computers

MDPI

Article

Scaling National Cybersecurity Assessment with LLM-Assisted Coding: A Validated Comparative Documentary Baseline of Four Andean Countries (2020–2026)

Luis J. Serpa-Andrade [1,*], Cristhian H. Flores-Urgilés [2,3], Mario Ron-Egas [3,4], Fernanda M. Narváez-Ochoa [2], Cristina M. Flores-Urgilés [2] and Dario Piccirilli [3]

[1] Grupo de Investigación GIHEA, Universidad Politécnica Salesiana, Cuenca 010105, Ecuador
[2] Unidad Académica de Informática, Ciencias de la Computación e Innovación Tecnológica, Universidad Católica de Cuenca, Cuenca 010107, Ecuador; chfloresu@ucacue.edu.ec (C.H.F.-U.); fernanda.narvaez@ucacue.edu.ec (F.M.N.-O.); cmfloresu@ucacue.edu.ec (C.M.F.-U.)
[3] Facultad de Informática, Universidad Nacional de La Plata, La Plata 1900, Argentina; mbron@espe.edu.ec (M.R.-E.); dpiccirilli@unlp.edu.ar (D.P.)
[4] Departamento de Ciencias de la Computación, Universidad de las Fuerzas Armadas ESPE, Sangolquí 171103, Ecuador
* Correspondence: lserpa@ups.edu.ec

Abstract

National cybersecurity assessment through documentary coding faces a scalability bottleneck: manually coding institutional documents across countries consumes weeks of expert effort, confining comparative studies to single nations. This paper makes a dual contribution. Methodologically, it validates whether a Large Language Model (LLM) can reproduce a structured documentary coding rubric—the Evidence Density Index of the DAM–NC methodology—enabling scalable comparative assessment. Substantively, it applies the validated approach to produce the first comparative cybersecurity documentary baseline of four Andean countries (Ecuador, Colombia, Peru, Chile) from a corpus of 55 verified institutional documents. The validation is three-layered: LLM coding was compared against an independent rule-based reference, replicated across three independent LLMs, and anchored against a human gold standard built by two independent expert coders with excellent inter-coder reliability. The LLM tracked expert human judgment more closely than the rule-based reference, and all substantive regional conclusions proved robust across coding methods and models. The baseline reveals a consistent Govern-to-Recover documentary gradient (Kendall's $W = 0.893$) and a regional gap in documented recovery capability that survives multi-label coding: no Andean country records post-incident recovery as an institutionalized national instrument, either as a primary or a secondary document function. We interpret this as a documentary gap—a potential indicator of, but not equivalent to, weak operational recovery capability. LLM-assisted coding, embedded in a human-in-the-loop calibration protocol, is a viable accelerator for comparative cybersecurity assessment.

Keywords: LLM-assisted coding; comparative cybersecurity; evidence density index; NIST CSF 2.0; Andean region; inter-method robustness; human-in-the-loop

Check for updates

Academic Editor: Phivos Mylonas

Received: 6 August 2026
Revised: 17 September 2026
Accepted: 2 October 2026
Published: 8 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Evidence-based assessment of national cybersecurity capabilities increasingly relies on the systematic coding of institutional documents—strategies, laws, regulations, incident reports, and international indices [1–3]. As cybersecurity has evolved from a technical concern

into a multidimensional governance challenge [4,5], the documentary record has become a primary, externally verifiable source of evidence about national capability. Documentary maturity methodologies such as the Design–Action Maturity for National Cybersecurity (DAM–NC) operationalize this through trained human coders who assign structured scores aggregated into composite indices such as the Evidence Density Index (EDI) [6], drawing on content-analysis [7] and grey-literature systematic-review traditions [8–10].

This manual coding is simultaneously the backbone and the bottleneck of documentary assessment. Coding dozens of documents across multiple coders with inter-rater reconciliation consumes weeks of expert effort, which has confined documentary cybersecurity assessment largely to single countries [6,11]. Genuine comparative assessment—across several countries with a single instrument—multiplies this cost and has consequently remained rare in the Latin American region [12–14].

Two developments make this bottleneck addressable. First, Large Language Models (LLMs) have emerged as candidate automated annotators, reportedly matching or exceeding crowd workers and sometimes experts on text-annotation tasks [15–18]. If an LLM could reliably reproduce structured documentary coding, the scalability barrier would fall. Second, the NIST Cybersecurity Framework (CSF) 2.0 [19]—the de facto international standard for organizing cybersecurity capability into six functions (Govern, Identify, Protect, Detect, Respond, Recover)—provides a jurisdiction-neutral security scaffold onto which documents from any country can be mapped. Because each CSF function corresponds to an operational security capability rather than a reporting category, mapping national documents onto the CSF yields a security-meaningful profile: gaps in the profile mark capabilities for which no verifiable national documentary evidence exists, and therefore warrant scrutiny of the corresponding defensive posture. Yet LLM annotation is not a panacea: models hallucinate [20], are prompt-sensitive [21], embed biases [22], and—most importantly—their outputs require validation before being trusted as data [23–25].

The stakes extend beyond the academic. Under NIST CSF 2.0, the Recover function encompasses the capabilities a nation relies on to restore services after a major cyber incident—continuity plans, recovery exercises, post-incident communication. A documented void in Recover therefore warrants attention: it may indicate that, when a disruptive incident strikes critical services, no institutionalized national recovery playbook is verifiable in the public documentary record—while recognizing that documentary absence is not, by itself, proof of operational absence, a distinction we return to in the limitations. Identifying where such voids exist—and whether they are idiosyncratic to one country or structural across a region—is a core cybersecurity-governance problem that documentary assessment is uniquely positioned to answer at scale.

This paper unifies these two developments in a single design with a dual contribution. We first validate LLM-assisted coding against an independent reference, then apply it to produce a comparative documentary baseline of four Andean countries, and—critically—test whether the substantive regional findings are robust across coding methods. The research questions are:

> (RQ1) To what extent can an LLM reproduce a structured multi-dimensional documentary coding rubric for national cybersecurity? (RQ2) How is documentary evidence of cybersecurity capability distributed across functions and countries in the Andean region? (RQ3) Are the regional findings robust to the choice of coding method?

This study uses the EDI rubric and the Ecuadorian sub-corpus from the DAM–NC program [1,6], cited accordingly. The LLM validation design, the three additional national corpora (Colombia, Peru, Chile), the comparative analysis, and the inter-method robustness test are fully original.

2. Background and Related Work

2.1. LLMs as Automated Annotators

The use of LLMs for annotation accelerated sharply after 2023. Gilardi et al. [15] reported ChatGPT (gpt-3.5-turbo) outperforming crowd workers; Törnberg [16] found GPT-4 matching experts on political text; Ziems et al. [17] surveyed the potential for computational social science; Rathje et al. [18] extended results to multilingual analysis. A validation-first counter-current is equally important: Pangakis et al. [23] showed automated annotation quality varies by task and demands validation; Reiss [24] found reliability sensitive to prompts; Kuzman et al. [26] and Thapa et al. [27] questioned whether manual annotation is obsolescing while Huang et al. [28] probed LLM limits on nuanced phenomena; Ollion et al. [25] warned against uncritical use of proprietary models for specialized, non-English text. Broader reflections caution that AI tools can create illusions of understanding if unvalidated [29,30]. Our design follows this validation-first tradition: we never treat LLM output as ground truth without comparison.

2.2. Comparative Cybersecurity Assessment

Cross-country cybersecurity comparison is dominated by composite indices—GCI [3] and NCSI [31]—and maturity models such as the Oxford CMM [32]. These produce aggregate rankings that compress sectoral and functional heterogeneity into a single figure. Luiijf et al. [33], comparing nineteen national strategies, found that they share common elements while differing markedly in definitions, scope, and implementation focus, a divergence invisible to aggregate scoring. The EDI offers a finer-grained, cell-level alternative grounded in Design Science [34], and belongs to a broader family of cybersecurity maturity instruments [35–38]; its manual application has limited prior comparative use—precisely the bottleneck LLM coding could relieve. Comparative research designs that hold the instrument constant across cases follow the most-similar-systems logic of comparative method [39].

2.3. The EDI Coding Rubric and Measuring Agreement

The verified evidence summaries were prepared in a single manual pass per document: a researcher read each source and recorded, in a structured template, the document's primary and secondary CSF functions together with verbatim or lightly condensed excerpts substantiating each of the three EDI dimensions (quality/recency, control criticality, coverage). This manual step—the only labor-intensive stage of the workflow—required on the order of 15–25 min per document; all subsequent coding (rule-based and LLM) operates automatically on these summaries, and the independent human gold standard described below instead works from the complete source texts, so the summary-preparation effort does not affect the validity of the reported agreement. This manual-then-automated division is what makes the approach scale: summary preparation is linear and one-time, whereas re-coding the entire corpus under a new rubric, a new model, or a new density setting is essentially free.

The EDI [6] scores each document on three anchored 0–10 dimensions—quality/recency (C_q), control criticality (C_c), coverage (C_o)—combined as

$$\mathrm{EDI} = \left(0.4\,C_q + 0.3\,C_c + 0.3\,C_o\right) \times 10 \times \delta(n), \tag{1}$$

where $\delta(n) = 1 - e^{-0.45n}$ is a documentary density factor preventing a single source from inflating a cell. Bands: scarce (0–49), moderate (50–69), abundant (70–100). Validating reproduction of such an ordinal-to-continuous scheme calls for correlation [40], intraclass [41], and Bland–Altman [42] analyses rather than a single categorical κ [43–46].

2.4. Cybersecurity Governance in the Four Andean Countries

The four countries present distinct trajectories. Chile enacted the region's most comprehensive framework, Law 21.663 (2024), creating the ANCI and a National CSIRT with binding authority, mandatory 3 h notification, and sanctions up to 20,000 UTM [47–51]. Colombia built its posture incrementally through CONPES documents (3854/2016, 3995/2020) and a 2025–2027 strategy, with an active ColCERT and sectoral financial-resilience circulars [52–57]. Ecuador consolidated a 2022–2025 strategy, strengthened EcuCERT, and enacted a 2026 Organic Cybersecurity Law, recording the largest GCI gain worldwide 2020–2024 [58–62]. Peru established its CNSD by emergency decree (2020) and published a 2026–2028 strategy in 2025, but remains comparatively nascent without a comprehensive organic law [63–66]. Budapest Convention accession varies: Chile (2017), Colombia (2020), Peru (2024), Ecuador pending [67]. ISO/IEC 27001 [68] and ITU strategy guidance [69] recur as documentary anchors across all four.

3. Materials and Methods

3.1. Corpus

The corpus comprises 55 verified institutional documents (2009–2026): 12 Ecuador, 13 Colombia, 10 Peru, 12 Chile, and 8 multilateral references (GCI, NCSI, OAS/IDB, ENISA, CEPAL, Budapest) [70–72]. The four Andean countries constitute a deliberate regional case-study design rather than a sample intended to support inferential generalization: they share civil-law tradition, comparable institutional architectures, and parallel CSIRT trajectories, which makes a single coding instrument meaningfully portable across them; the comparative analysis is accordingly framed as structured descriptive comparison. Construction followed a PRISMA-guided grey-literature protocol [8–10] with uniform inclusion criteria. Each document was assigned to its primary (and where applicable secondary) NIST CSF 2.0 function. The corpus is predominantly Spanish-language, providing a meaningful test of LLM coding in a non-English specialized domain. We acknowledge that the smaller Peruvian sub-corpus (10 documents) could in principle depress Peru's EDI through the density factor independently of its actual documentary capability; we return to this density-versus-search-exhaustiveness confound in the limitations.

3.2. Two Independent Codings

The corpus was coded by two independent methods. The rule-based coding derived C_q, C_c, C_o deterministically from verifiable attributes (type, recency, authority, function, scope) via fixed mappings calibrated to rubric anchors—transparent and fully reproducible. The LLM coding used Claude Opus 4.8 (Anthropic, San Francisco, CA, USA) with a structured prompt presenting the anchored rubric, document metadata, and a verified evidence summary, instructed to output three integers plus a one-line justification; the prompt followed established practice (explicit anchors, structured output, justification to elicit reasoning) [73–76]. To support reproducibility, coding was performed at low temperature to minimize stochastic variation, with the model version and the exact prompt archived verbatim in the public repository (Supplementary Material); all LLM codings reported here were executed in June 2026 as single runs per model. We note that systematic prompt- and temperature-sensitivity analysis remains future work [24]. The verified evidence summaries supplied to the LLM were compiled by the research team from the primary sources. To provide an independent human reference, two expert coders additionally coded a stratified 47% sample by hand, reading each full source document rather than the evidence summaries used by the automated methods (Section 4.7); this ensures the human gold standard is methodologically independent of the shared summary inputs.

Their agreement and the anchoring of both automated methods against this gold standard are reported in Results.

3.3. Analysis

Methodological axis (RQ1). Agreement between LLM and rule-based codings was assessed at the dimension level (Pearson r, Spearman ρ, MAE, bias, within-one-point agreement) and composite level (EDI r, MAE, bias, Bland–Altman limits) [40,42].

Substantive axis (RQ2). Per-cell EDI was computed via Equation (1) aggregating national documents (full weight) and multilateral documents (half weight) with the density factor on the effective count.

Robustness axis (RQ3). Consistent with the four-country case-study design, the between-country Kruskal–Wallis statistic [77] is reported as a descriptive index of dispersion rather than as an inferential test of convergence; the Friedman test across functions [78] operates on within-country profiles and retains conventional interpretation. Both were run independently under both coding methods, and their qualitative conclusions compared. Convergence of conclusions across methods constitutes inter-method robustness. All statistical analyses were performed in Python 3 (Python Software Foundation, Wilmington, DE, USA) with the NumPy and SciPy libraries; the analysis script is provided in the Supplementary Material.

4. Results

4.1. RQ1: LLM Coding Agreement

One document (a multilateral reference lacking a rule-based coding for its primary CSF function) could not be matched across both methods and was excluded from this document-level comparison, yielding $n = 54$ of the 55 corpus documents; it is retained in all corpus-level and aggregated analyses.

Table 1 reports dimension-level agreement. All three dimensions correlated strongly and significantly; coverage achieved the highest rank-order agreement ($\rho = 0.833$), criticality the lowest ($r = 0.644$), consistent with its requiring more contextual judgment. All exhibited a small negative bias.

At the composite EDI level (Figure 1), the codings correlated strongly ($r = 0.729$, $p < 0.001$) but with a systematic -7.9-point bias (MAE 8.4; Bland–Altman limits $[-20.0, +4.1]$; exact band agreement 59%). Inspection of LLM justifications attributed the conservative bias to steeper age penalties, stricter criticality for contextual sources (indices, statistics), and sharper discrimination between binding and declarative instruments. The LLM thus reliably orders documents but applies a stricter absolute scale. Figure 2 summarizes the dimension-level concordance.

Table 1. Dimension-level agreement: LLM versus rule-based coding ($n = 54$).

Dimension	Pearson r	Spearman ρ	MAE	Bias	Within ± 1 pt
C_q (quality/recency)	0.781	0.764	0.86	-0.77	69%
C_c (criticality)	0.644	0.686	1.03	-0.79	65%
C_o (coverage)	0.769	0.833	0.85	-0.83	74%

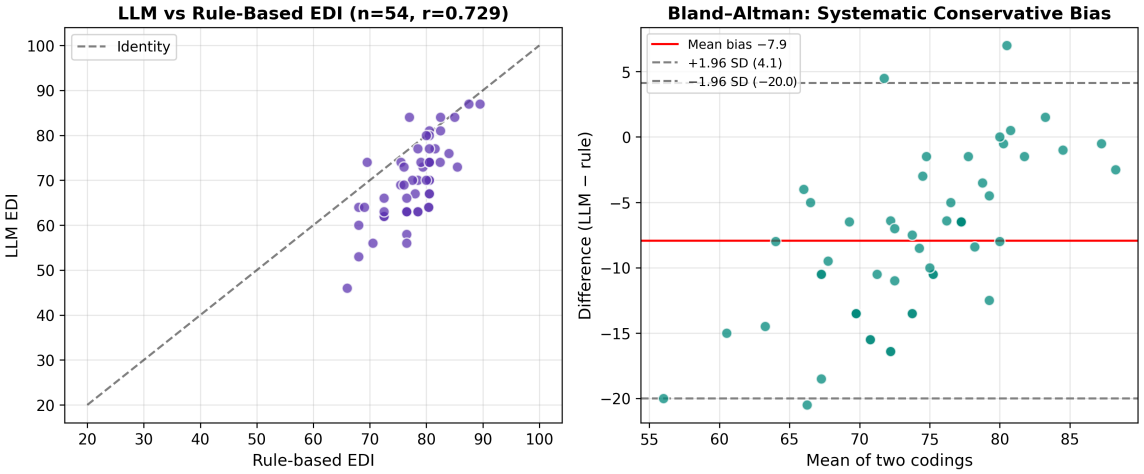


Figure 1. Document-level EDI agreement. (**Left**): LLM versus reference scatter ($r = 0.729$); points fall consistently below the identity line. (**Right**): Bland–Altman plot showing −7.9-point bias, limits $[-20.0, +4.1]$.

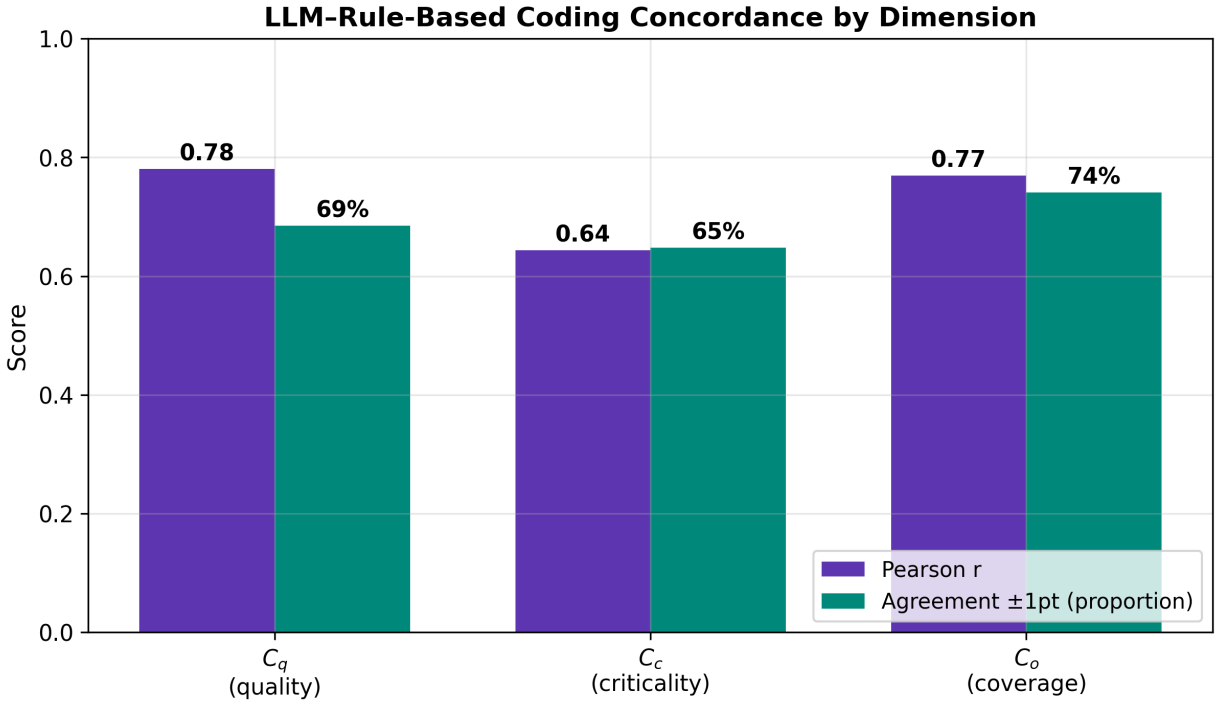


Figure 2. LLM–rule-based coding concordance by dimension: Pearson correlation and within-one-point agreement.

4.2. From Document Scores to the Cell Index: EDI Aggregation

Each document d receives, for the CSF function(s) it addresses, a dimension triple (C_q, C_c, C_o) on a 0–10 scale, combined into a per-document, per-function score $\text{EDI}_d = 10\,(0.4\,C_q + 0.3\,C_c + 0.3\,C_o)$. A document contributes to its primary function with full weight ($w = 1$) and, where it substantively addresses a secondary function, to that function with reduced weight ($w = 0.5$); documents addressing a single function contribute only there. Multilateral references (GCI, NCSI, OAS/IDB, ENISA, CEPAL, Budapest), which are shared across countries rather than national, enter each country's cells with weight $w = 0.5$ to avoid overstating national documentary density on the strength of external sources; robustness to their inclusion is verified separately in RQ3 by re-running the analysis on national documents only. The cell value reported in Table 2 is the weighted

mean of the contributing document scores for that country–function pair, scaled by the density factor evaluated at the effective (weighted) document count:

$$\mathrm{EDI}_{c,f} = \delta\left(n^{\mathrm{eff}}_{c,f}\right) \frac{\sum_{d \in \mathcal{D}_{c,f}} w_d \, \mathrm{EDI}_{d,f}}{n^{\mathrm{eff}}_{c,f}}, \qquad n^{\mathrm{eff}}_{c,f} = \sum_{d \in \mathcal{D}_{c,f}} w_d,$$

where $\mathcal{D}_{c,f}$ is the set of documents of country c contributing to function f. A cell with no contributing document takes the value 0 by construction, denoting absence of documentary evidence rather than a low-quality score; this is the case for every Recover cell (see below). The density factor $\delta(n) = 1 - e^{-0.45n}$ of Equation (1) is thus applied once per cell, to the effective count $n^{\mathrm{eff}}_{c,f}$, so that multilateral and secondary contributions increase density with half weight; for example, a cell with two national documents and one multilateral reference has $n^{\mathrm{eff}} = 2.5$ and $\delta = 0.68$. The cell-level document counts (national and multilateral) and density factors are reported in the Supplementary File edi_by_country_function.csv, and the influence of the coefficient is examined in the sensitivity analysis.

Table 2. Evidence Density Index by country and CSF 2.0 function (rule-based coding).

Country	Govern	Identify	Protect	Detect	Respond	Recover
Ecuador	78.4	71.8	58.3	60.8	53.1	0.0
Colombia	77.4	69.8	65.2	50.6	57.2	0.0
Peru	76.5	62.0	42.7	50.1	60.2	0.0
Chile	78.0	67.9	58.3	54.2	63.2	0.0
Mean	77.6	67.9	56.1	53.9	58.4	0.0

4.3. RQ2: The Comparative Andean Baseline

Figure 3 presents the EDI heatmap (rule-based coding) across the four countries and six CSF functions. All four countries show abundant Govern documentation (mean 77.6) declining through operational functions to an absence of Recover-primary documentation. Figure 4 plots the convergent "Govern-to-Recover" gradient.

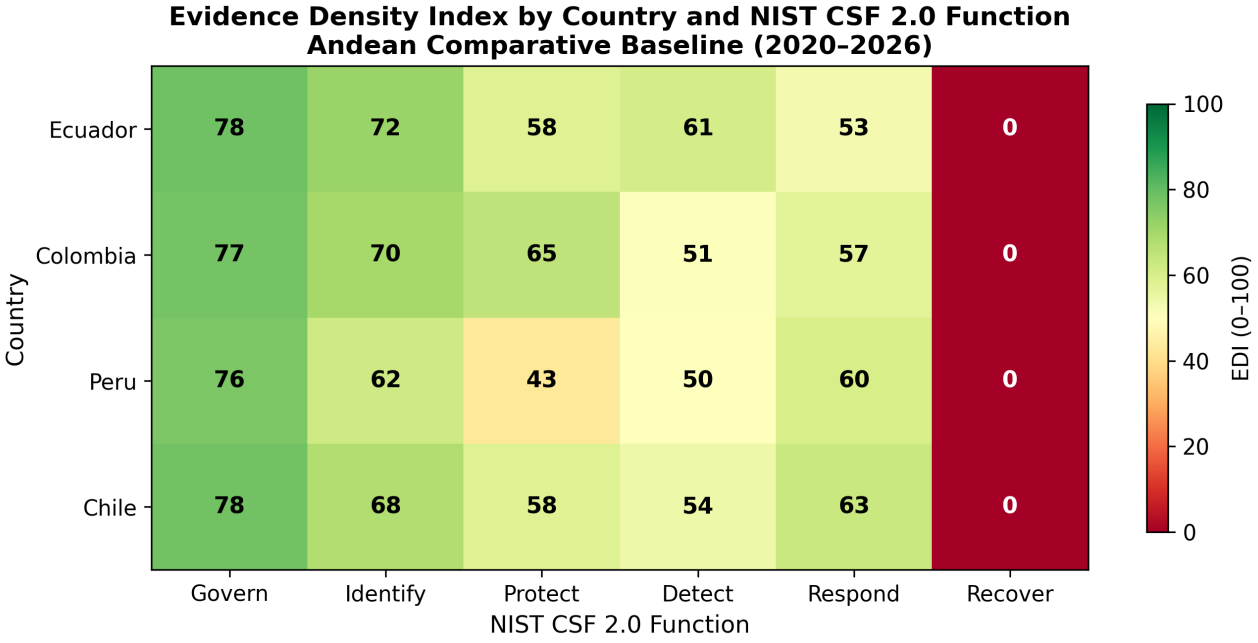


Figure 3. Evidence Density Index by country and NIST CSF 2.0 function (rule-based coding). Red (scarce, 0–49), yellow (moderate, 50–69), green (abundant, 70–100).

Table 2 reports the values. The inter-country spread is narrowest in Govern (1.9 points) and widest in Protect (22.5 points): governance documentation is uniformly abundant, while protective documentation varies with each country's sectoral regulatory maturity. Chile leads Respond (operational CSIRT), Colombia leads Protect (financial-sector circulars), Ecuador leads Identify and Detect, and Peru trails in Protect (scarce band).

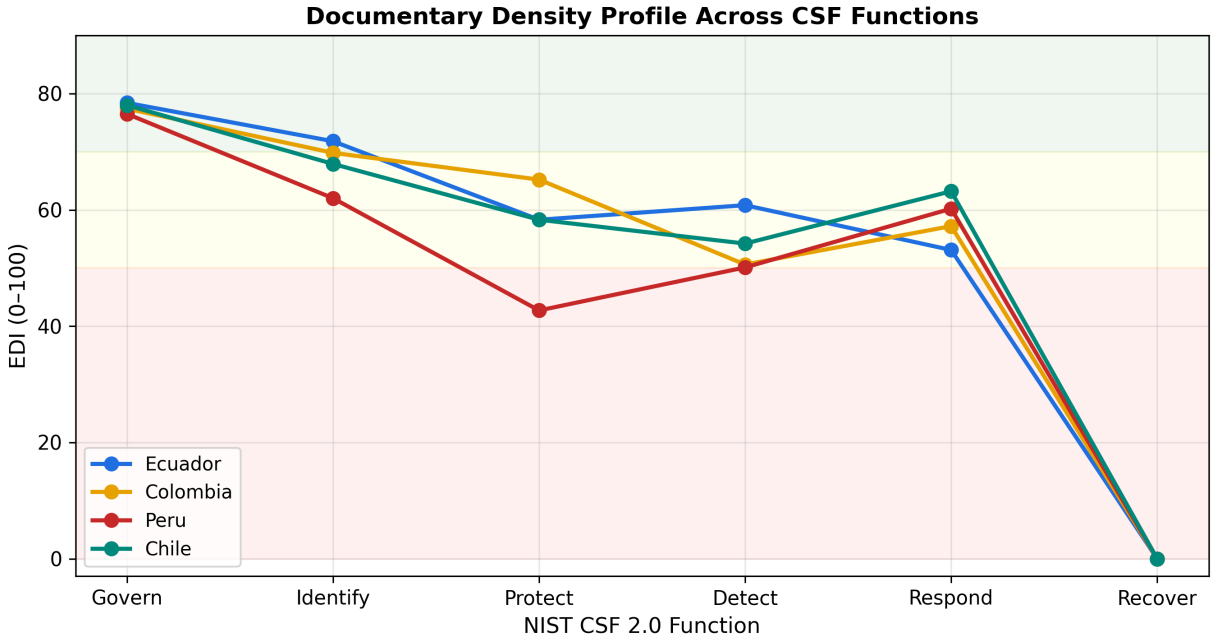


Figure 4. Documentary density profile across CSF functions. The convergent Govern-to-Recover gradient appears in all four countries. Background shading marks the EDI bands: red, scarce (0–49); yellow, moderate (50–69); green, abundant (70–100).

4.4. The Recover Cell Under Multi-Label Coding

A reviewer reasonably asked whether the zero Recover values in Table 2 are an artifact of assigning each document only to its primary function, with recovery content possibly present as secondary material. We tested this directly. Our coding scheme records both a primary and a secondary CSF function for every document, and we re-computed the baseline under a multi-label rule in which each secondary assignment also contributes (at $w = 0.5$) to its function's cell.

The result is unambiguous: the Recover column remains exactly zero. Across the 55-document corpus, Recover appears neither as a primary function (0 documents) nor as a secondary function (0 documents); secondary assignments distribute over Identify (14), Govern (7), Detect (6), Protect (6) and Respond (6), with 16 documents having no secondary function. The Govern-to-Recover gradient, the between-country similarity, and every other substantive conclusion are therefore identical under single-label and multi-label coding. The multi-label analysis does not weaken the finding; it strengthens it, because the void survives the most permissive coding rule our scheme permits.

This does not contradict the fact—which we report in the verification search below—that recovery provisions exist inside some Andean instruments, for example Ecuador's 2022 National Strategy and 2026 Cybersecurity Law, and Chile's Framework Law 21.663 (2024), which mandates continuity and recovery plans for operators of vital services, as well as entity-level operational-continuity plans issued under Peru's all-hazards disaster-risk framework. The distinction is one of unit of analysis: those are clauses within documents whose institutional purpose is governance, not documents that constitute a recovery capability. Passage-level coding would surface such clauses, but it would answer a different

question—whether the word appears—rather than the one this baseline asks, namely whether a country has institutionalized recovery as a documented national capability. We therefore state the finding precisely: no Andean country documents recovery as an institutionalized, national-level capability with a dedicated instrument; recovery survives only as subordinate clauses within governance documents or as mandates delegated to private operators. Extending the rubric to passage-level multi-label coding is a natural next step and is noted in the limitations.

4.5. RQ3: Inter-Method Robustness

Two features of the design bear directly on statistical inference and shape how we report these tests. First, country–function observations are not fully independent: the six functional cells within a country share documents and institutional context. Second, with only four countries the asymptotic χ^2 approximations underlying Kruskal–Wallis and Friedman are unreliable and statistical power is very low. We therefore (i) treat the between-country statistic as descriptive rather than inferential, (ii) recompute both tests using permutation distributions (10^5 resamples; country relabelings for the between-country test, within-country function permutations for Friedman), and (iii) report effect sizes alongside each test. These permutation analyses use the full six-function matrix of Table 2, including the all-zero Recover column, whereas the asymptotic tests in Tables 3 and 4 were computed on the five functions with documentary evidence (Govern–Respond); this explains the different p-values (e.g., Friedman $p = 0.003$ with six functions versus $p = 0.011$ with five). Because Recover ranks last in every country by construction, the six-function Friedman statistic and Kendall's W partly reflect the void itself; restricted to the five evidenced functions, concordance remains high ($W = 0.81$) and the gradient remains significant ($p = 0.011$).

The permutation results agree closely with the asymptotic ones. Between countries, Kruskal–Wallis gives $H = 0.71$ with an asymptotic $p = 0.870$ and a permutation $p = 0.83$; the coefficient of variation across country means is 4.8%, indicating tight clustering. We stress that this non-significance is not evidence of homogeneity—with four countries the test has almost no power—which is precisely why we report the dispersion measure and frame between-country similarity as a descriptive observation throughout. Across functions, Friedman gives $\chi^2 = 17.86$ with an asymptotic $p = 0.003$ and a permutation $p < 0.001$, and Kendall's coefficient of concordance is $W = 0.893$, indicating that the four countries order the six CSF functions almost identically. The functional gradient is thus supported both by a test that remains valid under permutation and by a large effect size, whereas the between-country claim rests deliberately on description rather than inference.

The decisive test is whether the substantive conclusions survive the choice of coding method. Table 3 reports the comparative tests run independently under both codings. Both methods agree on every qualitative conclusion: no significant between-country difference, a dominant functional gradient, and a documentary Recover void. The Friedman result is significant under rule-based coding ($p = 0.011$); under LLM coding it falls just short of the conventional threshold ($p = 0.061$, i.e., not statistically significant), although the rank order of functions is identical under both methods. We therefore describe the functional gradient as consistent in rank order across methods rather than claiming preserved statistical significance.

Table 3. Inter-method robustness: comparative tests under both coding methods.

Test	Rule-Based	LLM	Consistent?
Kruskal–Wallis (countries)	$p = 0.749$	$p = 0.536$	Yes: both n.s.
Friedman (functions)	$p = 0.011$	$p = 0.061$	Rank order yes; significance differs
Recover void	0 docs	0 docs	Yes: void in both
Weakest country	Peru	Peru	Yes

4.5.1. A Note on Statistical Power

With only four countries, the Kruskal–Wallis test has very low power, so a non-significant result cannot be interpreted as positive evidence of convergence. We therefore treat between-country similarity as a descriptive observation: the country mean EDI values span a narrow range (58.3–64.5), with a between-country standard deviation of 3.0 points and a coefficient of variation of only 4.8%. This low dispersion is the empirical basis for the convergence observation; the non-significant test is consistent with it but does not, on its own, establish it.

4.5.2. Sensitivity to the Density Coefficient

Because the density factor $\delta(n) = 1 - e^{-\kappa n}$ scales all EDI values, we tested the sensitivity of the conclusions to the coefficient κ. Table 4 varies κ from 0.30 to 0.60. While the absolute mean EDI rises with κ (as expected), the qualitative conclusions are invariant: the between-country test remains non-significant ($p \approx 0.74$–0.80) and the functional gradient remains significant ($p = 0.011$) throughout, and the Recover void is mechanically robust (zero documents yield zero EDI for any κ).

Table 4. Sensitivity of conclusions to the density coefficient κ.

κ	Kruskal p	Friedman p	Mean EDI	Recover
0.30	0.798	0.011	54.2	0
0.40	0.749	0.011	60.4	0
0.45	0.749	0.011	62.8	0
0.50	0.749	0.011	64.7	0
0.60	0.740	0.011	67.8	0

4.5.3. Excluding Shared Multilateral Documents

Because the eight multilateral documents are shared across all four countries (each contributing with weight 0.5), they could artificially inflate between-country homogeneity. We therefore recomputed the analysis using only national documents. The between-country similarity persists (Kruskal–Wallis $p = 0.315$, still non-significant) and the country ordering is unchanged (Peru remains lowest: 51.8 versus 59.9–61.1 for the others). The convergence observation is thus not an artifact of the shared multilateral corpus, though the absolute EDI values are naturally lower without the corroborating international evidence.

Figure 5 shows cell-level confidence (rule-based), with high-confidence cells concentrated in Govern and operational functions carrying medium-to-low confidence.

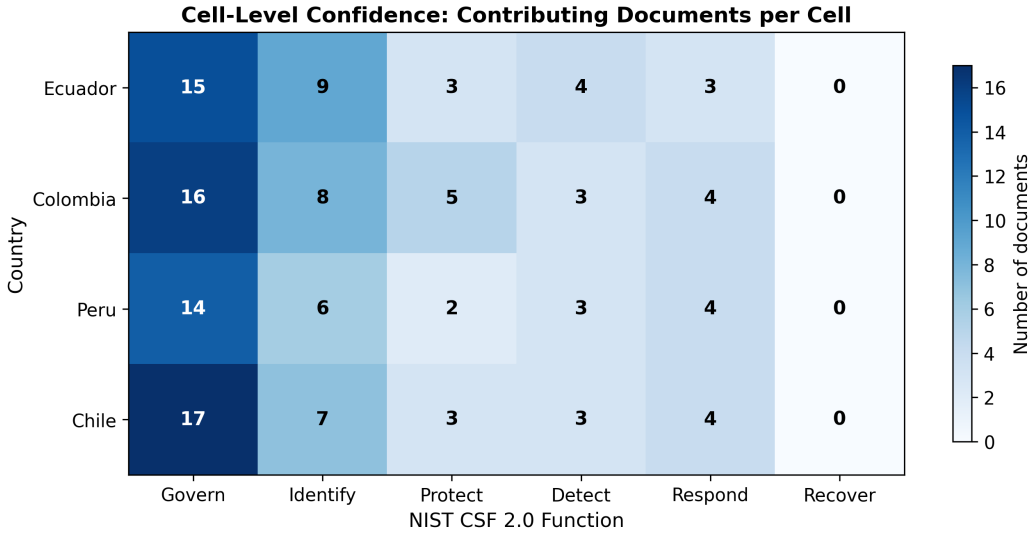


Figure 5. Cell-level confidence based on documentary density (number of contributing documents per cell).

4.6. Multi-Model Validation: Three Independent LLMs

To test whether the LLM coding behavior is specific to a single model or general across LLMs, the identical coding protocol (same rubric, same evidence, same prompt) was applied with three independent models from three providers: Claude Opus 4.8 (Anthropic), Gemini 3.1 Pro (Google LLC, Mountain View, CA, USA), and GPT-5.5 Pro (OpenAI, San Francisco, CA, USA). The Gemini and GPT-5.5 Pro codings were obtained through each provider's public web interface rather than the API, so sampling temperature could not be fixed and the runs are not deterministic: an identical re-run may yield slightly different item-level scores. Each model coded the corpus in a single run in June 2026, using the model versions stated above, which were the providers' current releases at that time. This is acknowledged as a reproducibility limitation; however, the inter-model agreement is measured on the aggregated country-by-function structure, which is robust to per-item stochasticity, and the exact prompts are archived verbatim in the Supplementary Material to allow independent re-execution. The analysis proceeds at three levels, summarized in Tables 5 and 6.

4.6.1. Document Level

At the level of individual documents, Claude and Gemini agreed almost perfectly (EDI Pearson $r = 0.961$, mean bias -0.1), whereas GPT-5.5 Pro diverged more from both ($r = 0.612$ with Claude, $r = 0.568$ with Gemini). The divergence is traceable to a single dimension: GPT-5.5 Pro assigned systematically lower quality/recency scores (C_q mean 5.37 versus 7.06 for Claude and 7.04 for Gemini), penalizing document age more aggressively. Agreement on control criticality (C_c) remained high across all three models (mean pairwise $r = 0.863$).

Table 5. Inter-model agreement at document and aggregated cell levels (EDI Pearson r).

Model Pair	Document Level	Cell Level (Country × Function)
Claude–Gemini	0.961	0.944
Claude–GPT-5.5 Pro	0.612	0.943
Gemini–GPT-5.5 Pro	0.568	0.918

4.6.2. Cell Level

When document codings are aggregated into the country × function grid through the full EDI pipeline (density factor and multilateral weighting), the three models converge strongly ($r = 0.918$–0.944 for all pairs, including GPT-5.5 Pro). The reason is that GPT-5.5 Pro's C_q bias is systematic and roughly uniform: it lowers absolute values but preserves the relative structure, so aggregation recovers a near-identical pattern. The scale bias affects magnitude, not shape.

4.6.3. Conclusion Level

Most importantly, the four substantive conclusions of the study hold identically under all three models, as Table 6 shows. The between-country test is non-significant under every model (regional similarity), the functional gradient is significant under every model, the documentary Recover void appears under every model, and Peru is the weakest country under every model.

This three-level pattern directly addresses the single-model limitation: the substantive findings are model-independent, while the absolute-scale bias is model-specific. The choice of LLM affects calibration but not the comparative conclusions—a result only obtainable by testing more than one model.

Table 6. Robustness of the four core conclusions across three independent LLMs (full EDI pipeline).

Conclusion	Claude	Gemini	GPT-5.5 Pro	Robust?
Between-country (Kruskal p)	0.406	0.452	0.261	Yes: all n.s.
Functional gradient (Friedman p)	0.009	0.009	0.019	Yes: all sig.
Recover void (sum EDI)	0	0	0	Yes: all zero
Weakest country	Peru	Peru	Peru	Yes

4.7. Human Gold-Standard Validation

To resolve the circularity concern that both automated codings (rule-based and LLM) operate partly on shared observable attributes, two independent expert coders manually coded a stratified sample of 26 documents (47% of the corpus, covering all CSF functions and all countries). The two coders worked independently, each reading the full source document of every sampled item—not the evidence summaries used by the automated methods—and applied the anchored rubric after a joint calibration session on non-sample documents; the gold standard is the mean of their two codings. Because the human reference derives from full-text reading, it is independent of the summary inputs shared by the rule-based and LLM methods, and provides a genuine third reference against which both automated methods can be anchored.

4.7.1. Inter-Coder Reliability

The two coders showed substantial-to-excellent agreement (Table 7). At the composite EDI level, the intra-class correlation coefficient was ICC $(2, 1) = 0.914$ (Pearson $r = 0.915$), with a near-zero mean bias of -0.5 points and narrow Bland–Altman limits $[-8.0, +7.0]$. Within-one-point agreement was 100% on all three dimensions, and band agreement (scarce/moderate/abundant) was 88% (Cohen $\kappa = 0.769$). These values confirm that the EDI rubric is applied reliably by independent experts and is not idiosyncratic. Notably, human–human band agreement (88%) exceeds the LLM–rule band agreement reported above (59%), reinforcing that threshold-based classifications are where calibration matters most.

Table 7. Inter-coder reliability on the stratified human-coded sample ($n = 26$).

Metric	Pearson r	Weighted κ	ICC (2, 1)	Within ± 1 pt
C_q (quality)	0.933	0.801	0.923	100%
C_c (criticality)	0.828	0.657	0.833	100%
C_o (coverage)	0.915	0.741	0.889	100%
EDI composite	0.915	0.769 †	0.914	—

† Cohen κ on maturity bands (scarce/moderate/abundant).

4.7.2. Anchoring Against the Human Gold Standard

Constructing the gold standard as the mean of the two coders, both automated methods were compared against it (Table 8 and Figure 6). The LLM tracked the human judgment closely (Pearson $r = 0.948$, 95% CI $[0.886, 0.977]$, mean bias -1.2), substantially better than the rule-based coding ($r = 0.756$, 95% CI $[0.522, 0.884]$, mean bias $+6.8$). Because both correlations share the human gold standard, the contrast was tested with Steiger's test for dependent correlations (Williams' t): the LLM's advantage is statistically significant, $t(23) = 3.62$, $p = 0.001$. Crucially, this reframes the earlier LLM-versus-rule divergence: relative to human experts, it is the rule-based method that overestimates, while the LLM is nearly calibrated. Far from being unreliable, the LLM is a well-calibrated proxy for expert human coding.

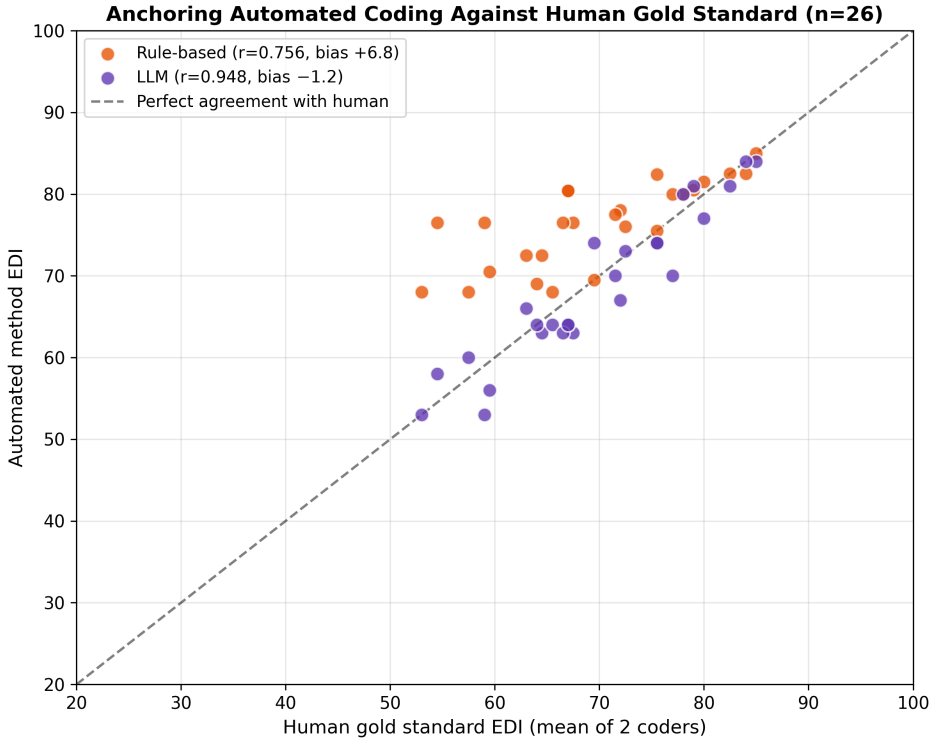


Figure 6. Automated codings anchored against the human gold standard ($n = 26$). LLM points (purple) cluster on the perfect-agreement diagonal; rule-based points (orange) sit systematically above it, evidencing overestimation relative to human judgment.

Table 8. Anchoring automated codings against the human gold standard ($n = 26$).

Automated Method	Pearson r	MAE	Bias vs. Human
Rule-based	0.756	6.9	+6.8 (overestimates)
LLM (Claude)	0.948	2.6	−1.2 (near-calibrated)

5. Discussion

5.1. Reliable Ordering, Biased Scale—And Why It Suffices Here

The methodological finding (RQ1) is a dissociation between relative and absolute reliability: strong correlation ($r = 0.729$) with a consistent -7.9-point bias relative to the rule-based reference. The human gold standard (Section 4.7) refines the interpretation of this bias: against expert human judgment, the LLM is in fact near-calibrated ($r = 0.948$, bias -1.2), whereas the rule-based method overestimates ($+6.8$). The earlier LLM-versus-rule gap therefore largely reflects rule-based overestimation, not LLM unreliability. The strategic value of the combined design is that RQ3 demonstrates this bias does not threaten the substantive conclusions. Because the comparative analysis depends on ordering (which countries and functions rank higher), not on absolute thresholds, the LLM's reliable ordering is sufficient: both codings yield the same regional verdict. This is a generalizable lesson—for comparative and ranking tasks, LLM-assisted coding can be trusted even with a biased absolute scale, provided the bias is roughly uniform [15,23]. The three-model validation reinforces this: although GPT-5.5 Pro applied a stricter quality scale than Claude and Gemini, all three produced the same aggregated structure and the same four substantive conclusions, because each model's bias was internally consistent. For threshold-based tasks (e.g., certifying a sector as "abundant"), the bias must still be calibrated per model, since 41% of documents shifted band under LLM coding relative to the rule-based reference.

5.2. The Andean Documentary Asymmetry Is Structural

The substantive finding (RQ2) is regional similarity: despite divergent institutional designs—Chile's centralized agency versus Peru's decree-based center—the four countries produced documentary records of low between-country dispersion (coefficient of variation 4.8%), all following the same Govern-to-Recover gradient. That this holds under two independent coding methods, across density-factor settings, and after excluding shared multilateral documents (RQ3) strengthens the claim that the Govern-to-Recover documentary asymmetry is structural rather than an artifact of coding, parameterization, or any single country—though we are careful to present between-country similarity as a descriptive observation rather than a hypothesis-test result, given the limited power of a four-country comparison. The pattern is consistent with institutional isomorphism in the diffusion of national strategies [33,79]: countries adopt similar formal templates, often encouraged by the same indices and international actors, producing similar documentary footprints. The Recover void—no country documents post-incident recovery as an institutionalized, national-level capability with a dedicated instrument—is a regional documentary gap with potential national-security implications, amplified by persistent digital divides [71,80] and by the disciplinary fragmentation that has historically impeded cyber-risk research [81].

A targeted post hoc verification search (July 2026) corroborates that the Recover void reflects documentary reality rather than a corpus-construction artifact. No Andean country has published a national-level instrument whose primary purpose is cyber recovery—no national cyber-recovery plan, no national continuity plan for cyber incidents, and no published recovery-exercise reports. The closest artifacts are of three kinds: recovery objectives embedded within Govern-primary instruments (Ecuador's 2022 national strategy and its 2026 Cybersecurity Strengthening Law [82]); entity-level operational-continuity plans issued under Peru's all-hazards disaster-risk framework, which are neither national in scope nor cyber-primary; and recent legal mandates requiring operators of vital services to maintain continuity and recovery plans (Chile's Framework Law 21.663 of 2024 [83]). The mandates in fact reinforce the interpretation: states are delegating recovery planning to operators precisely because national playbooks do not yet exist. The contrast with

neighboring Argentina—whose National Cybersecurity Center was mandated in 2025–2026 to produce a Disaster Recovery Plan for the national public sector, explicitly aligned with NIST SP 800-184 [84]—shows that Recover-primary national instruments are attainable in the region, sharpening the significance of their absence across the four Andean countries.

5.3. A Human-in-the-Loop Protocol for Scalable Comparative Assessment

Combining both findings yields a practical protocol, now empirically grounded by the human validation. **Stage 1:** the LLM codes the full multi-country corpus in minutes. **Stage 2:** expert coders code a stratified subsample, estimate the LLM's bias and limits of agreement, and derive a calibration offset—as done here, where two coders achieved ICC = 0.914 and the LLM tracked their gold standard at $r = 0.948$. **Stage 3:** the offset is applied and band-boundary cells are flagged for human review. This preserves the efficiency that makes regional comparison feasible while bounding the risk from systematic bias [23,29,85]. The approach transforms documentary cybersecurity assessment from a single-country, multi-week effort into a scalable regional capability.

5.4. Limitations

Five limitations apply. First, the human gold standard reported in Section 4.7 is based on a stratified 47% sample rather than the full corpus; extending independent human coding to all 55 documents would further tighten the anchoring estimates, though the strong inter-coder reliability (ICC = 0.914) and the close LLM–human agreement ($r = 0.948$) already substantially mitigate the original circularity concern. Second, three LLMs from three providers (Claude Opus 4.8, Gemini 3.1 Pro, GPT-5.5 Pro) were tested, which substantially strengthens generality over a single-model design; however, all three are proprietary, and open-weight models plus systematic prompt- and temperature-sensitivity analysis remain future work [21,24,85]. The observed scale bias was model-specific (most pronounced in GPT-5.5 Pro's quality scoring), confirming that calibration must be per-model even though the comparative conclusions proved model-independent. Third, the automated methods coded from verified evidence summaries, not full texts; if such summaries embed human interpretation, the scalability claim must be discounted by the summarization cost. Notably, the human gold standard was built from full source documents, so the close LLM–human agreement ($r = 0.948$) holds despite the automated methods working from summaries—evidence that the summaries preserved the information relevant to the rubric—though full-text automated coding remains a worthwhile robustness check. Fourth, with four countries the between-country test has minimal power; the non-significant result is reported as descriptive low dispersion, not as a hypothesis-test proof of convergence, and the smaller Peruvian sub-corpus may depress its EDI via the density factor independently of capability (a density-versus-search-exhaustiveness confound). Fifth, documentary density may partly reflect governmental transparency rather than capability, and the EDI measures documentation, not operational practice; moreover, coding operates at the document level, so recovery clauses embedded within governance instruments are not scored as Recover evidence, and passage-level multi-label coding remains future work.

5.5. Contribution

This study contributes: (1) the first validation of LLM-assisted coding for cybersecurity documentary assessment, replicated across three independent LLMs; (2) the first comparative documentary baseline of the Andean region with a single instrument; (3) evidence—robust across two coding methods and three LLMs—that the regional Govern-to-Recover documentary asymmetry is structural; (4) identification of a regional documentary Recover void; and (5) a human-in-the-loop protocol enabling scalable comparative assessment.

6. Conclusions

This paper unified a methodological validation and a substantive comparative study. Methodologically, an LLM reproduced a structured three-dimensional cybersecurity coding rubric with strong correlation ($r = 0.729$) but a systematic -7.9-point conservative bias relative to rules—reliable ordering, biased scale—and replication across three independent LLMs (Claude Opus 4.8, Gemini 3.1 Pro, GPT-5.5 Pro) showed the comparative conclusions to be model-independent while the scale bias is model-specific. Two independent human coders established a gold standard (ICC $= 0.914$) against which the LLM proved near-calibrated ($r = 0.948$), better than the rule-based reference, resolving the circularity concern. Substantively, four Andean countries exhibit a convergent documentary profile dominated by governance and culminating in a documentary Recover void. Decisively, these regional findings held under two coding methods and three LLMs, demonstrating that the asymmetry is structural and that LLM-assisted coding, suitably calibrated, is a viable accelerator for comparative cybersecurity assessment.

Future work should (a) extend independent human coding from the present stratified 47% sample to the full corpus and add a third coder to triangulate; (b) extend the multi-model validation to open-weight LLMs and add systematic prompt- and temperature-sensitivity analysis; (c) expand to the full Latin American region to increase statistical power and test whether Andean convergence generalizes; and (d) complement documentary EDI with operational indicators where available.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/computers15100687/s1: File S1: Full corpus of 55 verified institutional documents with metadata, sources, and primary/secondary NIST CSF 2.0 functions (corpus_andean_completo.csv); File S2: Rule-based coding (C_q, C_c, C_o, EDI) of each document (coded_documents.csv); File S3: LLM (Claude Opus 4.8) coding of each document with one-line justifications (llm_coded_documents.csv); File S4: Three-model coding (Claude Opus 4.8, Gemini 3.1 Pro, GPT-5.5 Pro) with justifications (comparacion_3modelos.csv); File S5: Cell-level EDI by country and CSF function, with national and multilateral document counts, density factors, and confidence levels (edi_by_country_function.csv); File S6: Human gold-standard coding sheet of the two expert coders ($n = 26$) with agreement summary (Hoja_Codificacion_Humana.xlsx); File S7: Exact LLM prompts used for each model, verbatim (folder LLM_Prompts); File S8: Python script for the permutation tests and effect sizes (permutation_analysis.py).

Author Contributions: Conceptualization, L.J.S.-A. and C.H.F.-U.; methodology, C.H.F.-U., M.R.-E. and D.P.; software and formal analysis, C.H.F.-U. and F.M.N.-O.; validation, C.H.F.-U., C.M.F.-U. and M.R.-E.; investigation, C.H.F.-U. and F.M.N.-O.; data curation, C.M.F.-U.; writing—original draft preparation, C.H.F.-U.; writing—review and editing, L.J.S.-A., M.R.-E. and D.P.; supervision, L.J.S.-A. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was developed within research project PIC5P23-53 of Universidad Católica de Cuenca, in collaboration with Universidad Politécnica Salesiana, Universidad Nacional de La Plata, and Universidad de las Fuerzas Armadas ESPE.

Institutional Review Board Statement: Not applicable. This study used only publicly available institutional documents and did not involve human participants, human data, or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The corpus, both codings, the three-model codings with justifications, the human gold-standard coding sheet, the LLM prompts, and the analysis script are openly available in the Supplementary Materials of this article (Files S1–S8).

Acknowledgments: The authors used a generative AI tool for language refinement and for implementing analysis and visualization code, all reviewed by the authors. All research design, the validation framework, corpus construction, interpretation, and conclusions are the authors' own. No AI tool is listed as an author.

Conflicts of Interest: The authors declare no conflicts of interest. The use of generative AI tools is disclosed in the Acknowledgments; the three LLMs evaluated in this study are the research object of the methodological validation, not authorship tools.

References

1. Flores-Urgiles, C.H.; Ron-Egas, M.B.; Serpa-Andrade, L.J.; Chasi-Pesantez, P.A.; Santacruz-Espinoza, J.J.; Piccirilli, D.; Fuertes-Diaz, W.M. Design–Action Maturity for National Cybersecurity (DAM–NC): An Adaptive Methodology for Context-Constrained Environments. In *Trends in Artificial Intelligence, and Computer Engineering (ICAETT 2025)*; Lecture Notes in Networks and Systems; Springer: Berlin/Heidelberg, Germany, 2026; Volume 1868, pp. 19–34. [CrossRef]
2. Marican, M.N.Y.; Razak, S.A.; Selamat, A.; Othman, S.H. Cyber Security Maturity Assessment Framework for Technology Startups: A Systematic Literature Review. *IEEE Access* **2023**, *11*, 5442–5452. [CrossRef]
3. International Telecommunication Union. *Global Cybersecurity Index 2024*; Technical Report; ITU: Geneva, Switzerland, 2024.
4. von Solms, R.; van Niekerk, J. From information security to cyber security. *Comput. Secur.* **2013**, *38*, 97–102. [CrossRef]
5. Schatz, D.; Bashroush, R.; Wall, J. Towards a More Representative Definition of Cyber Security. *J. Digit. Forensics Secur. Law* **2017**, *12*, 53–74. [CrossRef]
6. Serpa-Andrade, L.J.; Flores-Urgilés, C.H.; Narváez-Ochoa, F.; Chasi-Pesantez, P.; Ron-Egas, M.; Piccirilli, D. An Evidence-Based Documentary Baseline of National Cybersecurity Capabilities in Ecuador (2020–2025): An Exploratory Descriptive Study within the DAM–NC Research Program. In *Proceedings of the Future Technologies Conference (FTC) 2026, Berlin, Germany, 15–16 October 2026*; Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2026; in press.
7. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 4th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2018.
8. Adams, R.J.; Smart, P.; Huff, A.S. Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews. *Int. J. Manag. Rev.* **2017**, *19*, 432–454. [CrossRef]
9. Garousi, V.; Felderer, M.; Mäntylä, M.V. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf. Softw. Technol.* **2019**, *106*, 101–121. [CrossRef]
10. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
11. Ron, M.; Rivera, O.; Fuertes, W.; Toulkeridis, T.; Díaz, J. Cybersecurity baseline: An exploration which permits to delineate national cybersecurity strategy in Ecuador. In *Information Technology and Systems*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 847–857. [CrossRef]
12. Cano Martínez, J.J. Modelos de madurez en ciberseguridad. In *Revista Sistemas (ACIS)*; Asociación Colombiana de Ingenieros de Sistemas: Bogotá, Colombia, 2018; pp. 36–44.
13. Sabillón, R.; Cavaller, V.; Cano, J.; Serra-Ruiz, J. Cybersecurity Awareness Programs in Latin American Countries. *Int. J. Inf. Secur. Cybercrime* **2017**, *6*, 21–30.
14. Organization of American States; Inter-American Development Bank. *Cybersecurity Report 2022: Risks, Progress, and the Way Forward in Latin America and the Caribbean*; Technical Report; OAS: Washington, DC, USA, 2022.
15. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120. [CrossRef] [PubMed]
16. Törnberg, P. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages. *arXiv* **2023**, arXiv:2304.06588.
17. Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; Yang, D. Can Large Language Models Transform Computational Social Science? *Comput. Linguist.* **2024**, *50*, 237–291. [CrossRef]
18. Rathje, S.; Mirea, D.M.; Sucholutsky, I.; Marjieh, R.; Robertson, C.E.; Bavel, J.J.V. GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2308950121. [CrossRef] [PubMed]
19. National Institute of Standards and Technology. *The NIST Cybersecurity Framework (CSF) 2.0*; Technical Report; U.S. Department of Commerce: Washington, DC, USA, 2024. [CrossRef]
20. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
21. Wang, J.; Hu, X.; Hou, W.; Chen, H.; Zheng, R.; Wang, Y.; Yang, L.; Huang, H.; Ye, W.; Geng, X.; et al. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *arXiv* **2023**, arXiv:2302.12095.

22. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM FAccT*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 610–623. [CrossRef]
23. Pangakis, N.; Wolken, S.; Fasching, N. Automated Annotation with Generative AI Requires Validation. *arXiv* **2023**, arXiv:2306.00176.
24. Reiss, M.V. Testing the Reliability of ChatGPT for Text Annotation and Classification. *arXiv* **2023**, arXiv:2304.11085.
25. Ollion, E.; Shen, R.; Macanovic, A.; Chatelain, A. The dangers of using proprietary LLMs for research. *Nat. Mach. Intell.* **2024**, *6*, 4–5. [CrossRef]
26. Kuzman, T.; Mozetič, I.; Ljubešić, N. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? *arXiv* **2023**, arXiv:2303.03953.
27. Thapa, S.; Naseem, U.; Nasim, M. From Humans to Machines: Can ChatGPT-like LLMs effectively Replace Human Annotators in NLP Tasks? In *Workshop Proceedings of the ICWSM*; AAAI Press: Menlo Park, CA, USA, 2023.
28. Huang, F.; Kwak, H.; An, J. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference*; Association for Computing Machinery (ACM): New York, NY, USA, 2023; pp. 294–297. [CrossRef]
29. Messeri, L.; Crockett, M.J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **2024**, *627*, 49–58. [CrossRef] [PubMed]
30. Birhane, A.; Kasirzadeh, A.; Leslie, D.; Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **2023**, *5*, 277–280. [CrossRef]
31. e-Governance Academy. National Cyber Security Index (NCSI). 2024. Available online: https://ncsi.ega.ee/ (accessed on 1 October 2026).
32. Global Cyber Security Capacity Centre. *Cybersecurity Capacity Maturity Model for Nations (CMM)*; Technical Report; University of Oxford: Oxford, UK, 2021.
33. Luiijf, E.; Besseling, K.; de Graaf, P. Nineteen national cyber security strategies. *Int. J. Crit. Infrastruct.* **2013**, *9*, 3–31. [CrossRef]
34. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design Science in Information Systems Research. *MIS Q.* **2004**, *28*, 75–105. [CrossRef]
35. Becker, J.; Knackstedt, R.; Pöppelbuß, J. Developing Maturity Models for IT Management. *Bus. Inf. Syst. Eng.* **2009**, *1*, 213–222. [CrossRef]
36. Mateus-Coelho, N.; Ávila, P.; Bastos, J. Maturity Models for Cybersecurity: An Evaluation Approach. *Procedia Comput. Sci.* **2023**, *219*, 1015–1022. [CrossRef]
37. Mbanaso, U.M.; Abrahams, L.; Apene, O.Z. Conceptual Design of a Cybersecurity Resilience Maturity Measurement Framework. *Afr. J. Inf. Commun.* **2019**, *23*, 1–26. [CrossRef]
38. Kandasamy, K.; Srinivas, S.; Achuthan, K.; Rangan, V.P. IoT cyber risk: A holistic analysis of cyber risk assessment frameworks. *EURASIP J. Inf. Secur.* **2020**, *2020*, 8. [CrossRef]
39. Lijphart, A. *Comparative Politics and the Comparative Method*; American Political Science Review; Cambridge University Press: Cambridge, UK, 1971.
40. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef] [PubMed]
41. Shrout, P.E.; Fleiss, J.L. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [CrossRef] [PubMed]
42. Bland, J.M.; Altman, D.G. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet* **1986**, *327*, 307–310. [CrossRef]
43. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
44. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
45. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Medica* **2012**, *22*, 276–282. [CrossRef]
46. Krippendorff, K. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Hum. Commun. Res.* **2004**, *30*, 411–433. [CrossRef]
47. Congreso Nacional de Chile. *Ley 21.663: Ley Marco Sobre Ciberseguridad e Infraestructura Crítica de la Información*; Technical Report; Diario Oficial; Congreso Nacional de Chile: Santiago, Chile, 2024.
48. Agencia Nacional de Ciberseguridad. *Resolución ANCI N^{o} 024/2025: Norma técnica Vinculante*; Technical Report; ANCI: Santiago, Chile, 2025.
49. CSIRT Nacional de Chile. *Operación del CSIRT Nacional bajo la Ley Marco de Ciberseguridad*; Technical Report; CSIRT.gob.cl: Santiago, Chile, 2025.
50. Congreso Nacional de Chile. *Ley 21.459: Delitos Informáticos*; Technical Report; Diario Oficial; Congreso Nacional de Chile: Santiago, Chile, 2022.
51. Ministerio del Interior y Seguridad Pública. *Política Nacional de Ciberseguridad 2017–2022*; Technical Report; Ministerio del Interior y Seguridad Pública: Santiago, Chile, 2017.

52. Departamento Nacional de Planeación. *CONPES 3854: Política Nacional de Seguridad Digital*; Technical Report; DNP: Bogotá, Colombia, 2016.
53. Departamento Nacional de Planeación. *CONPES 3995: Política Nacional de Confianza y Seguridad Digital*; Technical Report; DNP: Bogotá, Colombia, 2020.
54. Ministerio de Tecnologías de la Información y las Comunicaciones. *Estrategia Nacional de Seguridad Digital 2025–2027*; Technical Report; MinTIC: Bogotá, Colombia, 2025.
55. Grupo de Respuesta a Emergencias Cibernéticas de Colombia. *Informe de Gestión de Incidentes de Seguridad Digital*; Technical Report; ColCERT/MinTIC: Bogotá, Colombia, 2025.
56. Superintendencia Financiera de Colombia. *Circular Externa 029 de 2024: Requerimientos de Ciberresiliencia*; Technical Report; SFC: Bogotá, Colombia, 2024.
57. Congreso de la República de Colombia. *Ley 1581 de 2012: Protección de Datos Personales*; Technical Report; Congreso de la República de Colombia: Bogotá, Colombia, 2012.
58. Ministerio de Telecomunicaciones del Ecuador. *Plan de Fortalecimiento del EcuCERT*; Technical Report; MINTEL: Quito, Ecuador, 2022.
59. Ministerio de Telecomunicaciones del Ecuador. *Estrategia Nacional de Ciberseguridad del Ecuador 2022–2025*; Technical Report; MINTEL: Quito, Ecuador, 2022.
60. Ministerio de Telecomunicaciones del Ecuador. *Esquema Gubernamental de Seguridad de la Información (EGSI)*, Versión 3.0; Technical Report; MINTEL: Quito, Ecuador, 2024.
61. Asamblea Nacional del Ecuador. *Ley Orgánica de Ciberseguridad*; Technical Report; Registro Oficial; Asamblea Nacional del Ecuador: Quito, Ecuador, 2026.
62. Garzón, V. Análisis de la incidencia de ataques de ingeniería social en Ecuador. *J. Comput. Sci.* **2024**, *20*, 1330–1338. [CrossRef]
63. Presidencia del Consejo de Ministros. *Decreto de Urgencia 007-2020: Creación del Centro Nacional de Seguridad Digital*; Technical Report; PCM: Lima, Peru, 2020.
64. Secretaría de Gobierno y Transformación Digital. *Estrategia Nacional de Ciberseguridad del Perú 2026–2028 (ESNACIB)*; Technical Report; PCM-SGTD: Lima, Peru, 2025.
65. Centro Nacional de Seguridad Digital. *Informe de Gestión de Alertas de Seguridad Digital*; Technical Report; PCM: Lima, Peru, 2024.
66. Congreso de la República del Perú. *Ley 30999: Ley de Ciberdefensa*; Technical Report; Congreso de la República del Perú: Lima, Peru, 2019.
67. Council of Europe. Convention on Cybercrime (Budapest Convention), ETS No. 185. 2001. Available online: https://www.coe.int/en/web/cybercrime (accessed on 1 October 2026).
68. *ISO/IEC 27001:2022*; Information Security Management Systems. International Organization for Standardization: Geneva, Switzerland, 2022.
69. International Telecommunication Union. *Guide to Developing a National Cybersecurity Strategy*; Technical Report; ITU: Geneva, Switzerland, 2023.
70. Inter-American Development Bank; Organization of American States. *Cybersecurity: Are We Ready in Latin America and the Caribbean?* IDB: Washington, DC, USA, 2020. [CrossRef]
71. Economic Commission for Latin America and the Caribbean. *Data and Facts on Digital Transformation in Latin America and the Caribbean*; Technical Report; United Nations: Santiago, Chile, 2022.
72. European Union Agency for Cybersecurity. *ENISA Threat Landscape 2024*; Technical Report; ENISA: Athens, Greece, 2024.
73. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
74. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. [CrossRef]
75. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22199–22213. [CrossRef]
76. Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; Chadha, A. A Systematic Survey of Prompt Engineering in Large Language Models. *arXiv* **2024**, arXiv:2402.07927.
77. Kruskal, W.H.; Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
78. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
79. Pawlak, P.; Barmpaliou, N. Politics of cybersecurity capacity building: Conundrum and opportunity. *J. Cyber Policy* **2017**, *2*, 123–144. [CrossRef]
80. Creese, S.; Dutton, W.H.; Esteve-Gonzalez, P. The social and cultural shaping of cybersecurity capacity building: A comparative study of nations and regions. *Pers. Ubiquitous Comput.* **2021**, *25*, 941–955. [CrossRef] [PubMed]

81. Falco, G.; Eling, M.; Jablanski, D.; Weber, M.; Miller, V.; Gordon, L.A.; Wang, S.S.; Schmit, J.; Thomas, R.; Elvedi, M.; et al. Cyber risk research impeded by disciplinary barriers. *Science* **2019**, *366*, 1066–1069. [CrossRef] [PubMed]
82. Asamblea Nacional del Ecuador. *Ley Orgánica para el Fortalecimiento de la Ciberseguridad*; Establishes Mandatory Incident-Management Obligations Including Containment and Recovery; Registro Oficial; Asamblea Nacional del Ecuador: Quito, Ecuador, 2026.
83. Congreso Nacional de Chile. *Ley 21.663, Ley Marco de Ciberseguridad*; Diario Oficial de la República de Chile; Requires Operators of Vital Importance to Maintain Operational-Continuity and Cybersecurity Recovery Plans; Congreso Nacional de Chile: Valparaíso, Chile, 2024.
84. Centro Nacional de Ciberseguridad, República Argentina. *Disposición 1/2026: Reglamento Técnico de Políticas de Planes de Contingencia para el Sector Público Nacional*; Boletín Oficial de la República Argentina; Centro Nacional de Ciberseguridad, República Argentina: Buenos Aires, Argentina, 2026.
85. Spirling, A. Why open-source generative AI models are an ethical way forward for science. *Nature* **2023**, *616*, 413. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.