A Defamation-Safe, Evidence-Anchored Communication Protocol Between AI Agents in Workers' Compensation SIU Workflows

Design Principles for Auditability, Human Accountability, and Inter-Agent Trust

Upendra Belhe, Ph.D.
Independent Research
January 2026

Category: Enterprise AI Systems • Governance • Insurance Operations
Keywords: Workers' Compensation, SIU, Artificial Intelligence, Agentic Systems, Evidence Provenance, Defamation Risk, Auditability, Claims Governance

Abstract. *Artificial intelligence is increasingly deployed in workers' compensation claims operations to surface inconsistencies, anomalies, and patterns warranting Special Investigations Unit review. However, many AI-enabled fraud detection systems fail under regulatory examination, discovery, or litigation—not due to model inaccuracy, but because their outputs are not governable, reproducible, or legally defensible. Unstructured communication between automated systems can embed implicit accusations, omit evidence provenance, and obscure accountability for decisions affecting statutory benefits. This paper introduces a formal, defamation-safe, evidence-anchored communication protocol governing interactions between two AI agents: a Fraud Recommendation Agent and an SIU Triage Agent. Designed for workers' compensation claims, the protocol defines message contracts, evidence bundle requirements, language constraints, decision boundaries, and human-in-the-loop escalation points. Rather than automating fraud adjudication, the protocol establishes a systems-level governance framework preserving investigative integrity while enabling continuous learning. The contribution is a citable, enterprise-grade design pattern reframing AI-enabled SIU workflows as a coherence and accountability problem rather than a prediction accuracy problem.*

1. Introduction

Workers' compensation claims administration operates within a uniquely constrained environment: statutory benefits, regulated timelines, extensive documentation requirements, and frequent retrospective scrutiny. Unlike many commercial decision workflows, the claim file is not merely an operational record; it is an evidentiary artifact that may later be reviewed by regulators, opposing counsel, arbitrators, or courts.

As artificial intelligence is introduced to assist claim handlers—especially in fraud detection—organizations face a mismatch between what machine outputs provide (scores, flags, embeddings, summaries) and what regulated environments demand (coherent narratives, traceable evidence, and accountable decision pathways).

In practice, many AI deployments fail not because models are incapable of learning patterns, but because the organization cannot explain, reproduce, or defend the chain from signal to action. The most persistent weakness is not analytics; it is the absence of a standardized communication protocol governing how automated systems share suspicion, evidence, and decisions.

This paper proposes a protocol-level solution: a formal, defamation-safe, evidence-anchored communication contract between a Fraud Recommendation AI Agent (FRA) and an SIU Triage AI Agent, with explicit human decision boundaries. The protocol is implementation-agnostic—independent of model choice, vendor platform, or underlying data stack.

2. Problem Context

2.1 Legal and Regulatory Sensitivity

Workers' compensation differs materially from other insurance workflows. Benefit eligibility, medical treatment, return-to-work accommodations, and indemnity payments are governed by statute and administrative rules. Investigative actions may be constrained by jurisdictional requirements, privacy constraints, and labor considerations.

Claim notes, investigative referrals, and internal communications may be discoverable. Language implying intent or guilt introduces defamation risk or creates adverse inferences. AI outputs that are unstructured, overly confident, or phrased as determinations materially increase legal exposure.

2.2 Common Failure Modes

Observed failure modes in naïve AI-to-SIU integration include: (i) implicit accusations embedded in referral summaries; (ii) missing evidence provenance ("high risk" without artifact references); (iii) probability scores without narrative grounding; (iv) automation bias, where investigators over-weight opaque scores; and (v) inability to replay decisions after model or data drift.

These are governance failures. Even high-performing models become operational liabilities if the system cannot

prove what was known, when it was known, and why escalation occurred.

3. Design Principles

Principle 1: No Autonomous Accusation. The protocol prohibits any agent from labeling a person, provider, or employer as fraudulent. Outputs may only express "indicators consistent with potential misrepresentation" and must be framed as recommendations for review.

Principle 2: Evidence Before Inference. Recommendations must be supported by an evidence bundle with artifact pointers (claim notes, forms, billing records, transcriptions), each accompanied by a brief explanation of relevance.

Principle 3: Reproducibility Over Accuracy. The ability to replay a decision under audit is often more important than marginal predictive improvements. The protocol requires strict versioning and data snapshotting.

Principle 4: Human Authority at Decision Boundaries. Case opening, adverse action, law enforcement contact, and surveillance decisions must be gated behind explicit human review and documented sign-off.

Principle 5: Language Safety as System Constraint. Defamation-safe language is a technical requirement. The protocol mandates neutral phrasing templates and disallows prohibited terms within automated summaries.

4. System Roles and Responsibilities

Fraud Recommendation AI Agent (FRA). The FRA monitors claim signals and generates structured referrals when predefined indicators are triggered. It must: (i) assemble an evidence bundle with provenance, (ii) include countervailing factors, and (iii) propose verification steps—never adverse actions.

SIU Triage AI Agent. The SIU Agent evaluates referrals against jurisdictional thresholds, organizational policies, and resource constraints. It may request additional evidence, classify disposition (open, monitor, decline, escalate), and return structured feedback to the FRA.

Human Supervisor. Human supervisors retain authority over consequential actions. The protocol makes this explicit by requiring conditional escalation and capturing approvals in the audit log.

5. Communication Protocol

5.1 Message Envelope

All inter-agent messages include a standardized envelope ensuring traceability, auditability, and replay. Required fields: `message_id`, `correlation_id`, `claim_id`, `jurisdiction`, `policy_id`, `sender_agent`, `receiver_agent`, `timestamp_utc`, `purpose`, `privacy_scope`, `legal_hold_flag`, `confidence_band`, `model_version`, and `ruleset_version`.

This envelope acts as a compliance boundary: downstream actors verify scope, identify model provenance, and map messages into an audit trail without relying on unstructured text.

5.2 Referral Packet (FRA to SIU)

A referral packet contains: (i) a neutral referral summary, (ii) standardized indicator codes, (iii) an evidence bundle with artifact pointers, (iv) a risk assessment expressed as calibrated bands, (v) countervailing factors, and (vi) recommended verification steps with explicit questions the SIU Agent should resolve.

Risk assessments must avoid implying intent. The objective is not to determine fraud, but to justify why the claim merits SIU attention, grounded in concrete artifacts.

5.3 Evidence Bundle Requirements

Each evidence item includes: `evidence_id`, `source_system`, `artifact_pointer`, `excerpt_or_feature`, `why_it_matters`, and `reliability_grade` (A/B/C) based on provenance. Bundles should include both supporting and countervailing artifacts to reduce confirmation bias.

Evidence bundles must be designed for discovery: reviewers must retrieve the original artifact and verify that the summary did not distort context.

5.4 Triage Decision (SIU to FRA)

The SIU Agent returns a disposition: `OPEN_CASE`, `MONITOR`, `DECLINE`, or `ESCALATE_TO_HUMAN`. The response includes `decision_rationale` mapped to evidence IDs, `next_steps`, human-review flags, and structured feedback for continuous learning.

6. Indicator Taxonomy

Table 1 provides an illustrative taxonomy of neutral indicator codes with their meanings and typical evidence sources.

Indicator	Meaning and Evidence
`TIMELINE`	Reported timeline conflicts with documented events. Evidence: First report of injury, RTW forms, adjuster notes.
`TREATMENT`	Treatment cadence deviates from peer norms. Evidence: Bills, UR notes, provider records.
`MECHANISM`	Mechanism narrative varies across sources. Evidence: Recorded statements, ER notes, incident report.
`ACTIVITY`	Restrictions conflict with work/activity records. Evidence: Employer logs, wage statements, RTW notes.
`DOCUMENT`	Metadata suggests alteration or mismatch. Evidence: Form versions, timestamps, submission logs.

Table 1: Illustrative indicator taxonomy with neutral phrasing and typical evidence artifacts.

7. Workflow Diagrams

7.1 Sequence Diagram

Figure 1 illustrates the protocol's primary interaction: structured referral, optional evidence request, conditional human escalation, and disposition feedback.

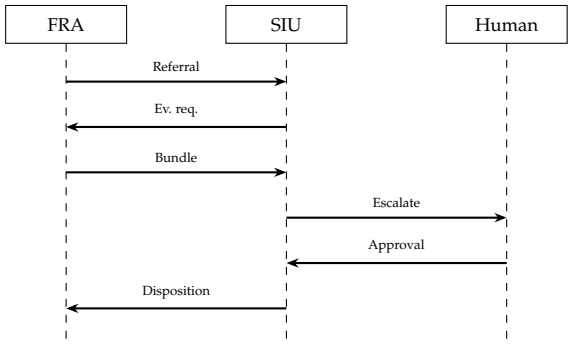

Figure 1: Inter-agent sequence for SIU referral, evidence request, escalation, and disposition.

7.2 Finite-State Workflow

Figure 2 summarizes the finite-state lifecycle. `REQ_EV` is optional; `ACK` may transition directly to `TRIAGE`.

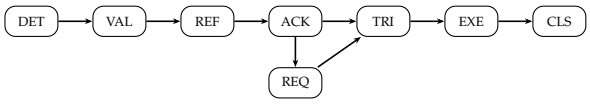

Figure 2: Finite-state lifecycle. DET=Detect, VAL=Validate, REF=Refer, ACK=Acknowledge, REQ=Request Evidence, TRI=Triage, EXE=Execute, CLS=Close.

8. Auditability and Decision Replay

Auditability requires more than log persistence. The protocol mandates that every message be traceable to a fixed evidence set, a model and ruleset version, and a data snapshot sufficient to reproduce the agent's output at the time of decision.

Operationally, the system must store payload hashes, evidence pointers, and feature snapshots so a reviewer can reconstruct what the agent "saw." Without this, the organization cannot reliably answer examination questions: what evidence supported referral, what counterevidence existed, and who approved escalation.

Replay is particularly important in workers' compensation because disputes arise long after initial actions. Versioning is therefore a first-class requirement: `model_version` and `ruleset_version` are part of the message envelope, and data snapshots are referenced by immutable identifiers.

9. Safety and Legal Considerations

The protocol treats defamation risk as a systems design constraint. Automated summaries are restricted to neutral, review-oriented language. Prohibited terms include "fraudster," "faking," "lying," "scam," "guilty," or any phrasing asserting intent.

To minimize discovery risk, the protocol enforces data minimization within evidence bundles. Evidence items cite artifact pointers and include minimal excerpts sufficient for triage, rather than duplicating extensive PHI across systems.

Jurisdictional differences are addressed by parameterizing thresholds and permissible actions. The SIU Agent's decision logic references an external jurisdiction policy table so policy changes do not require model retraining.

10. Continuous Learning Without PHI Leakage

The protocol supports continuous learning through structured outcomes rather than raw investigative narratives. The SIU Agent returns feedback signals: disposition label, confirmed indicators, negated indicators, missing evidence classes, and whether the referral improved investigative efficiency.

This enables performance improvement while minimizing PHI propagation. Learning signals can be stored and analyzed at the indicator level without transporting full medical records or investigative notes.

11. Non-Goals and Limitations

The protocol does not automate fraud determinations, replace investigators, produce "probability of guilt" scores, or authorize surveillance or adverse action. These constraints are deliberate and reduce legal exposure.

Effectiveness depends on organizational adoption. If staff bypass the protocol via informal channels, auditability degrades. Implementation assumes policy enforcement and training alongside technical controls.

12. Conclusion

As AI systems enter workers' compensation SIU workflows, the dominant risk is not model error but incoherent decision narratives that fail under scrutiny. Litigation does not test intelligence; it tests coherence. Formal inter-agent communication enables defensible, auditable AI-assisted investigations aligned with regulated claims operations.

`corpXiv:2601.00009v1 [ai-systems] 10 Jan 2026`

A. Message Templates

A.1 Referral Packet (FRA to SIU)

```
{
  message_id: UUID,
  correlation_id: UUID,
  claim_id: <internal>,
  jurisdiction: <STATE>,
  sender_agent: FRA,
  receiver_agent: SIU,
  timestamp_utc: <ISO-8601>,
  purpose: REFERRAL_SUBMIT,
  privacy_scope: SIU_CONFIDENTIAL,
  legal_hold_flag: true|false,
  confidence_band: LOW|MED|HIGH,
  model_version: <v>,
  ruleset_version: <v>,
  referral_reason_summary: "...",
  indicator_taxonomy: [...],
  evidence_bundle: [
    {evidence_id: "...",
     source_system: "...",
     artifact_pointer: "...",
     excerpt_or_feature: "...",
     why_it_matters: "...",
     reliability_grade: "A"}
  ],
  risk_assessment: {
    risk_band: "MED",
    key_drivers: [...],
    countervailing_factors: [...]
  },
  recommended_actions: [...],
  handoff_questions: [...]
}
```

A.2 Triage Decision (SIU to FRA)

```
{
  message_id: UUID,
  correlation_id: UUID,
  claim_id: <internal>,
  jurisdiction: <STATE>,
  sender_agent: SIU,
  receiver_agent: FRA,
  timestamp_utc: <ISO-8601>,
  purpose: TRIAGE_DECISION,
  privacy_scope: SIU_CONFIDENTIAL,
  model_version: <v>,
  ruleset_version: <v>,
  decision: OPEN_CASE | MONITOR |
            DECLINE | ESCALATE,
  decision_rationale: [
    {evidence_id: "...",
     rationale: "..."}
  ],
  next_steps: [...],
  human_review_required: true|false,
  feedback_to_model: {
    confirmed_indicators: [...],
    negated_indicators: [...],
    missing_evidence: [...]
  }
}
```