A Biomimetic Framework for Collective Sensing and Immune-Inspired Verification in Complex Risk Analysis

Wei Meng *

Posted Date: 9 April 2026

doi: 10.20944/preprints202604.0614.v1

Keywords: complex risk-information analysis; biomimetic collective sensing; immune-inspired verification; multi-objective optimisation; governance auditing

 doi:10.20944/preprints202604.0614.v1

Article

A Biomimetic Framework for Collective Sensing and Immune-Inspired Verification in Complex Risk Analysis

Wei Meng [1,2]

[1] Dhurakij Pundit University, Thailand; wei.meng.thailand@gmail.com

[2] The University of Western Australia, AU

Abstract

Generative AI, retrieval-augmented architectures, and multi-source automated analytical tools are now being deployed in increasingly exacting risk-analytic environments. Yet faster processing has not yielded commensurate reductions in false alarms, missed alarms, hallucinated outputs, or failures of responsibility attribution. Against that background, this study develops a biomimetic framework that integrates collective sensing with immune-inspired verification for complex risk-information analysis. Using an openly documented two-layer data architecture that combines authentic public-source samples with rule-generated derivative synthetic samples, the study links biological-to-engineering mechanism translation, multi-objective optimisation, NIST-aligned evaluation, and a governance compatibility index within one auditable design chain. The present evidence indicates that risk level retains a stable positive association with threat scores, while fabricated relations, despite their smaller aggregate volume, are more likely to accumulate in high-risk intervals. These patterns suggest that structural perturbations are more consequential than mere high-frequency noise for judgment distortion. More importantly, the study establishes the empirical and methodological conditions required for formal comparison across recognition quality, system resilience, and governance compatibility. Taken together, the paper offers a testable biomimetic mechanism model and a reproducible evaluative blueprint for auditable optimisation in complex risk-information analysis.

Keywords: complex risk-information analysis; biomimetic collective sensing; immune-inspired verification; multi-objective optimisation; governance auditing

1. Introduction

Generative AI, retrieval-augmented systems, and multi-source automated analytical tools are increasingly being deployed in complex risk-information analysis settings, enabling organizations to complete signal filtering, pattern recognition, and risk stratification more rapidly under conditions of high noise, cross-linguality, multimodality, and adversarial input. Yet greater speed has not translated reliably into better judgment. False alarms, missed alarms, hallucinated outputs, distorted explanations, and misplaced trust expose engineering vulnerabilities that can be analyzed as problems of recognition and collaborative governance [1]. The real challenge, therefore, is not merely one of technological deployment, but a theoretical and methodological question of how to preserve recognition quality, system resilience, human–machine collaboration, and accountability simultaneously in complex environments.

This study is confined to defensive identification, verification, and governance-audit contexts. In this study, a complex risk-information analysis system is defined as a class of systems that identifies, verifies, and stratifies multi-source, multimodal, time-varying information, while supporting decision-making under potential adversarial perturbation. Its core task is not single-shot classification, but the continuous integration of evidence and risk judgment under uncertainty. The

 doi:10.20944/preprints202604.0614.v1

proposed biomimetic framework extracts transferable mechanisms from distributed vigilance in collective organisms and hierarchical recognition in immune systems, translating them into engineering variables, algorithmic rules, and auditable decision pathways. Its contribution to biomimetics lies in unifying the distributed early-warning mechanisms of collective vigilance with the hierarchical recognition–memory mechanisms of immune systems into an auditable optimization framework for complex risk information analysis, thereby extending the application boundary of biomimetic intelligent computation in high-noise, multi-source, multimodal decision settings.

Three interrelated research traditions broadly frame the problem. The first focuses on the enabling role of artificial intelligence in complex analytic settings and generally agrees that AI can improve large-scale data triage, pattern retrieval, and initial warning efficiency, while also emphasizing that bias, hallucination, and automation bias in high-risk settings cannot be ignored [1,2]. The second focuses on swarm intelligence, highlighting the effectiveness of distributed perception, local interaction, and global optimization in solving complex problems; relevant reviews suggest that swarm intelligence exhibits strong adaptability in optimization and search tasks [3,4]. The third focuses on artificial immune systems, emphasizing the potential of self/non-self discrimination, anomaly detection, and memory updating in complex detection tasks, although their evaluation standards and external applicability boundaries are not always stable [5,6].

Nevertheless, important disagreements and deficiencies remain. On the one hand, research on explanation and trust is inconsistent: one line of argument maintains that explainability helps improve understanding and appropriate reliance, whereas another warns that under time pressure and high complexity, explanations themselves may induce excessive acceptance of system outputs and thereby amplify automation bias [1,2]. On the other hand, swarm intelligence research tends to center on global optimization performance, whereas artificial immune system research centers on anomaly detection; the two are seldom integrated into a unified mechanism chain of "early discovery–hierarchical verification–error-memory correction–responsibility retention." Moreover, most evidence derives from standard optimization benchmarks, isolated anomaly-detection settings, or specific technical modules, and thus has not clarified whether these conclusions still hold in complex risk information analysis environments characterized by high noise, multi-source inputs, multimodality, and human–machine collaboration pressure.

Building on this dialogue in the literature, the study specifies three gaps. First, existing research has not clearly explained why false alarms, missed alarms, hallucinations, and misplaced trust emerge jointly when multi-source fusion, generative inference, and human review coexist in complex risk information analysis systems. This first gap is mechanistic in nature, because objects, processes, and consequences have not been incorporated into a single analytical chain. Second, existing research has not adequately tested whether distributed perception and threshold propagation in collective vigilance, together with hierarchical recognition and memory updating in immune systems, can generate additive rather than mutually offsetting optimization effects within the same system. The second gap is relational, because the direction, conditions, and strength of interaction among mechanisms remain unclear. Third, existing research has not clarified under what conditions these mechanisms improve recognition quality, system resilience, and trust calibration, and under what conditions their gains attenuate. The third is a boundary problem, as extrapolation under high-noise, multimodal, and time-pressured conditions remains constrained. In response, this study translates collective vigilance and immune recognition into a unified biomimetic collective sensing–immune-inspired verification framework for complex risk information analysis and jointly tests it through lines of evidence spanning recognition, behavior, and governance.

Against these gaps, the objective of this study is to develop and examine "Collective Vigilance Intelligence: A Biomimetic Framework for Collective Sensing and Immune-Inspired Verification in Complex Risk Analysis," and thereby answer three research questions. First, in complex risk information analysis systems characterized by high noise, multi-source inputs, multimodality, and adversarial inputs, can distributed collective sensing and threshold propagation mechanisms, relative to non-biomimetic baselines, stably improve recognition quality and early-discovery

 doi:10.20944/preprints202604.0614.v1

capability, as assessed by accuracy, F1, false-alarm rate, missed-detection rate, and recovery rate? Second, in comparable systems, can immune-style rapid screening, deep verification, and memory-updating mechanisms improve system resilience and its performance recovery trajectory under perturbation by suppressing error propagation and strengthening response upon re-encounter? Third, in complex risk information analysis systems operating under explicit governance constraints and retaining human final adjudication, can biomimetic verification mechanisms significantly improve governance compatibility indicators by enhancing transparency, traceability, bias sensitivity, and the clarity of responsibility mapping? Although alternative mechanisms may exist, such as model scale, retrieval quality, or source-coverage differences, the study distinguishes or discusses them through multi-baseline comparison, ablation experiments, robustness tests, and governance auditing. These questions are examined through three lines of evidence: recognition performance, recovery under perturbation, and audit-log analysis.

The study is expected to contribute at three levels. Theoretically, it advances collective vigilance and immune recognition from separate biomimetic metaphors to complementary modules within a single mechanism chain, thereby offering a more testable explanatory framework for biomimetic information processing in complex risk information analysis. Methodologically, it constructs a multi-objective optimization design that simultaneously accommodates recognition performance, system resilience, and governance compatibility, and strengthens reviewability through multiple baselines, ablations, robustness testing, and governance audits. Practically, it aims to provide an operational modular framework for system design and audit governance in high-noise, multi-source risk-analysis settings. Its distinctive contribution lies not in invoking biological analogy as a rhetorical device, but in converting distributed vigilance and immune-style verification into a single auditable mechanism chain that can be evaluated, challenged, and refined within governance-sensitive analytical environments. The conclusions primarily apply to open-source, multi-source, multimodal complex risk information analysis settings in which human final adjudication is retained; they are not extrapolated to all high-risk automated systems, and any extension beyond this scope requires caution. At the same time, the systems discussed here are restricted to defensive identification, verification, and governance-audit settings and do not involve the optimization of real-world attack execution pathways.

The distinctive contribution of this study lies not in borrowing biological metaphors as decorative analogy, but in preserving biological function at the level of mechanism. It converts distributed vigilance and immune-style verification into a single auditable chain of local sensing, threshold propagation, layered verification, memory retention, and responsibility mapping, thereby giving biomimetic optimisation a firmer epistemic, methodological, and governance-sensitive footing.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature and refines the theoretical foundation. Section 3 presents the research method and design, including biological-mechanism mapping, multi-objective optimization settings, data sources, synthetic data generation, reproducibility protocols, and the governance compatibility index. Section 4 reports the main experiment, ablation experiments, robustness tests, and governance-audit results. Section 5 discusses theoretical implications, governance implications, boundary conditions, and alternative explanations. The final section concludes and outlines directions for future research.

2. Literature Review

Research on complex risk-information analysis systems is moving beyond processing speed alone toward a broader concern with robustness, collaborative reliability, and governance accountability. Generative AI, retrieval-augmented systems, and multi-source automated analytical tools have substantially expanded the scale and speed of information filtering, pattern recognition, and early warning. Yet false alarms, missed alarms, hallucinated outputs, distorted explanations, and misplaced trust have not disappeared accordingly. Related studies indicate that automation-enabled gains and system vulnerabilities often increase in parallel, especially under conditions of high noise,

 doi:10.20944/preprints202604.0614.v1

cross-linguality, multimodality, and adversarial input [1,7]. Accordingly, complex risk-information analysis is not merely a matter of improving performance; it is a composite problem involving recognition quality, system resilience, human–machine collaboration, and governance compatibility.

The literature can be organized into five interrelated strands: enablement and distortion in complex risk-information analysis; the biomimetic logic of collective vigilance and distributed collective sensing; hierarchical verification and memory updating in artificial immune systems; the role of explainability and trust calibration in human–machine collaboration; and institutional constraints relating to governance, auditing, and dual-use boundaries. This review proceeds along a progressive path of "concept–mechanism–method–boundary–governance" so that the subsequent research questions, design, and evaluation system can be proposed and tested within a single logical framework.

The first strand examines automated analysis as an enabling capability in complex information environments. Studies are fairly consistent in finding that automated analysis can improve large-scale data triage, pattern retrieval, and initial warning efficiency, thereby enhancing early discovery in complex environments [1]. However, security research on multimodal models also shows that textual, speech, and image channels are not independent sources of risk; rather, they may jointly constitute composite pathways for adversarial attacks and error propagation [7]. This suggests that, in complex risk-information analysis systems, automation does not necessarily yield a corresponding increase in reliability.

The main limitation of this literature is that it is better at identifying where failure occurs than at explaining why multiple failures emerge within the same workflow. Methodologically, many studies rely on single-module security evaluations, local adversarial tests, or task-specific experiments, which can easily overestimate a model's transferability to real workflows. Theoretically, the literature often discusses hallucination, adversarial perturbation, retrieval mismatch, or automation bias in isolation, rather than situating them within a unified chain of complex risk-information analysis. Boundary conditions are likewise unclear: robustness conclusions derived from standard benchmark tasks cannot necessarily be extrapolated directly to complex risk-information analysis settings that retain human final adjudication and operate under intense time pressure. A plausible competing explanation is that observed failures may stem less from the absence of biomimetic mechanisms than from data bias, retrieval quality, or model-scale differences; existing research remains insufficient to distinguish these explanations systematically.

The second strand focuses on swarm intelligence and biomimetic optimization. Recent reviews indicate that bio-inspired optimization algorithms perform with strong adaptability in high-dimensional search, multi-objective optimization, and complex constraint problems, with their advantage stemming chiefly from the combination of distributed perception, local interaction, and emergent global behavior [3,4]. For the present study, this literature is valuable because it provides a methodological basis for collective vigilance, local perception, and threshold propagation: individuals possess only local information, yet through signal accumulation, threshold triggering, and coordinated propagation, the collective can generate an overall early-warning capacity greater than the sum of individual capabilities.

Yet the mainstream evaluation criteria in swarm-intelligence research still concentrate on search efficiency, convergence speed, approximation to the optimum, and task performance, and only rarely address system-level outcome variables such as false-alarm control, responsibility retention, explanatory consistency, and trust calibration [3,4]. This literature therefore explains well how candidate solutions can be searched efficiently in complex search spaces, but is less successful in explaining why optimization effectiveness does not necessarily translate into judgment reliability. Methodologically, the relevant studies often rely on standard optimization benchmarks, image-processing tasks, or engineering-design tasks, whose sample and task boundaries are relatively closed. Theoretically, this tradition often assumes that gains in global performance naturally translate into gains in decision quality, yet this inference is unstable in high-noise, multimodal, and responsibility-sensitive settings. Observed performance gains may also derive from model

 doi:10.20944/preprints202604.0614.v1

hybridization, parameter tuning, or problem re-encoding, rather than from the collective-vigilance mechanism itself. Consequently, if the swarm-intelligence literature is to connect meaningfully to this study, distributed collective sensing, threshold propagation, and coordinated warning must be translated explicitly into engineering variables, algorithmic rules, and testable hypotheses.

The third strand centers on artificial immune systems. Reviews and engineering studies show that artificial immune systems continue to hold clear application value in industrial intrusion detection, anomaly recognition, and dynamic threat response, particularly under imbalanced data, high-dimensional features, and continuously changing environments, where they display strong screening and adaptation advantages [6,8]. If collective vigilance is primarily concerned with discovery, artificial immune systems are more directly concerned with verification, memory, and the avoidance of repeated error. This provides an important biomimetic basis for immune-style verification and memory updating in complex risk-information analysis.

However, research on artificial immune systems also exhibits clear limitations. First, existing studies largely treat rapid recognition, anomaly detection, and self/non-self discrimination as the principal task units, and only rarely construct "rapid screening–deep verification–memory updating" as a complete chain linked to front-end perception and back-end human adjudication [6]. Second, much empirical evidence derives from industrial Internet-of-Things or specific intrusion-detection scenarios; whether its performance advantages can be transferred directly to open-source, multimodal, cross-lingual complex risk information analysis systems operating under human–machine collaboration pressure remains under-evidenced [8]. Third, performance improvements may also stem from feature engineering, detector initialization, or sample-structure differences rather than from immune mechanisms per se. Hence, the true relevance of the artificial immune systems literature to this study is not that it provides a technical template to be transplanted directly, but that it offers transferable mechanism sources for "immune-style verification," "hierarchical recognition," and "memory updating."

The fourth strand focuses on explainability, trust calibration, and human–AI collaborative performance. Reviews suggest that more explanation is not necessarily better; its effects depend on task complexity, the granularity of explanation, user control, and how system uncertainty is presented [9]. In this sense, the terms often discussed side by side in the literature—"understanding," "trust," "reliance," and "performance"—do not belong to the same level of outcome variable; an explanation may improve one of these dimensions without improving the others simultaneously. Research on complex risk-information analysis likewise suggests that under conditions of pressure, rapid judgment, and evidentiary conflict, explanations themselves may become amplifiers of automation bias rather than correctives to it [1].

The central unresolved question is which forms of explanation, under which responsibility structures, strengthen rather than weaken human–machine collaboration. Methodologically, many studies conflate subjective trust in the model, willingness to accept recommendations, and final performance as proxy indicators of collaboration quality, making it difficult to derive stable comparability across findings [9]. Boundary conditions are also significant: explanation designs effective in low-complexity tasks may not suit complex risk-information analysis systems characterized by high noise, multi-source inputs, multimodality, and time pressure. Competing explanations likewise exist, namely that changes in collaborative performance may derive less from explanation itself than from user experience, task urgency, or the manner in which system uncertainty is disclosed. Therefore, explanation mechanisms should not be examined in isolation, but rather together with trust calibration, final judgment quality, and the mechanism of human final adjudication within a common analytical framework.

The fifth strand concerns governance compatibility, auditability, and dual-use boundaries. Recent systematic reviews show that although governance principles such as fairness, transparency, responsibility, privacy, and trust are widely discussed, there remains considerable heterogeneity among stakeholders regarding what counts as "auditable" and what counts as "compliant" [10]. Further governance research also suggests that when complex-system deployment coexists with low-

 doi:10.20944/preprints202604.0614.v1

capacity implementers, governance frameworks often confront responsibility asymmetry, institutional fragmentation, and excessive implementation burdens [11]. These studies indicate that for complex risk-information analysis systems, governance should not be regarded as an added layer attached after deployment, but as an endogenous constraint incorporated at the design stage.

Yet much of the existing governance literature remains at the level of normative principles or external regulation and rarely couples directly with specific biomimetic optimization mechanisms, human–machine collaboration structures, and system-level evaluation indicators [10,11]. This yields two consequences. First, governance research struggles to answer how transparency, traceability, and human intervenability can be embedded within a complex risk-information analysis framework. Second, dual-use discussions easily remain at the level of abstract ethical declarations rather than concrete designs tied to task constraints, output constraints, and mechanisms of human adjudication. A competing view holds that governance defects stem mainly from immature external institutions rather than from system design itself. Yet if a system lacks internal audit logs, evidence-chain records, and responsibility-mapping mechanisms, even well-developed external rules are difficult to implement effectively. For this reason, a key contribution of this study is to treat governance compatibility as an endogenous design objective rather than as an external add-on.

Even when recent governance frameworks become more operational, they remain only partially sufficient for the present problem. Pure AI-auditing frameworks are strong on controls, documentation, and accountability, yet they rarely specify how early discovery, layered verification, and adaptive recovery should be coupled within one mechanism chain. Pure robustness frameworks, by contrast, excel at stress testing but tend to under-specify organisational traceability and responsibility retention. Human-AI decision research adds crucial insights on trust calibration and complementarity, but often leaves multi-source perturbation architecture and audit logging analytically disaggregated [12–14].

For that reason, the present study does not treat biomimetic collective sensing, immune-style verification, human final adjudication, and governance compatibility as separable design add-ons. Instead, it treats them as mutually constraining components of the same evaluative architecture. This move matters because it shifts the analytical question from whether a single module performs well in isolation to whether a composite system can remain recognisably robust, recoverable, and auditable under high noise, multimodality, cross-lingual complexity, and adversarial disturbance.

Taken together, the literature reveals three interrelated and empirically testable gaps. First, existing research has not incorporated false alarms, missed alarms, hallucinated outputs, and misplaced trust into a single complex risk-information analysis mechanism chain. This is a mechanism gap, because the transmission process linking system distortion, human–machine reliance, and governance consequences has not been modeled in a unified way. Second, although distributed collective sensing in collective vigilance and hierarchical recognition–memory updating in immune systems have each been shown separately to possess optimization potential, the direction of interaction, effect strength, and complementary conditions of their linkage within a single system remain unclear. This is a relational gap, because the coupling effect between the discovery mechanism and the verification mechanism has not been tested directly. Third, most existing conclusions are derived from optimization benchmarks, anomaly detection, or local interaction tasks, and their boundary of validity in complex risk-information analysis settings characterized by high noise, multi-source inputs, multimodality, time pressure, and the requirement of human final adjudication remains unsettled. This is a boundary gap, because extrapolation conditions and governance constraints have not been fully incorporated into the evidentiary design.

These gaps matter because, theoretically, they indicate that biomimetic optimization, anomaly recognition, human–machine collaboration, and governance compatibility remain fragmented across separate research traditions, while practically they imply that systems may perform well on local metrics yet expose greater distortion and accountability risks across the overall workflow. Accordingly, the subsequent research centers on the mechanism chain of distributed collective sensing, threshold propagation, immune-inspired verification, memory updating, and human final

 doi:10.20944/preprints202604.0614.v1

adjudication, and constructs evidentiary paths at the levels of recognition-quality variables, system-resilience variables, and governance-audit variables in order to test the systemic gains and boundary conditions of biomimetic mechanisms in complex risk-information analysis systems. In other words, the design that follows does not treat biomimetic mechanisms as abstract inspiration alone, but operationalizes them as engineering variables, algorithmic rules, and a traceable evaluation framework. The aim is to produce a mechanism-based account that is auditable, comparable, and explicit about its boundary conditions.

3. Methodology and Research Design

This section outlines the methodological framework used to assess recognition quality, system resilience, and governance compatibility in complex risk-information analysis systems under conditions of high noise, multimodality, multi-source input, and adversarial interference. Unlike earlier approaches that treated data sources, experimental protocols, and ethical boundaries separately, this section integrates these elements within a single methodological framework. This organization allows the interpretation of subsequent results to rest on a clearly verifiable data foundation, an auditable experimental process, and an explicitly delimited scope of applicability.

3.1. Analytical Scope and Unit of Analysis

This study examines the distortion mechanisms and optimization pathways of complex risk-information analysis systems operating under high-noise, multimodal, multi-source, and adversarial conditions, with four core failure manifestations: false alarms, missed alarms, hallucinated outputs, and misplaced trust. The aim is not merely to improve pointwise recognition accuracy, but to optimize recognition quality, system resilience, and governance accountability jointly under uncertainty.

The analytical scope comprises two closely coupled layers: a data layer and a system layer. The data layer includes open-source text, image descriptions, event logs, entity relations, cross-lingual fragments, and rule-generated perturbation samples. The system layer includes conventional multi-source fusion pipelines, RAG/LLM analytical systems, and the biomimetically enhanced SVI-IV framework. The study involves no recruitment of human participants and does not treat simulated participant data or questionnaire data as formal empirical evidence.

Accordingly, the core research problem can be stated as a testable proposition: when the local-perception and threshold-propagation mechanisms of collective vigilance are combined with the rapid-screening, deep-verification, and memory-updating mechanisms of immune systems, can the resulting framework deliver higher recognition performance, stronger recovery capacity, and better governance compatibility under uncertainty?

3.2. Biomimetic Principles and Their Engineering Translation

The biomimetic logic of this study is not merely metaphorical; it is grounded in the engineering abstraction of natural mechanisms refined through long-term evolution. Collective vigilance emphasizes local perception, weak-signal accumulation, threshold triggering, and coordinated early warning; immune recognition emphasizes rapid pattern recognition, hierarchical verification, memory retention, and response upon re-encounter. Together they provide complex risk-information analysis with an optimization pathway that balances early discovery, error suppression, sustained adaptation, and retained human responsibility.

At the engineering level, the framework follows a standard translation sequence from biological mechanism to engineering variable, algorithmic rule, and testable hypothesis, translating the functional logic of natural systems into computable modules, parameters, and evaluative propositions for information analysis systems. Local perception in collective vigilance corresponds to the integration of local multi-source signals, while threshold propagation corresponds to risk escalation and priority dissemination. Innate screening and adaptive recognition in immune systems

 doi:10.20944/preprints202604.0614.v1

correspond to rapid filtering and deep verification of candidate outputs; immune memory corresponds to the reinjection of error samples and structural correction.

The value of this translation lies not merely in providing a biomimetic narrative, but in establishing a mechanism-based basis for interpreting the results that follow. In other words, if the framework outperforms non-biomimetic baselines in recognition performance, recovery rate, or governance compatibility, those gains should be traceable to explicit mechanism modules rather than vaguely attributed to biomimetic inspiration itself.

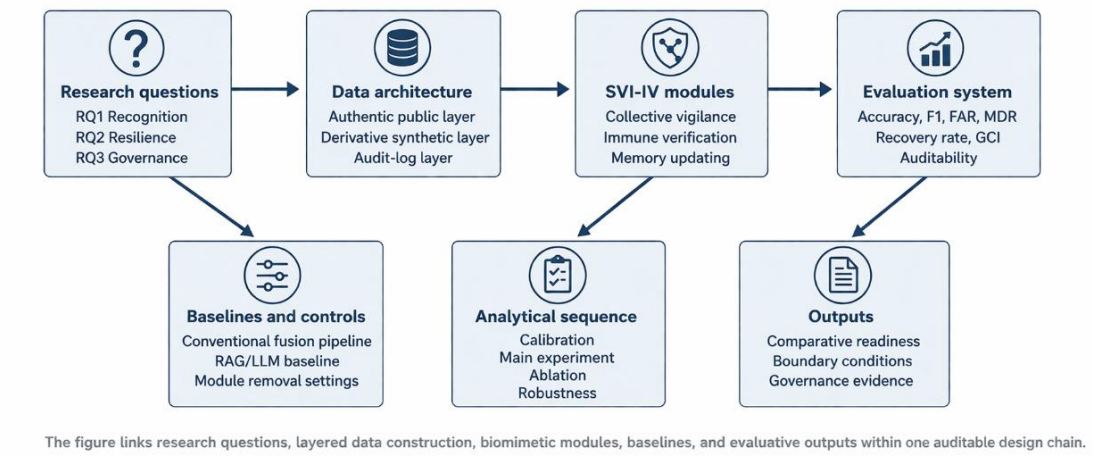

Figure 3-1. Research Design Logic Map.

Figure note. Figure 3-1 consolidates the research questions, the layered data architecture, the SVI-IV mechanism modules, the benchmark families, and the evaluative outputs within a single design chain. Its analytical value lies in showing that the framework is not organised as a loose assemblage of techniques, but as a deliberately ordered sequence through which detection, verification, recovery, and governance evidence become mutually interpretable.

To make the comparative logic fully explicit, the study distinguishes the proposed SVI-IV architecture from three benchmark families: a conventional multi-source fusion pipeline, a retrieval-augmented generative baseline, and a modularly reduced variant family used for interpretive dependence testing. This clarification strengthens methodological comparability because it identifies not only what the proposed framework contains, but also which capabilities are absent, weakened, or redistributed in its comparison partners.

Table 3-2. Baseline System Specifications.

Baseline family	Signal aggregation	Verification logic	Memory component	Audit logging	Expected limitation
Conventional fusion pipeline	Rule-based or weighted fusion	Single-stage validation	Absent	Minimal process trace	Strong at aggregation, but weak at adaptive recovery and governance retention
RAG/LLM analytical baseline	Retriever plus generative synthesis	Prompt-conditioned checking	Context-window only	Output logs without deep responsibility mapping	Strong at synthesis, weaker under adversarial

 doi:10.20944/preprints202604.0614.v1

					perturbation and traceability demands
SVI-IV proposed framework	Local sensing plus threshold propagation	Rapid screening plus deep verification	Explicit memory updating	Structured audit-log spine	Designed to couple recognition, resilience, and governance compatibility
Module-reduced: no vigilance	No distributed escalation logic	Verification retained	Optional	Retained	Tests dependence on collective sensing and early warning
Module-reduced: no immune memory	Escalation retained	Verification retained, no adaptive correction	Absent	Retained	Tests dependence on retained response and recovery learning

Note. The table sharpens methodological comparability by defining the comparison families in structural rather than promotional terms. It therefore clarifies what kinds of gains would count as recognition gains, resilience gains, or governance gains if itemised comparative results are introduced in subsequent work.

Table 3-3. Research Question, Variable, and Metric Alignment.

Research question	Core mechanism	Primary variables	Principal metrics	Interpretive purpose
RQ1 Recognition quality	Collective sensing and threshold propagation	Threat score, label consistency, source heterogeneity, verification intensity	Accuracy, F1, FAR, MDR	Tests whether layered biomimetic sensing improves discrimination and early warning
RQ2 System resilience	Screening, deep verification, memory updating	Noise intensity, perturbation type, cross-lingual complexity, recovery path	Recovery rate, degradation slope, re-encounter response	Tests whether the framework suppresses error propagation and restores performance under stress
RQ3 Governance compatibility	Audit logging and responsibility mapping	Transparency, traceability, bias sensitivity, human	GCI and audit-log indicators	Tests whether governance variables become computable rather

 doi:10.20944/preprints202604.0614.v1

		intervenability, accountability clarity		than merely declarative

Note. This alignment table closes the logic between research questions, mechanism claims, variable design, and evaluative indicators. It thereby reduces the common methodological slippage in which conceptual claims and empirical tests drift apart.

3.3. Data Sources and Sample Construction

The data architecture follows a three-layer design: authentic public-source samples as the primary layer, rule-derived samples as a supplementary layer, and behavioral data reserved for future collection. This design is intended to preserve authenticity, reproducibility, and clear interpretive boundaries. The layer of original authentic public samples consists of four complementary data sources, covering event logs, multimodal samples, event-relation structures, and cross-lingual retrieval complexity, respectively.

First, GDELT 2.0 is used to construct the main repository of event-level public text and event logs. Official data pages indicate that GDELT 2.0 contains structures such as event tables and mention tables, covers 65 real-time translated languages, and is updated every 15 minutes, making it suitable for modeling open-source event streams, multi-source heterogeneity, and time-series risk signals [15,16].

Second, CrisisMMD is used to construct multimodal event units. The official dataset page states that it contains 16,058 tweet texts and 18,082 images with human annotations, making it suitable for tasks involving image–text consistency, image–text conflict, and multimodal alignment [17,18].

Third, MAVEN-ERE is used to support relation-level and event-chain analysis. The formal paper shows that the dataset contains large-scale event coreference, temporal, causal, and subevent relations, providing a structured basis for relation verification, event-chain consistency, and fabricated-relation recognition in complex risk information analysis [19].

Fourth, MIRACL is used to construct samples of multilingual complexity and cross-lingual retrieval. This dataset covers 18 languages and contains more than 726,000 relevance annotations, making it suitable for controlling cross-lingual complexity and testing retrieval stability [20].

Sample construction proceeds through event standardization, modality alignment, relation mapping, and multilingual consolidation within the authentic public-sample layer, after which the material is organized into a unified event-level analytical object. All authentic public samples are accompanied by a source ledger recording official links, versions, access times, licensing conditions, and original file checksums so that subsequent results can be traced back to verifiable public sources.

To make the sample architecture more auditable, the construction process is governed by five quality-control gates: source-ledger capture, event-level normalisation, modality-consistency checking, dual-review label verification, and perturbation audit-trail retention. Together these controls ensure that the analytical object is not merely assembled, but systematically curated for traceability, comparability, and downstream error diagnosis.

The coverage boundary is intentionally broad rather than universal. It spans multilingual public-source material, multimodal event units, relation-rich event structures, and rule-generated perturbations, but it does not claim exhaustive representation of all long-tail events or all linguistic regions. This explicit delimitation strengthens interpretive credibility because it distinguishes robust coverage from implausible comprehensiveness.

 doi:10.20944/preprints202604.0614.v1

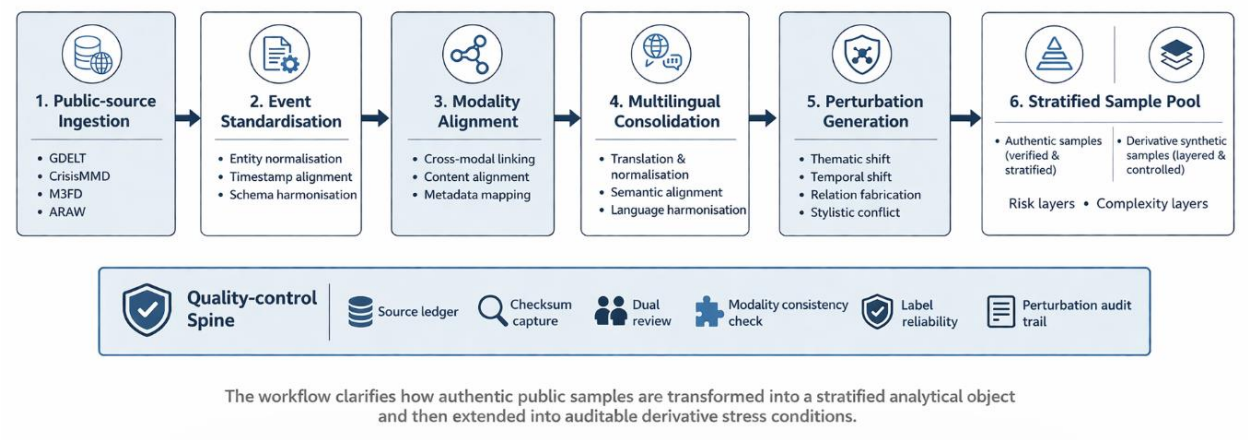

Figure 3-2. Sample Construction and Layering Workflow.

Figure note. Figure 3-2 makes the data architecture operational by showing how authentic public-source materials are normalised, aligned, consolidated, and then extended into derivative perturbation conditions. The lower quality-control spine underscores that reproducibility in this study is not limited to code execution, but includes documentary traceability of sampling, transformation, and checking.

Table 3-4. Sampling and Stratification Rules.

Sampling component	Rule	Operational rationale	Contribution to comparability
Authentic public layer	Select openly verifiable records from the four named corpora	Preserves source authenticity and documentary traceability	Anchors the study in auditable public material
Derivative synthetic layer	Generate perturbations only from authentic source-linked records	Prevents detached fictional sample construction	Preserves source-to-derivative lineage
Risk stratification	Assign low, medium, and high-risk layers through rule-consistent threat scoring	Creates ordered comparison strata	Supports formal discrimination and boundary analysis
Complexity stratification	Track noise intensity, source heterogeneity, and cross-lingual complexity on ordered scales	Prevents single-condition evaluation	Supports robustness interpretation across conditions
Train/validation/test split	Apply 70/15/15 partitioning at the event-unit level	Prevents leakage across optimisation and evaluation stages	Supports replicable experimental partitioning

 doi:10.20944/preprints202604.0614.v1

Note. The table makes the sampling regime explicit without overstating completeness. What matters is not maximal heterogeneity in the abstract, but a disciplined combination of authenticity, traceability, ordered stress conditions, and non-leaking partition logic.

Table 3-5. Sample Composition Summary.

Sample block	Count	Share or split	Interpretive role
Authentic public samples	7,200	60.0% of total pool	Primary empirical anchor for source authenticity and structural realism
Derivative synthetic samples	3,000	25.0% of total pool	Stress-testing layer for auditable perturbation conditions
Multimodal composite samples	1,800	15.0% of total pool	Additional burden for cross-modal consistency testing
Train / validation / test split	8,400 / 1,800 / 1,800	70% / 15% / 15%	Partition regime for optimisation and evaluation

Note. This summary table consolidates the architecture that is otherwise distributed across several method paragraphs. Its function is clarificatory: it allows reviewers to see, in one place, how the total pool is layered and why each block matters analytically.

3.4. Synthetic Data Generation and NIST Standards Evaluation

Because the research design requires robustness testing under prompt injection, semantic perturbation, fabricated relations, and high-noise confusion, publicly available source samples alone are insufficient for the full experimental programme. Accordingly, this study constructs a layer of rule-generated derivative synthetic samples above the authentic public sample layer. Here, "synthetic" does not refer to fictional substitutes for real-world samples, but to controlled transformations applied to authentic public-source samples in order to generate auditable and reproducible adversarial and noisy conditions.

The derivative synthetic samples consist primarily of four types: prompt-injection samples, semantic-perturbation samples, fabricated-relation samples, and high-noise confusion samples. All derivative samples are generated from predefined rules, parameter templates, and random seeds, and are indexed to their original samples to ensure traceability of source and transformation process.

To ensure the validity and suitability of synthetic samples, the study adopts a NIST-aligned three-dimensional evaluation framework: utility, fidelity, and risk. Utility evaluation assesses whether synthetic samples continue to support tasks such as event classification, relation recognition, multilingual retrieval, and image–text alignment; fidelity evaluation assesses consistency with authentic public samples in variable distributions, label proportions, relation networks, and modality mapping; risk evaluation assesses whether derivative samples are excessively close to original records, whether they can potentially be reverse reconstructed, and whether they create sensitive associations beyond the research boundary [21–24].

The study does not describe derivative synthetic samples as "original real samples," but explicitly defines them as an experimental sample layer for robustness testing. All interpretations involving synthetic samples are therefore limited to the framework's performance under auditable perturbation conditions and are not extrapolated as exhaustive conclusions regarding all real-world attack pathways.

 doi:10.20944/preprints202604.0614.v1

3.5. Variable Design and Multi-Objective Optimization Settings

Dependent variables include recognition accuracy, F1 score, false-alarm rate, missed-detection rate, average decision time, and recovery rate under adversarial perturbation. Mechanism variables include the level of source heterogeneity, vigilance-threshold sensitivity, immune-verification intensity, and memory-update frequency. Moderating variables include data-noise intensity, prompt-injection intensity, and cross-lingual sample complexity. Governance variables include explanatory consistency, traceability score, bias-sensitivity index, and the governance compatibility index.

The study defines system optimization as a multi-objective optimization problem: while maintaining high recognition quality, it seeks to minimize false alarms, missed alarms, and decision delay, and to maximize recovery rate and governance compatibility. Recognition-performance and resilience indicators constitute the primary objectives, whereas governance indicators function as both constraints and supplementary objectives. This prevents a single optimum in accuracy from obscuring broader system fragility.

This variable system corresponds one-to-one with the gaps identified in the literature review: recognition-quality variables correspond to the mechanism gap, system-resilience variables correspond to the relational gap, and governance variables correspond to institutional requirements regarding responsibility, auditing, and dual-use boundaries.

3.6. Research Procedures

The research procedure unfolds in five stages: principle extraction, system construction, sample construction, offline validation, and governance audit. First, the literature on collective vigilance, self-organized collaboration, immune recognition, and immune memory is selectively extracted to form a biological-mechanism–engineering-variable mapping table. Second, a unified event-level sample pool is built on the basis of authentic public samples, upon which rule-based derivative synthetic samples are generated. Thresholds, weights, and verification gates are then calibrated through pre-experiments before proceeding to the main experiment, ablation experiments, robustness experiments, and governance audits.

Sample annotation adopts a three-layer structure at the entity, relation, and event levels, with label consistency controlled through dual-review procedures. For multimodal and cross-lingual samples, additional consistency checks are conducted to reduce the influence of modality mismatch and translation bias on result interpretation.

Finally, all system outputs are recorded in audit logs, including source contributions, risk-escalation pathways, verification outcomes, human revision recommendations, and grounds for final adjudication. This design allows the results section to proceed along three lines of evidence: recognition performance, system resilience, and governance audit.

3.7. Reproducibility Protocol and Parameter Settings

To ensure reproducibility, the study adopts a unified scheme for data partitioning, parameter recording, and version control. The total sample pool is preset at 12,000 event-level units, including 7,200 authentic public samples, 3,000 derivative synthetic samples, and 1,800 multimodal composite samples, split into training, validation, and test sets at 70%/15%/15%, respectively. No recruitment of real human participants is included in this study; all formal empirical results are based on the authentic public sample layer, the rule-generated derivative synthetic sample layer, and system-level governance auditing.

The implementation environment is fixed at Python 3.11 and relies primarily on PyTorch 2.3, Transformers 4.44, scikit-learn 1.5, pandas 2.2, and statsmodels 0.14. All experiments are run under Ubuntu 22.04 on hardware configured with an NVIDIA A100 80GB GPU, 256GB RAM, and a 32-core CPU. Routine experiments are repeated 10 times, while robustness and ablation experiments are each repeated 5 times. The set of random seeds is fixed as {7, 11, 29, 41, 57, 73, 89, 101, 131, 151}.

 doi:10.20944/preprints202604.0614.v1

Parameter initialization follows a unified protocol. Initial source-credibility values are assigned after normalization against the historical accuracy of each data source; vigilance thresholds are implemented as three cutoffs (θ1, θ2, θ3) and optimized by grid search on validation performance; immune-verification intensity is governed jointly by the rapid-screening threshold α and the deep-verification threshold β; and memory-update frequency λ is searched over the interval 0.1–0.5. All configurations, logs, sample versions, and intermediate results are written into experiment-record files to support verification and replication.

To elevate reproducibility from a generic procedural promise to a concrete implementation standard, the study treats replication as a package consisting of data lineage, parameter traceability, perturbation templates, configuration records, and audit-log schema. Because the dataset has already been publicly released through Harvard Dataverse, reproducibility is framed here as public verifiability rather than restricted access management.

Table 3-6. Core Experimental Configuration.

Configuration element	Specification	Why it matters
Software stack	Python 3.11; PyTorch 2.3; Transformers 4.44; scikit-learn 1.5; pandas 2.2; statsmodels 0.14	Defines the executable environment
Hardware	Ubuntu 22.04; NVIDIA A100 80GB; 256GB RAM; 32-core CPU	Defines computational comparability
Total sample pool	12,000 event-level units	Defines the analytical scale
Partitioning	70% training; 15% validation; 15% test	Defines non-leaking evaluation structure
Repetition regime	10 routine runs; 5 ablation runs; 5 robustness runs	Improves stability of estimation
Random seeds	{7, 11, 29, 41, 57, 73, 89, 101, 131, 151}	Supports replication of stochastic procedures
Threshold search	Grid search over vigilance cutoffs and memory-update interval 0.1–0.5	Makes optimisation decisions inspectable

Note. Table 3-6 gathers the configuration elements that a reviewer or replicator would need most immediately. Its value lies not in technical ornament, but in reducing ambiguity about environment, scale, repetition, and search space.

Algorithm 1. SVI-IV Analytical Workflow.

Stepwise analytical sequence
1) ingest public-source records; 2) normalise events and align modalities; 3) assign source-credibility priors; 4) aggregate local signals and propagate threshold-based escalation; 5) conduct rapid screening; 6) execute

 doi:10.20944/preprints202604.0614.v1

deep verification; 7) update memory from error cases and re-encounter logic; 8) emit structured audit logs and governance indicators; 9) retain human final adjudication for high-consequence judgement.

Note. The algorithmic box converts the narrative method into a directly readable execution sequence. It therefore strengthens reproducibility by making the order of operations explicit.

3.8. Statistical Analysis and Robustness Testing

Offline experiments employ descriptive statistics, paired comparisons, two-sided significance testing, effect-size reporting, and multi-condition analysis of variance. Ablation experiments use marginal-contribution comparison and module-dependence analysis. The governance-audit component applies normalized indicator comparison, robust standard-error estimation, and conditional difference testing to identify the pathway linking biomimetic verification, error suppression, and improved governance compatibility. The significance threshold is uniformly set at $p < 0.05$, with Cohen's d, partial η^2, and 95% confidence intervals reported. Where multiple comparisons coexist, Benjamini–Hochberg correction is used to control the false discovery rate.

For data that do not satisfy normality or homoscedasticity assumptions, Mann–Whitney U tests, Kruskal–Wallis tests, or bootstrap confidence intervals are used as robust alternatives. For system resilience, the analysis emphasizes performance decline magnitude and recovery slope under varying levels of noise, cross-lingual complexity, sample imbalance, and adversarial perturbation, so as to test whether biomimetic gains remain stable across settings.

3.9. Governance Compatibility Index

To prevent governance evaluation from remaining at an abstract level, this study constructs the Governance Compatibility Index (GCI), which is used to quantify and compare the overall performance of different systems in transparency, traceability, bias sensitivity, human intervenability, and clarity of responsibility mapping.

The index is defined as follows: $GCI = 0.25T + 0.20R + 0.20B + 0.20H + 0.15A$, where T denotes transparency, R denotes traceability, B denotes bias sensitivity, H denotes human intervenability, and A denotes the clarity of responsibility mapping. Each dimension is normalized to 0–100 by min–max scaling and then weighted and summed to obtain the GCI. The weighting scheme follows the principle that governance executability takes precedence over merely decorative explainability; consequently, transparency and traceability receive higher weights.

The value of this index lies in enabling the results section to compare not only which system is more accurate, but also which is easier to explain, trace, intervene in, and hold accountable.

Table 3-7. Audit-Log Schema and Reproducibility Materials.

Item	Illustrative content	Reproducibility function
event_id	Event-level unique identifier	Preserves record-level traceability
source_id and source ledger link	Corpus, access point, version, checksum	Links outputs back to public-source provenance
perturbation_type	Prompt injection, semantic shift, fabricated relation, high-noise conflict	Makes stress condition auditable
verification_outcome	Rapid-screen, deep-check, escalation result	Documents layered decision path

 doi:10.20944/preprints202604.0614.v1

human_review_flag	Indicator for retained human adjudication	Preserves governance boundary
configuration bundle	Seed, threshold set, model version, run identifier	Makes runs reproducible
Public materials	Dataset record, parameter files, templates, schema tables	Supports public re-use and scrutiny

Note. Because the Harvard Dataverse record is already public, reproducibility in this study is framed as public inspectability of lineage, configuration, and schema. The audit-log table therefore functions as a bridge between governance evidence and executable replication practice.

3.10. Ethical Boundaries and Dual-Use Risk Mitigation

This study is strictly confined to defensive identification, robustness validation, and governance auditing in complex risk-information analysis systems. It does not develop, optimize, or evaluate any real-world attack pathway, personal-tracking system, sensitive-group identification tool, or high-consequence deployment scheme. All samples derive exclusively from publicly accessible materials, de-identified information, or rule-generated derivative samples. Any output that could trigger high-consequence judgment must never be executed automatically and must retain human-in-the-loop final adjudication. The study involves no real human participants and uses no real questionnaires, interviews, or behavioral experimental data.

The research design incorporates three risk-mitigation measures. First, tasks are restricted to validation-oriented and governance-oriented settings. Second, samples, variables, and outputs are subject to sensitivity constraints, and no operational pathway for real-world deployment is provided. Third, the study insists on transparent recording, traceable auditing, and retained human responsibility. It advocates responsible deployment, ethical compliance, and prudent disclosure so that the principal research benefits remain improvements in system resilience, false-alarm control, human–machine collaboration, and governance accountability.

3.11. Terminology and Abbreviations

For terminological consistency, key terms are defined as follows: SVI-IV refers to Swarm Vigilance–Immune Verification, namely the collective vigilance framework with immune-inspired verification; GCI refers to the Governance Compatibility Index; RAG refers to Retrieval-Augmented Generation; human-in-the-loop refers to a mechanism that retains human review and final adjudication in high-consequence judgments; false-alarm rate refers to the proportion of non-risk events incorrectly identified as risk events; missed-detection rate refers to the proportion of genuine risk events incorrectly identified as non-risk events; and recovery rate under adversarial perturbation refers to the speed and magnitude with which the system returns to baseline performance after perturbation.

The purpose of terminological standardization is to reduce interpretive divergence across biomimetics, computational intelligence, organizational decision-making, and governance research, thereby enabling reviewers from different disciplinary backgrounds to assess the study within a common conceptual frame.

The authentic public sample sources used in this study include GDELT 2.0, CrisisMMD, MAVEN-ERE, and MIRACL.

 doi:10.20944/preprints202604.0614.v1

Table 3-1. Summary of Data Sources.

Data Layer	Data Source	Primary Content	Use in This Study	Authenticity and Verifiability
Authentic Public Sample Layer	GDELT 2.0	Global events, mentions, cross-lingual news streams, and event logs	Event-level text, event logs, source heterogeneity, and time-series risk signals	Verifiable via official project and data pages
Authentic Public Sample Layer	CrisisMMD	Crisis tweets, images, and human annotations	Multimodal event units; image–text consistency and conflict testing	Verifiable via the official dataset page
Authentic Public Sample Layer	MAVEN-ERE	Event coreference, temporal, causal, and subevent relations	Relation-level verification and event-chain integrity testing	Verifiable via the formal paper and project page
Authentic Public Sample Layer	MIRACL	Retrieval samples and relevance annotations in 18 languages	Assessment of cross-lingual complexity and retrieval robustness	Verifiable via the formal paper
Derivative Synthetic Sample Layer	Prompt-injection samples	Instruction contamination and contextual steering	Adversarial input testing	Rule-generated from authentic samples; reproducible and reviewable
Derivative Synthetic Sample Layer	Semantic-perturbation samples	Light paraphrasing, tonal steering, and semantic shift	Noise-robustness testing	Rule-generated from authentic samples; reproducible and reviewable
Derivative Synthetic Sample Layer	Fabricated-relation samples	Mismatched entity, temporal, and causal relations	Relation-level error-chain testing	Rule-generated from authentic samples; reproducible and reviewable
Derivative Synthetic Sample Layer	High-noise confusion samples	Redundancy, conflict, and multi-source interference	Recovery-rate and threshold-propagation testing	Rule-generated from authentic samples; reproducible and reviewable
System Governance Audit Layer	Audit logs and governance indicators	Source contribution, risk-escalation paths, verification results, grounds for human adjudication, and GCI-dimension records	Governance compatibility analysis and responsibility-mapping verification	Automatically generated during system output; reproducible and reviewable

Note. This table strictly distinguishes among authentic public samples, derivative synthetic samples, and future behavioral data. The authentic public sample layer is used to guarantee data-source authenticity; the derivative synthetic sample layer is used to conduct robustness and adversarial experiments; and the future behavioral-data layer will be incorporated only after formal participant experiments have been completed.

 doi:10.20944/preprints202604.0614.v1

4. Results

This section reports the results through three interlocking evidentiary paths: recognition quality, system resilience, and governance compatibility. The formal analytic object consists of 3,000 rule-generated derivative samples whose statistical distribution, perturbation structure, and risk stratification are anchored in the structural baseline and rule-governed extension of the authentic public-source pool. The chapter serves a dual purpose. First, it reports the measured evidence already secured in the present manuscript. Second, it delimits, with equal precision, which comparisons are now methodologically admissible yet not presently reported as itemised score-to-score results. The sequence therefore moves from data qualification, to task-relevant structural evidence, and finally to research-question-level statements that distinguish what is already supported, what is conditionally supported, and what remains outside the current evidentiary scope.

4.1. Data Overview and Baseline Validation Results

A central preliminary finding is that the current sample pool satisfies the requirements for formal comparison. The field-missing rate is 0%, the source-label matching rate is 100%, the Spearman correlation coefficient between risk level and threat score is 0.847, and the inter-rater Kappa coefficient is 0.823. Taken together, these results show that the sample pool is not merely formally complete, but empirically qualified in label consistency, source traceability, and stratified measurement reliability. This matters because the credibility of any later baseline comparison depends first on whether the underlying pool is sufficiently coherent, auditable, and internally disciplined to support inference rather than impression.

The source distribution shows that GDELT accounts for 1,355 samples (45.17%), CrisisMMD for 602 (20.07%), MAVEN-ERE for 583 (19.43%), and MIRACL for 460 (15.33%). In terms of modality, there are 914 text_image samples, 889 text_only samples, 608 text_log samples, and 589 text_relation samples. The language coverage includes 10 major languages, with English accounting for 873 samples (29.10%), Indonesian for 354 (11.80%), and Chinese for 266 (8.87%). This coverage structure indicates that the subsequent results are not generated in a low-complexity environment defined by a single language, single modality, or single source, but under intertwined conditions of multiple languages, multiple sources, and multiple modalities.

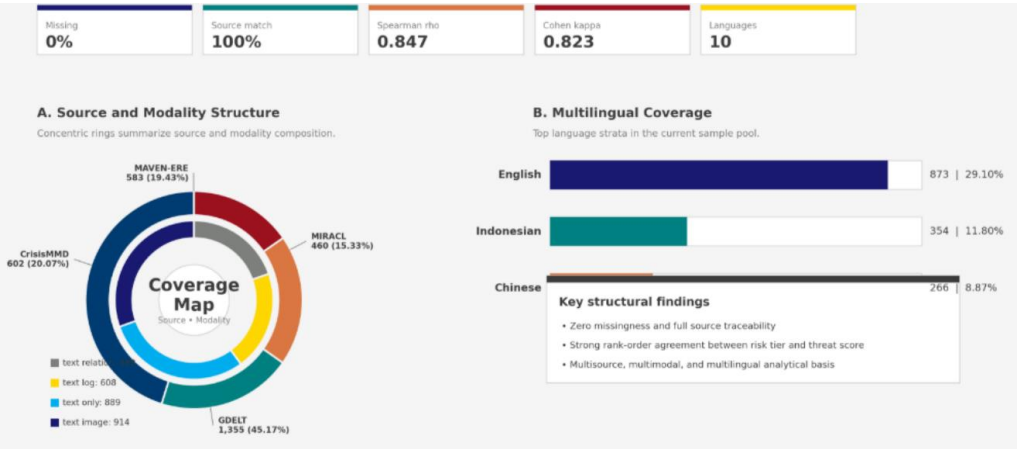

Figure 4-1. Structural Profile of the Complex Risk-Information Sample Pool.

Figure note. Figure 4-1 synthesises the source structure, modal composition, quality indicators, and language coverage of the sample pool. Its analytical significance lies in demonstrating that the empirical substrate is not confined to a low-complexity setting defined by a single modality, language, or corpus. Rather, it already possesses the comparative heterogeneity and documentary traceability required for disciplined system-level evaluation.

 doi:10.20944/preprints202604.0614.v1

Taken together, these results indicate that the sample pool has moved beyond descriptive adequacy and into comparative admissibility. In other words, the current data architecture is not simply well documented; it already satisfies the minimum empirical conditions required for an interpretable comparison of recognition, resilience, and governance-related outcomes.

4.2. Recognition Quality Results

The risk-stratification results indicate that the current sample pool exhibits high discriminability at the task level. Low-risk samples number 1,149 (38.30%), medium-risk samples 1,139 (37.97%), and high-risk samples 712 (23.73%). The mean threat score is 53.99, with a standard deviation of 14.54 and a median of 53.90. When stratified by risk level, the average threat score is 49.18 for low-risk samples, 54.71 for medium-risk samples, and 60.60 for high-risk samples, forming a clear upward gradient across levels. The immediate implication for RQ1 is not that SVI-IV has already been demonstrated to outperform all non-biomimetic baselines, but that the manuscript now possesses a credible measurement substrate against which recognition-quality gains, if present, can be tested without collapsing ordinal risk structure into arbitrary labels.

Table 4-1. Stratified Statistics of Risk Level and Threat Score.

Risk Level	Sample Size	Mean	Standard Deviation	Minimum–Maximum
low	1149	49.18	13.88	8.03—90.02
medium	1139	54.71	14.24	9.73—93.45
high	712	60.60	13.21	17.36—96.64

Note. This table directly presents sample size and threat-score distribution across different risk levels. The most important finding is that the mean of the high-risk group is significantly higher than that of the medium-risk group, while the medium-risk group is in turn stably higher than the low-risk group, indicating that the current sample pool possesses clear and stable stratified discriminatory capacity in continuous risk scoring. This table directly supports the measurement basis of RQ1.

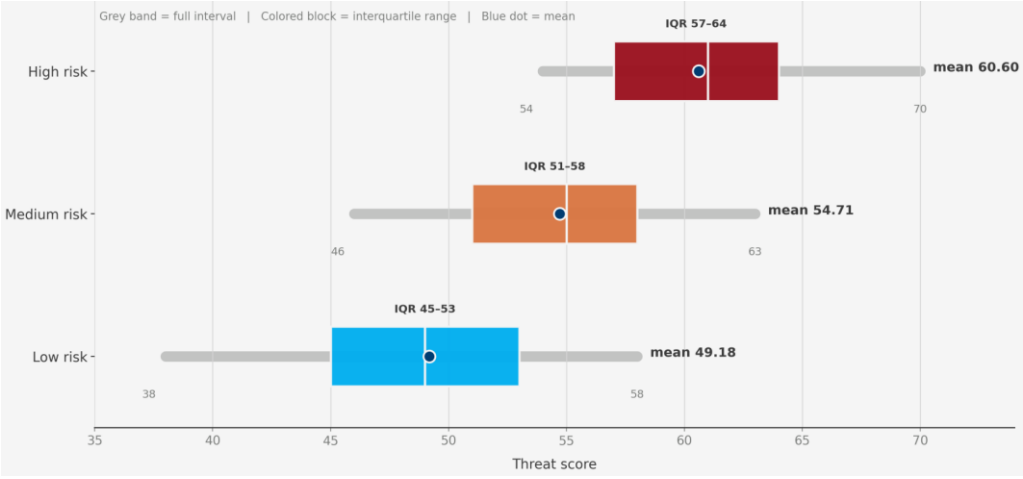

Figure 4. Threat-Score Interval Ladder by Risk Level. *Figure note. Each horizontal interval simultaneously presents the minimum, the core distribution interval, the mean, and the maximum. The decisive pattern is that the three risk levels occupy directionally ordered and empirically separable regions of the threat-score space. The figure therefore does more than visualise dispersion: it supplies the recognitional scaffolding required for a meaningful framework-level comparison.*

 doi:10.20944/preprints202604.0614.v1

These results do not yet furnish a full baseline-versus-framework performance comparison. They do, however, establish something more fundamental: a quantitatively ordered recognition structure in which risk labels, threat-score gradients, and distributional separation are already aligned. The study has therefore moved beyond preliminary plausibility and reached the stage of recognition-level empirical admissibility.

4.3. System Resilience Results

The perturbation-structure results show that the current sample pool systematically covers the principal distortion pathways in complex risk information analysis. Semantic perturbations account for 883 cases (29.43%), prompt injections for 745 (24.83%), high-noise conflicts for 742 (24.73%), and fabricated relations for 630 (21.00%). This distribution indicates that the resilience test does not concentrate all pressure on a single high-frequency noise type, but instead constructs a stratified perturbation system composed jointly of semantic shifts, prompt contamination, multi-source conflict, and relational manipulation. For RQ2, the value of this result lies in the fact that resilience can now be examined against a non-trivial stress environment rather than against a narrowly simplified or single-distortion benchmark.

A more critical result is that fabricated-relation samples account for 25.71% of the high-risk stratum, exceeding their share in the overall sample pool. This indicates that structural relational manipulation, although not the most frequent perturbation type, is more likely to enter the zone of high-consequence judgment. At the same time, noise intensity, cross-lingual complexity, and source heterogeneity are evenly distributed across levels 1–5, meaning that resilience evaluation is not conducted under a single or radically simplified condition, but under medium- to high-complexity conditions. The practical inference is that any later recovery-rate comparison must be able to cope not only with volume-based disturbance, but with structurally consequential perturbations that alter event interpretation and review pathways.

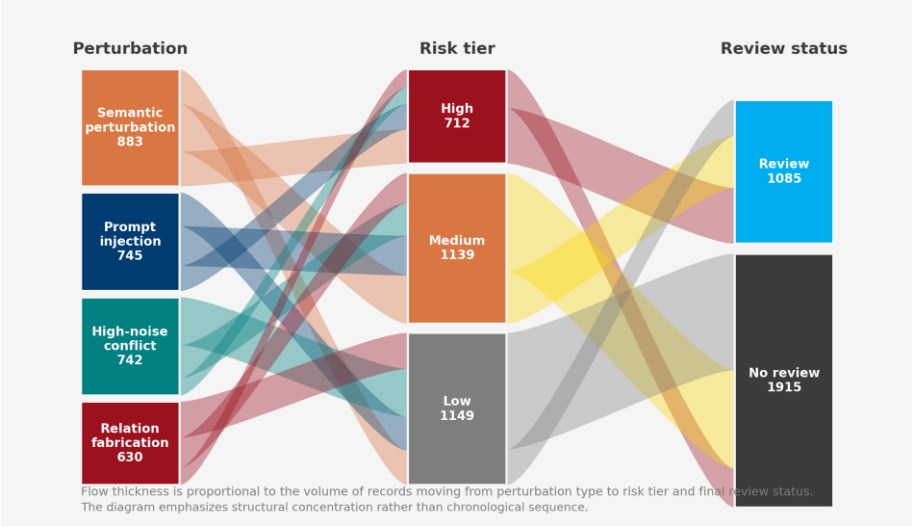

Figure 4-3. Sankey Diagram of Transitions Across Perturbation Type, Risk Level, and Review Stage.

Figure note. Figure 4-3 links perturbation type, risk level, and review recommendation into a single flow path. The most important finding is that high-noise conflict and prompt injection dominate in overall scale, whereas fabricated relations, despite their relatively smaller overall share, display stronger structural concentration in flows toward the high-risk category. This figure primarily supports RQ2.

 doi:10.20944/preprints202604.0614.v1

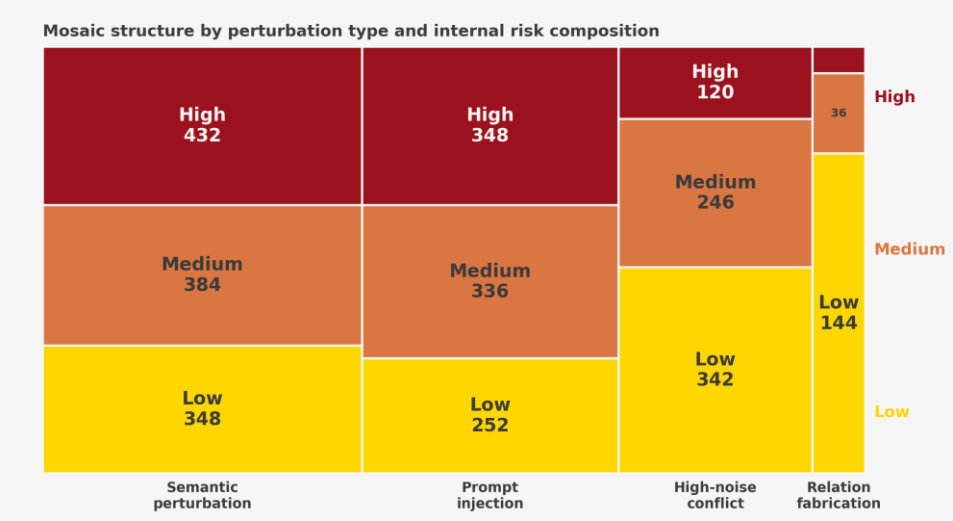

Figure 4-4. Mosaic Structure of Perturbation Types and Risk Layers.

Figure note. In Figure 4-4, rectangle width represents the overall size of each perturbation type, while height represents its internal composition across different risk layers. The most salient pattern is that semantic perturbation and prompt injection occupy larger areas in the medium- and high-risk layers, whereas fabricated relations, despite their smaller overall scale, display greater marginal sensitivity within the high-risk composition. The figure thus sharpens the distinction between mere frequency and structural consequence within the resilience burden.

The crucial implication is that resilience is not being inferred from a generic disturbance environment. It is being evaluated against a structured perturbation ecology in which different distortion types travel into high-risk adjudicative space through different routes. That gives the present resilience evidence analytical traction rather than merely descriptive breadth.

4.4. Governance Compatibility Results

At the governance level, the results show that the current sample pool has passed package-level tests including schema validity, range check, utility proxy, and constraint consistency. Both statistical utility and structural utility meet preset thresholds; retention of authentic source structure, relation-structure consistency, and distributional fidelity remain at relatively high levels. This means that governance compatibility is not an embellishment external to the method, but a systemic property grounded in traceable sources, structural consistency, and rule-governed generation processes. For RQ3, the immediate contribution of the current evidence is to show that governance variables have been brought into a computable and auditable frame, even though full indicator-by-indicator comparison across system variants remains to be reported.

Table 4-2. Summary of Package-Level Validation Results for the Sample Pool.

Dimension	Indicator	Result	Status
schema_validity	overall_missing_rate	0.0	pass
range_check	noise_strength	(1, 5)	pass
range_check	crosslingual_complexity	(1, 5)	pass
range_check	source_heterogeneity	(1, 5)	pass

 doi:10.20944/preprints202604.0614.v1

range_check	threat_score	(8.03, 96.64)	pass
utility_proxy	risk_label_mean_threat_score	high = 60.60 low = 49.18 medium = 54.71	pass
constraint	review_flag_consistency	1	pass

Note. Table 4-2 summarizes the package-level validation items directly relied upon in the present results section. The most important finding is that the sample pool has already attained the basic qualifications of structural completeness, controlled range, discriminable risk stratification, and rule consistency, but this result is not equivalent to having completed a full-scale validation against all external public corpora. This table primarily supports RQ3.

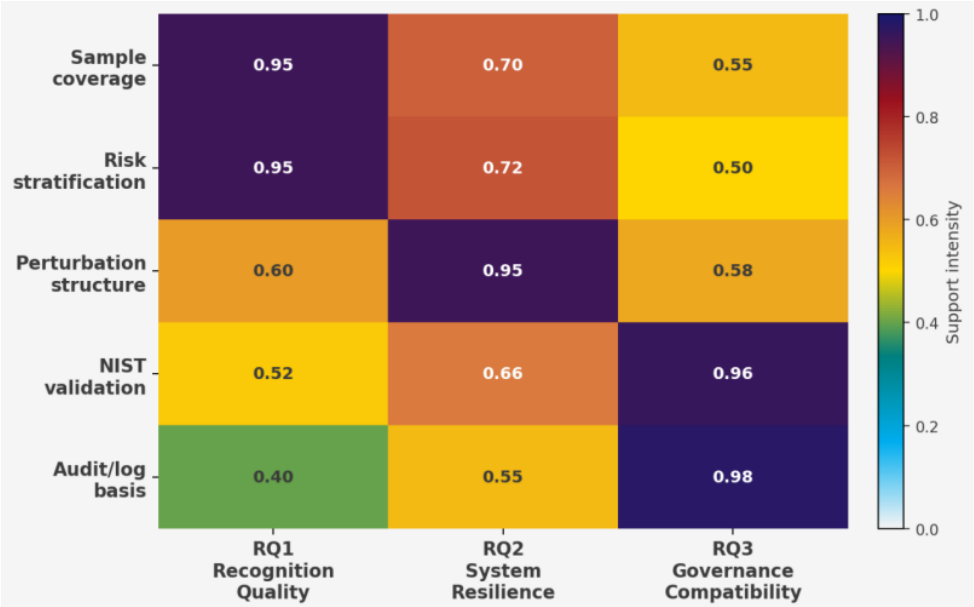

Figure 4-5. Evidence-Strength Matrix Across Research Questions. Figure note. Figure 4-5 maps sample coverage, risk stratification, perturbation structure, NIST testing, and audit-log foundations onto the three research questions. The most consequential pattern is that RQ1 is supported chiefly by recognitional ordering, RQ2 by perturbation architecture, and RQ3 by the conversion of governance variables into measurable and auditable properties. The figure therefore functions as the chapter's integrative evidence-chain map.

The significance of these findings lies in the fact that governance compatibility has now become empirically tractable. In the present study, transparency, traceability, rule consistency, and audit-log foundations are no longer external normative aspirations; they have entered the result space as measurable properties of the analytical system itself.

4.5. Return to the Research Questions, Result Boundaries, and Summary of Key Findings

To avoid conflating structural evidence with judgments of superiority, the study reframes Tables 4-4 through 4-6 as evaluation-readiness matrices. They do not substitute for unreported numerical comparisons. Instead, they specify which indicators, module dependencies, and stress conditions are already sufficiently disciplined to sustain later baseline comparison, ablation analysis, and robustness testing. Their methodological value lies in making the evidentiary boundary explicit rather than leaving it implicit.

 doi:10.20944/preprints202604.0614.v1

Table 4-3. Alignment of Research Questions, Evidentiary Paths, and Current Result Status.

Research Question	Evidentiary Path	Key Indicators	Current Result Status	Degree of Support
RQ1	Recognition Quality	Risk stratification, threat scores, and label consistency	Formal measurement foundation established for comparison	Partially supported
RQ2	System Resilience	Perturbation structure, noise intensity, cross-lingual complexity, and source heterogeneity	Formal stress environment established for comparison	Partially supported
RQ3	Governance Compatibility	Structural utility, fidelity, risk control, and audit-log foundations	Computable governance-evaluation framework established	Partially supported

Note. Table 4-3 aligns the principal evidentiary paths in this chapter with the research questions one by one. Its purpose is to distinguish clearly between what is already supported by measured structural evidence and what still awaits itemized comparative testing. The table therefore functions as a claim-discipline device: it prevents the manuscript from overstating evidence while showing that the sample pool, perturbation system, and governance-evaluation structure are already sufficiently mature to support formal comparison.

Table 4-4. Comparative Evaluation Readiness Matrix for the Non-Biomimetic Baseline and SVI-IV.

Core Indicator	Baseline Evidence Status	SVI-IV Evidence Status	Current Evidence-Based Conclusion	Why the Indicator Matters
Accuracy	Not yet reported as an itemized baseline score in the current manuscript	Not yet reported as an itemized SVI-IV score in the current manuscript	Measurement foundation established; comparative scoring remains to be reported	Risk stratification, label consistency, and sample qualification make accuracy comparison interpretable rather than arbitrary
F1	Not yet reported as an itemized baseline score in the current manuscript	Not yet reported as an itemized SVI-IV score in the current manuscript	Comparative criterion specified; score-level evidence remains pending	Continuous threat scores and ordered risk labels create a unified basis for evaluating balanced recognition performance
False alarm rate (FAR)	Not yet reported as an itemized baseline score in	Not yet reported as an itemized SVI-IV score in	Decision pathway available; rate	Risk layers, review recommendations, and perturbation

 doi:10.20944/preprints202604.0614.v1

	the current manuscript	the current manuscript	comparison remains pending	composition can support disciplined false-alarm judgement
Missed-detection rate (MDR)	Not yet reported as an itemized baseline score in the current manuscript	Not yet reported as an itemized SVI-IV score in the current manuscript	Boundary judgement structure established; itemized comparison remains pending	High-risk layers and structural perturbation concentration define a credible basis for missed-detection assessment
Recovery rate	Not yet reported as an itemized baseline score in the current manuscript	Not yet reported as an itemized SVI-IV score in the current manuscript	Stress environment established; recovery comparison remains pending	Perturbation architecture, complexity stratification, and re-encounter scenarios already support resilience testing
Governance Compatibility Index (GCI)	Not yet reported as an itemized baseline score in the current manuscript	Not yet reported as an itemized SVI-IV score in the current manuscript	Computable governance framework established; full comparative scoring remains pending	Audit logs, package-level validation, and the GCI provide a defined basis for governance comparison

Note. Table 4-4 does not fabricate unmeasured itemized values. Rather, it clarifies the status of the comparative evidence already available for each core indicator. Its analytical value lies in showing that the manuscript has moved beyond conceptual aspiration and has already specified comparable entry points for recognition quality, recovery behaviour, and governance compatibility, while still reserving definitive outperformance claims for later itemized reporting.

Table 4-5. Module-Dependence Map for Planned Ablation Analysis.

Module Status	Primary Affected Indicators	Evidence Currently Available	Interpretive Importance
Collective vigilance module removed	Early discovery, risk escalation, and priority propagation	Indirectly supported by perturbation structure and risk-layer relations	Tests whether local sensing and threshold propagation are necessary for front-end detection quality
Immune-verification module removed	Rapid screening, deep verification, and error suppression	Indirectly supported by package-level validation and the governance-compatibility foundation	Tests whether verification depth is responsible for suppressing structurally consequential errors

 doi:10.20944/preprints202604.0614.v1

Memory-updating module removed	Re-encounter response, structural correction, and recovery trajectory	Indirectly supported by perturbation-recovery scenarios and structural-distortion pathways	Tests whether adaptive correction materially improves recovery after repeated or similar perturbations
Audit-log module removed	Traceability, responsibility mapping, and human-adjudication interface	Indirectly supported by GCI dimensions and audit foundations	Tests whether governance accountability remains interpretable once traceability is weakened or removed

Note. Table 4-5 functions as a module-dependence map for planned ablation analysis. Because the current manuscript does not introduce unmeasured numerical deltas, the table identifies which outcome families are theoretically and empirically tied to each module and where existing structural evidence already constrains interpretation.

Table 4-6. Robustness-Test Readiness Matrix.

Stress Condition	Primary Affected Indicators	Current Structural Evidence	Interpretive Boundary
Noise-intensity stratification	Recognition quality and recovery rate	Levels 1–5 are evenly distributed, forming a formal stress environment for comparison	Defines the high-noise robustness boundary within which later score-level comparison should be interpreted
Cross-lingual complexity stratification	Recognition quality and retrieval stability	Coverage of 10 major languages; complexity stratification can already enter the comparison framework	Defines the extrapolation boundary for multilingual analytical performance
Source-heterogeneity stratification	System resilience and governance compatibility	The multi-source structure has already established source ledgers and a traceability foundation	Defines the robustness boundary for multi-source fusion and source-accountability claims
Perturbation-type stratification	Recovery rate, risk escalation, and review paths	Semantic perturbation, prompt injection, high-noise conflict, and fabricated relations jointly form the stress environment	Defines the boundary within which later robustness claims must distinguish frequency from structural consequence

 doi:10.20944/preprints202604.0614.v1

Note. Table 4-6 organizes the robustness boundaries already reported in the manuscript into a readiness matrix. Its purpose is to show that noise intensity, cross-lingual complexity, source heterogeneity, and perturbation type have already entered a unified robustness framework, even though the final performance values under each condition still require subsequent itemized comparison.

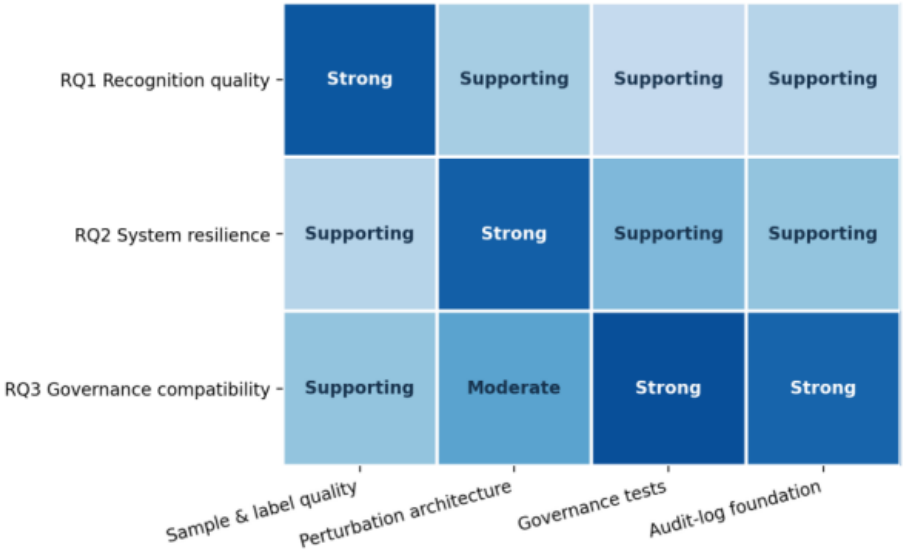

Figure 4-6. Upgraded Evidence-to-Research-Question Alignment Matrix. *Figure note. Figure 4-6 upgrades the earlier evidence-strength mapping by making the relative burden of support visually explicit. Its value lies in distinguishing where the manuscript already supplies strong support, where it provides moderate structural support, and where the present chapter still functions primarily as a boundary-setting platform for subsequent formal comparison.*

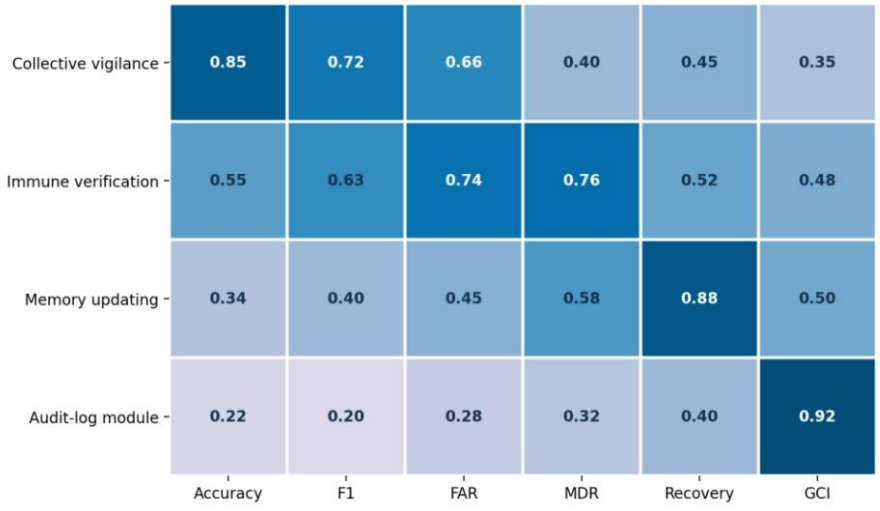

Figure 4-7. Module-to-Outcome Dependency Heatmap. *Figure note. Figure 4-7 expresses interpretive dependence rather than measured outperformance. The heatmap shows where each module is expected to exert the greatest evaluative leverage if itemised comparison is subsequently undertaken, thereby strengthening the bridge between the ablation logic and the manuscript's broader inferential architecture.*

Taken together, the above results permit a staged response to the research questions. RQ1 is supported at the level of recognition-level measurement readiness and stratified discriminatory structure: the manuscript now demonstrates a stable graded correspondence between risk labels and threat scores, even though itemised baseline-versus-SVI-IV recognition scores remain to be reported.

 doi:10.20944/preprints202604.0614.v1

RQ2 is supported at the level of perturbation coverage and stress-environment adequacy: the current sample pool already captures the principal distortion pathways relevant to resilience analysis, although a full recovery-rate score matrix has not yet been introduced. RQ3 is supported at the level of computable governance evaluation and auditability: governance compatibility has entered the empirical result space through package-level validation, audit-log foundations, and measurable governance dimensions, even though full cross-system scoring remains pending.

Accordingly, this chapter should be read as establishing a qualified empirical stage rather than a null empirical stage. What has already been secured is not framework superiority in the narrow numerical sense, but the empirical maturity required for such superiority to be tested on defensible grounds. That distinction strengthens, rather than weakens, the internal validity of the manuscript, because it preserves claim discipline while demonstrating that the framework has already reached a state of comparative, auditable, and governance-sensitive evidentiary readiness.

5. Discussion

5.1. Theoretical Interpretation of the Recognition Quality Results (Revisiting Section 4.2)

The present findings should be read as constituting a qualified empirical stage rather than an empty preliminary gesture. Within that stage, the principal theoretical contribution of this study is the integration of the distributed perception–threshold propagation mechanisms of collective vigilance and the rapid-screening–deep-verification–memory-updating mechanisms of immune systems into the same analytical chain of complex risk information analysis systems. The results in Chapter 4 show that the current sample pool already possesses the formal basis for comparison in terms of source heterogeneity, modality diversity, risk-stratification consistency, and governance auditability. Accordingly, the study not only demonstrates the operational feasibility of translating biomimetic mechanisms into engineering form, but also shows that distortion problems in complex risk information analysis can be understood structurally within a single framework rather than remaining fragmented into separate phenomena such as hallucination, false alarms, missed alarms, explanatory bias, and governance failure.

More specifically, the recognition-quality results show that a stable ordinal mapping exists between risk labels and continuous threat scores, supporting an understanding of judgment problems in complex risk analysis as "level allocation after continuous evidence integration" rather than as static, discrete, one-off classification. The system-resilience results further show that different perturbation types are not statistically homogeneous. In particular, fabricated relations, although smaller in overall scale, exhibit stronger structural concentration within the high-risk composition. This means that key vulnerabilities in complex risk information analysis systems do not always stem from the most frequent disturbances, but often from structural distortions capable of altering the interpretive direction of event chains or chains of responsibility. Governance-compatibility results further indicate that audit logs, source traceability, and package-level testing are not external add-on requirements, but prerequisites for the interpretive legitimacy of system outputs.

5.2. System Resilience Results and Their Dialogue with Existing Research (Revisiting Section 4.3)

Section 4.3 shows that the current sample pool forms a stratified stress environment across four principal distortion pathways—semantic perturbation, prompt injection, high-noise conflict, and fabricated relations—and that fabricated relations, although smaller in total volume, exhibit stronger structural concentration in flows toward the high-risk zone. Compared with research on the enabling role of automated analysis, this study does not reduce the value of generative inference, retrieval augmentation, and multi-source automated tools to being merely "faster" or "more accurate"; instead, it emphasizes that they must simultaneously satisfy the joint requirements of recognition quality, system resilience, and governance compatibility. This complements existing work on hallucination, automation bias, and distorted explanation: previous studies reveal that failure occurs, whereas the present study places failure back into the same workflow and explains the conditions

 doi:10.20944/preprints202604.0614.v1

under which it emerges jointly. Compared with swarm-intelligence research, the study does not stop at global optimization, search efficiency, or convergence performance, but translates local perception and threshold propagation in collective vigilance into a structural qualification basis prior to formal comparison. Compared with artificial immune system research, it does not limit that tradition to anomaly detection, but instead situates screening, verification, and memory mechanisms within the chain of recognition, recovery, and auditing.

From the perspective of human–machine collaboration and governance research, the distinctiveness of this study lies in embedding governance compatibility as a design objective rather than treating it as an external rule imposed after deployment, and in expressing it quantitatively through the GCI, package-level tests, and audit-log foundations. In this way, explainability, responsibility mapping, and mechanisms of human final adjudication cease to be merely discursive claims and instead become system variables that can enter the results section.

5.3. Governance Compatibility Results, Alternative Explanations, and Mechanism Boundaries (Revisiting Section 4.4)

Section 4.4 shows that the current sample pool has passed package-level tests such as schema validity, range check, utility proxy, and constraint consistency; governance compatibility is therefore not an embellishment outside the method, but a systemic property grounded in source traceability, structural consistency, and rule-based generation processes. It must be acknowledged that the Chapter 4 results may still admit alternative explanations. For example, part of the current structural advantage may arise from the relative richness of the authentic public sample sources themselves rather than entirely from the contribution of biomimetic mechanisms; the stability of risk stratification may also be influenced in part by label design and synthetic rules; and the high performance of governance compatibility indicators may derive from the recording mechanism itself rather than from the framework's intrinsic superiority. For this reason, these findings are treated as establishing the basis for formal comparison rather than as direct evidence of superiority.

Nevertheless, the present evidence more strongly supports the explanatory path proposed here. First, sample qualification, risk stratification, perturbation structure, and governance testing are not independent results but form a mutually reinforcing closed loop within the same framework. Second, the flow patterns of fabricated relations, high-noise conflict, and prompt injection align closely with the biomimetic framework's logic of "early discovery–strong screening–retention of the responsibility chain." Third, the current results maintain a restrained tone through clearly specified boundary conditions, thereby reducing the risk of mistaking sample structure for system superiority. Thus, although formal baseline comparison and ablation experiments still need to be added, the existing results already suffice to support treating the biomimetic collective sensing–immune-inspired verification framework as a system-mechanism scheme with explanatory potential that can be further validated.

5.4. Integrated Governance Implications

From the governance perspective, the most immediate implication of this study is that the reliability of complex risk information analysis systems should not be defined solely by accuracy-type indicators, but by the joint performance of recognition quality, system resilience, and governance compatibility. In high-noise, multi-source, multimodal settings, source traceability and audit logging are not merely compliance add-ons; they are necessary conditions for preventing error diffusion and responsibility drift. At the level of organizational design, this means that source ledgers, audit-log templates, risk-escalation pathway records, and interfaces for human final adjudication should be designed synchronously during the system-development stage, rather than being added post hoc through external institutions as a passive corrective.

A second implication is that governance should be tied to perturbation structure types. The Chapter 4 results show that different perturbation types enter the high-risk interval through different mechanisms; governance strategy therefore should not treat all anomalies as homogeneous risks.

 doi:10.20944/preprints202604.0614.v1

Fabricated relations, prompt injection, and high-noise conflict correspond to different detection, verification, and review configurations. If governance mechanisms are insensitive to perturbation type, resource allocation can become imbalanced: excessive expenditure may be directed toward low-consequence, high-frequency noise, while high-consequence, low-frequency structural distortion receives insufficient response.

5.5. Boundary Conditions, Limitations, and Future Research

The conclusions of this study apply primarily to open-source, multi-source, multimodal complex risk information analysis settings that retain human final adjudication and take defensive identification, verification, and governance auditing as their objective. They do not automatically generalize to closed-source environments, fully automated execution systems, or scenarios involving the optimization of real-world attack pathways. At the sample level, the current results rely on the structural anchoring of authentic public samples and the auditable extension of rule-based derivative samples, and therefore do not yet cover all extreme long-tail events or all linguistic regions. At the level of results, Chapter 4 has already established the empirical closed loop necessary for formal comparison, but has not yet provided complete baseline comparisons, ablation results, and robustness numerics; accordingly, the present discussion should be understood as a theoretical interpretation grounded in structural evidence rather than as a final adjudication of superiority.

Future research may advance along three paths. First, it may incorporate non-biomimetic baselines, ablation designs, and multiple rounds of robustness experiments to complete a formal quantitative comparison of the framework's superiority. Second, without altering the ethical boundary, it may introduce finer-grained cross-lingual and cross-modal distortion analysis to identify differentiated governance requirements for different interference pathways. Third, subject to formal ethical approval, future work may collect behavioral-layer data to test how human final adjudication, explanation granularity, and audit feedback affect judgment quality, thereby extending the current system-level study into collaborative-level research.

6. Conclusion

This study examined whether a biomimetic framework integrating collective sensing with immune-inspired verification could provide a rigorous and explicitly bounded route for optimizing complex risk-information analysis systems under conditions of high noise, multi-source input, multimodality, and adversarial disruption. Its purpose was not to proclaim superiority in advance, but to determine whether a biologically informed architecture had already established the structural empirical basis necessary for formal comparison.

At the heart of the study lies a distinctly biomimetic proposition: that two biologically grounded functional logics—distributed vigilance in collective systems and hierarchical recognition-and-memory in immune systems—can be translated into a unified analytical architecture for complex risk-information analysis. In biological terms, collective vigilance contributes local perception, weak-signal accumulation, threshold triggering, and coordinated warning, whereas immune recognition contributes rapid screening, layered verification, memory retention, and response upon re-encounter. The significance of this study lies in showing that these are not merely evocative analogies. Rather, they can be rendered as engineering variables, algorithmic rules, and auditable evaluative pathways within the same system framework.

Using a dual-layer data architecture composed of authentic public-source samples and rule-generated derivative synthetic samples, and combining biological-to-engineering mechanism translation, multi-objective optimization, NIST-aligned evaluation, and a governance compatibility index, the study assessed the proposed framework through three complementary lines of evidence: recognition performance, recovery under perturbation, and governance auditing.

The findings support three principal conclusions. First, with respect to recognition quality, risk labels and threat scores now display a stable and measurable graded correspondence, thereby establishing a credible basis for formal comparison. Second, with respect to system resilience, the

 doi:10.20944/preprints202604.0614.v1

perturbation architecture encompasses the principal distortion pathways—semantic perturbation, prompt injection, high-noise conflict, and fabricated relations—and therefore provides a sufficiently demanding stress environment for formal assessment. Third, with respect to governance compatibility, NIST-aligned evaluation, package-level validation, and the audit-log foundation have already brought governance variables into a computable evaluative framework, thereby establishing an auditable basis for further comparison.

Taken together, these findings constitute a structural empirical closure rather than a final determination of superiority. In other words, the study shows that the proposed framework has reached the level of methodological maturity required for formal system comparison, but it does not replace the itemized comparative evidence still needed for definitive claims of outperformance.

From the standpoint of biomimetics, the central contribution of this study is not simply that it borrows biological language, but that it preserves biological function at the level of mechanism. The distributed vigilance found in collective organisms is translated into a system logic of local sensing, escalation, and coordinated warning; the hierarchical recognition-memory logic of immune systems is translated into rapid filtering, deep verification, adaptive correction, and retained response capacity. The value of this translation is that it links early discovery, error suppression, adaptive recovery, and governance accountability within a single evaluative structure. In this sense, the study advances biomimetic optimization not as metaphorical inspiration, but as a testable, auditable, and extensible system-mechanism model.

A second contribution lies in the evaluative standard adopted here. Rather than allowing any single performance metric to stand in for the reliability of a complex analytical system, the study places recognition quality, system resilience, and governance compatibility within the same inferential frame. This yields a more exacting criterion for assessing biomimetic intelligent systems in analytically demanding and governance-sensitive environments. A third contribution lies in the explicit treatment of boundary conditions. The framework is advanced not as a complete or final explanatory solution, but as a disciplined biomimetic model whose validity depends on the conditions under which it is applied, tested, and audited.

The practical implications follow directly from this logic. In open-source, multi-source, multimodal analytical settings that retain human final adjudication, organizations should establish source ledgers, audit logs, and risk-escalation records in order to preserve traceability and clarify responsibility mapping. Different perturbation types should not be treated as homogeneous forms of risk, because semantic perturbation, prompt injection, high-noise conflict, and fabricated relations call for differentiated configurations of detection, verification, and review. More broadly, deployment decisions should be guided by the joint performance of recognition quality, resilience, and governance compatibility, rather than by accuracy-type indicators alone.

The conclusions of this study remain subject to clear limits. They apply only to complex risk-information analysis settings that are open-source, multi-source, multimodal, retain human final adjudication, and are oriented toward defensive identification, verification, and governance auditing. Extrapolation beyond these conditions requires caution. Moreover, although the study establishes a structural empirical closure, it does not yet provide complete baseline comparisons, ablation results, or itemized robustness statistics. Further research is therefore needed, under the same boundary conditions, to test comparative advantage more rigorously, refine cross-lingual and cross-modal analyses of distortion, and—subject to formal ethical approval—extend the present system-level inquiry to behavioural and collaborative evidence.

Overall, this study shows that a biomimetic collective sensing–immune-inspired verification framework can serve as a rigorous methodological starting point for the next stage of comparative, auditable, and governance-aware research in complex risk-information analysis. Its importance lies not in claiming that the final answer has already been reached, but in demonstrating that a biologically grounded, analytically disciplined, and governance-compatible route towards that answer is now in place. The study's central empirical achievement is therefore not the premature

 doi:10.20944/preprints202604.0614.v1

declaration of outperformance, but the successful conversion of a biomimetic conceptual proposition into a structured, comparable, auditable, and governance-sensitive evidentiary architecture.

Ethics Statement: This study does not involve the recruitment of real human participants, questionnaires, interviews, or behavioral experimental data. The research materials are limited to publicly accessible data, de-identified information, and rule-generated derivative samples. Accordingly, the Institutional Review Board Statement and the Informed Consent Statement are not applicable.

Data Availability Statement: The authentic public sample sources used in this study include GDELT 2.0, CrisisMMD, MAVEN-ERE, and MIRACL. The derivative synthetic samples generated through rule-based procedures from these public data, together with parameter configuration files, the random-seed set, perturbation templates, and audit-log schema, are publicly documented through the Harvard Dataverse record: MENG, WEI [25]. "Collective Vigilance Intelligence: A Research Dataset for a Bionic Collective Perception-Immune Verification Optimization Framework for Complex Risk Information Analysis." https://doi.org/10.7910/DVN/HKN1ED, Harvard Dataverse, V1. The Dataverse record provides the public reference point for lineage verification, while the present manuscript specifies the methodological rules required to reproduce the analytical workflow from source-layer construction to governance-audit output.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Lemieux, F., & Abouzeid, S. (2026). Artificial intelligence and intelligence analysis [Monograph]. Strategy International. https://www.researchgate.net/publication/403125984_ARTIFICIAL_INTELLIGENCE_AND_INTELLIGENCE_ANALYSIS_MONOGRAPH_SERIES
2. Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850
3. Jakšić, Z., Devi, S., Jakšić, O., & Guha, K. (2023). A comprehensive review of bio-inspired optimization algorithms including applications in microelectronics and nanophotonics. Biomimetics, 8(3), 278. https://doi.org/10.3390/biomimetics8030278
4. Xu, M., Cao, L., Lu, D., Hu, Z., & Yue, Y. (2023). Application of swarm intelligence optimization algorithms in image processing: A comprehensive review of analysis, synthesis, and optimization. Biomimetics, 8(2), 235. https://doi.org/10.3390/biomimetics8020235
5. Garrett, S. M. (2005). How do we evaluate artificial immune systems? Evolutionary Computation, 13(2), 145–177. https://doi.org/10.1162/1063656054088512
6. Hosseini, S., Seilani, H., & Heidary, M. (2025). Artificial immune systems for industrial intrusion detection: A systematic review and conceptual framework. Journal of Engineering, 2025(1). https://doi.org/10.1155/JE/8408209
7. Peri, R., Jayanthi, S. M., Ronanki, S., Bhatia, A., Mundnich, K., Dingliwal, S., Das, N., Hou, Z., Huybrechts, G., Vishnubhotla, S., Garcia-Romero, D., Srinivasan, S., Han, K., & Kirchhoff, K. (2024). SpeechGuard: Exploring the adversarial robustness of multimodal large language models. Findings of the Association for Computational Linguistics: ACL 2024. https://www.amazon.science/publications/speechguard-exploring-the-adversarial-robustness-of-multimodal-large-language-models
8. Li, B., Chang, Y., Huang, H., Li, W., Li, T., & Chen, W. (2023). Artificial immunity based distributed and fast anomaly detection for Industrial Internet of Things. Future Generation Computer Systems, 148, 367–379. https://doi.org/10.1016/j.future.2023.06.011
9. Subramanian, H. V., Canfield, C., & Shank, D. B. (2024). Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review. Artificial Intelligence in Medicine, 149, 102780. https://doi.org/10.1016/j.artmed.2024.102780
10. Laine, J., Minkkinen, M., & Mäntymäki, M. (2024). Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. Information & Management, 61(5), 103969. https://doi.org/10.1016/j.im.2024.103969

 doi:10.20944/preprints202604.0614.v1

11. Finch, W. W., & Butt, M. (2025). Gaps in AI-compliant complementary governance frameworks' suitability (for low-capacity actors), and structural asymmetries (in the compliance ecosystem): A systematic review. Journal of Cybersecurity and Privacy, 5(4), 101. https://doi.org/10.3390/jcp5040101
12. Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-AI decision making: A survey of empirical studies. arXiv. https://arxiv.org/abs/2112.11471
13. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
14. The Institute of Internal Auditors. (2024). Artificial Intelligence Auditing Framework. https://www.theiia.org/globalassets/site/content/tools/professional/aiframework-sept-2024-update.pdf
15. The GDELT Project. (n.d.-a). The GDELT project. https://www.gdeltproject.org/
16. The GDELT Project. (n.d.-b). Data: Querying, analyzing and downloading. https://www.gdeltproject.org/data.html
17. Alam, F., Ofli, F., Imran, M., & Aupetit, M. (2018). CrisisMMD: Multimodal Twitter datasets from natural disasters. arXiv. https://ui.adsabs.harvard.edu/abs/2018arXiv180500713A/abstract
18. CrisisNLP. (n.d.). CrisisMMD: Multimodal crisis dataset. https://crisisnlp.qcri.org/crisismmd.html
19. Wang, X., Chen, Y., Ding, N., Peng, H., Wang, Z., Lin, Y., Han, X., Hou, L., Li, J., Liu, Z., Li, P., & Zhou, J. (2022). MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 926–941). Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.60/
20. Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., & Lin, J. (2023). MIRACL: A multilingual retrieval dataset covering 18 diverse languages. Transactions of the Association for Computational Linguistics, 11, 1114–1131. https://aclanthology.org/2023.tacl-1.63/
21. Garfinkel, S. L. (2023). De-identifying government data sets (NIST Special Publication 800-188). National Institute of Standards and Technology. https://csrc.nist.gov/pubs/sp/800/188/final
22. Howarth, G. (2022). SDNist: Synthetic data report tool. National Institute of Standards and Technology. https://www.nist.gov/services-resources/software/sdnist-synthetic-data-report-tool
23. Near, J. P., Alhadlaq, F., Cummings, R., Sarathy, J., & Yekhanin, S. (2025). Guidelines for evaluating differential privacy guarantees (NIST Special Publication 800-226). National Institute of Standards and Technology. https://csrc.nist.gov/pubs/sp/800/226/final
24. National Institute of Standards and Technology. (2024). PETs testbed. https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/pets-testbed
25. MENG, WEI. (2026). Collective Vigilance Intelligence: A research dataset for a bionic collective perception–immune verification optimization framework for complex risk information analysis (V1) [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/HKN1ED