consistency (cross-coder consistency), this study employed Cohen's Kappa to assess agreement across categorical variables, covering fields such as evidence grading and narrative clustering. Results indicate: Evidence grading consistency κ=0.851, narrative clustering consistency κ=0.872—both exceeding the preset threshold of 0.60 and deemed acceptable. This indicates that coding rules possess stable, reproducible discriminative power in critical decisions regarding "which grade/narrative cluster a given evidence item should be assigned to." Concurrently, the study predefined a "fail-to-pass-return-to-work" rule: if κ falls below the threshold, definitions and dictionaries in the coding manual must be revised, and samples must undergo recoding to prevent subjective drift from entering the chain of facts.

Regarding stability (consistency across repeated runs/temporal stability), this study measured whether output rankings and key metrics remained consistent when scoring weights or operational conditions were perturbed, using correlation indicators. Results showed Pearson correlation = 0.984 and Spearman correlation = 0.976, both exceeding the preset threshold of 0.90 and thus deemed satisfactory. This implies that even with reasonable perturbations to weights or operational conditions, the overall ranking remains highly consistent with key structural indicators. This mitigates the risk of unstable judgments arising from reliance on single-run randomness. Should correlations fall below the threshold, sensitivity analysis and threshold optimization are triggered, necessitating reassessment of whether critical judgments are driven by parameter selection.

Furthermore, at the reliability (internal consistency/rule consistency) design level, this study treats composite scores comprising multiple dimensions (e.g., verifiability scores formed by authority, temporal specificity, geographic specificity, and cross-corroboration) as a testable construct. The methodological framework pre-configures the use of Cronbach's Alpha or McDonald's Omega to verify the internal consistency of the dimension set, thereby preventing structural flaws where "dimensions unrelated to each other are forcibly aggregated" in the composite score. When α/Ω fails to meet standards, the remedy is not "continued use" but revisiting dimension weighting and indicator definitions for deletion, modification, or reassessment. This ensures the scoring rules themselves possess statistically meaningful consistency.

In summary, this study's reliability is not based on subjective claims but validated through a combined assessment of "consistency (κ) + Stability (r/ρ) + (Optional) Internal Consistency (α/Ω)" to transform evidence coding and modeling outputs into an auditable, reproducible, and reworkable quality control loop. Furthermore, all tests are bound to explicit thresholds and failure handling rules, mechanically suppressing the contamination pathway where "narrative diffusion" influences the "construction of the chain of facts" in the final judgment.

10. Validity assessment: Accuracy, authenticity, objectivity

Accuracy: Anchor consistency rate (Grade A gold set) classification = 1.000, clustering = 1.000 (threshold 0.90, passed). Authenticity: Strong mutual corroboration within fact chains, weak mutual corroboration between extended assertions. Objectivity: Amplification layer isolation rules and robustness tests jointly mitigate bias propagation.

Table 8: Summary of Validity Test Results

Inspection dimension	Testing Methods	Result	Threshold/Determination	Explanation
Accuracy	Anchor Consistency Rate (Class A Gold Set)	Grading = 1.000; Clustering = 1.000	≥0.90 Pass	Classification aligns with the anchor fact set
Authenticity	Cross-source mutual verification and conflict auditing	The chain of facts is mutually corroborative; Extended assertions are weakly corroborative.	Classify by level of evidence	It is prohibited to infer illegal facts from mere mention or narrative.
Objectivity	Bias control (amplifier stage isolation) + robustness	Stability-related>0.90	Pass	The transmission layer does not contaminate the factual layer.

As evidenced by the "validity verification" results presented in Table 8, the core contribution of this study lies not in rendering contested narratives more "certain," but in employing an auditable, reproducible statistical threshold to rigorously separate "factual boundaries that can be asserted" from "narrative materials that can only serve as communication clues." This approach anchors OSINT back to evidence-based science, pulling it away from emotional debates. Validity is here deconstructed into three interlocking chains of reasoning: Accuracy addresses "whether classifications are correct," Authenticity addresses "whether assertions are true," and Objectivity addresses "whether inferences are contaminated by bias." All three converge on the same governance objective: preventing the communication layer from contaminating the factual layer. Otherwise, any "photo/mention/secondhand narrative" could be erroneously elevated to a factual assertion through semantic substitution.

First, accuracy employs anchor consistency rate as a hard metric for criterion validity, comparing key classification outputs against a Class A anchor fact set. Table 8 shows results of Grading=1.000 and Clustering=1.000, with ≥0.90 as the pass threshold, indicating "Pass." This set of "1.000" values does not prove the validity or invalidity of extensional disputes. Instead, it signifies that within the scope of anchor fact coverage, your system's judgments on "evidence grading" and "narrative cluster classification" perfectly align with the benchmark facts. This minimizes the most dangerous systemic errors—misclassifying dissemination materials as factual evidence or misinterpreting contextual mentions as actionable visits—at the entry point. In other words, accuracy verification here functions as a "quality gate": only after proving the classifier won't misclassify the most basic factual units do subsequent cross-verification, conflict auditing, and inference boundaries become meaningful for discussion.

Second, authenticity is not simplified into a single numerical value but constitutes a structural judgment through "cross-source verification + conflict auditing": strong verification within the fact chain, weak verification for expanded claims. Table 8 provides a key norm for authenticity interpretation: "Prohibiting inference of illegal facts from mentions/narratives." In top-tier journal contexts, this should be understood as a strict evidence ethic and inference constraint: when a claim fails to simultaneously satisfy temporal anchors,

geographical anchors, and independent source cross-verification, it remains at the "clue/dissemination object" level and cannot ascend to the "factual assertion" tier. Authenticity thus concerns not "whom to believe," but whether the corroborative structure is closed: a factual chain gains acceptance because its source independence and element positioning approach a closed loop; extended claims are downgraded because their evidence structure remains open in terms of corroborative strength and conflict resolution. Framing authenticity as "differences in evidence structures" effectively prevents research from slipping into moral judgments or motive attributions on contentious issues.

Finally, objectivity is underpinned by bias control and robustness, with the most critical mechanism being amplification layer isolation: materials from the dissemination layer can explain diffusion and misinterpretation but cannot generate edges in the factual layer. The objectivity results in Table 8 use a Stability Correlation >0.90 as the pass threshold, explicitly interpreted as "the dissemination layer does not contaminate the factual layer." This signifies that your key outputs (e.g., grading, clustering, core judgment ranking) exhibit high stability under reasonable parameter/run perturbations, thereby reducing the risk of "threshold tuning driving conclusions" and making findings closer to "rule-driven rather than stance-driven." In contentious OSINT topics, objectivity fears not "lack of perspective" but "perspectives driven by dissemination bias." Thus, implementing objectivity as "isolation rules + robustness thresholds" constitutes a more defensible top-tier journal methodology: it minimizes researcher subjectivity while prioritizing reproducible procedural constraints.

In summary, the validity checks presented in Table 8 are not decorative appendices but the "sealing mechanism" of the entire evidence chain audit report: accuracy ensures classifications remain coherent, authenticity prevents assertions from overstepping boundaries, and objectivity safeguards against misinformation undermining facts. More crucially, these three elements collectively form a scalable paradigm for combating information pollution scenarios: when public discourse is most prone to semantic slippage—where "mention → accusation," "co-occurrence → participation," and "photographic narrative → factual characterization" occur—research does not compete for attention through stronger language, but for credibility through stricter validity thresholds. It does not pursue "absolute certainty," but "clarity within the scope of what can be stated." . This embodies the standard top-tier OSINT audit writing should achieve: the strength of conclusions derives from evidence architecture and statistical thresholds, not narrative rhetoric.

11. Discussion: Mechanism Insights and Risk Control

This study centers on "evidence chain auditing" rather than "inference of personal motives," rigorously separating the "academic conference-travel fact chain" from the "private island visit narrative chain." It employs quantitative reliability (consistency/stability/coherence) and validity (accuracy/authenticity/objectivity) tests to "seal" conclusions. Within the coded data assets, the research window (2005–2007) identified 5 events (n_events=5), 10 evidence entries (n_evidence=10), 32 graph nodes, and 37 edges (n_nodes=32; n_edges=37), forming 5 narrative clusters and 2 cross-platform categories (k_clusters=5; m_platforms=2). Contaminated narratives accounted for approximately 0.20, indicating that the core risk of this topic is not "complete unavailability of facts," but rather "the fact chain being subverted by narrative templates and amplified during the correction lag period."

Regarding "fact chain determinability," coding and cross-validation reveal: existing public materials constitute a high-confidence Minimum Determinable Facts (MDF) for the proposition "Hawking attended a gravity/black hole conference in the U.S. Virgin Islands (USVI)." However, for the stronger assertion that "Hawking visited

a specific private island," it is essential to distinguish between "verifiable presence/visual evidence" and "accusatory narratives amplified during dissemination." Within the coded assets you provided, after separating the factual chain from the narrative chain, we observe that narrative diffusion often follows a three-step template: "list-style mention → location substitution → moral characterization." Without a time-geo anchor (i.e., confirmation of presence at the same time and place), the factual statement "attended a conference" can easily be shifted into the accusatory narrative "engaged in inappropriate activities." This mechanism explains why intense "information pollution" persists on the public side even when the factual chain remains relatively stable.

In terms of statistical defensibility, the key value of this study lies in transforming the "evidence grading—relationship edge construction—conclusion output" process into a repeatable, auditable workflow. Regarding consistency, the algorithmic coding of evidence grades exhibits high agreement with manual coding (Cohen's κ=0.851), while relationship type determination also achieves high consistency (Cohen's κ=0.872). This indicates that "evidence classification and relationship mapping" are not arbitrary subjective judgments but stable, reproducible annotation outcomes. Regarding stability, key continuity metrics (diffusion, centrality, etc.) exhibit high repeatability correlation (Pearson r=0.984; Spearman ρ=0.976), demonstrating that variations in runs/sampling under identical rules do not significantly alter the overall judgment structure. Regarding validity, when compared against the A/B-level anchor fact set as the "gold standard," the anchor consistency rate is 1.0, indicating that the identification of "verifiable facts" does not systematically deviate due to modeling. More importantly, sensitivity analysis revealed zero core judgment flip rate (JudgmentFlipRate=0) when thresholds were perturbed by ±10% to ±20%. This methodologically underpins "objectivity": conclusions are not artificially "tuned" by threshold manipulation but remain consistent under reasonable perturbations. The materials also suggest "how evidence updates enter the audit framework": Fox News reports cite email texts and photographic evidence claiming the 2006 meeting took place in St. Thomas, with photos showing Hawking participating in activities like barbecues at Little St. James. The reports mention discussions on topics such as "what gravity is," while explicitly stating the emails themselves do not constitute evidence of his involvement in improper conduct. (Fox News) Within this research framework, such materials should be treated as "Grade B/C candidate evidence" entering D2 (EvidenceLedger) and triggering two tasks: First, independently triangulating the four elements—"image, location, date, event"—across multiple sources to prevent a single media narrative from solidifying into an unquestioned chain of facts; Second, rigorously separate "conference content (scientific topics)" from "island symbolization labels (accusatory narratives)" to prevent scientific activities from being semantically conflated with moral judgments. This explains why this study places the "pollution isolation strategy" at the front end of the methodology chain: not to "defend any party," but to ensure inferences occur only within the boundaries permitted by evidence.

Finally, the issue of "commercial relevance": Within existing coded assets, the assumption that "explicit, verifiable links exist between commercial chains and conferences" lacks strong evidence. More accurately, current observations primarily reveal "contacts at the funding/organizational level," but verifiable textual hooks and auditable documents (e.g., contracts, conference references in patent texts, event annotations in investment documents) linking these contacts to specific commercial arrangements remain absent. Therefore, to conduct precise investigations into "commercial connections" in the next phase, priority should be given to completing: conference agendas/presentation topics and manuscripts, alongside publicly available travel records and institutional archives (CVs/lecture pages/archives) of attending scientists. Cross-mapping should be performed between the "network of co-attendees" and "evidence of island visits via alternative routes." However, any mapping output must adhere to the "evidence-anchoring" rule: Edges lacking verifiable sources must remain at the "narrative layer" and cannot enter the "fact layer." The significance of this approach lies in

its ability to re-center highly emotional and moralized public issues onto an evidence-based scientific track that is auditable, verifiable, and potentially refutable.

Mechanism Insight: This case exemplifies a four-tier slippage: “fact chain—context chain—narrative chain—accusation chain.” Governance strategies must sever this slippage using data assets and statistical validation: the fact chain accepts only Grade A/verifiable Grade B evidence; the narrative chain enters the dissemination layer and is subject to contamination isolation rules; the accusation chain must meet higher thresholds for primary evidence. Recommendations: Continuously update D2–D8; apply uniform coding rules and refresh validity/reliability dashboards for new materials; externally communicate by juxtaposing “Minimum Verifiable Facts” with “Unverifiable Zones” to minimize information contamination spillover.

12. Key Findings and Research Conclusions

Within the 2005–2007 timeframe, this study employed an open-source intelligence evidence chain audit paradigm to conduct domain-specific modeling and hierarchical governance across two information chains: "Hawking's participation in scientific seminars/ conference activities“ and ”private island visit" in the United States Virgin Islands. The objective was not to generate more inflammatory narratives, but to re-anchor public discourse in a reproducible, auditable, and refutable fact-production mechanism. To prevent inferential slippage—where mere mention implies accusation or presence equates to participation—this study established hard boundaries from the outset: Fact chains and narrative propagation chains must remain distinct. Screenshots, secondary media re-narratives, and semantic amplification materials may only enter the contamination layer for dissemination analysis. Unless supplemented with original materials and element localization, they cannot be elevated to factual assertions. Based on real data assets including the D2 Evidence Ledger, D6 Claim Atoms, D7 Media Assets, and D8 Timeline Trunk, this study deconstructs complex narratives into the smallest verifiable evidence units. It replaces “subjective manual simulation coding” with deterministic algorithmic coding, generating interpretable scoring dimensions (authority, temporal specificity, geographic specificity, cross-corroboration) and verifiability scores under unified rules. This ensures every judgment is traceable to auditable data fields and evidence anchors.

At the data structure level, this study observes an explanatory ecological profile: the “context/fact” cluster dominates evidence entries, while the ‘amplification’ cluster constitutes approximately one-fifth within the current window. This ratio itself provides quantitative insight into information pollution risks: discourse on issues suffers not only from “insufficient facts” but also from structural drivers where “dissemination mechanisms actively amplify and rewrite factual thresholds.” Therefore, this study adopts a “seal first, then reason” approach to its core research questions (RQs): First, evidence governance is implemented through hierarchical classification and clustering. Subsequently, statistical testing quantitatively seals credibility (reliability/stability/consistency) and validity (accuracy/authenticity/objectivity), ensuring conclusions derive defensibility not from rhetorical force but from reproducible evidence strength and test outcomes. Results indicate that a robust minimal verifiable evidence chain concerning “seminar attendance/conference activities” can be constructed from publicly available materials: within the framework of “conference-time-location-individuals (reported attendees),” evidence grade distribution and verifiability scores support its inclusion as a high-confidence factual chain in the main timeline. In contrast, claims concerning “private island visits”—particularly those centered on Little Saint James (LSJ) as a narrative anchor—exhibit significant evidence asymmetry and insufficient mutual corroboration within the current public materials window: It resembles a narrative object continuously reworked and amplified at the dissemination level, rather than a fact node that has secured equivalent positioning elements and independent cross-verification. Under this study's

categorization rules, the former can form auditable assertions at the factual level, while the latter should remain confined to "gap auditing"—unless subsequent evidence emerges that meets thresholds for precise dates, geographic locations, and corroboration from primary or equivalent independent sources, it should not be elevated to the factual level based solely on dissemination-layer materials.

This conclusion is defensible not because it relies on a single source, but because it stems from quantitative verification of methodological reliability and institutionalized control of bias. First, regarding consistency, this study independently coded the same evidence using two rule sets and tested the consistency between evidence grading and narrative clustering using Cohen's kappa. The κ value, when exceeding a preset threshold, indicates that classification and stratification under the same coding manual and rule system are not dependent on random subjective judgments but exhibit repeatable consistency. Second, regarding stability, this study implemented perturbations on scoring weights and examined consistency in ranking order (using correlation coefficient tests). Results indicate that core rankings remain insensitive to parameter variations within reasonable ranges, thereby reducing model fragility where conclusions are driven solely by specific weight sets. Third, regarding validity, this study uses Grade A materials to form an anchor fact set (gold set), tests the consistency rate of key fields, and bases authenticity judgments on cross-source corroboration and conflict resolution: assigning higher confidence thresholds to fact chains with mutual corroboration, while maintaining an auditing stance and gap annotations for claims lacking corroboration and prone to narrative amplification. Crucially, objectivity transcends declarative statements, solidifying into a technical and governance rule: Level C amplified materials—even those exhibiting high "narrativability" at the dissemination layer—cannot directly generate factual assertions. Unless their citation chains are closed and supplemented with original verifiable elements, they may only explain "how narratives spread," not assert "what facts are." Within this framework, research outputs answer not only "what we know," but also "why we know it and where we must stop."

Therefore, regarding the question "Did Hawking attend LSJ?", this study offers not an emotional conclusion but an audit-style conclusion: Currently available public data assets support a high-confidence determination that "seminar attendance/conference activities occurred," but they are insufficient to elevate assertions like "name mentioned/photo narrative/secondhand reports" to definitive claims of "visiting LSJ and participating in misconduct" at the factual level. The latter should remain classified within the "uncertain zone" under the current evidence structure, requiring ongoing updates through the "claim-evidence-gap" methodology rather than substituting moral innuendo for evidence. Simultaneously, this study offers more explanatory mechanism insights: controversy stems not solely from factual scarcity, but from a four-tier slippage mechanism involving "fact chains—context chains—narrative chains—accusation chains." Within this mechanism, "mentions" and "co-presence" are continuously recoded by dissemination systems into 'visits' and "participation," solidifying into public perceptions through correction lag and cross-platform migration. The core contribution lies in deconstructing and engineering this slippage mechanism: modeling fact production and narrative diffusion as distinct computable entities, leveraging data assets D2–D8 to implement evidence integration, conflict governance, and contamination isolation. A credibility/validity dashboard serves as the sealing device, enabling conclusions to be recalculated with new materials and audited with rule revisions—thus transforming "controversy" into a manageable evidence engineering problem.

Table 9: Summary of Key Judgements (Audit Style)

Judgement Number	Judgment Content (Summary)	confidence level	Primary Evidence Structure	Falsifiable conditions
J1	In March 2006, a relevant seminar was held in Saint Thomas, USVI, and received coverage in the local media.	High	Class B fact chain as primary; κ passed; anchor passed	Authoritative correction refutes the existence of the meeting or errors in its time and location
J2	There is a description of the conference/travel experience, but it is insufficient to pinpoint LSJ.	Medium	Textual corroboration; geographical landing gap	Verifiable transport/landing records are available, accurate to LSJ.
J3	The corrective material emphasises that 'mention does not equate to accusation' and highlights the risk of misrepresentation.	High	Grade A correction material	Authoritative materials have overturned the corrective conclusions through counter-analysis.
J4	Certain media outlets have forcibly linked photographic narratives with private island visits, constituting a Class C amplification chain.	Low-Medium	Class C amplification materials as the primary component	Where Class A judicial/official materials explicitly allege and substantiate
J5	No evidence of commercial association meeting the explicit threshold has been identified in the current window.	Medium	Lack of direct contractual/patent/investment financing ties	Discover Verifiable Commercial Documents: Connecting Workshops

13. Limitations and Future Work

Limitations primarily stem from two gaps: content gaps in presentations and geographical location gaps. Future work should leverage a reproducible retrieval repository to systematically collect archival materials from organizers, attendee resumes, academic institution press releases, and web archives. Each newly added piece of evidence should be logged in a change log with version numbers.

This study proposes a reproducible completion pathway for future work: prioritize retrieval of key materials from the seminar content chain (official agendas, presentation lists, manuscripts/videos/minutes, lecture records from attendees' personal pages and resumes) to further refine the “thematic and discussion ecosystem” into “verifiable individual statements”; Simultaneously, impose stricter thresholds on any expansive claims involving “private island visits,” requiring precise temporal and geographic anchors alongside independent corroboration of equivalent rigor before elevating them to factual status. All new evidence must undergo identical deterministic algorithmic coding and validity/reliability verification to prevent consistency collapse or

contamination backflow during expansion. By unifying evidence governance, statistical verification, and dissemination mechanism analysis under a single argumentative framework, this research ultimately achieves not “endorsing a particular narrative,” but establishing a reproducible evidence order for public disputes: providing definitive conclusions where verifiable, marking gaps and outlining completion pathways where unverifiable, and setting strict isolation and audit thresholds where contamination risks exist. This enables discourse to return to the fundamental principles of fact production.

Appendix: Reproducible Coding and Verification Process

All data assets, coding manuals, algorithmic coding specifications, and reliability/validity verification procedures from this study have been packaged and archived as a dataset. This ensures third parties can reproduce the entire process—"evidence chain audit → judgment sealing"—without accessing any non-public materials. This appendix provides only key rules, field definitions, and calculation steps that are highly consistent with the main text; the complete document list is in Appendix C.

Appendix A: Coding Manual and Variable Dictionary

The coding manual standardizes evidence grading, narrative clustering, and conflict flagging to prevent "different interpretations of the same entry under varying rules." This study employs a three-tier core coding system: Evidence Grade (A/B/C), Narrative Cluster (fact/context/amplification/correction), and Conflict Flag (conflict_flag).

Table A1 Core Coding Dimensions and Determination Criteria

Coding dimension	value	Criteria for Determination (Necessary Conditions)	Typical source types	Corresponding chapter/table in the main text
Evidence Grade	A/B/C	A: Authoritative primary records or verifiable first-hand publication pages, containing explicit details of time, location and responsible entity; B: Authoritative secondary reports/institutional records, capable of cross-verification; C: Paraphrased accounts, screenshots, secondary collations or materials lacking key elements.	A: Court/Government/Institutional Archives/Original Web Pages; B: Mainstream Media In-Depth Reports, Institutional Announcements; C: Secondary Dissemination on Social Platforms, Unverified Screenshots.	Chapter 4 (Evidence Governance); Table 5 (Evidence Ledger); Table 6 (Reliability)
Narrative Cluster	fact/context/amplification/correction	fact: Directly describes verifiable facts; context: Explains background without carrying personal assertions; amplification: Secondary dissemination and template-based rhetoric; correction: Fact-checking and rebuttal evidence.	News reports, academic conference pages, social media posts, fact-checking organisations	Chapters 10–11 (Diffusion and Correction); Table 8 (Validity)
Conflict Flag	No conflicts (0) / Resolvable conflicts	Mark inconsistencies in date, location, subject, or level of evidence for the same assertion; assign a value of 1 if explainable by 'discrepancies in accounts', otherwise assign a value of 2.	Cross-media reporting discrepancies, secondary citation temporal drift, screenshot	Chapter 6 (Conflict Management); Quality Audit Master Schedule

	(1) / Unresolved conflicts (2)		reprocessing	(Chapter 13)

Appendix B: Algorithm Coding Specifications

To mitigate subjective fluctuations in manual coding, this study employs reproducible algorithmic coding across three critical stages: "evidence classification", "narrative clustering", and "claim extraction". Two independent rule sets (g1/g2) are utilised to construct a 'dispute pool', thereby quantifying coding consistency and stability (see Table 6). Algorithmic coding does not replace the evidence itself, but rather translates 'how the same evidence is utilised' into auditable, deterministic rules.

Table B1 Algorithmic Coding Process (corresponding to Chapter 4 and Tables 6 and 8 in the main text)

Steps	Input/Output	Core Algorithms/Rules	Corresponding to the dataset file
1) Evidence cleansing and deduplication	Input: Original evidence entries; Output: D2_cleaned (after deduplication)	Deduplication: URL standardisation, title similarity, publication date window; retain the earliest and most authoritative version	data/D2_EvidenceLedger_Cleaned_AlgoCoded.csv; process_docs/changelog.csv
2) Grading of evidence (G1/G2)	Input: D2_cleaned; Output: grade_g1, grade_g2, and grade_consensus	A weighted rule based on verifiable elements (time/location/subject), source authority, and the number of corroborating references; Entries with discrepancies are placed in the dispute pool.	data/D2_EvidenceLedger_Cleaned_AlgoCoded.csv; data/RV_Dashboard_AlgoCoded.csv
3) Narrative Clustering	Input: Evidence text/title; Output:cluster_consensus	Keyword rules + template detection + bias correction indicators; prioritise fact-checking texts for bias correction. (correction)	data/D2_EvidenceLedger_Cleaned_AlgoCoded.csv
4) Advocates atomic extraction	Input: Evidence text; Output:D6_ClaimAtoms	Rule-based segmentation based on syntactically triggered words and entity slot extraction; Each assertion is bound to a source_id (evidence ID).	data/D6_ClaimAtoms.csv
5) Scoring and Dashboard Generation	Input: D2/D6/D7/D8; Output: Reliability/Validity Dashboard and Quality Audit Sheet	Consistency: Cohen's κ; Stability: Pearson/Spearman; Validity: Anchor consistency rate, bias control stability	data/RV_Dashboard_AlgoCoded.csv; data/D8_TimeBackbone.csv

Algorithm implementation utilizes reproducible scripts for execution, archiving rule thresholds, field mappings, and change logs together. Any rule adjustments must synchronously update the dispute pool ratio and reliability metrics to prevent "post-fitting caused by parameter tuning."

Appendix C: Reliability/Validity Verification and Reproducibility Workflow

The reproducibility process follows the sequence "Data → Coding → Scoring → Verification → Sealing Decision," with each update locked via version numbers and change logs. The table below outlines the minimal reproducible workflow and deliverables.

Table C1: Minimal Reproducible Workflow and Quality Control Gates (QC Gates)

stage	Key operation	Output (file/table)	Quality Control Gate
A Data Loading	Read D2/D6/D7/D8; verify field integrity and ID uniqueness	D2_EvidenceLedger_Cleaned_AlgoCoded；D6_ClaimAtoms；D7_MediaAssets；D8_TimeBackbone	Missing key fields (date/source/ID) will result in direct rejection or downgrading to C.
B Code Execution	Execute rule sets g1/g2 to complete classification and clustering; generate dispute pool.	grade_g1、grade_g2、grade_consensus；Dispute Pool Statistics	If the dispute pool ratio exceeds 0.20, rework is triggered: tighten thresholds or introduce additional anchor rules.
C Grading and Conflict Management	Calculate verifiability scores and mutual verification counts; generate conflict matrices and annotate unresolved causes.	VerifiabilityScore；conflict_flag；List of Conflict Entries	Unresolved conflicts shall not be subject to 'affirmative factual assertions'.
D Reliability Test	Consistency (κ); Stability (correlation coefficient); Disputed pool ratio	RV_Dashboard: Corresponding Metrics for Table 6	$\kappa \geq 0.80$ and correlation ≥ 0.90 constitutes a pass; otherwise, only a 'low confidence' judgement may be output.
E Validity Testing	Anchor consistency rate (Class A gold set); Bias control stability; Cross-source mutual verification	RV_Dashboard: Corresponding Metrics for Table 8	Anchor consistency rate ≥ 0.90 and stability correlation ≥ 0.90 constitutes a pass.
F Determining Sealing and Falsifiability Conditions	Generate J1–J5 judgements and bind evidence/claim IDs; write to Table 7	Table 9 (Key Judgements); Table 7 (Judgement Traceability Matrix)	Any judgement must be accompanied by conditions under which it can be disproved; inference must not be substituted for 'lack of

			evidence'.

Appendix D: Dual-Algorithm Coding Evidence

This appendix supplements the names, principles, and outputs of two independent coding algorithms, providing formulas for calculating reliability (consistency/stability/coherence) and validity (accuracy/authenticity/objectivity) based on both coding systems, along with key calculation steps and verifiable results. The terms “Algorithm A/Algorithm B” correspond one-to-one with fields in the dataset: Algorithm A = (g1, c1); Algorithm B = (g2, c2).

Table D1: Dual-Algorithm Coding Schemes (Algorithm A vs Algorithm B)

Algorithm	Algorithm Name	Type	Input	Output	Core Determination Mechanism
Algorithm A	Rule Scoring - Threshold Encoder	Explainable Rule Systems	Authority/TimeSpecificity/GeoSpecificity/CrossCorroboration Structural features; evidence type; publisher authority level	Evidence level g1 ∈ {A, B, C}; Narrative cluster c1 ∈ {factual, corrective, contextual, amplified}	First calculate the VerifiabilityScore, then map the level according to the threshold; narrative clusters employ the "fact/allegation/correction/amplification" rule dictionary for event alignment gating.
Algorithm B	Weakly Supervised Consistency Encoder	Weakly supervised/ constrained learning	Evidence text + structured features; Anchor fact set constraints; Cross-source consistency constraints	Evidence level g2 ∈ {A, B, C}; Narrative cluster c2 ∈ {factual, corrective, contextual, amplified}	Without altering the definitions of scoring dimensions, calibrate classification boundaries using anchor constraints and consistency constraints; outputs remain independent of Algorithm A, with discrepancies entering a 'dispute pool' for audit.

(1) Consistency Reliability: Cohen's kappa (κ) was used to assess consistency between the two coding sets. Formula: $\kappa = (P_o - P_e) / (1 - P_e)$, where P_o is the observed agreement rate and P_e is the random agreement rate (calculated from marginal distribution).

(2) Stability Reliability: After applying controlled perturbations to scoring weights, compare linear correlation (Pearson r) and rank correlation (Spearman ρ) between original and perturbed scores. If both r and $\rho \geq 0.90$, the scoring is deemed stable to minor weight variations.

(3) Accuracy Validity: Using the A/B-level "anchor fact set (gold set)" as the benchmark, calculate the Anchor Accuracy = 1 − (#mismatches / #anchor entries). The threshold for this study is set at ≥0.90.

Table D2: Comparative Outputs and Dispute Pool Summary (Algorithm A vs Algorithm B)

Project	AlgorithmA（g1/c1）	AlgorithmB（g2/c2）	Note
Total number of evidence entries n	10	10	Each item of evidence within the same batch shall be assigned a distinct code to ensure comparability.
Evidence Level Distribution	A=5, B=3, C=2	A=4, B=4, C=2	Levels are used for fact chain gating (A/B) and risk warnings (C).
Narrative Cluster Distribution	Facts=3, Correction=2, Context=3, Amplification=2	Facts=3, Correction=2, Context=4, Amplification=1	Narrative clusters are employed for the isolation of the 'fact layer/dissemination layer' and the identification of contamination.
Dispute Pool Ratio	0.100	0.100	Dispute Pool = Entries where the two algorithms disagree; enters conflict resolution and review.
Example of a disputed pool entry	E005：g1=A g2=B(Differences in authority and textual framework result in differing boundaries)	E010：c1=Magnify c2=Background (Enlargement determination requires cross-platform transition/depth gating)	The example illustrates the mechanism only; the full list of disputed entries can be found in the dataset.

Table D3: Inter-Algorithm Agreement (Cohen's κ) and Confusion Matrices

object	Confusion Matrix	P_o	P_e	κ	Judgement （threshold=0.60）
Evidence Level（g1 vs g2）	[[4,1,0],[0,3,0],[0,0,2]] (A,B,C)	0.900	0.360	0.843750	Pass
narrative cluster （c1 vs c2）	[[3,0,0,0],[0,2,0,0],[0,0,3,0],[0,0,1,1]] (Facts,	0.900	0.270	0.863014	Pass

	Corrections, Context, Amplification)				
Validity dimension	Anchor point scale n	Number of consistent entries	Consistency rate	Threshold	Conclusion
Accuracy: Consistency rate of evidence level anchors	6	6	1.000	≥0.90	Through (to demonstrate that the encoding has not deviated from the verifiable anchor point facts)
Accuracy: Consistency rate of narrative cluster anchor points	10	10	1.000	≥0.90	Through (to demonstrate that clustering does not conflate the propagation layer with the factual layer)
Supplement: Consistency of single algorithms with respect to anchor points	Level: n=6; Cluster: n=10	Algorithm A: Level 1.000; Cluster 1.000	Algorithm B: Level 1.000; Cluster 0.900	≥0.90	Algorithm B exhibits one boundary discrepancy; this is resolved through the application of the consensus layer (cluster_consensus) and gated rework rules.
Subject of Stability Testing	Original weighting (orig)	Disturbance weighting (alt1)	Pearson r	Spearman ρ	Judgement (threshold=0.90)
Verifiability Score	0.35/0.25/0.20/0.20 (Authority/Time/Geo/Cross)	0.40/0.25/0.20/0.15 (Authority/Time/Geo/Cross)	0.984001	0.975610	Pass
Calculation Notes	orig = Σ w_i x_i	alt1 = Σ w'_i x_i	The correlation coefficient is calculated using all n=10	Rank correlation is used to test for stability in ordering.	Satisfy stability gating

			evidence ratings.		

14. Data and Materials Availability

The study's reproducible artefacts (D2 Evidence Ledger, D6 Claim Atoms, D7 Media Assets Register, D8 Timeline Backbone), coding manuals, rule definitions, and reliability/validity dashboards are packaged as a dataset to enable third-party reproduction of the full pipeline from evidence ingestion to judgement sealing. Versioning and change logs record rule updates, dispute pool changes, and any reclassification events triggered by newly located primary sources.

15. Ethics and Responsible Communication

This paper adopts a harm-minimising OSINT posture: it does not infer motives, does not conflate mention/co-occurrence with misconduct, and treats contested claims as auditable propositions that require primary, time- and geo-anchored evidence. Secondary screenshots, paraphrased documents, and high-salience narratives are constrained to the amplification layer and may only be used to analyse diffusion and correction delay. All claims are presented with explicit evidence grades, gating rules, and falsifiability conditions to reduce interpretive latitude and to support responsible downstream use.

Limitations and Validity Boundaries

While the proposed framework demonstrates strong reliability and validity performance under the reported conditions, its effectiveness remains bounded by the construction scope of the Class A anchor set and the current sampling window. The gold-standard anchor set prioritizes high-certainty sources, which strengthens internal validity but may reduce coverage of emergent or weak-signal narratives. Consequently, conclusions derived from materials outside the anchor coverage are deliberately constrained in confidence level and are not escalated to strong factual judgments.

Reproducibility and Audit Extension

To further reduce analytical discretion, future iterations will publish a minimal reproducibility package, including input data hashes, coding scripts, parameter thresholds, and expected output schemas. This extension will allow independent auditors to reproduce the full Data–Coding–Scoring–Verification–Sealing workflow and to quantitatively assess deviation risks under alternative parameterizations.

Contribution Clarification

This study contributes beyond conventional text-analytic approaches by operationalizing evidentiary reasoning as a gated, auditable mechanism rather than a post-hoc interpretive exercise. By binding conclusions to explicit reliability thresholds, anchor-based validity checks, and mandatory re-coding rules, the framework materially reduces interpretive drift, analyst variance, and non-replicable inference in intelligence-oriented text analysis.

References

ABC News Australia. (2024, January 3). What we found in the Jeffrey Epstein documents. ABC News. https://www.abc.net.au/news/2024-01-04/what-we-found-in-the-jeffrey-epstein-documents/103285072

Associated Press (via CityNews). (2024, January 4). Fact Focus: False claims swirl around release of Jeffrey Epstein court documents. CityNews Halifax. https://halifax.citynews.ca/2024/01/04/fact-focus-false-claims-swirl-around-release-of-jeffrey-epstein-court-documents/

Edge.org. (2006, July 5). The energy of empty space that isn't zero. Edge Foundation. https://www.edge.org/conversation/lawrence_m_krauss-the-energy-of-empty-space-that-isnt-zero

Fox News. (2024, January 6). Jeffrey Epstein name-dropped Stephen Hawking, photos show physicist on sex trafficker's island. Fox News. https://www.foxnews.com/us/jeffrey-epstein-name-dropped-stephen-hawking-photos-show-physicist-sex-traffickers-island

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Snopes. (2023, May 20). Fact check: Photo shows Stephen Hawking at a conference funded by Jeffrey Epstein. Snopes. https://www.snopes.com/fact-check/stephen-hawking-epstein-conference/

St. Thomas Source. (2006, March 15). World's Top Physicists to Meet in the Virgin Islands. St. Thomas Source. https://stthomassource.com/content/2006/03/15/worlds-top-physicists-meet-virgin-islands/

St. Thomas Source. (2006, March 17). Physicists Debate Gravity at St. Thomas Symposium. St. Thomas Source. https://stthomassource.com/content/2006/03/17/physicists-debate-gravity-st-thomas-symposium/