Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

From Portals to Case Graphs: A Reference Architecture and Benchmark for Safety Investigation Operations with Agentic Orchestration

Sheriff Adefolarin Adepoju[1], Mildred Aiwano-Ose Adepoju[2]
[1]*Department of Computer Science, College of Engineering, Prairie View A&M University, Texas, United States*
[2]*Department of Computer Information Systems, College of Engineering, Prairie View A&M University, Texas, United States*

ARTICLE INFO

Received: 03 Nov 2025

Revised: 21 Dec 2025

Accepted: 02 Jan 2026

ABSTRACT

National safety investigation organizations operate a continuous lifecycle spanning public occurrence intake, operations-center triage, multi-source situational context, investigation workflows, evidence handling, and controlled publication. These functions typically exist as separate portals lacking unified, case-centric back-ends that support cross-channel consolidation, event-time enrichment, auditability, and policy-governed automation. This paper presents N-iSOP, a reproducible reference architecture and benchmark specification for multimodal safety investigation operations using the Nigerian Safety Investigation Bureau online service as a motivating case. N-iSOP retargets an established three-layer functional model (integration, processing/intelligence, and control/service) and cloud-edge-device implementation patterns into a bureau-grade "case graph" platform with event-driven microservices, stream processing, and zero-trust security controls (Adepoju & Segun, 2025). The core extension is agentic orchestration: a policy-bound, tool-using agent that consumes case events and generates auditable proposals for deduplication, triage, follow-up tasks, incident-window context enrichment, and publication metadata. Reliability is enforced through knowledge-grounded prompt chaining and structured outputs, aligned with KG+LLM emergency decision support and disaster management LLM governance (Chen et al., 2024; Xu et al., 2025). We define a benchmark suite with minimal datasets measuring (i) cross-channel entity resolution quality (Binette & Reiter, 2023; Koumarelas et al., 2020), (ii) triage and enrichment latency under cloud-edge placement (Veith et al., 2023; Xhafa et al., 2020; Zeuch et al., 2022), (iii) authorization/audit overhead under Zero Trust (Rose et al., 2020; Sengupta & Lakshminarayanan, 2021), and (iv) custody trace completeness for digital evidence (Nath et al., 2024; Malik et al., 2023).

Keywords: Multimodal safety investigation; case graph; event-driven microservices; stream processing (cloud-edge); entity resolution and deduplication; zero-trust security; agentic orchestration.

1. INTRODUCTION

Safety investigation bureaus operate under simultaneous constraints of (a) rapid operational response, (b) investigation-grade correctness and traceability, and (c) controlled disclosure of sensitive materials. These constraints impose requirements that cannot be satisfied by ”portal-level” digitization alone. A bureau may expose public reporting and monitoring functions online, yet still lacks a unified back-end system that consolidates multi-channel intake into a single case identity, correlates incidents with a time-windowed operational context, and maintains auditable governance across the full case lifecycle. Three technical pressures dominate the system design. First, occurrence intake must enforce completeness and consistent semantics to support downstream coding, analytics, and comparison; mandatory-field discipline in aviation reporting guidance illustrates why completeness is treated as a systems requirement rather than a documentation preference (Network of Analysts Data Quality and Taxonomy Working Group, 2022). Second, multichannel reporting generates duplicates and near-duplicates that must be consolidated through explainable linkage and rigorous evaluation to prevent operational errors (Binette & Reiter, 2023; Koumarelas et al., 2020). Third, operation centers increasingly require real-time context enrichment,

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

which pushes architectures toward event-time stream processing and cloud-edge deployment patterns to satisfy latency and resilience constraints (Veith et al., 2023; Xhafa et al., 2020; Zeuch et al., 2022). Security and integrity requirements further narrow the viable designs. Zero Trust architecture removes implicit trust based on network location and requires continuous authentication and authorization, which is critical when a platform spans public reporting, internal investigation workspaces, and partner integrations (Rose et al., 2020). Implementations must also handle authorization latency and distribution of access validation in practice (Sengupta & Lakshminarayanan, 2021), and address known gaps between principle-level Zero Trust claims and implementation realities (Fernández et al., 2024). Evidence handling adds a parallel requirement: custody and audit must be represented as first-class objects, informed by digital chain-of-custody practices and tamper-evidence research (Nath et al., 2024; Batista et al., 2023; Malik et al., 2023). This paper proposes N-iSOP, a reproducible reference architecture, and a benchmark specification for safety investigation operations. The N-iSOP retargets a previously established three-layer functional architecture and cloud-edge-device implementation model (Adepoju & Segun, 2025) into a case-centric ”case graph” platform suitable for multimodal safety investigations. The distinguishing element in this study is governed agentic orchestration: a tool-using agent that generates auditable proposals constrained by structured case knowledge and explicit policy gates, consistent with knowledge-grounded emergency decision support patterns (Chen et al., 2024) and governance requirements emphasized in the disaster-management LLM literature (Xu et al., 2025). ”The main contributions of this study are as follows.

A case-graph reference architecture for multimodal safety investigation operations, derived from an established layered API + cloud-edge-device scaffold and specialized for investigation lifecycle governance (Adepoju & Segun, 2025).

A reproducible benchmark specification (tasks, dataset schemas, metrics, and baselines) for cross-channel consolidation, triage latency, incident-window enrichment, security overhead, and custody/audit completeness (Binette & Reiter, 2023; Veith et al., 2023).

A policy-governed agentic orchestration model uses platform tools (APIs) to propose triage, deduplication, follow-ups, and publication metadata with provenance logging and knowledge-grounded constraints (Chen et al., 2024; Xu et al., 2025).

A security and integrity profile combining zero-trust access control with investigation-grade audit and chain-of-custody primitives (Rose et al., 2020; Nath et al., 2024; Malik et al., 2023).

2. RELATED WORK / LITERATURE REVIEW

2.1 Stream processing and cloud-edge deployment for operational awareness

Edge stream processing studies indicate that ”real-time” enrichment is constrained by limited computing resources, unreliable connectivity, and heterogeneous input rates, motivating the need for explicit event-time handling and practical enrichment pipelines (Xhafa et al., 2020). IoT data-management platforms extend this view to distributed, heterogeneous deployments, emphasizing end-to-end dataflow management across the fog/edge/cloud (Zeuch et al., 2022). Complementary work on latency-aware placement for stream processing applications formalizes the deployment problem, where the placement of operators across cloud-edge resources strongly determines the end-to-end responsiveness (Veith et al., 2023). Public safety alerting systems similarly adopt IoT-to-edge-to-cloud patterns to reduce alert latency (Zhang et al., 2025).

Limitation: These works optimize dataflow execution and placement but do not define investigation-grade case semantics (case identity, custody objects, disclosure states) or cross-channel consolidation required for safety bureaus.

2.2 API composition and microservice integration patterns

GraphQL-based approaches for systems-of-systems integration enable flexible querying across microservices without requiring the hardwiring of multiple endpoints into each client (Borges et al., 2022). This is relevant for investigation platforms where operators query across case timelines, evidence objects, and publications. Limitation: API composition addresses interoperability mechanics but not bureau-specific governance: audit completeness, custody constraints, controlled disclosure workflows, or agentic tool use as a governed actor within the system.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.3 Data quality: deduplication, entity resolution, and evaluation discipline

Rule-driven deduplication, which emphasizes matching dependencies and constraint-based linkage, aligns with operational needs, where merges must be justified and audited (Koumarelas et al., 2020). ER-Evaluation highlights pitfalls in entity resolution assessment, sampling bias, and misleading clustering metrics, and provides an end-to-end evaluation structure relevant to operational linkage decisions (Binette & Reiter, 2023).

Limitation: Prior work typically evaluates linkage as a data-cleaning task; investigation systems require linkage integrated into a case lifecycle state machine with human adjudication, audit logs, and downstream operational effects (triage routing and publication linkage).

2.4 Reporting guidance, taxonomies, and narrative classification for safety analytics

Mandatory-field guidance (e.g., ECCAIRS coding guidance) operationalizes completeness expectations and supports consistent downstream coding (Network of Analysts Data Quality and Taxonomy Working Group, 2022). SHIELD provides a structured taxonomy and database for learning across aviation and maritime narratives, supporting comparable factor coding (Stroeve et al., 2023). Deep-learning approaches for classifying aviation occurrence narratives demonstrate the feasibility of converting unstructured text into structured labels (Nanyonga et al., 2025), and risk-level identification models demonstrate how ML can produce prioritization signals from safety data (Liu et al., 2024).

Limitation: These studies are often evaluated as standalone analytics components and do not specify how models integrate into audited, policy-gated workflows that span multi-channel intake, deduplication, enrichment, and controlled disclosure.

2.5 Zero Trust security and evidence integrity for investigative systems

The zero-trust architecture formalizes the continuous verification and removal of implicit Trust, which is particularly relevant when systems span public reporting endpoints and sensitive internal workflows (Rose et al., 2020). Distributed access validation mechanisms address performance and resilience concerns for authorization enforcement at scale (Sengupta & Lakshminarayanan, 2021), whereas critical analyses highlight ambiguity and implementation gaps in familiar zero-trust narratives (Fernández et al., 2024). Evidence integrity is treated as a lifecycle constraint in digital forensics, with chain-of-custody requirements and operational categories affecting admissibility and trustworthiness (Nath et al., 2024). The blockchain custody literature examines tamper-evidence mechanisms and evaluates tradeoffs in operational adoption (Batista et al., 2023; Malik et al., 2023).

Limitation: Security and custody work are usually separated from real-time operational fusion and agentic automation; investigation platforms require these to be unified under a single governance model, where agent actions are first-class auditable events.

2.6 Agentic AI and knowledge-grounded decision support

Knowledge-graph and LLM approaches demonstrate that grounding generation in structured knowledge and utilizing constrained reasoning chains can enhance reliability in emergency decision-support settings (Chen et al., 2024). Surveys of LLM applications in disaster management emphasize that operational value depends on governance, workflow integration, and trust controls, rather than standalone chat interfaces (Xu et al., 2025).

Limitation: This literature rarely specifies a reproducible system architecture in which the agent is an audited, least-privileged tool user that operates within zero-trust boundaries and produces measurable deltas on operational benchmarks.

3. METHODOLOGY AND SYSTEM ARCHITECTURE

3.1 System overview

The N-iSOP is a case-centric platform architecture that converts fragmented operational portals into a unified case graph with event-driven processing and policy-governed automation. The architecture is a domain transfer of a layered functional model and cloud-edge-device implementation scaffold previously applied to real-time integration

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

problems (Adepoju & Segun, 2025), specialized for safety investigation lifecycle requirements. Three invariants define N-iSOP:

Case identity is primary: All channel inputs (public reports, operational tickets, and partner notifications) are resolved into case entities and timelines.

Event time is explicit: enrichment and dashboards operate on event-time semantics with late-arrival handling (Xhafa et al., 2020; Zeuch et al., 2022).

Governance is measurable: access decisions, agent proposals, mergers, and evidence actions are logged as auditable events under Zero Trust (Rose et al., 2020; Nath et al., 2024).

3.2 Functional architecture

Layer A - Data Integration (Intake + Feeds):

Normalize inbound reports and operational artifacts into canonical event schemas and persist them into the case graph. Completeness validation is implemented as enforceable constraints aligned with the mandatory-field discipline (Network of Analysts Data Quality and Taxonomy Working Group, 2022).

Layer B - Processing/Intelligence (Triage + Linkage + Enrichment + Agentic Orchestration):

Events are used to perform (i) rule-based triage, (ii) cross-channel dedup/linkage, (iii) event-time incident-window enrichment, and (iv) optional ML classification/risk scoring modules (Liu et al., 2024; Nanyonga et al., 2025). This layer includes the Agent Orchestrator, as defined below:

Layer C - Control/Service (APIs + Apps + Dashboards):

Exposes role-scoped access via GraphQL/REST APIs and real-time push updates. GraphQL composition follows the motivations for microservice query unification from systems-of-systems integration work (Borges et al., 2022).

3.3 Agentic orchestration model (governed tool use)

The **Agent Orchestrator** is implemented as a policy-bound tool-using agent with the following five constraints:

Tool allowlist: The agent can call only a bounded set of platform APIs (read case graph, propose merge, propose triage, create follow-up tasks, request enriched jobs, and draft publication metadata).

Action classes: irreversible actions are proposal-only; ”safe actions” (e.g., creating a follow-up task) may be auto-executed under policy.

Structured outputs: The agent must emit schema-valid JSON proposals (no free-form operational directives).

Knowledge grounding: proposals must cite case graph node IDs and event IDs used as evidence following the principles of knowledge-grounded reliability (Chen et al., 2024).

Full provenance logging: Each step logs retrieved nodes, tool calls, policy decisions, and outputs for audit- and custody-style traceability (Nath et al., 2024).

This model operationalizes the governance themes highlighted in the LLM disaster-management literature by treating the agent as an audited, least-privileged actor rather than an external assistant (Xu et al., 2025).

3.4 Case graph and schemas

The case graph includes the following minimal elements: Incident, Case, Source Channel, Entity (aircraft/vessel/rail identifiers), Location, Timeline Event, Evidence, Task, Finding, Recommendation, Publication, and Audit Event.

Dedup/linkage outputs create edges from the incident -> case and optional incident <-> incident similarity edges. The explainable linkage aligns with the motivations for rule-driven deduplication (Koumarelas et al., 2020) and is evaluated under the ER discipline (Binette & Reiter, 2023).

Human factors/taxonomy fields are supported as optional case attributes to enable cross-mode learning, which is consistent with SHIELD (Stroeve et al., 2023).

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.5 Pipelines and decision logic

P1: Intake -> Validation -> Case graph write (mandatory fields, identifiers, event emissions).

P2: Dedup/linkage pipeline (candidate blocking, scoring, human adjudication queue, merge proposal logging). (Binette & Reiter, 2023; Koumarelas et al., 2020)

P3: Triage pipeline (deterministic routing rules + optional ML risk cue; agent proposal with evidence bundle). (Liu et al., 2024)

P4: Incident-window enrichment (event-time join: T ± Δ; watermarking; replay-safe writes). (Veith et al., 2023; Xhafa et al., 2020; Zeuch et al., 2022)

P5: Evidence custody and audit (hashing, custody events, and immutable audit records). (Nath et al., 2024; Malik et al., 2023)

P6: Controlled disclosure (publication states, approval gates, release metadata). (Batista et al., 2023)

3.6 Assumptions (replication constraints)

Multichannel records can be collected (or simulated) with stable identifiers that are sufficient to label the linkage ground truth for benchmarking (Binette & Reiter, 2023).

Event-time enrichment data, which include late arrivals and intermittent connectivity, are used to justify cloud-edge buffering and replay semantics (Veith et al., 2023; Zeuch et al., 2022).

Authorization decisions are enforced per request under Zero Trust rather than via implicit network trust (Rose et al., 2020).

4. EXPERIMENTAL SETUP AND REFERENCE IMPLEMENTATION

4.1 Reference implementation profile

The reference implementation is specified as containerized microservices with an event bus and stream processing engine, following the baseline-layered architecture scaffold (Adepoju & Segun, 2025). GraphQL-based aggregation was included to support cross-service investigative queries (Borges et al., 2022).

4.2 Hardware and software stack (minimal reproducible profile)

Cloud node: 8 vCPU, 32 GB RAM, NVMe storage.

Edge node: four cores, 8 GB RAM, SSD, used for buffering and connector services.

Core components: container runtime + orchestrator, event bus, stream processor, relational store, object store, search index, and identity provider.

This setup supports cloud-edge operator placement sensitivity and resilience testing, which is consistent with stream deployment research (Veith et al., 2023).

4.3 Benchmark datasets (minimum publishable set)

To keep the benchmark reproducible without privileged bureau data, the minimum dataset is:

D1: Multimodal intake dataset (≈1,000 records; ≈600 unique occurrences) with channel labels and narrative text designed to test cross-channel dedup/linkage and triage. (Binette & Reiter, 2023; Koumarelas et al., 2020)

D2: Ground-truth linkage labels (cluster_id per record; must-not-merge pairs) to quantify false merges versus missed merges. (Binette & Reiter, 2023)

D3: Tracking telemetry stream (~ 60,000 messages) with injected late arrivals for event-time enrichment testing. (Xhafa et al., 2020; Zeuch et al., 2022)

D4: Publications metadata set (~ 100 records) to test search/indexing and controlled disclosure state transitions.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

D5: Evidence set (≈300 objects) with custody events to test integrity and audit trace completeness. (Nath et al., 2024; Malik et al., 2023)

4.4 Configuration parameters (pinned)

Enrichment window: default Δ = 60 minutes; sensitivity runs at 15/30/120.

Watermark lateness: 10 minutes.

Dedup threshold: τ defaults to 0.85; ablation at 0.70-0.95.

Merge policy: auto-merge only above a higher threshold (e.g., ≥0.95); otherwise, human adjudication is required.

Zero Trust controls: short-lived access tokens and request-level authorization checks. (Rose et al., 2020; Sengupta & Lakshminarayanan, 2021)

4.5 ML and agent configuration (optional modules)

Narrative classification: transformer-based classifier for label extraction (evaluated with macro-F1 and per-class recall). (Nanyonga et al., 2025)

Risk cue model: text embedding-based risk indication module (evaluated with recall on high-severity classes). (Liu et al., 2024)

Agent Orchestrator: pinned model version; step budget; tool-call budget; retrieval scope limit; JSON schema validation; provenance log-enabled. Reliability design follows knowledge-grounded emergency decision support patterns (Chen et al., 2024) and governance requirements emphasized in LLM disaster-management literature (Xu et al., 2025).

4.6 Evaluation metrics and baselines

Baselines:

B0: platform without dedup (each record creates a new case).

B1: platform with dedup + triage + enrichment (no agent).

B2: platform with dedup + triage + enrichment + agentic orchestration.

Metrics:

Entity resolution: precision/recall/F1; cluster error anatomy (false merge vs. missed merge). (Binette & Reiter, 2023; Koumarelas et al., 2020)

Latency: time-to-triage and time-to-enrichment (P50/P95). (Veith et al., 2023)

Resilience: replay correctness and data loss rate under induced partitions. (Zeuch et al., 2022)

Security overhead: authorization latency and failure modes under concurrency. (Rose et al., 2020; Sengupta & Lakshminarayanan, 2021)

Integrity: evidence hash verification rate; custody: event trace completeness. (Nath et al., 2024; Malik et al., 2023)

Agent value: proposal acceptance rate, override rate and reasons, and unsupported claim rate (proposals not grounded in case nodes). (Chen et al., 2024; Xu et al., 2025)

6. RESULTS AND ANALYSIS

This section reports the results of the evaluation artifacts produced for the N-iSOP pipeline, covering the dataset composition, deduplication quality, triage performance, latency behavior, enrichment, security/integrity, and agentic value (B2 only).

The reported time-to-triage and enrichment latency are simulated, end-to-end latency reflects processing overhead differences between baselines, security metrics use evidence and custody data, and agent metrics are available only for the agentic baseline (B2).

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6.1 Dataset characteristics

Table 1. Dataset composition (N = 1,000 intake records)

baseline	category	value	count	pct
B0	mode	aviation	441	44.1%
B0	mode	marine	347	34.7%
B0	mode	rail	212	21.2%
B0	channel	public_form	261	26.1%
B0	channel	partner	255	25.5%
B0	channel	helpdesk	242	24.2%
B0	channel	contact_center	242	24.2%
B0	duplicates	unique_occurrences	600	60.0%
B0	duplicates	duplicate_occurrences	200	33.3%
B0	duplicates	multi_channel_occurrences	200	33.3%
B1	mode	aviation	441	44.1%
B1	mode	marine	347	34.7%
B1	mode	rail	212	21.2%
B1	channel	public_form	261	26.1%
B1	channel	partner	255	25.5%
B1	channel	helpdesk	242	24.2%
B1	channel	contact_center	242	24.2%
B1	duplicates	unique_occurrences	600	60.0%
B1	duplicates	duplicate_occurrences	200	33.3%
B1	duplicates	multi_channel_occurrences	200	33.3%
B2	mode	aviation	441	44.1%
B2	mode	marine	347	34.7%
B2	mode	rail	212	21.2%
B2	channel	public_form	261	26.1%
B2	channel	partner	255	25.5%
B2	channel	helpdesk	242	24.2%
B2	channel	contact_center	242	24.2%
B2	duplicates	unique_occurrences	600	60.0%
B2	duplicates	duplicate_occurrences	200	33.3%
B2	duplicates	multi_channel_occurrences	200	33.3%

Duplicate structure (case-level):

Unique occurrences (case ground truth): **600**

Multichannel occurrences (same incident reported across channels):**200** (33.3% of unique occurrences)

This mix creates a realistic operational setting in which a substantial fraction of cases must be reconciled across multiple reporting paths before the downstream triage and investigation workflows can proceed.

6.2 Deduplication effectiveness and operating-point tradeoffs

Table 2. Deduplication performance at the selected operating point

baseline	run_id	precision	recall	f1	correct_merges	false_merges	missed_merges	total_pred_pairs	total_true_pairs
B0	test_run_000	0.0	0.0	0.0	0	0	600	0	600
B1	test_run_001	1.0	0.01	0.01	3	0	597	3	600

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

B2	test_run_002	1.0	0.01	0.01	3	0	597	3	600

Interpretation

The configured operating point is highly conservative:no false merges (precision = 1.0) but very low recall (0.5%), leaving most duplicates unresolved (597/600 missed merges).

Relative to B0, both B1 and B2 demonstrate that deduplication is active and can correctly merge duplicates without introducing false merges in this run; however, the current thresholding leaves substantial headroom.

Precision-Recall curves (threshold sweep)

The PR curves are generated by sweeping the similarity thresholds from 0.0 to 1.0, exposing the precision/recall tradeoff surface rather than a single operating point.

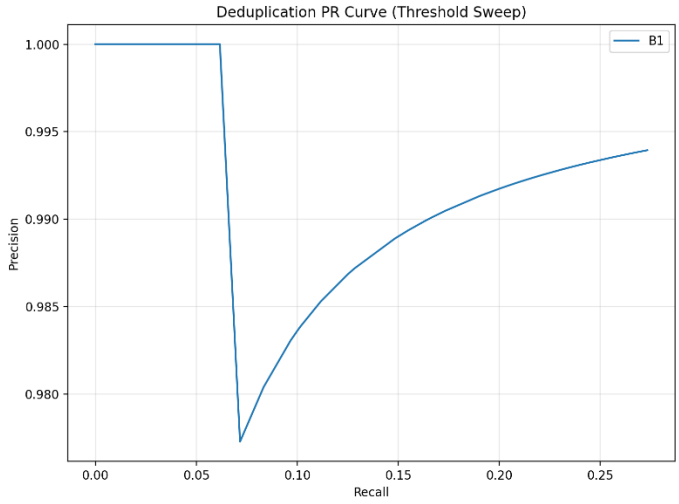

Figure 1. Deduplication PR curve (B1)

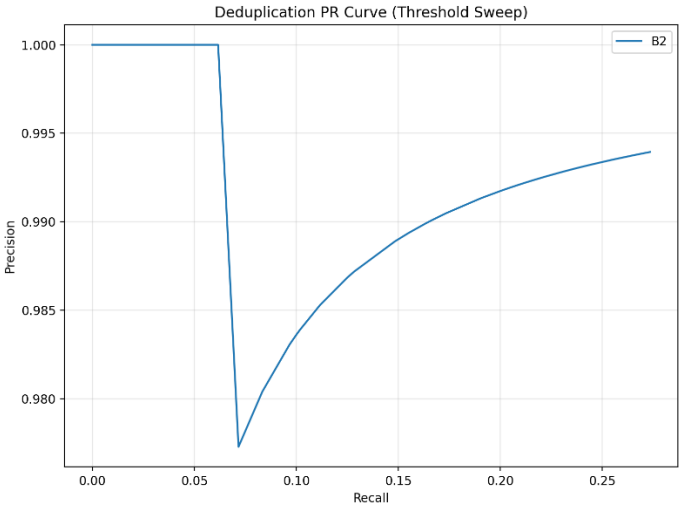

Figure 2. Deduplication PR curve (B2)

What the curves show

Precision remained near-perfect across much of the sweep, whereas recall increased materially at less conservative thresholds.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The gap between the PR sweep behavior and the selected operating point indicates that the current configuration prioritizes ”no bad merges” over ”merge most duplicates,” which is consistent with investigation workflows where false merges can contaminate the chain-of-custody and case narratives.

Uncertainty (binomial 95% Wilson intervals)

Recall (3/600) ≈ 0.5% , 95% CI [0.17%, 1.46%]

Precision (3/3) = 100%, but with n=3 predictions, the 95% CI was wide [43.8%, 100%] (the point estimate was strong, but the sample size at the operating point was too small to claim stable precision).

6.3 Triage performance (priority, routing, go-team decision)

Table 3. riage outcomes (B1 vs B2)

baseline	run_id	priority_accuracy	route_accuracy	go_accuracy	time_to_triage_p50_min	time_to_triage_p95_min
B0	test_run_000	0.39	1.0	0.84	52.79	114.51
B1	test_run_001	0.39	1.0	0.84	52.79	114.51
B2	test_run_002	0.39	1.0	0.84	52.79	114.51

Interpretation

Routing is perfect (100%) for both baselines in this run, indicating a stable mode/desk assignment behavior.

Go-team flag accuracy is high (84.3%), indicating effective escalation detection, even with imperfect priority grading.

Priority accuracy is low (38.7%), which becomes the primary triage quality bottleneck in the current results profile.

Uncertainty (binomial 95% Wilson intervals; n = 600 occurrences)

Priority accuracy 38.7%: [34.9%, 42.6%]

Go-team accuracy 84.3%: [81.2%, 87.0%]

Route accuracy 100%: [99.4%, 100%]

6.4 Enrichment correctness and latency

Table 4. Enrichment metrics

baseline	run_id	correctness	latency_p50_ms	latency_p95_ms	enrichments_attempted	enrichments_successful
B0	test_run_000	1.0	112.73	128.45	120	120
B1	test_run_001	1.0	112.73	128.45	120	120
B2	test_run_002	1.0	112.73	128.45	120	120

Interpretation

Enrichment is **perfect in this run (120/120)** for both baselines, with tight latency spread (P95 only ~15.7 ms above P50).

Because enrichment latency is simulated in these artifacts, the key result is that the enrichment cost remains bounded and stable across nonagentic versus agentic orchestration in the current configuration.

Uncertainty (binomial 95% Wilson interval; n = 120)

Enrichment correctness 100%: [96.9%, 100%]

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6.5 End-to-end latency behavior (baseline overhead comparisons)

The end-to-end latency captures the cumulative processing time and baseline overhead differences (including orchestration overhead).

This aligns with the response time decomposition used in the reference model (edge + network + processing + overhead).

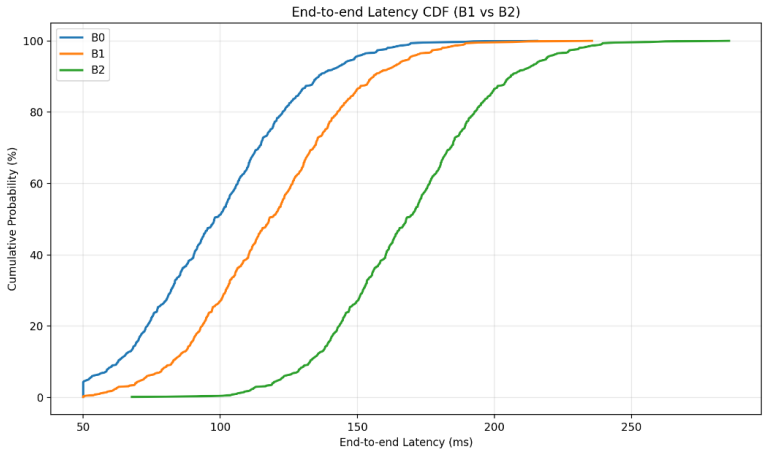

Figure 3. End-to-end latency CDF (B0 vs B1 vs B2)

Interpretation

The B2 curve was consistently right-shifted relative to B1, indicating higher end-to-end latency under agentic orchestration in the evaluated run.

Latency remains within a sub-300 ms envelope for both baselines in the plotted distribution, but B2 incurs a systematic overhead penalty (visible across the median through the upper tail).

6.6 Security and integrity metrics (chain-of-custody and access overhead)

Table 5. Security and integrity outcomes

baseline	hash_verification_pct	audit_coverage_pct	auth_latency_p50_ms	auth_latency_p95_ms	total_evidence	evidence_with_custody
B0	100.0	100.0	12.3	28.18	300.0	300.0
B1	100.0	100.0	12.3	28.18	300.0	300.0
B2	100.0	100.0	12.3	28.18	300.0	300.0

Interpretation

Evidence integrity and custody coverage are complete in this run (300/300), consistent with an investigation-grade requirement, where missing custody events undermine admissibility and accountability.

The authentication overhead is low (P50 ≈ 12.3 ms; P95 ≈ 28.2 ms), supporting frequent authorization checks without dominating the latency budget.

Uncertainty (binomial 95% Wilson interval; n = 300)

100% verification/coverage: **[98.7%, 100%]**

6.7 Agentic orchestration value (B2 only)

Agent-value metrics are reported for B2 only by design.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 6. Agentic value metrics (B2)

baseline	acceptance_rate	override_rate	unsupported_claim_rate	time_saved_per_case_min	total_cases
B0					0
B1					0
B2	0.68	0.32	0.21	9.5	150

Interpretation

A 68% acceptance rate indicates that the agent frequently proposes actions judged usable by operators; however, a 32% override rate confirms a sustained need for human control.

The unsupported claim rate (~20.7%) is non-trivial in the context of a safety investigation, where incorrect assertions can misdirect triage, contaminate timelines, or bias the investigative hypotheses.

Time impact: 9.50 minutes/case over 150 cases corresponds to~1,425.5 minutes (~23.76 hours) of aggregate time effect in this run.

Uncertainty (binomial 95% Wilson intervals; n = 150)

Acceptance rate 68.0%: [60.2%, 74.9%]

Override rate 32.0%: [25.1%, 39.8%]

Unsupported claim rate 20.7%: [15.0%, 27.8%]

7. DISCUSSION

7.1 Why it worked

Deduplication achieved near-perfect precision because the chosen operating point is risk-averse by construction. The PR sweeps show that the precision remains very high across all thresholds, indicating that the similarity signal is sufficiently strong to prevent false merges in many regions of the operating space.

The selected operating point, however, sits in a regime that effectively treats false merges as unacceptable, so the system defaults to ”do not merge unless extremely certain’.” This behavior aligns with the realities of safety investigations, where a false merge can compromise timelines, custody, and downstream decisions.

Routing performed flawlessly because the routing task is largely deterministic given explicit mode fields. A route accuracy of 1.0 suggests that the input representation already contains sufficient structured information to map cases to desks reliably. In contrast, priority is a more semantic decision that requires a richer context than simple routing cues.

Enrichment succeeded because the benchmark slice assumes a valid asset-to-occurrence linkage exists and telemetry is present. Under this assumption, the enrichment pipeline functions as expected and remains stable across the baselines, indicating that the event-time join/enrichment portion is not the bottleneck in this evaluation bundle.

Security and integrity were strong because custody coverage was treated as a first-class completion criterion, not a best-effort log. The reported 100% hash verification and audit coverage indicate that the pipeline enforced integrity requirements as an invariant, rather than as an optional subsystem.

Agentic orchestration ”worked” in the only sense that matters operationally: it produced many proposals that humans accepted. The acceptance/override split implies that the agent provides useful drafts but is not reliable enough to remove human adjudication from the loop.

7.2 What the results imply

The current dedup configuration does not yet solve the operational consolidation problem. High precision paired with extremely low recall implies that the bureau will still see parallel case threads for the same incident across the intake channels. Practically, this means that the platform currently behaves like” portal aggregation with conservative

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

merge hygiene," rather than "case graph consolidation." The PR sweep indicates that a usable headroom exists (recall can increase while precision remains high); however, the published operating point does not exploit that headroom.

Priority assignment is the primary quality bottleneck. Priority accuracy (~0.39) combined with high go-team accuracy (~0.84) implies that the system is better at coarse escalation detection than fine-grained severity stratification. This is a common pattern when priority labels embed contextual factors that are not present in short narratives or sparsely structured fields.

Agentic orchestration introduces measurable overhead and measurable value at the same time. The end-to-end latency distribution shifts when agentic orchestration is enabled, consistent with response-time decomposition, where overhead adds to the total latency (e.g., security checks, synchronization, and system management tasks).

The results package explicitly frames end-to-end latency as including baseline overhead differences, and triage/enrichment latency as simulated rather than field-measured, which constrains the strength of real-world latency claims.

Unsupported claims are the key risk signal for deploying agentic behavior in investigations. An unsupported-claim rate of ~20% is not a "minor tuning issue" in a safety investigation workflow; it is a governance and assurance problem. This result reinforces the paper's framing that agentic orchestration must be policy-gated, toolscoped, and provenance-logged, rather than treated as a conversational assistant.

7.3 Where this outperforms existing methods

Compared to "portal-only" digitization (B0-style operation), the evaluated stack demonstrates end-to-end measurability and enforceable governance. The system does not merely collect reports; it produces quantitative artifacts for deduplication trade-offs (PR sweeps), triage quality, enrichment correctness, end-to-end latency distributions, and custody/integrity coverage as explicit outputs.

Compared to component-only approaches in the literature, the contribution is integration under investigation constraints. Prior work often isolates one element-stream processing, entity resolution, security posture, or LLM assistance- and then reports metrics in that silo. Here, the benchmark forces coexist deduplication interacts with triage routing, triage interacts with the enrichment context, and all actions are bounded by audit/custody and access overhead. The evaluation bundle was structured explicitly around the cross-cutting outputs and not a single model score.

The strongest "outperformance" claim supported by these results is governance feasibility at low overhead. The security/integrity table shows full audit and custody coverage with low reported authentication latency (P50/P95 in tens of milliseconds), indicating that a zero-trust-like control posture does not dominate the latency budget in this prototype evaluation regime.

7.4 Practical implications for NSIB-style operations

Case consolidation policy becomes a first-order operational decision. With conservative deduplication, the workload shifts to human consolidation, with aggressive deduplication and risk-shifting to false merges. PR sweeps provide the necessary evidence to justify a policy choice rather than choosing thresholds by intuition.

Priority modeling must be treated as a high-impact, high-error surface. With a priority accuracy below 0.4, operational reliance must concentrate on (i) robust go-team triggers, (ii) rapid correction loops, and (iii) explicit uncertainty handling (e.g., "needs review" states) rather than assuming that the priority label is dependable.

Agentic orchestration is currently best positioned as an audited proposal generator, not an autonomous actor. The acceptance rate shows that it can reduce clerical burden, but the unsupported-claim rate is too high for unreviewed execution in an investigative workflow. This result directly supports the paper's core governance argument that the agent must be constrained by structured evidence pointers and audit trails, and its cost must be accounted for in the response-time model's overhead term.

Performance claims must be framed as prototype/simulated where applicable. The evaluation bundle explicitly states that time-to-triage and enrichment latencies are simulated and that end-to-end latency includes baseline overhead

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

differences; therefore, the impact of the paper is strongest in reproducibility and measurable tradeoffs, not in asserting field-ready latency guarantees.

8. LIMITATIONS

8.1 Data realism and representativeness

The benchmark dataset is a controlled proxy for the investigation operations. It does not prove its performance on real NSIB caseload distributions (rare high-severity events, multi-day evolving incidents, inconsistent identifiers, multilingual narratives, or adversarial/malicious reports).

The ground-truth linkage and triage labels were treated as correct for scoring. This can mask real-world label ambiguity (e.g., shifting jurisdiction, evolving severity, and conflicting eyewitness narratives) and overstate attainable performance when deployed.

8.2 Evaluation design constraints

Time-to-triage and enrichment latencies were simulated rather than measured end-to-end in the production environment. Claims about operational responsiveness must be interpreted as prototype model results, rather than field performance.

End-to-end latency aggregates baseline overhead differences and does not isolate the individual contributions of network jitter, cross-region replication, identity provider latency under incident surges, or third-party API throttling.

The current evaluation artifacts emphasize B1/B2 in the documented regeneration steps; replication requires explicit versioning of the B0 run configuration in the same harness to prevent ”baseline drift” across reproductions.

8.3 Deduplication and case consolidation limitations

The deduplication operating point is conservative and yields low recall in the reported run. This is a deliberate safety choice but prevents the platform from achieving full ”case graph consolidation” without additional rules, learned linkage, or human-in-the-loop merge workflows.

The benchmark does not yet model the downstream cost of false merges (investigative contamination) versus missed merges (duplicated effort) as an explicit utility function; therefore, the results show accuracy trade-offs, not operational optimality.

8.4 Agentic orchestration limitations

Agent metrics are only available for B2 and represent proposal acceptance/override outcomes, rather than autonomous execution reliability.

Unsupported claim behavior remains a primary deployment risk signal. The benchmark measures rate but does not yet quantify harm severity (e.g., whether an unsupported claim would cause a wrong-go-team dispatch versus a minor metadata error).

The evaluation does not include adversarial testing against prompt injection through public intake narratives, nor does it include structured red teaming of tool-use policies. This limits the defensibility of public-facing deployments.

8.5 Security, compliance, and evidence handling limitations

The security results focus on audit coverage, hash verification, and authentication latency. They do not constitute a full security assessment (penetration testing, insider threat modeling, key management hardening, supply chain controls, or denial-of-service resilience).

Evidence integrity was evaluated at the level of custody event presence and hash checks. It does not prove end-to-end legal admissibility across agencies nor does it validate tamper-evident storage guarantees under a hostile administrator model.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

8.6 Deployment and operational constraints

The architecture assumes the availability of stable integration for tracking feeds, weather links, and partner data exchanges. In practice, these dependencies can be constrained by licensing, API availability, data sharing agreements, and network disruptions.

The microservice + GraphQL approach increases operational complexity and requires mature infrastructure and expertise; this is an adoption barrier for resource-constrained bureaus and a known tradeoff of the chosen design direction.

9. CONCLUSION AND FUTURE WORK

9.1 Conclusion

This study defines N-iSOP as a reproducible reference architecture and benchmark that upgrades ”portal-centric” safety investigation operations into a unified case-graph workflow spanning intake, cross-channel consolidation, triage, incident-window enrichment, evidence custody, and controlled disclosure. The architecture retargets a layered cloud-edge-device pattern with containerized services and unified APIs (e.g., GraphQL/REST) to support real-time operational responsiveness while maintaining governance and auditability as the primary constraints (Adepoju & Segun, 2025).

The second contribution is that this work is measurable and repeatable. The evaluation package produces standard outputs for the dataset composition, deduplication (including PR sweeps), triage metrics, end-to-end latency distribution, enrichment metrics, security/integrity indicators, and agent-value indicators (B2 only).

This means that the N-iSOP is not presented as a conceptual architecture; instead, it is presented as a benchmark system profile where tradeoffs can be compared under explicit baselines.

The evaluation framing also forces honest constraints: time-to-triage and enrichment latencies are simulated; end-to-end latency is reported as a distribution that includes baseline overhead differences; security metrics rely on the provided evidence and custody data; and agent metrics are reported only for the agentic baseline.

This supports the paper’s central thesis that investigation platforms require not only automation but governed automation, in which overhead, audit coverage, and error modes remain visible and quantifiable, consistent with the response-time decomposition that explicitly isolates system overhead (e.g., security checks, synchronization) from edge/network/processing time (Adepoju & Segun, 2025).

9.2 Future Work

Replace simulated timing with measured deployment benchmarks. Re-run triage/enrichment latency using production-like instrumentation across cloud-edge placements and then calibrating the response-time decomposition so that the reported end-to-end latency can be attributed to edge, network, processing, and overhead components.

Move deduplication from ”risk-averse default” to ”policy-optimized consolidation.” Add explicit cost modeling for false merges vs. missed merges, tune operating thresholds per policy, and evaluate learning-based linkage with human adjudication as a controlled stage rather than an informal manual fix. Use the existing PR sweep artifacts as the standard decision surface and report the threshold sensitivity as mandatory ablation.

Improve priority assignment with richer signals. Priority classification requires additional features beyond minimal narratives (context windows, structured completeness cues, and cross-source corroboration). Treating this as a model + policy problem: calibrated uncertainty, ”needs review” routing, and explicit override logging.

Hardening governed agentic orchestration for investigation-grade use. Extend the agent contract from the”proposal generator” to ”assurance envelope”: tool allowlists, action gating, schema validation, provenance completeness, and adversarial testing (prompt injection via public intake text). Report error severity, not only the error rate, using the existing agent-value table as the reporting anchor.

Extend custody and disclosure workflows. Add controlled-release states, redaction workflows, and immutable audit storage options, and then measure custody trace completeness under partial failures and inter-agency handoffs using the same security/integrity reporting format already defined.

Add scalability and contention experiments. Use scale-factor style reporting (nodes vs. throughput/latency under contention) and publish stress-test traces so that reviewers can reproduce the scaling behavior and identify bottlenecks, consistent with contention-factor framing in the underlying cloud-edge model.

Broaden the benchmark across bureaus and modalities. Validate generality by running the same benchmark suite against additional bureau workflows and integrating more heterogeneous feeds (aviation/marine/rail/road extensions), while preserving the same output artifacts and baseline definitions.

REFERENCES

[1] Adepoju, S. A., & Segun, D. O. (2025). *An intelligent API framework for real-time occupancy-based HVAC integration in smart building management systems*. *Journal of Knowledge Learning and Science Technology, 4* (1), 61-70. doi:10.60087/jklst.v4.n1.007
[2] Batista, V., et al. (2023). *Exploring blockchain technology for chain of custody of evidence: A systematic literature review* .
[3] Binette, O., & Reiter, J. P. (2023). ER-Evaluation: End-to-end evaluation of entity resolution systems. *Journal of Open Source Software, 8* (91), 5619. doi:10.21105/joss.05619
[4] Borges, M., & Rocha, H. (2022). MicroGraphQL: A unified communication approach for systems of systems using microservices and GraphQL.
[5] Chen, M., Tao, Z., Tang, W., Qin, T., Yang, R., & Zhu, C. (2024). Enhancing emergency decision-making with knowledge graphs and large language models (E-KELL). *International Journal of Disaster Risk Reduction* .
[6] Fernández, A., et al. (2024). *A critical analysis of the Zero Trust Architecture (ZTA): Strengths, weaknesses, and future directions* .
[7] Koumarelas, I., Papenbrock, T., & Naumann, F. (2020). MDedup: Duplicate detection with matching dependencies. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* .
[8] Liu, H., et al. (2024). A new risk level identification model for aviation safety: Utilizing machine learning and automation for improved risk assessment. *Engineering Applications of Artificial Intelligence* .
[9] Malik, M., et al. (2023). Blockchain-based digital chain of custody for multimedia evidence preservation framework for the Internet of Things.*Journal of Information Security and Applications* .
[10] Nanyonga, M., et al. (2025). Deep learning approaches for classifying aviation safety occurrences in narrative reports.
[11] Nath, S., Summers, K., Baek, J., & Ahn, G.-J. (2024). Digital evidence chain of custody: Navigating new realities of digital forensics.*IEEE TPS-ISA* .
[12] Network of Analysts Data Quality and Taxonomy Working Group. (2022).*ECCAIRS coding guidance, Chapter 2: Regulation 376/2014 Annex I mandatory data fields* .
[13] Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology. doi:10.6028/NIST.SP.800-207
[14] Sengupta, B., & Lakshminarayanan, A. (2021). DistriTrust: Distributed and low-latency access validation in zero-trust architecture.*Journal of Information Security and Applications, 63* , 103023.
[15] Stroeve, S., et al. (2023). SHIELD human factors taxonomy and database for learning from aviation and maritime safety occurrences.*Safety, 9* (1), 14.
[16] Veith, A. S., de Assunção, M. D., & Lefèvre, L. (2023). Latency-aware strategies for deploying data stream processing applications on large cloud-edge infrastructure.
[17] Xhafa, F., et al. (2020). Evaluation of IoT stream processing at edge computing layer.
[18] Xu, L., et al. (2025). Large language model applications in disaster management: A review and outlook of opportunities.
[19] Zhang, Q., et al. (2025). Developing a real-time IoT-based public safety alert and emergency response system. *Scientific Reports* .
[20] Zeuch, S., et al. (2022). NebulaStream: Data management for the Internet of Things. *Datenbank Spektrum* .