Compliance-Aware, Human-Gated Cloud Remediation: An LLM-Augmented, Specificity-Scored Framework for Multi-Jurisdictional Cloud Security Posture Management

An Empirical Evaluation Using the ConfigHawk Research Dataset

Sushant Santosh Khot[1], Guna Dhondwad[2]
[1]*Design, Analytics and Cyber Security, MIT Art, Commerce and Science College*
doi.org/10.64643/ijirt.208799-459

***Abstract*—Cloud Security Posture Management (CSPM) systems can produce large queues of configuration findings, yet a nominal severity label alone does not capture internet reachability, attainable privilege, asset importance, blast radius, attack-path contribution, or the regulatory relevance of the affected control. Organizations operating across India and South Korea also need a defensible way to relate one technical cloud condition to several differently structured security frameworks.**

This paper presents *ConfigHawk*, a compliance-aware and human-gated CSPM framework with three components: a 24-control canonical AWS catalogue mapped to CERT-In, ISMS-P, ISO/IEC 27001:2022, and NIST CSF 2.0 through atomic obligations and explicit relationship types; the locked CH-SPEC-v1.0 contextual prioritization model; and a rule-first remediation workflow in which deterministic actions remain approval-gated and an LLM is restricted to optional, on-demand explanation.

The mapping dataset contains 96 reviewer-accepted mappings. In a development-set evaluation of 98 findings retained from a 100-finding AWS scan, CH-SPEC-v1.0 achieved a Spearman rank correlation of 0.885 with the provisional expert-proxy priority order, compared with 0.405 for severity-only ranking; Top-10 overlap was 9/10 versus 6/10.

These results indicate that contextual factors can improve prioritization on the studied dataset, but they are not independent validation because the expert-proxy labels and model were developed within the same research programme.

***Index Terms*—Cloud Security Posture Management, compliance mapping, atomic obligations, CERT-In, ISMS-P, ISO/IEC 27001, NIST CSF, risk prioritization, human-gated remediation, large language models.**

I. INTRODUCTION

CSPM combines continuous configuration assessment, compliance checking, and remediation support for cloud environments [1]. Existing evaluations show that tools differ in the misconfigurations they detect and, in the coverage, offered by native, open-source, and custom scanners [2]. Detection, however, does not resolve which finding should be addressed first. Findings with the same vendor severity can present materially different risk when one is directly reachable, affects a privileged identity, or connects to a credible attack path, while another is isolated or already partly mitigated.This paper formalizes ConfigHawk as a research and engineering framework rather than treating the product interface itself as evidence of effectiveness. The contributions are: (1) a canonical cloud-control layer that preserves the separate semantics of four frameworks while mapping one AWS finding to all of them; (2) CH-SPEC-v1.0, a locked seven-factor scoring model evaluated against a severity-only baseline; and (3) a rule-first remediation architecture that confines cloud-changing operations to deterministic, allow-listed actions with explicit human approval and post-change scan verification. An LLM may explain a finding on request, but it has no authority over scoring, compliance status, approval, or execution.

II. RELATED WORK

Cloud Security Posture Management. CSPM literature describes continuous discovery of insecure cloud configurations, compliance drift, identity and access

weaknesses, and monitoring gaps, supported by cloud-native and third-party tools [1]. Moss compares native, open-source, and custom scanning approaches and focuses on differences in detection coverage [2]. This body of work establishes the value of continuous configuration assessment, but it gives less attention to ranking heterogeneous findings with resource and organizational context or to preserving semantics across several national and international frameworks.

Control crosswalks and compliance semantics. NIST IR 8278 Rev. 1 defines a structured approach for publishing element-to-element informative-reference mappings and relationships between cybersecurity documents [3]. Its value lies in repeatable, machine-readable crosswalks, while it also makes clear that a mapping does not itself certify compliance. ConfigHawk adopts this element-level discipline but uses a project-specific relationship vocabulary and decomposes each cloud control into atomic obligations so that covered and uncovered portions of a broader requirement remain visible. The target frameworks differ in purpose: NIST CSF 2.0 states outcome-oriented, non-prescriptive risk-management objectives [4]; CERT-In Directions emphasize incident, logging, time-synchronization, and reporting duties [5]; the CERT-In audit guidelines define evidence, independence, and risk-based audit expectations [6]; ISMS-P combines management and technical safeguards within a certification scope [7]; and ISO/IEC 27001:2022 links Annex A control selection to risk treatment and the Statement of Applicability [8].

Vulnerability and finding prioritization. CVSS v4.0 separates Base, Threat, Environmental, and Supplemental metrics and explicitly expects consumers to add environment-specific and organizational considerations that fall outside the score [9]. EPSS estimates the probability that a published CVE will be exploited in the wild during the next 30 days, which is valuable for likelihood but does not represent business impact or configuration-specific cloud context [10]. Koscinski et al. found substantial divergence among commonly used vulnerability-scoring and decision systems, showing that their outputs are not interchangeable [11]. Shimizu and Hashimoto combine KEV, EPSS, and CVSS signals to reduce CVE-remediation workload [12]. Their chaining approach is the closest prioritization precedent, but it remains vulnerability-feed-centric; ConfigHawk instead ranks cloud misconfigurations using reachability, privilege, blast radius, asset criticality, attack paths, compliance relevance, and evidenced mitigations.

LLM-assisted security operations. A ten-month study of 3,090 analyst queries found that security analysts used LLMs primarily as on-demand aids for interpretation, context gathering, and communication rather than as autonomous decision makers [13]. A broader survey identifies applications across alert triage, threat intelligence, incident response, reporting, and vulnerability management, while emphasizing governance and human oversight [14]. Experiments on LLM-based vulnerability triage further report over-prediction and difficulty with context-sensitive judgments [15]. These findings support a narrow advisory role: ConfigHawk uses deterministic rules for security decisions and permits an LLM only to explain evidence or remediation reasoning when an analyst requests it.

Research gap. Prior work separately addresses cloud misconfiguration discovery, framework crosswalks, vulnerability scoring, and LLM-assisted analyst work. The combined problem studied here is how to preserve multi-framework meaning, rank configuration findings with environment-specific evidence, and govern remediation without allowing probabilistic language-model output to become executable authority.

III. PROBLEM STATEMENT

Given a stream of AWS CSPM findings, (a) how should those findings be ranked so that the ranking reflects real-world risk rather than vendor severity alone, and (b) how can a single technical finding be connected to its obligations under multiple, independently-structured compliance frameworks, without requiring a separate manual mapping exercise for each framework and each audit cycle?

IV. RESEARCH OBJECTIVES AND QUESTIONS

4.1. Objectives

- O1: Design a canonical cloud-control layer that maps a bounded set of AWS controls to CERT-In, ISMS-P, ISO/IEC 27001:2022, and NIST CSF 2.0

without treating a technical check as proof of full compliance.

- O2: Define and lock a specificity-weighted scoring formula that reprioritizes findings relative to a severity-only baseline while retaining unknown context rather than silently replacing it with worst-case values.
- O3: Evaluate the score against a provisional expert-proxy ranking of a fixed AWS scan using rank correlation and top-k overlap metrics, and document the resulting validity constraints.
- O4: Specify a rule-first, human-gated remediation workflow in which deterministic actions are allow-listed, approved, audited, and verified by a later scan, while LLM assistance remains optional and non-executable.

4.2. Research Questions

- RQ1: On the fixed evaluation set, to what extent does CH-SPEC-v1.0 improve agreement with the expert-proxy priority order relative to native severity alone?
- RQ2: Can one canonical control layer represent the relationship between an AWS finding and four independently structured frameworks while retaining each framework's uncovered obligations and applicability conditions?
- RQ3: What proportion of the 24 canonical controls admits deterministic remediation, and how should context-dependent controls be handled without giving an LLM ranking or execution authority?

V. HYPOTHESES

H1: On the development dataset, a specificity-weighted score using exposure, privilege, exploitability, blast radius, asset criticality, attack-path relevance, compliance relevance, and evidenced mitigation will show higher rank agreement with the expert-proxy order than severity-only ranking.

H0: On the same dataset, the specificity-weighted ranking will not improve rank agreement over the severity-only baseline.

VI. SCOPE AND LIMITATIONS OF SCOPE

The technical scope is AWS and covers IAM, S3, CloudTrail, VPC/network controls, EC2, EBS, GuardDuty, AWS Config, and KMS through 24 canonical controls. The compliance experiment is limited to CERT-In Directions 2022, ISMS-P 2023.11, ISO/IEC 27001:2022 Annex A, and NIST CSF 2.0. GDPR and PCI DSS are extension targets but are not included. The prioritization experiment uses one 100-finding scan from one AWS account and a provisional expert-proxy ordering created within the project. Remediation is evaluated as an architecture and catalogue property; this paper does not claim a controlled production trial of cloud-changing actions or human-LLM collaboration.

VII. SYSTEM ARCHITECTURE

The architecture separates evidence collection, semantic mapping, prioritization, governed action, and validation. AWS scanner modules first create evidence-backed findings. A deterministic matcher links eligible findings to the 24-control catalogue, after which the control's pre-computed framework relationships are attached. CH-SPEC-v1.0 then computes a contextual score as a research and shadow-mode profile. At the issue level, analysts may fix, defer, accept risk, mark not applicable, or classify a false positive. Only approved deterministic actions in Fix Center may change cloud state; optional LLM explanation is advisory. A completed post-remediation scan, not an API success response, determines whether the finding has been removed. Figure 1 shows the governed path and keeps the advisory LLM branch outside approval and execution.

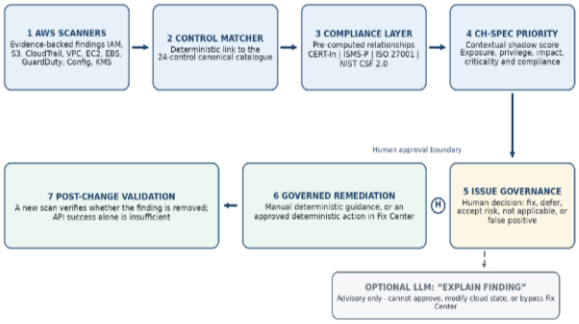

Figure 1. ConfigHawk governed scan-to-validation architecture. The dashed LLM branch is advisory and has no approval or execution authority.

Source: Author's architecture based on the ConfigHawk control catalogue and locked prioritization artifacts [16], [17].

VIII. METHODOLOGY

8.1. Canonical Control Catalogue

The catalogue contains 24 controls across nine AWS service areas: identity and access management, S3 storage, CloudTrail logging, network security, compute metadata protection, EBS encryption, threat detection, configuration recording, and key management. Each record contains a stable control identifier, machine-checkable condition, default severity, blast-radius classification, atomic obligations, framework mappings, and remediation metadata. The reviewer-validated catalogue is versioned as 1.2.0-reviewer-validated [16].

Table 1. ConfigHawk canonical control catalogue (24 controls).

Control ID	Title	Category	Severity
CH-IAM-001	Root account MFA is enabled	Identity & Access	Critical
CH-IAM-002	Root account access keys are absent	Identity & Access	Critical
CH-IAM-003	IAM identities do not receive unnecessary administrator privileges	Identity & Access	High
CH-IAM-004	IAM policies do not contain unjustified wildcard permissions	Identity & Access	High
CH-IAM-005	Inactive IAM access keys are disabled or removed	Identity & Access	Medium
CH-IAM-006	Unused IAM roles are reviewed and removed	Identity & Access	Medium
CH-S3-001	Amazon S3 Block Public Access is enabled	Storage Security	High
CH-S3-002	S3 bucket default encryption is enabled	Storage Security	High
CH-S3-003	S3 bucket versioning is enabled	Storage Security	Medium
CH-S3-004	S3 server access logging is enabled	Logging & Monitoring	Medium
CH-S3-005	S3 bucket policies do not permit anonymous access	Storage Security	Critical
CH-CT-001	A multi-region CloudTrail trail is enabled	Logging & Monitoring	Critical
CH-CT-002	CloudTrail log-file validation is enabled	Logging & Monitoring	Medium
CH-CT-003	CloudTrail logs use approved encryption	Encryption & Key Mgmt	High
CH-CT-004	CloudTrail logs are stored in a restricted central bucket	Logging & Monitoring	High
CH-CT-005	CloudTrail is integrated with CloudWatch Logs	Logging & Monitoring	Medium
CH-NET-001	SSH is not exposed to the public internet	Network Security	Critical
CH-NET-002	RDP is not exposed to the public internet	Network Security	Critical
CH-NET-003	Default security groups do not allow unrestricted traffic	Network Security	High
CH-EC2-001	EC2 instances require IMDSv2	Compute Security	High
CH-EBS-001	EBS volumes are encrypted	Encryption & Key Mgmt	High
CH-GD-001	Amazon GuardDuty is enabled	Logging & Monitoring	High
CH-CFG-001	AWS Config is enabled with required recording scope	Logging & Monitoring	High
CH-KMS-001	Customer-managed KMS keys use automatic rotation	Encryption & Key Mgmt	Medium

Source: ConfigHawk Canonical Cloud Security Control Catalogue, version 1.2.0-reviewer-validated [16].

8.2. Compliance Mapping Methodology

Mapping is performed at the smallest obligation that can be tested or evidenced. A canonical control is decomposed into atomic obligations, and each framework is assessed independently against those obligations. The project uses the relationship values satisfies, partially_satisfies, supports, not_applicable, no_relationship, and uncertain.

For every mapping, the dataset records applicability conditions, covered atomic obligations, broader framework obligations that remain uncovered, source evidence, interpretation notes, and confidence.

This adapts OLIR's element-level and relationship-aware discipline without claiming that the project vocabulary is identical to OLIR or that a crosswalk demonstrates certification [3].

Source basis and attribution. Framework interpretation is grounded in the official or primary documents: NIST CSF 2.0 [4], CERT-In Directions [5] and audit guidelines [6], the ISMS-P certification criteria [7], and ISO/IEC 27001:2022 [8]. NIST OLIR [3] informs the relationship-aware mapping discipline. CSPM, prioritization, and LLM-security literature [1], [2], [9]-[15] is used to position the contribution rather than to replace framework text. The Korean reform update [18] is used only as future-policy context. Source language is paraphrased throughout; formal control identifiers, document names, and standard terminology are retained only where necessary for traceability.

Atomic obligations are necessary because framework requirements and technical checks have different granularity. For example, restricting a public SSH rule can evidence one part of secure network management, but it does not establish organization-wide approval, monitoring, exception, or review processes. Recording only a binary mapped/not-mapped flag would overstate coverage; recording both covered and uncovered obligations allow the relationship to remain useful for engineering while preserving the auditor's view of what still requires evidence.

CERT-In. The 2022 Directions concentrate on operational duties such as reporting specified incidents, maintaining logs, synchronizing system clocks, and preserving information needed for response [5]. The 2025 audit guidelines add expectations for independent and evidence-based assessment, context-aware findings, risk-based planning, and remedial decisions approved by responsible management [6]. Therefore, ConfigHawk records no_relationship for many preventive AWS checks rather than forcing a mapping, while logging, monitoring, and incident-readiness controls may support or partially satisfy relevant duties.

ISMS-P. The 2023.11 criteria and detailed inspection items cover account and access-right lifecycle management, administrative privileges, authentication, remote access, encryption and key management, log protection, cloud security, and abnormal-behaviour monitoring [7]. An AWS control normally implements only a resource-specific portion of those criteria; policy, approval, inventory, monitoring, evidence retention, and periodic review remain certification-scope obligations.

ISO/IEC 27001:2022. The Annex A controls represented in Table 2 are A.5.16, A.5.17, A.5.33, A.8.2, A.8.3, A.8.5, A.8.9, A.8.13, A.8.15, A.8.16, A.8.20, and A.8.24 [8]. The mappings use short paraphrases and treat each AWS check as supporting or partially satisfying a control objective. They do not establish conformity with Clauses 4-10, which depends on the organization's management system, risk treatment, and Statement of Applicability.

NIST CSF 2.0. CSF 2.0 provides outcome-oriented guidance that is designed to be tailored to an organization's risk, sector, technology, and maturity rather than prescribing one implementation [4]. The catalogue maps to the outcomes actually relevant to the 24 controls, including identity and authorization, data protection, configuration management, infrastructure resilience, and continuous monitoring. A cloud configuration check is consequently treated as evidence toward an outcome, not as proof that the outcome is fully achieved.

Validation. The completed dataset contains 96 mappings (24 controls multiplied by four frameworks). Each mapping was reviewed and accepted by Reviewer-01 on 18 July 2026, and structural checks confirmed unique control and mapping identifiers and exactly four mappings per control [16]. This is reviewer validation, not multi-rater external validation; the absence of a second reviewer is retained as a threat to validity.

8.3. Full Compliance Mapping Table

Table 2 reports the primary reference and relationship assigned to each control. PS denotes partially_satisfies, S denotes supports, and NR denotes no_relationship. The dash is used where the source framework contains no directly corresponding obligation at the technical specificity of the AWS check.

Table 2. Compliance mapping summary across all four frameworks (96 mappings).

Control ID	NIST CSF 2.0	CERT-In 2022	ISMS-P	ISO/IEC 27001:2022
CH-IAM-001	PR. AA-03 (PS)	— (NR)	2.5.3 (PS)	A.8.5 (PS)
CH-IAM-002	PR. AA-01 (PS)	— (NR)	2.5.5 (PS)	A.8.2 (PS)

Control ID	NIST CSF 2.0	CERT-In 2022	ISMS-P	ISO/IEC 27001:2022
CH-IAM-003	PR. AA-05 (PS)	— (NR)	2.5.5 (PS)	A.8.2 (PS)
CH-IAM-004	PR. AA-05 (PS)	— (NR)	2.5.5 (PS)	A.8.3 (PS)
CH-IAM-005	PR. AA-01 (PS)	— (NR)	2.5.1 (PS)	A.5.17 (PS)
CH-IAM-006	PR. AA-01 (PS)	— (NR)	2.5.1 (PS)	A.5.16 (PS)
CH-S3-001	PR. AA-05 (PS)	— (NR)	2.10.2 (PS)	A.8.3 (PS)
CH-S3-002	PR.DS-01 (PS)	— (NR)	2.7.1 (PS)	A.8.24 (PS)
CH-S3-003	PR.DS-11 (S)	— (NR)	2.9.3 (S)	A.8.13 (S)
CH-S3-004	PR.PS-04 (PS)	Dir.(iv) (PS)	2.9.4 (PS)	A.8.15 (PS)
CH-S3-005	PR. AA-05 (PS)	— (NR)	2.10.2 (PS)	A.8.3 (PS)
CH-CT-001	PR.PS-04 (PS)	Dir.(iv) (PS)	2.9.4 (PS)	A.8.15 (PS)
CH-CT-002	PR.DS-01 (S)	Dir.(iv) (S)	2.9.4 (S)	A.5.33 (S)
CH-CT-003	PR.DS-01 (PS)	Dir.(iv) (S)	2.7.1 (PS)	A.8.24 (PS)
CH-CT-004	PR.DS-01 (PS)	Dir.(iv) (PS)	2.9.4 (PS)	A.5.33 (PS)
CH-CT-005	PR.PS-04 (PS)	Dir.(iv) (PS)	2.9.5 (PS)	A.8.15 (PS)
CH-NET-001	PR.IR-01 (PS)	— (NR)	2.6.6 (PS)	A.8.20 (PS)
CH-NET-002	PR.IR-01 (PS)	— (NR)	2.6.6 (PS)	A.8.20 (PS)
CH-NET-003	PR.IR-01 (PS)	— (NR)	2.6.1 (PS)	A.8.20 (PS)
CH-EC2-001	PR. AA-03 (S)	— (NR)	2.10.2 (PS)	A.8.5 (S)
CH-EBS-001	PR.DS-01 (PS)	— (NR)	2.7.1 (PS)	A.8.24 (PS)
CH-GD-001	DE.CM-09 (PS)	Dir.(ii) (S)	2.11.3 (PS)	A.8.16 (PS)
CH-CFG-001	PR.PS-01 (PS)	Dir.(iv) (S)	2.10.2 (PS)	A.8.9 (PS)
CH-KMS-001	PR.DS-01 (S)	— (NR)	2.7.2 (PS)	A.8.24 (PS)

Source: Author's reviewer-validated cross-framework mapping based on NIST OLIR [3], NIST CSF 2.0 [4], CERT-In [5], ISMS-P [7], and ISO/IEC 27001:2022 [8]; mapping artifact [16].

Across the reviewer-validated dataset, NIST CSF 2.0 contains 20 partially_satisfies and 4 supports relationships; CERT-In contains 16 no_relationship, 4 partially_satisfies, and 4 supports relationships; ISMS-P contains 22 partially_satisfies and 2 supports relationships; and ISO/IEC 27001:2022 contains 21 partially_satisfies and 3 supports relationships. No canonical control is recorded as fully satisfying a framework requirement by itself, because each technical check covers only a subset of broader governance, process, evidence, and lifecycle obligations.

8.4. Specificity Scoring Model (CH-SPEC-v1.0, frozen)

The severity-only baseline converts Critical, High, Medium, and Low labels to 4, 3, 2, and 1. CH-SPEC-v1.0 instead combines seven contextual factors normalized to [0,1]. Unknown is retained as a distinct conservative category rather than being promoted to the maximum. Evidenced mitigating controls are applied as a bounded discount, not as an additive risk factor [17].

*Raw Score = 100 * (0.20*Exposure + 0.18*Privilege + 0.18*Exploitability + 0.14*BlastRadius + 0.10*AssetCriticality + 0.16*AttackPathRelevance + 0.04*ComplianceRelevance)*

*Final Score = round (Raw Score * (1 - 0.15*MitigationStrength), 1)*

Internet exposure has weight 0.20; privilege and exploitability each have 0.18; attack-path relevance 0.16; blast radius 0.14; asset criticality 0.10; and compliance relevance 0.04. Mitigation strength is none, weak, moderate, or strong, normalized to 0, 0.25, 0.5, or 0.75, so the maximum possible reduction is 11.25% of the raw score under the locked scale. Scores are grouped as Critical (80-100), Very High (65-79.9), High (50-64.9), Moderate (35-49.9), Low (20-34.9), and Minimal (below 20).

The specification was locked on 22 July 2026 with SHA-256
ad0b7d63fb19ad1b9aea205466d37e2d624d7cef7924

995822dbec05b2ac458d [17]. The lock prevents further post-result tuning, but it is not a preregistration: provisional expert-proxy labels already existed when the model was frozen. Operationally, CH-SPEC-v1.0 remains a shadow profile and does not yet replace production priority, ordering, SLAs, or automation decisions.

The frozen normalization is: internet exposure Yes/Unknown/No = 1.0/0.4/0.0; privilege root/administrative/privileged/service/standard/unknown/none = 1.0/1.0/0.8/0.4/0.3/0.4/0.0; exploitability critical/high/medium/low/unknown = 1.0/0.8/0.5/0.2/0.4; blast radius organization/account/region/VPC/service/resource/identity/unknown = 1.0/0.9/0.75/0.7/0.55/0.4/0.35/0.4; asset criticality critical/high/medium/low/unknown = 1.0/0.75/0.5/0.25/0.5; attack-path relevance critical/high/medium/low/none/unknown = 1.0/0.8/0.5/0.2/0.0/0.4; and compliance relevance high/medium/low/none/unknown = 1.0/0.6/0.3/0.0/0.4 [17].

8.5. Remediation Engine Design

The remediation catalogue distinguishes deterministic technical actions from controls that require business or architectural judgment. Fifteen of the 24 canonical controls have deterministic remediation templates; nine are context-dependent and provide governed manual guidance. Every catalogue entry is suggest-only by default and requires human approval. Cloud-changing execution is confined to allow-listed Fix Center actions that validate organization, account, region, resource identity, prerequisites, and action parameters, and record an auditable plan, approval, execution result, and validation state.

The optional LLM component is analyst-triggered and may explain the finding, summarize the evidence, or clarify why a deterministic rule recommends a step. It cannot change the canonical mapping, CH-SPEC score, issue decision, approval state, or executable action, and it receives no standing cloud credentials. This boundary follows empirical evidence that analysts use LLMs most effectively for sensemaking and communication [13] and responds to observed weaknesses in context-sensitive automated triage [15]. Validation is scan-backed. A provider API success confirms that a request was accepted, but it does not prove that the security condition is gone. ConfigHawk therefore waits for the latest completed scan after the remediation baseline and matches the original resource and finding family. The result may be validated, validated_partial, failed, or pending_scan. This evidence-first and management-approved workflow is consistent with the CERT-In audit guidelines' emphasis on verifiable findings, corrective action, and responsible approval [6].

IX. EXPERIMENTAL SETUP

The source dataset is one ConfigHawk AWS scan containing 100 findings across ten services: S3 (24), ECR (17), EventBridge (17), CloudTrail (10), VPC (9), IAM (7), AWS Config (6), Secrets Manager (5), GuardDuty (4), and RDS (1). Native severity comprises 6 Critical, 40 High, 34 Medium, and 20 Low findings. Two composite S3 risk-score rows (F-003 and F-004) were marked Review and excluded from the ranking comparison because they could duplicate risk already represented by underlying S3 findings. The resulting evaluation set contains 98 findings. The exclusion was not based on canonical-control coverage: 75 of the 100 source rows do not directly match the bounded 24-control catalogue, but they remain eligible for contextual scoring when sufficient finding semantics are available.

Each retained finding was annotated for internet exposure, privilege level, exploitability, blast radius, asset criticality, attack-path relevance, mitigating controls, and compliance relevance. The same research programme assigned a 1-6 expert-proxy risk score and a tied priority rank using documented rationales.

For rank-correlation analysis, native severity and CH-SPEC-v1.0 scores were each converted to descending tied ranks over the 98 retained findings; equal scores retained equal ranks. Spearman rho and Kendall tau-b were then computed against the expert-proxy priority rank. For exact Top-10 and Top-20 set comparisons, ties at a cutoff were resolved deterministically by Finding ID solely to keep the compared set size equal to k. Because the annotations, labels, and scoring design are not independent, the experiment measures development-set agreement rather than generalizable predictive accuracy.

X. RESULTS

Table 3 compares native severity and CH-SPEC-v1.0 against the provisional expert-proxy order for the 98 retained findings. Higher positive rank correlation indicates closer ordering agreement; top-k overlap measures how many findings selected by a method also appear in the corresponding expert-proxy priority set.

Table 3. Ranking performance against expert-proxy ground truth (n = 98 findings).

Metric	Severity-only	ConfigHawk CH-SPEC-v1.0	Difference
Spearman rho vs expert priority rank	0.405	0.885	+0.480
Kendall tau-b vs expert priority rank	0.331	0.777	+0.446
Top-10 overlap with expert top-10	6 / 10	9 / 10	+3
Top-20 overlap with expert top-20	8 / 20	15 / 20	+7

Source: Author's analysis of the locked CH-SPEC-v1.0 prioritization experiment workbook [17].

CH-SPEC-v1.0 achieved Spearman rho = 0.885 and Kendall tau-b = 0.777, compared with 0.405 and 0.331 for severity-only ranking. It matched 9 of the expert-proxy Top 10 and 15 of the Top 20, while severity matched 6 and 8 respectively. The observed differences are therefore +0.480 in Spearman correlation, +0.446 in Kendall tau-b, +3 Top-10 findings, and +7 Top-20 findings.

These descriptive results support H1 and do not support H0 on this development dataset. They do not constitute an independent statistical validation of the weighting scheme, because the reference labels were created within the same project and were available before the model was locked. A held-out environment or independent expert panel is required before making population-level or production-effectiveness claims.

XI. DISCUSSION

RQ1 - prioritization. The large gap in rank agreement indicates that contextual factors distinguish findings that severity groups together. Direct public exposure, privileged identities, broad impact, and attack-path participation moved findings toward the top, while partial safeguards reduced scores in a bounded manner. This is consistent with CVSS v4.0's instruction to supplement standardized severity with local threat, environmental, and organizational information [9] and with empirical evidence that scoring systems produce different and sometimes conflicting signals [11]. The comparison with Vulnerability Management Chaining is complementary: that work integrates CVE-oriented signals [12], whereas CH-SPEC models configuration and resource context that may exist without a CVE.

RQ2 - compliance mapping. Atomic obligations allowed all 24 controls to be assessed against all four frameworks without collapsing them into one compliance flag.

The 16 CERT-In no_relationship outcomes are particularly informative: the model does not invent a preventive AWS requirement where the Directions focus on incident and operational duties. Conversely, ISMS-P and ISO mappings show that a technical control can partially satisfy access, logging, encryption, or cloud-security objectives while leaving policy, approval, scope, review, and evidence obligations visible. This supports defensible reuse of one control catalogue across jurisdictions without claiming equivalence among the frameworks.

RQ3 - remediation governance. Fifteen controls admit deterministic templates, while nine require contextual manual decisions such as privilege justification, role retirement, logging architecture, or exception handling. The appropriate division is therefore not deterministic versus autonomous LLM remediation. It is deterministic execution where the desired state can be validated, governed manual treatment where business context is required, and optional LLM explanation across both. This preserves analyst and approver authority and aligns with studies that position LLMs as support tools rather than high-stakes decision owners [13]-[15].

XII. LIMITATIONS AND THREATS TO VALIDITY

- Single-reviewer mapping validation: all 96 mappings were accepted by one reviewer. A second independent reviewer, inter-rater agreement statistic, and adjudication record are required for stronger external validity.
- Non-independent expert proxy: the prioritization labels were produced within the same research programme that designed CH-SPEC. Shared assumptions can inflate agreement between the model and the reference order.
- Timing of the model lock: provisional labels existed before CH-SPEC-v1.0 was frozen. The lock prevents subsequent edits but does not remove development-set leakage or substitute for preregistration.
- Single scan and account: the experiment uses one 100-finding AWS scan. Service mix, configuration patterns, and organizational context may differ in multi-account, larger, or multi-cloud environments.
- Composite-row exclusion: two derived S3 risk-score findings were removed to avoid potential double counting. A future protocol should define duplicate and aggregate-finding treatment before data collection.
- Bounded catalogue coverage: only 25 of the 100 source findings directly matched the current 24 controls. CH-SPEC can score additional findings, but validated compliance relationships are limited to catalogue-covered controls.
- Framework and jurisdiction scope: the mapping covers four sources and does not include GDPR, PCI DSS, sector-specific obligations, or all national requirements. Mapping applicability still depends on the assessed entity and certification or audit scope.
- Certification and legal interpretation: a technical mapping is an informative relationship, not legal advice, an audit opinion, or proof of ISO/IEC 27001 or ISMS-P conformity. Short ISO control paraphrases are used rather than reproducing licensed text.
- Remediation and LLM evaluation: the architecture, template coverage, and trust boundaries are documented, but no controlled study measures remediation success rate, analyst time, explanation quality, or human over-reliance on LLM output.
- Operational deployment status: CH-SPEC-v1.0 remains shadow-only and does not control production ordering, SLAs, or automation. Promotion requires independent or holdout evaluation and the documented acceptance gates.

Source currency: the Korean reform notice describes changes expected from 2027 [18]. Future work must re-check the enacted legal text and updated ISMS-P criteria rather than treating the announced design as current certification requirements.

XIII. CONCLUSION AND FUTURE WORK

This paper presented a bounded and auditable method for connecting AWS configuration evidence to multi-framework obligations, ranking findings with context, and governing remediation. The 24-control catalogue produced 96 reviewer-accepted mappings while preserving applicability and uncovered obligations. On the 98-finding development set, CH-SPEC-v1.0 aligned more closely with the provisional expert-proxy order than severity alone. The remediation design further separates deterministic, approval-gated execution from context-dependent manual decisions and confines LLM use to optional explanation. The immediate research priorities are an independently ranked holdout scan, a second compliance reviewer with adjudication, expansion of catalogue coverage, cross-account and multi-cloud evaluation, and a controlled human study of remediation and explanation quality. Until those steps are complete, the results support the feasibility of the framework but not a claim of externally validated production superiority.

REFERENCES

[1] FNU Jimmy, “Cloud security posture management: Tools and techniques,” *Journal of Knowledge Learning and Science Technology*, vol. 2, no. 3, pp. 614–622, Dec. 2023, doi: 10.60087/jklst.vol2.n3. p622.

[2] S. Moss, “Evaluating the effectiveness of CSPM tools in detecting cloud misconfigurations,” SSRN Electronic Journal, 2026, doi: 10.2139/ssrn.6244038. Preprint.

[3] N. Keller, S. Quinn, K. Scarfone, M. C. Smith, and V. Johnson, National Online Informative

References (OLIR) Program: Overview, Benefits, and Use, NIST IR 8278 Rev. 1, Feb. 2024, doi: 10.6028/NIST.IR.8278r1.

[4] National Institute of Standards and Technology, The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29, Feb. 26, 2024, doi: 10.6028/NIST.CSWP.29.

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

[6] Indian Computer Emergency Response Team, Comprehensive Cyber Security Audit Policy Guidelines, version 1.0, Jul. 25, 2025.

[7] Ministry of Science and ICT, Personal Information Protection Commission, and Korea Internet & Security Agency, ISMS-P Certification Criteria Guide (2023.11) and ISMS-P Detailed Inspection Items (2023.10.31), 2023.

[8] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection—Information security management systems—Requirements, 3rd ed., International Organization for Standardization and International Electrotechnical Commission, 2022.

[9] S. Devikar, M. Hinge, and S. S. Shevkari, "Human vs. machine: A review of AI's impact on language learning interaction or replacing human interaction," International Journal for Multidisciplinary Research, vol. 7, no. 3, 2025, doi: 10.36948/ijfmr. 2025.v07i03.49459.

[10] FIRST.Org, Inc., "Exploit Prediction Scoring System (EPSS)." [Online]. Available: EPSS. Accessed: Aug. 10, 2026.

[11] V. Koscinski, M. Nelson, A. Okutan, R. Falso, and M. Mirakhorli, "Conflicting scores, confusing signals: An empirical study of vulnerability scoring systems," arXiv:2508.13644, 2025, doi: 10.48550/arXiv.2508.13644.

[12] N. Shimizu and M. Hashimoto, "Vulnerability management chaining: An integrated framework for efficient cybersecurity risk prioritization," arXiv:2506.01220, 2025, doi: 10.48550/arXiv.2506.01220.

[13] R. Singh, S. Tariq, F. Jalalvand, M. B. Chhetri, S. Nepal, C. Paris, and M. Lochner, "LLMs in the SOC: An empirical study of human-AI collaboration in Security Operations Centres," arXiv:2508.18947, 2025, doi: 10.48550/arXiv.2508.18947.

[14] P. Shigvan, S. Shevkari, V. Auti, and G. Kumar, "Detecting and mitigating insider threats with artificial intelligence," International Journal of Innovative Inventions in Social Science and Humanities, 2025.

[15] S. Srinivas, B. Kirk, J. Zendejas, M. Espino, M. Boskovich, A. Bari, K. Dajani, and N. Alzahrani, "AI-augmented SOC: A survey of LLMs and agents for security automation," Journal of Cybersecurity and Privacy, vol. 5, no. 4, art. 95, 2025, doi: 10.3390/jcp5040095.

[16] O. Al Haddad, M. Ikram, E. Ahmed, and Y. Lee, "Prompting the priorities: A first look at evaluating LLMs for vulnerability triage and prioritization," arXiv:2510.18508, 2025, doi: 10.48550/arXiv.2510.18508.

[17] ConfigHawk, ConfigHawk Canonical Cloud Security Control Catalogue, version 1.2.0-reviewer-validated, Jul. 18, 2026. Internal research artifact.

[18] ConfigHawk, ConfigHawk Specificity Score CH-SPEC-v1.0 and confighawk_prioritization_experiment_dataset_v4_locked_specificity_score.xlsx, locked Jul. 22, 2026. Internal research artifacts.

[19] Yulchon LLC, "ISMS/ISMS-P Reform in Korea: Key Changes and Implications for Compliance," *Yulchon Legal Update*, Apr. 2026.