

## ENTERPRISE GRC PLATFORM: A TWO-STAGE RISK SCORING MODEL WITH AI-ASSISTED RISK IDENTIFICATION AND AUDIT WORKFLOW INTEGRATION

Saloni BHOSALE, Samson QUAYE, Maurice DAWSON

Illinois Institute of Technology, Chicago, USA

[sbhosale@hawk.iit.edu](mailto:sbhosale@hawk.iit.edu)

**Abstract:** *In enterprise security, risk scores determine what gets fixed and what gets ignored. Most organisations rely on a 5x5 likelihood-impact matrix, a method shown to generate misleading rankings that treats control effectiveness as little more than a rounding factor. Moving from no controls to best-in-class controls shifts a score by fewer than three points on a 25-point scale, a gap too small to justify any investment decision. This paper presents the Enterprise GRC Risk Intelligence Platform, which separates Inherent Risk from Residual Risk using a two-stage model. Asset criticality is weighted from 0.60 to 1.60 and Control Effectiveness is modelled as a calibrated reduction from 5% to 90%, producing a 22-point residual risk differential compared to two points under the conventional formula. The platform covers 22 asset types and 14 threat categories, combines AI-assisted risk identification with structured analyst input, role-based access control, audit workflow with evidence upload, and alignment across NIST CSF 2.0, NIST RMF, ISO 27001, and Zero Trust. Validation across all 625 scoring combinations confirms strict monotonicity.*

**Keywords:** GRC, Risk Scoring, Control Effectiveness, Residual Risk, Cybersecurity

### 1. Introduction

Risk scores sit at the centre of every GRC workflow. They determine which vulnerabilities get addressed this week and which are deferred to next quarter, which findings reach a board and which stay buried in a spreadsheet. Given how much depends on them, it is worth asking whether the formulas producing those scores are actually fit for purpose.

The most common approach multiplies a Likelihood rating by an Impact rating, each scored on a 1-5 ordinal scale, to produce a raw score on a 1-25 matrix. Two deficiencies motivate this work. First, asset criticality is either excluded from the formula or added as a linear multiplier of the raw score. This produces counter-intuitive results: a mission-critical system with strong controls can score similarly to a negligible system with no controls at all. Second, control effectiveness is modelled as a divisor over the same 1-5 scale. On a

maximum-severity risk, the difference between no controls (CE=1) and best-in-class controls (CE=5) under a naive divisor formula amounts to fewer than three points on the output scale. That is not a gap that changes any real decision.

This paper describes the Enterprise GRC Risk Intelligence Platform, a Streamlit application that tackles both problems. It separates the computation of Inherent Risk and Residual Risk into two distinct stages, and integrates AI-assisted risk identification, structured audit workflow, evidence tracking, and multi-framework alignment into a single risk register workflow.

This paper investigates the following research question: can a two-stage semi-quantitative scoring model, which separates inherent from residual risk and models control effectiveness as a calibrated percentage reduction, produce meaningfully greater score differentiation

across control quality levels than a conventional multiplicative divisor formula, while remaining deployable without requiring historical loss data? The work makes four contributions. The first is a two-stage scoring model in which asset criticality weights the impact dimension using a calibrated log-like scale, and control effectiveness is expressed as a percentage reduction drawn from a calibrated S-curve. This produces a 22-point residual risk differential between the weakest and strongest control postures. The second is an AI-assisted risk identification pipeline that parses free-text risk descriptions into structured asset-threat pairs, with a deterministic keyword-matching fallback when the API is unavailable. The third is an integrated audit workflow covering four evidence states, role-based access across three permission levels, duplicate detection, and three export formats. The fourth is multi-framework alignment across NIST CSF 2.0, NIST RMF, ISO 27001:2022, and Zero Trust, generated automatically for all 14 supported threat categories.

Section 2 reviews related work on risk scoring and AI-assisted GRC. Section 3 describes the platform architecture and scoring model. Section 4 covers the user experience design. Section 5 presents computational validation. Section 6 discusses implications and limitations. Section 7 concludes.

## **2. Related Work**

### **2.1 Qualitative Risk Matrices**

The 5x5 likelihood-impact matrix became the dominant risk scoring mechanism in cybersecurity GRC during the 1990s and 2000s, adopted by NIST, ISACA, and ISO largely because it is interpretable and requires no historical loss data. Cox [1] showed that these matrices can produce rankings that are anti-correlated with actual risk under certain conditions, meaning the method may consistently misidentify which risks are most serious. Hubbard and Seiersen [2] made a similar argument for information security, contending that

ordinal scales create false precision and that better estimates are achievable with calibrated probabilities.

Despite these critiques, ordinal matrices persist because they are explicitly permitted by PCI DSS, HIPAA risk analysis guidance, and ISO 27001 Annex A, and because practitioners find them easy to explain and defend. A further structural problem is that most implementations treat all five likelihood and impact values as equidistant, which is inconsistent with how risk actually scales. Semi-quantitative approaches that introduce weighted scales begin to address that inconsistency while retaining the interpretability that makes qualitative tools practical.

### **2.2 Control Effectiveness Modelling**

Control effectiveness has received surprisingly little treatment as a quantitative modifier in the literature. NIST SP 800-30 Rev 1 [3] addresses how to identify and evaluate controls but stops short of specifying how a control effectiveness rating should numerically reduce a risk score. ISO 27005:2022 [4] recommends treating inherent and residual risk as separate assessments and provides an asset-value-weighted impact scale, but the formula linking them is left to the implementer. Brotby [5] defines control effectiveness as the product of coverage, capability, and maturity, without specifying its mathematical relationship to risk output. The scoring model in this paper fills that gap with a calibrated S-curve that produces meaningful score separation across the full range of control maturity.

### **2.3 AI-Assisted Risk Identification**

Large language models have been applied to OSINT-driven threat intelligence collection by Soler, Dawson, and Colina [6], who demonstrate an LLM-based pipeline for automated target profiling from social media data. This platform applies a related capability, parsing a practitioner's free-text description of a risk incident into a structured asset-threat pair with associated likelihood and impact scores. The model is constrained to return a JSON object and

falls back to deterministic keyword matching when the API is unavailable or the response fails validation.

2.4 GRC Platform Design

Commercial GRC platforms such as ServiceNow GRC, Archer, and OneTrust provide mature workflow, evidence management, and reporting capabilities. They are also expensive, require significant configuration, and typically implement the conventional multiplicative scoring formula. Open-source alternatives offer more flexibility but rarely include validated scoring models or integrated AI capabilities. The platform described in this paper sits between these options: it provides production-grade workflow and reporting, a methodologically grounded scoring model, and AI-assisted risk entry, implemented as an open Streamlit application.

3. Platform Architecture

3.1 System Overview

The Enterprise GRC Risk Intelligence Platform is built in Python using the Streamlit framework and organises the risk assessment workflow across two tabs: Risk Assessment and Risk Register. Within the Risk Assessment tab, the analyst moves through five sequential form sections. The first is Risk Identification, which accepts either a free-text description in Smart Mode or a structured asset and threat selection in Analyst Mode. The second is Enterprise Context, capturing organisational details including industry, department, risk owner, review date, and status. The third is the Scoring Engine, where the analyst rates Asset Value and Control Effectiveness on

1-5 sliders. The fourth generates framework mappings to NIST CSF 2.0, NIST RMF, ISO 27001:2022, and Zero Trust. The fifth produces the risk register output and exports in CSV, plain-text, and PDF formats.

The platform covers 22 asset types across IT infrastructure, cloud environments, OT/ICS systems, physical security, and business applications. It supports 14 threat categories and holds 42 calibrated asset-threat baseline score pairs, so that common high-risk combinations such as Cloud Environment and Misconfiguration, or Database and Data Breach, receive scores grounded in observed threat frequency rather than generic defaults.

3.2 Two-Stage Scoring Model

The scoring model separates inherent risk from residual risk as two analytically distinct quantities, consistent with the guidance in ISO 27005:2022 §9 and the likelihood-impact determination tables in NIST SP 800-30 Rev 1.

In Stage 1, the impact dimension is adjusted by a criticality weight derived from the Asset Value rating. The weights follow a log-like progression calibrated against the five-level asset valuation scale in ISO 27005:2022 Annex C and the relative weighting principles of the Analytic Hierarchy Process [9]. Each step increases by a factor of approximately 1.25 to 1.30, consistent with geometric scaling ratios recommended in multi-criteria decision analysis for five-point ordinal scales. A mission-critical asset at AV=5 therefore carries 2.67 times the impact weight of a negligible asset at AV=1.

Table 1. Asset Value Criticality Weights

| Asset Value (AV) | Weight (w) | Interpretation            |
|------------------|------------|---------------------------|
| 1                | 0.60       | Negligible business value |
| 2                | 0.75       | Low business value        |
| 3                | 1.00       | Moderate — baseline       |
| 4                | 1.30       | High business value       |
| 5                | 1.60       | Mission-critical          |

The adjusted impact is multiplied by likelihood and clamped to the [1, 25] output range:

```
Impact_adj = Impact x w(AV)
Inherent    = clamp(Likelihood x
                    Impact_adj, 1, 25)
```

In Stage 2, the Control Effectiveness rating is translated into a percentage reduction using a calibrated S-curve shown in Table 2. The five reduction values (5%, 20%,

45%, 70%, 90%) were derived by mapping each CE level to the corresponding risk reduction implied by CMMI maturity tiers and NIST SP 800-53 control effectiveness categories. CE=1 corresponds to approximately 5% reduction, consistent with near-zero mitigation of ad-hoc, undocumented controls. CE=5 produces 90% reduction, consistent with CMMI Level 5 and fully certified ISO 27001 implementations.

Table 2. Control Effectiveness S-Curve Reduction

| CE Rating | Reduction (r) | Maturity Description                |
|-----------|---------------|-------------------------------------|
| 1         | 5%            | Minimal / no controls in place      |
| 2         | 20%           | Basic hygiene only                  |
| 3         | 45%           | Standard controls, partially tested |
| 4         | 70%           | Mature controls, regularly tested   |
| 5         | 90%           | Best-in-class, fully evidenced      |

Residual Risk is then computed as:

```
Residual = round(clamp(Inherent
                        x (1 - r(CE)), 1, 25))
```

Three properties follow from this formulation. Residual risk is strictly monotone decreasing in CE for any fixed inherent risk, which means improving controls never worsens the score. The full 1 to 24 output range is used, giving the scale genuine discriminatory power. And the score difference between CE=1 and CE=5 on a maximum-severity risk is 22 points, compared to 2 points under the conventional divisor formula.

3.3 Risk Level Thresholds and Treatment Logic

The residual risk output scale [1, 25] is partitioned into four severity bands aligned with NIST SP 800-30's Very High / High / Moderate / Low taxonomy: Critical (score 20 or above) requires immediate executive escalation; High (13 to 19) requires prioritised remediation without delay; Medium (7 to 12) is addressed within a planned remediation cycle; Low (1 to 6) is monitored and reviewed periodically. Treatment recommendations are generated automatically based on the residual score

and threat type. Critical and High risks receive a Mitigate recommendation. Medium-severity Supply Chain and DDoS risks receive Transfer, since a meaningful portion of that exposure can be shifted through vendor contracts or insurance. All other Medium risks receive Mitigate. Low risks receive Accept. Analysts can override any of the four options through a dropdown in the results panel.

3.4 AI-Assisted Risk Identification

Smart Mode sends the analyst's free-text description, capped at 2,000 characters, to an LLM REST API endpoint. The system prompt instructs the model to return a JSON object containing the asset type, threat category, likelihood and impact scores from 1 to 5, a plain-English summary, three prioritised action items, and a confidence rating. Every field is validated against the platform's canonical lists, and any value that does not match is replaced with the keyword-matching result.

The keyword-matching fallback covers 13 asset-type patterns and 11 threat patterns. It is designed to recognise the informal language practitioners use when describing incidents: phrases like laptop was stolen, someone shared their password, or a vendor

had a breach all map cleanly to structured asset-threat pairs without requiring the analyst to know the platform's taxonomy in advance.

Data confidentiality is a genuine concern when submitting risk descriptions to an external LLM API, as those descriptions may reference specific system vulnerabilities, asset names, or indicators of organisational security posture. Research on LLM privacy risks confirms that prompts submitted to cloud APIs can be retained in provider logs [10, 11]. The platform implements three mitigations: input is capped at 2,000 characters; a confidence field is displayed to prompt analyst review before saving; and organisations in regulated environments should consider a locally hosted LLM or a description sanitisation step before API submission.

## **4. User Experience Design**

### **4.1 Input Structure and Guidance**

One of the practical challenges in building a GRC tool is that the people who need to enter risks are not always security analysts. Department heads, compliance officers, and business unit managers all observe operational risks but may not know what "likelihood" means in a scoring context, or how to distinguish a threat from a vulnerability. The platform is designed with this in mind.

The assessment form covers five sequential sections, each opening with a short caption explaining what the section does and how it affects the output. The first is Enterprise Context, which captures company, industry, department, risk owner, status, and review date. The second is Business Impact Analysis, covering three dimensions rated Low through Critical: Financial, Operational, and Compliance. The third is Risk Factors, where Asset Value and Control Effectiveness are each rated on a 1-5 slider. The fourth is Vulnerability, Evidence and Audit Tracking, collecting the finding name, severity, source, audit status, evidence owner, and uploaded file. The fifth

is Risk Identification, offered either as a free-text Smart Mode entry or a structured Analyst Mode selection.

Three guidance mechanisms help non-specialist users without reducing precision for analysts. Every dropdown and slider carries a contextual tooltip, 21 in total, each providing two to five descriptions of the available options. Both risk factor sliders display a persistent inline legend showing all five level labels. Each section opens with a caption, 16 in total, explaining what it does and how it connects to the risk score or report.

The industry selector covers seven sectors: General, Technology, Finance, Healthcare, Education, Retail, and Manufacturing. Selecting an industry adds sector-specific context to the business impact narrative. Healthcare adds patient safety and HIPAA consequences; Finance adds PCI DSS, SOX, and fraud exposure.

### **4.2 Role-Based Access Control**

The platform supports three access roles with differentiated permissions. Admins can create, edit, export, and clear risks. Managers can create, edit, and export but cannot clear the register. Viewers can export reports but cannot create or modify entries. Permissions are checked before every write operation. In a production environment these would be backed by server-side authentication rather than client-side session state.

### **4.3 Input Validation and Duplicate Detection**

Four validation rules run before the analysis button becomes active. Asset Value and Control Effectiveness must fall within the 1 to 5 range. Selecting Audit Ready without uploading an evidence file triggers a warning. Linking a Critical-severity vulnerability to an asset rated AV=1 or AV=2 triggers a data consistency check. These rules exist because the most common errors in risk register data are not malicious but careless, and catching them early is far easier than correcting them after the fact. Duplicate detection computes an MD5 fingerprint of each asset-threat pair and

checks it against existing entries before saving.

#### 4.4 Audit Workflow

Each risk entry carries an Audit Status field with four states. Evidence Missing means no supporting documentation exists yet. Needs Review means evidence has been uploaded but not validated. Remediation In Progress means active remediation is underway. Audit Ready means the evidence is complete and current. For each state, the platform generates a specific action string that appears in the UI and in all three export formats, giving auditors a consistent record without requiring the analyst to write it by hand.

## 5. Computational Validation and Sensitivity Analysis

### 5.1 Comparative Analysis

The clearest way to show what the two-stage model changes is to run both formulas on the same inputs and compare the outputs. The scalar coefficient 1.15 in the conventional formula was introduced as a scaling correction to prevent score collapse at high CE values. It does not derive from any empirical dataset; it is an ad hoc adjustment to preserve score range, which itself illustrates the methodological limitations of divisor-based approaches.

$$\text{Score\_conv} = \text{clamp}(\text{round}((L \times I \times AV / CE) \times 1.15), 1, 25)$$

Table 3. Comparative Scoring — Conventional vs. Proposed Model

| Scenario                     | L/I/AV/CE | Inherent | Old Formula | New Residual |
|------------------------------|-----------|----------|-------------|--------------|
| Critical DB, strong CE=5     | 4/5/5/5   | 25.0     | 23          | 2            |
| Critical DB, weak CE=1       | 4/5/5/1   | 25.0     | 25          | 24           |
| Cloud misconfiguration, CE=1 | 5/5/5/1   | 25.0     | 25          | 24           |
| CEO laptop stolen, CE=3      | 1/5/5/3   | 8.0      | 10          | 4            |
| Phishing, low-AV, CE=3       | 5/3/1/3   | 9.0      | 6           | 5            |
| Ransomware on backup, CE=3   | 3/5/5/3   | 24.0     | 25          | 13           |

The critical database scenario illustrates the core problem. The conventional formula produces a two-point difference between CE=5 and CE=1, scores of 23 and 25 respectively. The proposed model produces a 22-point difference, scores of 2 and 24. The cloud misconfiguration scenario is even more revealing: the conventional formula assigns 25 in both cases, meaning it cannot distinguish between a fully unprotected cloud environment and one with mature, tested controls. The proposed model assigns 24 for CE=1 and 2 for CE=5.

### 5.2 Monotonicity Verification

For all 625 input combinations, the proposed model satisfies strict monotonicity: for any fixed combination of Likelihood, Impact, and Asset Value, increasing Control Effectiveness by one level never increases the residual risk score. This matters because a scoring model that could worsen a score when controls

improve would be unusable as a basis for investment decisions. The conventional divisor formula does not reliably satisfy this property at the lower end of the CE scale.

### 5.3 Scale Coverage and Distribution

Across all 625 input combinations, the proposed model generates 22 distinct output values spread across the 1 to 24 range. The conventional formula crowds too many results toward the upper end of the scale, which means risks that should be prioritised differently end up with scores that are too similar to distinguish. When a scale exists to support prioritisation decisions, it needs to use its range meaningfully.

### 5.4 Asset-Threat Baseline Validation

The 42 asset-threat baseline score pairs were drawn from published threat frequency data and calibrated for the 22 asset types in the platform. High-severity pairs such as Cloud Environment with Misconfiguration (L=5, I=5), Database with

Data Breach (L=4, I=5), and Email System with Phishing (L=5, I=4) reflect frequency and impact patterns that appear consistently in industry breach reports. When an analyst selects one of these combinations, the platform begins from a calibrated baseline rather than a neutral midpoint.

### **5.5 AI Component Evaluation**

The AI-assisted identification component was evaluated against 20 informal risk descriptions covering credential compromise, data breach, phishing, ransomware, misconfiguration, and supply chain compromise. The AI component produced correct asset-threat pair assignments in 17 of 20 cases (85% accuracy). The three failures involved ambiguous descriptions referencing multiple assets simultaneously. The rule-based fallback achieved 80% accuracy (16 of 20), confirming that the AI component provides a marginal improvement while introducing the risk of confident misclassification on edge cases. These findings are consistent with recent work evaluating LLMs in GRC contexts [12], which found that LLMs require human oversight on ambiguous scenarios.

## **6. Discussion**

### **6.1 Practical Implications**

The most direct consequence of the two-stage model is that it gives security teams a number they can take into a budget conversation. Moving from CE=2, basic hygiene with a 20% risk reduction, to CE=4, mature controls with a 70% reduction, on a critical asset with a high-severity threat reduces residual risk from approximately 20 to 8. That is a shift from Critical to Medium priority. It is the kind of argument that is difficult to make with a conventional risk matrix, where the same improvement might produce a two-point score change that looks like noise rather than genuine progress.

Tracking both Inherent and Residual Risk also changes how organisations can report on their security posture over time. It becomes possible to see whether the threat

environment is worsening even as controls improve, or whether residual risk is falling because of genuine control investment rather than a quieter threat period. Smart Mode addresses a different but equally real problem: risk registers fail not because analysts lack expertise but because the people closest to operational risks rarely have the vocabulary to complete a structured form. Being able to type a description and receive a structured risk entry lowers that barrier enough to matter.

### **6.2 Limitations**

Several honest limitations apply to this work. The Asset Value weights and Control Effectiveness reduction percentages are calibrated estimates grounded in the guidance tables of NIST SP 800-30 and ISO 27005, but they have not been validated against a large dataset of real incidents and losses. Organisations that maintain detailed incident records should treat these parameters as starting points and adjust them against their own data.

The model is semi-quantitative at its core. Likelihood and impact are ordinal ratings, not probabilities. For organisations that need to defend security investment in financial terms, this platform is a useful triage layer but not a substitute for quantitative methods such as FAIR. The AI identification component can misclassify ambiguous descriptions, as Section 5.5 demonstrates. Human review before saving remains important. Regarding data privacy, organisations in regulated environments should treat the API as an optional component that can be disabled in favour of the rule-based fallback without loss of core functionality.

## **7. Conclusion**

This paper set out to fix a specific and well-documented problem: that conventional GRC risk scoring formulas produce scores too insensitive to control quality to support real prioritisation decisions. The two-stage model addresses this directly. By separating the computation of Inherent Risk from Residual Risk, weighting asset criticality

through the impact dimension, and modelling control effectiveness as a calibrated percentage reduction, the platform produces a 22-point score differential between minimal and mature controls. Under the conventional formula, the same comparison produces two points. The scoring model is grounded in NIST SP 800-30 Rev 1 and ISO 27005:2022, validated across all 625 input combinations for monotonicity, and implemented in a complete GRC register workflow that includes AI-assisted risk identification, four-state audit tracking with evidence upload, role-based access control, multi-framework alignment, and export in three formats.

Several directions remain open. The calibration parameters require validation against real incident loss datasets such as the Verizon DBIR [8] and ENISA threat landscape reports. A controlled user study comparing risk register quality under the proposed model against a conventional 5x5 matrix approach would provide the empirical grounding that computational validation alone cannot supply.

What the platform demonstrates is that a more defensible risk score is achievable without abandoning the interpretability that makes qualitative GRC tools practical. A risk register that practitioners can genuinely argue from is more valuable than one that is technically correct but too blunt to guide decisions.

## References

- [1] Cox LA. What's Wrong with Risk Matrices? *Risk Anal.* 2008; 28(2):497-512.
- [2] Hubbard DW, Seiersen R. *How to Measure Anything in Cybersecurity Risk*. Hoboken, NJ: John Wiley & Sons; 2016.
- [3] National Institute of Standards and Technology. *Guide for Conducting Risk Assessments*. NIST SP 800-30 Rev 1. Gaithersburg, MD: NIST; 2012.
- [4] International Organization for Standardization. *Information Security, Cybersecurity and Privacy Protection: Guidance on Managing Information Security Risks*. ISO/IEC 27005:2022. Geneva: ISO; 2022.
- [5] Brothby WK. *Information Security Management Metrics*. Boca Raton, FL: CRC Press; 2009.
- [6] Soler R, Dawson M, Colina M. Automated Target Profiling: Leveraging Artificial Intelligence for Open-Source Intelligence Collection. *Int Conf Knowl-Based Organ.* 2025; XXXI(1).
- [7] National Institute of Standards and Technology. *Cybersecurity Framework 2.0*. Gaithersburg, MD: NIST; 2024.
- [8] Verizon. *2024 Data Breach Investigations Report*. Basking Ridge, NJ: Verizon Business; 2024.
- [9] Saaty TL. *The Analytic Hierarchy Process*. New York, NY: McGraw-Hill; 1980.
- [10] Tete SB. Threat Modelling and Risk Analysis for Large Language Model (LLM)-Powered Applications. *Proc IEEE Conf Cybersecurity*. 2024.
- [11] Zhang X, Lyu D, Li X. Risk Assessment and Security Analysis of Large Language Models. *arXiv*. 2025;2508.17329.
- [12] Kasri W, Himeur Y, Alkhazaleh HA, et al. From Vulnerability to Defence: The Role of Large Language Models in Enhancing Cybersecurity. *Computation*. 2025.13(2).