IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 February 2026, accepted 24 March 2026, date of publication 26 March 2026, date of current version 1 April 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3678044

RESEARCH ARTICLE

Automating Cyber Risk Assessment With Public LLMs: An Expert-Validated Framework and Comparative Analysis

NEZIH MAHMUT UNAL AND BARIS CELIKTAS
Computer Engineering Department, Isik University, 34398 Istanbul, Türkiye
Corresponding author: Nezih Mahmut Unal (23SIBE5011@isik.edu.tr)

ABSTRACT Traditional cyber risk assessment methodologies face a critical dilemma: they are either quantitative yet static and context-agnostic (e.g., CVSS), or context-aware yet highly labor-intensive and subjective (e.g., NIST SP 800-30). Consequently, organizations struggle to scale risk assessment to match the pace of evolving threats. This paper presents an automated, context-aware risk assessment framework that leverages the reasoning capabilities of publicly available Large Language Models (LLMs) to operationalize expert knowledge. Rather than positioning the LLM as the final decision-maker, the framework decouples semantic interpretation from risk scoring authority through a transparent, deterministic Dynamic Metric Engine. Unlike complex closed box machine learning models, our approach anchors the AI's reasoning to this expert-validated metric schema, with weights derived using the Rank Order Centroid (ROC) method from a survey of 101 cybersecurity professionals. We evaluated the framework through a comparative study involving 15 diverse real-world vulnerability scenarios ($C_1 - C_{15}$) and three supplementary sensitivity stress tests ($C_{16} - C_{18}$). The validation scenarios were independently assessed by a cohort of ten senior human experts and two state-of-the-art LLM agents (GPT-4o and Gemini 2.0 Flash). The results show that the LLM-driven agents achieve scoring consistency closely aligned with the human median (Pearson r ranging from 0.9390 to 0.9717, Spearman ρ from 0.8472 to 0.9276) against a highly reliable expert baseline (Cronbach's $\alpha = 0.996$), while reducing the assessment cycle time by more than 100× (averaging under 4 seconds per case vs. a human average of 6 minutes). Furthermore, a dedicated context sensitivity analysis ($C_{13} - C_{15}$) indicates that the framework adapts risk scores based on organizational context (e.g., SME vs. Critical Infrastructure) for identical technical vulnerabilities. Importantly, the system is designed not merely to replicate expert intuition, but to enforce bounded, policy-consistent risk evaluation under predefined governance constraints. Overall, these findings suggest that commercially available LLMs, when constrained by expert-validated metric schemas, can support reproducible, transparent, and real-time risk assessments.

INDEX TERMS Cyber risk assessment, large language models (LLMs), generative AI, rank order centroid (ROC), automated risk scoring, human–AI comparison.

I. INTRODUCTION

Cybersecurity risk management has become a cornerstone of organizational resilience in an era defined by increasing digital dependence and adversarial sophistication, where adaptive, data-driven approaches are widely recognized as essential enablers of effective security strategies [1]. However, many existing risk assessment methodologies struggle to capture unknown vulnerabilities and complex dependency structures, limiting their ability to produce comprehensive and reliable risk evaluations [2]. These limitations are further amplified in cloud computing environments, where dynamic resource allocation, multi-tenancy, and scalability introduce additional uncertainty and complexity into the risk assessment process [3].

Several methodologies have been proposed to address this challenge. Widely adopted scoring approaches such as the Common Vulnerability Scoring System (CVSS),

The associate editor coordinating the review of this manuscript and approving it for publication was Joao Bernardo Ferreira Sequeiros.

the Factor Analysis of Information Risk (FAIR), and the OWASP Risk Rating provide standardized mechanisms for quantifying vulnerabilities. Simultaneously, comprehensive risk assessment frameworks and standards including ISO/IEC 27005, ISO 31000, NIST SP 800-30, and OCTAVE offer structured governance processes. While these approaches have significantly advanced the field, they face a critical ''scalability-precision'' dilemma [1], [2]. Static scoring systems (e.g., CVSS) are fast but context-agnostic, often flagging low-priority issues as high-risk because they ignore organizational defenses [1], [3]. Conversely, qualitative frameworks provide deep contextual awareness but rely heavily on manual expert judgment. This reliance can lead to subjectivity, inconsistency, and slow turnaround in the face of automated threats. As a result, organizations often face a disconnect between risk assessment outputs and the real-time decisions required to prioritize mitigation [1].

The emergence of LLMs offers a potential path forward. Unlike traditional static algorithms, LLMs can interpret unstructured cyber threat intelligence (CTI), incorporate organizational context, and infer risk levels in a manner comparable to a human analyst. However, relying solely on ''closed box'' AI models introduces new risks, including hallucination, limited explainability, and inconsistency. To address this opacity, we introduce the concept of *Systemic Transparency*. In this proposed framework, the LLM is used solely as a semantic processor to interpret unstructured data, while the decision-making authority specifically, the risk scoring logic is decoupled and governed by a transparent, deterministic Metric Engine. Thus, to automate risk assessment safely, Generative AI is constrained by rigorous, expert-validated guardrails.

To address this gap, this paper presents an automated and expert-validated Risk Assessment Framework that leverages the reasoning capabilities of publicly available LLMs (specifically GPT-4o) within a strictly defined metric architecture. The objective of this study is not merely to replicate human expert intuition, but to construct a reproducible, policy-consistent risk scoring architecture validated against expert baselines. Rather than asking an AI to ''assess risk'' in an unconstrained manner, we anchor its reasoning to a precise ''Dynamic Metric Engine''. The parameters and weights of this engine were not arbitrarily chosen; they were derived using the ROC method from a comprehensive survey of 101 cybersecurity professionals. This approach helps ensure that while the AI performs the heavy lifting of data analysis, the decision logic remains grounded in empirically elicited expert judgments.

The proposed framework makes several contributions to the state of the art:

- **Expert-Validated Metric Design:** We introduce a risk scoring model grounded in empirical data, utilizing the ROC method to convert ordinal rankings from 101 industry experts into cardinal weights.
- **LLM-Driven Automation:** We describe how commercial LLMs can be engineered to function as automated risk analysts, parsing heterogeneous inputs (CTI reports, asset profiles) and mapping them to a standardized metric schema with high fidelity.
- **Comparative Validation (Human vs. AI):** Through a comparative case study of fifteen real-world scenarios ($C_1 - C_{15}$), evaluated by a cohort of ten senior human experts and two LLM agents (GPT-4o and Gemini 2.0 Flash), we provide empirical evidence of close alignment ($\sigma \approx 0.15 - 0.25$) while reducing the assessment cycle time by over 100x.

The remainder of this paper is organized as follows. Section II reviews related work and discusses the limitations of existing methodologies. Section III details the methodology, specifically the survey design and ROC weight derivation. Section IV introduces the proposed framework architecture and the AI agent workflow. Section V presents the comparative case study and evaluation results. Section VI analyzes the behavioral characteristics of the models, including algorithmic bias and limitations. Finally, Section VII concludes the paper and discusses directions for future research.

II. RELATED WORKS

Cybersecurity risk assessment has evolved along several methodological lines, ranging from governance-focused guidelines to standardized scoring frameworks and more recent efforts to incorporate contextual and AI-driven elements. In this section, we review representative approaches, summarize their limitations, and identify research gaps that motivate our work.

A. STANDARDIZED SCORING FRAMEWORKS

In parallel, quantitative scoring systems such as the Common Vulnerability Scoring System (CVSS) [4] and the Factor Analysis of Information Risk (FAIR) [5] have gained wide adoption. CVSS, across its v2, v3.1, and v4.0 versions, standardizes vulnerability scoring using parameters such as attack vector, privileges required, and user interaction. FAIR focuses on frequency and magnitude to estimate risk in financial terms. The Common Weakness Scoring System (CWSS) [6] extends assessment to software flaws. While these models provide more repeatable outcomes than qualitative approaches, they remain essentially static: they do not integrate live cyber threat intelligence (CTI), they neglect organizational context, and their weighting mechanisms are predefined rather than adaptive.

As analyzed in Table 1, while existing academic models cover traditional dimensions such as Exploitability and Impact, they show limited coverage of NLP/LLM integration and organizational-context awareness. Specifically, only a minority of studies incorporate CTI feeds directly, and few (if any) leverage Generative AI for automated reasoning, highlighting the gap this framework aims to address.

B. TRADITIONAL RISK ASSESSMENT METHODOLOGIES

Established methodologies such as NIST SP 800-30 [7], ISO/IEC 27005 [8], ISO 31000 [9], and OCTAVE [10] provide structured guidance for identifying, analyzing, and evaluating risks. They remain foundational in compliance and governance, serving as widely recognized standards for security and risk management programs. However, their reliance on expert judgment renders them largely subjective, resulting in inconsistencies and limited reproducibility across organizations. Tools such as 5×5 risk matrices exemplify these shortcomings: likelihood and impact are qualitatively defined and then mapped into high-, medium-, and low-risk categories without transparent sub-metrics. Such outputs are valuable for high-level reporting but insufficient for dynamic, data-driven decision-making in rapidly evolving threat environments.

C. CONTEXTUAL AND EMERGING APPROACHES

Recent research has sought to enrich vulnerability scoring by incorporating attacker modeling, dependency analysis, and business context. Examples include time-sensitive probabilistic risk assessment models [11] and adversary behavior modeling based on MITRE ATT&CK [12]. Despite these advances, many approaches remain siloed: they address specific gaps (e.g., temporal dynamics or adversary behavior) but do not provide a unified architecture that fuses threat intelligence, organizational context, automated computation, and human-in-the-loop (HITL) governance.

For instance, Wang et al. [13] proposed an ISA evaluation framework based on a hybrid AHP–TOPSIS method to assess and rank IoHT devices according to predefined security attributes. While their work illustrates the applicability of multi-criteria decision-making techniques in a healthcare-specific IoT context, it remains largely static and domain-bound. In contrast, the present framework generalizes this line of work beyond IoT-specific environments by integrating (i) LLM-assisted contextual extraction grounded in a survey-derived ROC weighting scheme, (ii) a HITL feedback loop with periodic recalibration, and (iii) cross-domain contextual data fusion to support dynamic cybersecurity risk assessment.

Table 2 provides a comparative overview of widely used frameworks and situates our proposed model in relation to them. Furthermore, in the domain of CTI sharing, approaches have emerged that model the process using cooperative game theory and the Shapley value, where the benefit distribution mechanism is adjusted by contextual parameters (e.g., a risk coefficient) to promote robust and fair outcomes [14].

In 2025, the literature indicates a growing emphasis on integrating Machine Learning (ML) techniques to enhance cybersecurity detection capabilities. Femi and Madu [15] reported that AI-driven security systems can reduce incident rates, highlighting cases where preemptive ransomware blocking achieved success rates of up to 98%. To operationalize such capabilities in cloud environments, Jamili et al. [16] proposed a hybrid framework combining Random Forest classifiers with autoencoder-based anomaly detection, reporting an accuracy of 95.3%. However, their approach relies on supervised learning components that require large volumes of labeled data, which are often difficult to obtain in real-time corporate settings.

A major challenge in these ML-based approaches is the closed box nature of the models. Islam et al. [17] attempted to address this issue by combining neural networks with Explainable AI (XAI) techniques such as SHAP and LIME, improving interpretability at the cost of increased computational complexity. Similarly, Malik et al. [18] employed Artificial Neural Networks (ANN) in conjunction with Interpretive Structural Modeling (ISM); however, their reliance on subjective expert surveys rather than objective metrics introduces potential bias into the training process. On the simulation front, Camacho et al. [19] proposed a framework based on Adversarial Risk Analysis (ARA). While effective in mitigating financial risks, their approach requires substantial modeling effort and extensive domain expert involvement, which limits its scalability for automated operations.

Recent studies have demonstrated the effectiveness of NLP- and LLM-based models for automating cyber threat extraction and contextual profiling. For instance, Hmimou et al. [20] introduced a multi-agent LLM framework that semantically correlates threat entities across unstructured data sources, while Mondal et al. [21] applied NLP-driven classification and entity recognition techniques to identify emerging threats and generate contextual profiles. These approaches align with the objectives of the proposed NLP/LLM layer by enabling dynamic, text-based intelligence generation within cybersecurity systems.

D. RESEARCH GAPS AND POSITIONING

From this survey, two persistent gaps emerge. First, most existing approaches face a *scalability-precision dilemma*: they are either static and fast yet context-agnostic (e.g., scoring systems), or context-aware yet labor-intensive and slow (e.g., qualitative frameworks) [22].

Second, while the potential of LLMs in cybersecurity is recognized, their integration into quantitative risk assessment remains largely exploratory. The current literature offers limited methodological guidance on how to constrain Generative AI using empirical expert knowledge to mitigate hallucinations. Furthermore, as observed in recent ML-based studies [16], [17], traditional models often exhibit closed box behavior or require extensive training datasets, while simulation-based frameworks [19] demand substantial modeling effort.

Our framework is designed to address these gaps by combining (i) multi-source data ingestion, (ii) automated reasoning via publicly available LLMs (GPT-4o), and (iii) a transparent, survey-derived metric engine. Unlike closed box approaches, we use the ROC method to anchor AI-driven

TABLE 1. Comparative summary of academic cyber risk assessment studies.

Study	Traditional & Contextual Risk Dimensions	CMW	CTI	NLP/LLM	AI/ML
[24]	EXP, ATT, IMP, BIZ, CTX	Yes	No	No	No
[25]	EXP, IMP, CTX	Yes	No	No	No
[26]	EXP, ATT, IMP, CTX	Yes	No	No	No
[27]	EXP, ATT, IMP, BIZ, CTX, ORG	Yes	Yes	No	No
[28]	EXP, ATT, IMP, CTX	Yes	No	No	No
[29]	EXP, ATT, IMP, CTX	Yes	No	No	No
[30]	EXP, ATT, IMP, CTX	No	No	No	No
[31]	EXP, ATT, IMP, CTX	Yes	No	No	No
[32]	EXP, ATT, IMP, CTX	No	Yes	No	Yes
[33]	IMP, BIZ, CTX, ORG	Yes	No	No	No
Our Framework	EXP, ATT, IMP, BIZ, CTX, ORG	Yes	Yes	Yes	Yes

Legend – Traditional & Contextual Risk Dimensions:
EXP: Exploitability, **ATT:** Attacker Modeling, **IMP:** Technical Impact, **BIZ:** Business/Regulatory Impact, **CTX:** Asset/Target Context Awareness, **ORG:** Organizational Factors.

Legend – AI-Driven & Data-Enriched Capabilities:
CMW: Customizable Metric Weighting, **CTI:** Threat Intelligence Integration, **NLP/LLM:** Natural Language Processing / Large Language Models, **AI/ML:** Machine Learning.

TABLE 2. Comparative overview of widely used risk assessment frameworks and the proposed adaptive model.

Criteria / Dimension	NIST SP 800-30	FAIR	OCTAVE	Proposed Framework
Scope & Purpose	Identify–Analyze–Evaluate	Quant. (freq×mag)	Scenario-based identification	Full lifecycle (Automated)
Calculation Approach	Qual./semi L×I	Freq×Mag	Qual. scenarios	LLM Extraction → ROC
Primary Parameters	Threat/vuln/impact	Actor/asset/controls	Asset/process/threats	30+ metrics + context
Context Awareness	Medium	High	Med–High	Very High (LLM-Driven)
Attacker Modeling	Limited	Partial	None	Detailed (13+ metrics)
Organizational Factors	Partial	Limited	High (process)	Fully integrated
CTI Integration	None	Indirect	None	Full integration
AI/ML Capabilities	None	None	None	Generative AI (GPT-4o)
NLP/LLM Capabilities	None	None	None	Core Engine
Adaptability Mechanism	None	None	None	Context-Aware (LLM)
Feedback Mechanism	None	None	None	HITL feedback loop
Automation Level	Manual	Partial	Manual	High (>100x Speedup)
Customizability	Low–Med	High	High	Very High
Treatment Layer Support	Indirect	Indirect	Indirect	Native

analysis to human-validated weights. As summarized in Tables 1 and 2, the proposed model provides an approach that is both *automated* and *expert-validated*, aiming to bridge the gap between static scoring and manual assessment.

III. METHODOLOGY: EXPERT-DRIVEN METRIC DESIGN

This section provides the methodological foundation of the proposed framework by describing its expert-informed quantification model. Rather than relying on "closed box" AI approaches with opaque decision logic, our framework uses a transparent, survey-derived metric engine. While the architecture in Section IV presents the system design, this section details the methodology for weight elicitation and parameter calibration to support reproducibility, explainability, and systematic improvement over static scoring approaches. To provide a standardized foundation for interpreting technical inputs, the framework integrates the MITRE ATT&CK knowledge base. MITRE ATT&CK is a globally recognized repository of adversary tactics, techniques, and procedures (TTPs) based on real-world observations [34]. In our methodology, it serves as a semantic anchor for the LLM to categorize vulnerability phases, ensuring that the input parameters for the Metric Engine are grounded in established cybersecurity taxonomies.

A. SURVEY DESIGN FOR PARAMETER AND METRIC WEIGHT ELICITATION

1) PURPOSE OF THE SURVEY

The primary objective of this survey was to shift the risk scoring methodology from a subjective, heuristic-based

approach toward a more objective, data-driven framework. Accordingly, the survey was designed to:

- Derive empirical importance rankings for the risk parameters defined in the framework (see Table 3), enabling the computation of baseline weights via the ROC method.
- Calibrate the severity scores of sub-metrics on a standardized scale to promote internal consistency within the risk calculation component.
- Establish an expert-derived baseline for initial deployment, reducing reliance on single-analyst judgment and supporting more consistent assessments.

2) SURVEY STRUCTURE AND QUESTION TYPES

The survey instrument was structured into three sections to capture different dimensions of expert knowledge (see Fig. 1):

- **Participant Profile:** Questions on the respondent's role (e.g., CISO, analyst), industry sector, and years of experience were collected to characterize expertise levels.
- **Parameter Ranking (for ROC):** Respondents were asked to rank parameters within the *Likelihood* and *Impact* groups from "Most Critical" to "Least Critical." These ordinal rankings served as direct input for ROC weight derivation.
- **Metric Scoring (0 to 10):** Respondents assigned a numerical severity score between 0 and 10 to specific sub-parameter conditions (e.g., "Exploit Available" vs. "No Patch").

B. PARAMETER GROUPS AND RANKING SCHEME

Building on our prior work [23], we employ a high-granularity metric structure, referred to as the "Metric Box," to represent the multidimensional nature of cyber risk. Accordingly, the survey design was divided into two primary dimensions: *Likelihood* and *Impact*. To derive weights via the ROC method, we used a hierarchical ranking strategy designed to promote consistent expert judgments:

- **Cognitive Priming (Intra-Group Ranking):** Participants first prioritized parameters within logical subgroups (e.g., *Attack Surface*, *Threat Actor Capabilities*). This step served as a preparatory phase, enabling experts to evaluate the relative importance of related factors in isolation before performing the more complex global ranking.
- **Global Ranking:** Following the priming phase, participants ranked all parameters within the *Likelihood* dimension (19 items) and the *Impact* dimension (12 items). The resulting rankings were used as the primary input for ROC weight calculation, as they capture each expert's holistic assessment of the risk landscape.

Box 1: Representative Questions from the Expert Survey

The survey instrument consisted of three distinct sections designed to elicit expert knowledge for the Metric Engine.

Type 1: Demographics & Expertise Profile
Q: "Which of the following best describes your primary role in cybersecurity?"

- ☐ CISO / Manager
- ☐ Security Architect / Engineer
- ☐ Risk Analyst / Consultant
- ☐ SOC Analyst / Incident Responder

Type 2: Parameter Ranking (For ROC Weights)
Q: "Please rank the following **Threat Actor Capabilities** from 'Most Critical' (1) to 'Least Critical' (6) based on their contribution to the **likelihood** of a successful breach."

1) Resources (Time, Funding)
2) Technical Skills
3) Motivation
4) Access Rights (Insider)

(Interface: Drag-and-drop orderable list)

Type 3: Metric Calibration (0-10 Scale)
Q: "On a scale of 0 to 10, how would you rate the severity of the following **Attack Vector** conditions?"

- **Network:** Exploitable remotely. *[Slider: 0-10]*
- **Local:** Requires shell access. *[Slider: 0-10]*
- **Physical:** Physical interaction. *[Slider: 0-10]*

FIGURE 1. Sample questions from the survey instrument illustrating the data collection strategy for demographics, parameter ranking (ROC), and metric calibration.

C. METRIC SCORING METHOD AND LOGIC

While the ranking process determines the weight of each parameter, quantitative risk calculation requires calibrated severity values for each sub-parameter condition. To obtain these values, experts assigned scores on a 0–10 scale based on calibration rules provided in the survey:

- **Baseline (Score 0):** Represents a condition in which risk is considered mitigated to the lowest level defined by the schema (e.g., *Patch Availability: Fully Patched*). Experts were instructed to assign 0 when a condition minimizes the corresponding Likelihood or Impact contribution.
- **Ceiling (Score 10):** Represents the theoretical worst-case condition, i.e., the maximum risk contribution for that parameter (e.g., *Patch Availability: No Patch / Zero-Day*).
- **Interval Calibration:** Intermediate scores reflect the relative magnitude of risk between these two extremes. This approach allows the framework to capture granular differences (e.g., the severity gap between "Local Access" and "Network Access") rather than relying on binary inputs.

D. DATA COLLECTION AND SURVEY ADMINISTRATION

Data collection was administered using the Jotform online survey platform, selected for its interface capabilities and available security features. The survey was distributed to the target expert group via a direct access link. To support data quality and reduce cognitive load during the ranking

tasks, a specialized "Orderable List" widget with a drag-and-drop interface was used. This mechanism offered several methodological advantages:

- **Holistic Visibility:** Participants could view the full set of parameters simultaneously on a single screen, reducing the memory decay and attentional drift often associated with multi-page or dropdown-based ranking formats.
- **Dynamic and Rapid Comparison:** The drag-and-drop interface enabled experts to adjust priorities by moving items up or down. This facilitated comparisons between adjacent parameters, allowing participants to refine the hierarchy until it reflected their professional judgment.

E. WEIGHT DERIVATION APPROACH

To translate expert judgments into a scoring model, we employed a dual-layer calibration approach: ROC for parameter weights and a normalization-and-anchoring technique for sub-parameter scores.

1) PARAMETER WEIGHTING (WF) VIA ROC

The weights (W_F) for the main parameters (e.g., Attack Vector, Confidentiality Impact) were derived using the ROC method based on the foundational ranking principles established by Barron and Barrett [35]. This approach converts ordinal rankings collected from experts into cardinal weights and is commonly used to reduce biases associated with direct scoring. The ROC weight w_k for the k-th most important item among M items is calculated as:

$$w_k = \frac{1}{M} \sum_{i=k}^{M} \frac{1}{i} \tag{1}$$

For reproducibility, parameters were categorized into two sets: Likelihood (L) and Impact (I), as denoted in the final column of Table 3. The ROC-derived weights (w_k) were normalized independently within these sets such that the sum of weights for each dimension equals unity ($\sum_{j\in\mathcal{L}} w_j \approx 1.0$ and $\sum_{j\in\mathcal{I}} w_j \approx 1.0$).

2) DETERMINISTIC RISK SCORE COMPUTATION

Let $s_j \in [0, 1]$ denote the *normalized* severity score selected for metric j, and let w_j denote the corresponding *ROC*-derived weight. Here, *Likelihood* refers to the likelihood of successful exploitation.

We compute the *Likelihood* (L) and *Impact* (I) components as:

$$L = \sum_{j\in\mathcal{L}} w_j\, s_j, \qquad I = \sum_{j\in\mathcal{I}} w_j\, s_j, \tag{2}$$

where $\sum_{j\in\mathcal{L}} w_j = 1$ and $\sum_{j\in\mathcal{I}} w_j = 1$.

The final risk score is computed by the Metric Engine as:

$$R = \frac{L + I}{2}. \tag{3}$$

While this study employs an equal-weighting strategy ($w_p = w_i = 0.5$) to establish a standardized baseline, the framework is designed to support adaptive weighting. Organizations may calibrate the contribution of Likelihood versus Impact to align with their specific risk appetite. For instance, a safety-critical entity may assign a higher coefficient to the Impact dimension, whereas an organization focused on threat prevention may prioritize Likelihood. Mathematically, the framework supports the generalized form $R = \alpha L + (1-\alpha)I$, enabling tailored governance.

Under the operational bounds defined by the metric schema (e.g., sector sensitivity, regulatory exposure, and baseline organizational factors), both dimensions can retain positive lower bounds; accordingly, $L > 0$ and $I > 0$, and thus $R > 0$ for the evaluated cases.

For readability, we also report the score on a 0–10 scale as $R_{10} = 10R$.

3) SUB-PARAMETER METRIC CALIBRATION

Scoring values for individual sub-parameters were calibrated using a survey-based dataset ($n > 100$) consisting of expert ratings on a 0–10 impact scale. The calibration process involved three steps:

- **Aggregation:** The arithmetic mean (μ) of expert ratings was computed for each metric option.
- **Normalization:** These mean values were mapped to the [0.0, 1.0] interval ($W = \mu/10$) to align with the framework's computational logic.
- **Theoretical Anchoring:** To support coverage of boundary conditions not captured in the survey data, "Anchor Points" were introduced (e.g., setting "Safety-Critical Impact" to 1.0).

The final calibrated weights, presented in Table 3, summarize the prioritization of risk factors in the expert cohort. For instance, within the Impact dimension, Confidentiality Impact ($W_F = 0.2586$) received the highest weight, exceeding Technical Impact ($W_F = 0.0321$) by nearly 8×. Similarly, in the Threat Actor category, Resources ($W_F = 0.1867$) was weighted higher than Motivation ($W_F = 0.0771$), suggesting that the cohort prioritized attacker capability over intent when scoring risk. These findings indicate that while the framework integrates multiple risk dimensions, the final scores are most strongly influenced by the Confidentiality Impact and Industry Sensitivity parameters, ensuring that the automated assessment remains aligned with organizational and sectoral priorities.

F. OPERATIONAL BOUNDS AND SAFETY-FIRST MODELING PRINCIPLES

A core design axiom of our framework is that, in operational environments, the residual risk of an active asset is modeled as positive ($R > 0$). This is a deliberate design choice aligned with the framework's "Safety-First" architecture rather than a purely theoretical assumption. In this setting, assigning a risk score of zero (0.00) would correspond to a "zero maintenance" or "zero threat" state, which is difficult to

justify given continuous monitoring requirements and the likelihood of unknown vulnerabilities (zero-days).

Let $S_{\min}(k)$ be the minimum possible severity score for parameter k. As defined in the expert-derived schema (see Table 3 and Appendix A), selected contextual parameters introduce a non-zero operational floor intended to quantify *residual risk*.

For the Impact (I) dimension, the "Industry Sensitivity" parameter introduces a base score reflecting the sector's criticality:

$$I_{\min} \geq w_{\text{sens}} \cdot S_{\min}(\text{Sens}) = 0.1753 \times 0.20 = 0.035 \quad (4)$$

where 0.20 corresponds to the "Low" sensitivity option defined in Table 3, representing baseline regulatory exposure.

Similarly, for the Likelihood (L) dimension, technical parameters such as "Patch Availability" retain a residual score of 0.25 even in the "Fully Patched" state. In this framework, the "Fully Patched" status is treated as a functional risk-equivalent state. It represents the baseline residual risk achieved when a vulnerability is effectively neutralized through software remediation or equivalent compensating controls, such as physical isolation or virtual patching, that reach the same operational floor. This design choice models the cost of verification and the potential for bypasses even in mitigated states:

$$L_{\min} \geq w_{\text{patch}} \cdot S_{\min}(\text{Patch}) = 0.1078 \times 0.25 = 0.027 \quad (5)$$

Accordingly, the global lower bound for the Risk Score (R) remains positive, establishing a minimum vigilance threshold:

$$\lim_{\text{mitigation} \to \infty} R = \frac{L_{\text{residual}} + I_{\text{residual}}}{2} > 0 \quad (6)$$

This operational floor helps prevent the model from driving risk to zero solely due to a single variable (e.g., classifying a scan result as a "False Positive"). Instead, it constrains the final score to reflect baseline risk associated with sector context and asset exposure, consistent with the principle of "Defense in Depth."

G. PARTICIPANT DEMOGRAPHICS AND EXPERTISE PROFILE

To support the reliability of the derived risk weights, the survey was administered to a cohort of cybersecurity professionals ($N = 101$). As detailed in Fig. 2, the participant distribution reflects the study's practitioner-focused sample:

- **Seniority:** As shown in Fig. 2(a), 85.1% reported more than five years of experience, and approximately 40% reported senior-level tenure (10+ years).
- **Decision-Making Power:** Fig. 2(b) indicates that 78.2% were technical specialists (engineers/analysts), while 21.8% held managerial or CISO roles, suggesting that the weights capture both tactical and strategic perspectives.
- **Sectoral Diversity:** Fig. 2(c) shows representation from highly regulated industries, with Finance & Banking (22.8%) and the Public Sector (16.8%) constituting the largest groups.

H. ETHICAL CONSIDERATIONS AND INFORMED CONSENT

Participation in the expert survey was entirely voluntary. Prior to data collection, all participants were informed about the purpose of the study and provided their informed consent. No personally identifiable information (PII) was collected, and all responses were recorded and analyzed anonymously. The survey focused exclusively on professional expertise and technical judgment related to cybersecurity risk assessment. Given the non-sensitive nature of the data and the absence of human subject experimentation, formal institutional ethics committee approval was not required.

IV. PROPOSED FRAMEWORK ARCHITECTURE

The proposed framework introduces a modular, layered design for dynamic cyber risk assessment. Unlike static methodologies that rely on manual inputs, the architecture supports automated ingestion, reasoning, and scoring using publicly available LLM agents (GPT-4o and Gemini 2.0 Flash) constrained by a survey-derived metric schema. As illustrated in Fig. 3, the system consists of five interoperable layers.

A. OVERALL ARCHITECTURE

As illustrated in Fig. 3, the proposed framework aligns with the ISO 31000 Risk Management Guidelines and defines a structured workflow from data ingestion to risk treatment. The system comprises five interoperable layers:

1) **Data Layer (Identification & Collection):** This foundational layer is designed to aggregate heterogeneous data from disparate sources to establish a comprehensive risk context. While the architecture supports the ingestion of organizational documents, compliance requirements, CTI reports, and classification schemas, for the purpose of this validation study, data collection was deliberately scoped to vulnerability reports and asset profiles. This intentional scoping serves to prioritize experimental control and enhance reproducibility by minimizing variable noise during the AI reasoning phase.
2) **Feed Layer (Validation & Normalization):** Serving as the pre-processing engine, this layer consolidates raw inputs into a structured, text based format suitable for the AI agent's context window. As detailed later in the workflow (see Fig. 4, Phase 1), this process transforms scattered data into normalized prompt contexts (e.g., `vuln.txt` and `org_context.txt`). This approach suggests a more token-efficient method by providing clean text based context injection, while structured JSON formats are strictly reserved for the agent's output and the deterministic metric schema definitions.

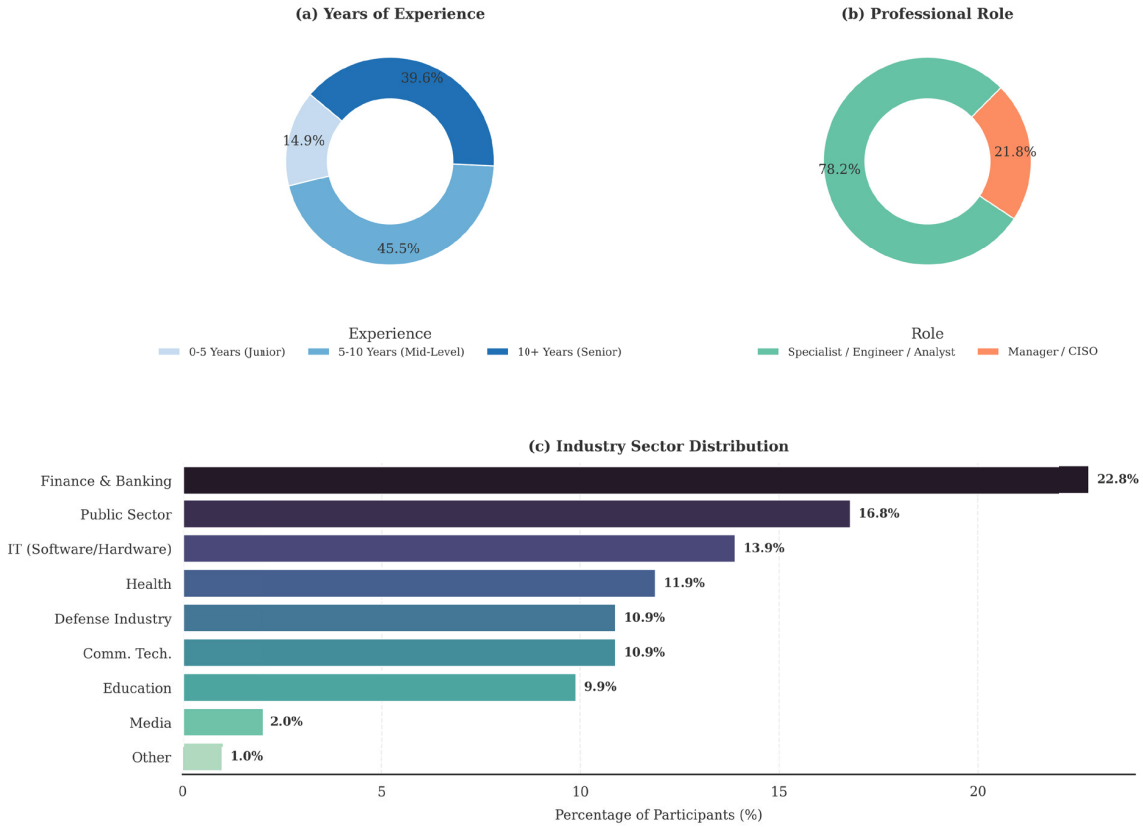

FIGURE 2. Demographic profile of the survey participants ($N = 101$). (a) Experience levels indicate a high degree of seniority. (b) The role distribution includes technical specialists and managerial decision-makers. (c) Sectoral distribution shows strong representation from highly regulated industries.

3) **Computation Layer (Risk Calculation):** At the core of the framework, the LLM reasoning agent analyzes the prepared context, selects appropriate sub-parameters, and generates a JSON output. The final risk score is then computed using the expert-derived weights defined in Table 3. This layer also incorporates a HITL mechanism: the Risk Owner can provide feedback to the agent and, when necessary, manually adjust metric selections or refine weights to address potential AI misinterpretations.
4) **Evaluation Layer (Comparison & Prioritization):** Once the calculation is finalized, the identified risks are registered in the central Risk Registry. The system prioritizes these risks based on the computed severity scores and the organization's defined risk appetite, supporting structured remediation planning.
5) **Treatment Layer (Response):** In the final phase, the framework presents prioritized risks to the Risk Owner, who selects the appropriate response strategy (Mitigate, Transfer, Accept, or Avoid). This decision triggers the generation of remediation workflows.

B. DATA LAYER: IDENTIFICATION AND COLLECTION

The Data Layer forms the foundation of the framework. It collects structured and unstructured sources relevant to risk assessment:

- **Context Documents:** Organizational policies, SLAs, and business-criticality factors.
- **CTI Feeds:** Threat intelligence reports (e.g., MISP, OpenCTI) providing indicators of compromise and adversary profiles.
- **MITRE ATT&CK:** Tactics, techniques, and procedures (TTPs) that inform behavioral threat modeling [12].
- **Incident Reports & Asset Inventory:** Dynamic records describing asset criticality, exposure, and active security events.

C. FEED LAYER: VALIDATION AND NORMALIZATION

The Feed Layer acts as a buffer between raw inputs and the AI engine. Its primary function is preprocessing. Before data are provided to the LLM, raw texts are cleaned and heterogeneous logs are converted into a unified JSON schema. This allows the LLM to consume structured context injection rather than noisy raw data, which can reduce the risk of hallucination.

D. COMPUTATION LAYER: LLM-DRIVEN REASONING

At the heart of the framework lies the Computation Layer, which substitutes traditional manual analysis with an AI reasoning agent. Rather than relying on end-to-end, closed box deep learning models, we employ a

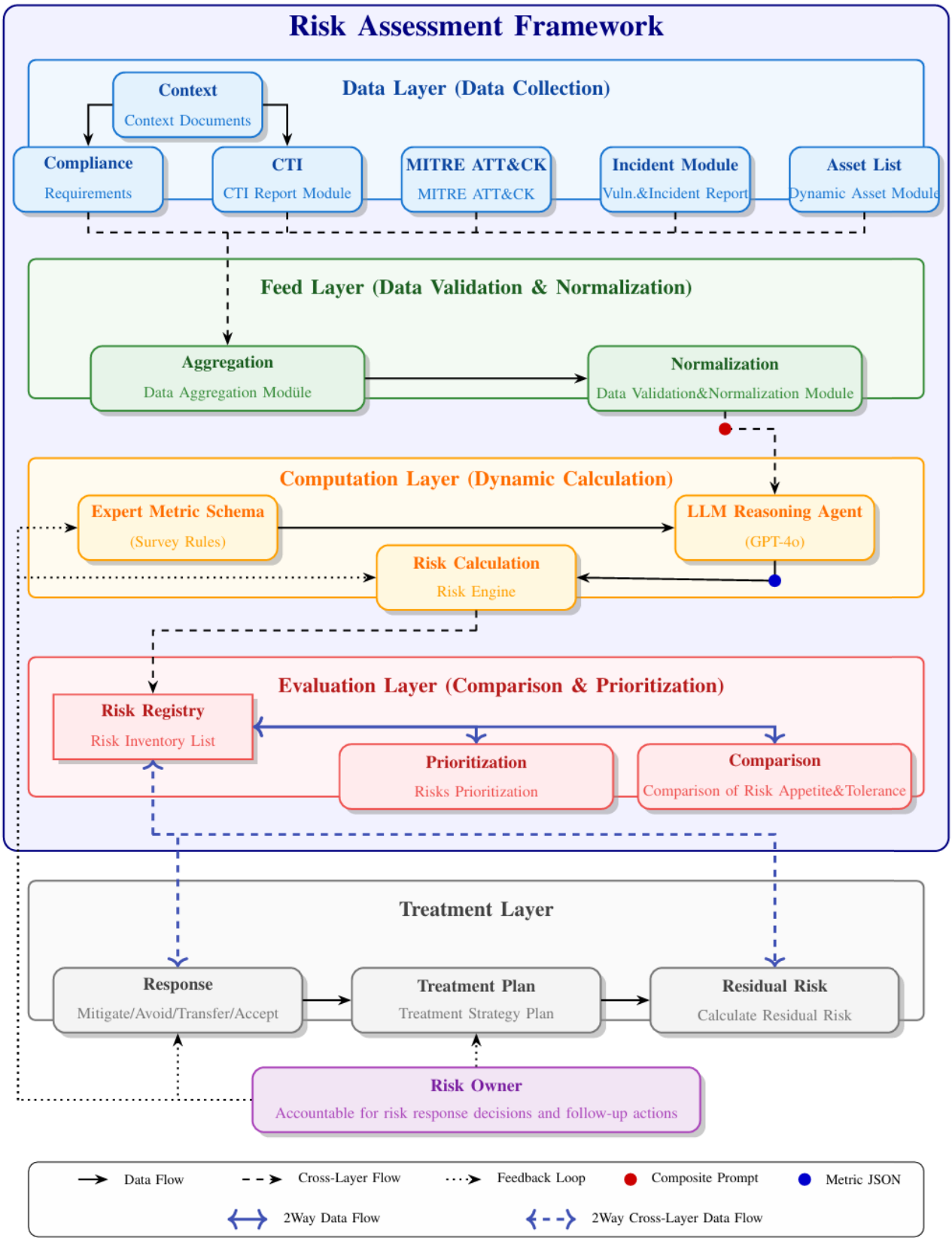

FIGURE 3. Overall Dynamic Risk Assessment Framework. The architecture integrates a Data Layer for collection, a Feed Layer for normalization, a Computation Layer featuring an LLM Reasoning Agent guided by an Expert Metric Schema, an Evaluation Layer for ROC-based scoring, and a Treatment Layer for governance.

context-injection workflow (akin to RAG in spirit), in which the model receives a fixed, pre-processed context package; notably, no external retrieval or tool use is enabled during evaluation.

A key distinction in this work is between *Algorithmic Opacity* and *Systemic Transparency*. While the underlying LLM (GPT-4o or Gemini 2.0 Flash) is a commercial ''closed box'' with respect to its internal weights, the framework confines its role to *semantic extraction*, i.e., mapping unstructured text to predefined categories. The risk calculation logic, however, remains transparent and deterministic, governed by the Metric Engine. This architecture separates the semantic processing component (the ''reader'') from the scoring logic (the ''judge''), which helps mitigate explainability limitations typically associated with end-to-end deep learning models.

The AI agent follows a structured four-phase pipeline that maps raw inputs to cardinal risk scores (see Fig. 4).

1) PHASE 1: PRE-PROCESSING (FEED LAYER)

Raw inputs are first normalized to mitigate token noise and facilitate formatting consistency. In this phase, heterogeneous vulnerability logs are converted into a standardized `vuln.txt` (see Appendix B, Fig. 14), while organizational assets are summarized in `org_context.txt` (see Appendix B, Fig. 15). This process indicates a targeted strategy for semantic extraction, providing the AI agent with structured text-based prompt contexts instead of raw data streams. To support reproducibility and maintain strict experimental control, the pre-processing scope in this study was deliberately restricted to these primary documents, suggesting a more reliable baseline for the subsequent AI reasoning phase.

2) PHASE 2: CONTEXT INJECTION

The system constructs a composite prompt by injecting four elements into the LLM context window:

- **System Persona:** Defining the AI as a ''Senior Cyber Risk Analyst.''
- **Metric Schema:** The JSON definition of the ''Metric Box'' (Table 3) and the scoring logic.
- **Context Data:** The pre-processed text files from Phase 1.
- **Logic Overrides:** A set of deterministic rules for handling ambiguous edge cases (e.g., air-gapped assets) to support safety-first behavior (see Appendix A).

3) PHASE 3: AI REASONING & CONSERVATIVE PROTOCOL

The LLM (GPT-4o or Gemini 2.0 Flash) analyzes the injected context and maps ambiguous information to structured sub-parameters. The agent follows a Conservative Estimation Protocol (see Fig. 4), under which it is constrained to select the higher-severity option in cases of ambiguity to reduce the risk of false negatives.

4) PHASE 4: OUTPUT & CALCULATION

The agent outputs a structured `Result JSON` containing only the selected metric keys. The final calculation is performed externally by the ''Math Engine'' using the ROC weights (Eq. 1), which avoids LLM-performed arithmetic and reduces the risk of numerical errors.

E. EVALUATION LAYER: SCORING AND PRIORITIZATION

Once the AI agent selects the appropriate sub-parameters (e.g., Attack Vector: Network, Impact: High), the Evaluation Layer computes the final risk score using the expert-derived ROC weights (Eq. 1). This computation is performed outside the LLM to avoid model-performed arithmetic.

F. TREATMENT LAYER: RESPONSE AND GOVERNANCE

Following assessment, the Treatment Layer supports risk management by presenting response options (Mitigate, Transfer, Accept). While the LLM may suggest treatment actions based on the identified vulnerability, the final decision remains under HITL governance. A designated Risk Owner reviews the AI-generated scorecard and authorizes the response, supporting accountability.

V. COMPARATIVE VALIDATION: HUMAN EXPERTS VS. AI AGENT

To assess the robustness and scalability of the proposed framework, we conducted a comparative study. Unlike prior work that relies solely on synthetic data, we evaluated the AI agent against a cohort of senior human experts using real-world vulnerability scenarios. The objective was to assess the framework across three dimensions: scoring accuracy, consistency, and operational efficiency.

A. EXPERIMENTAL SETUP AND DATASET

As detailed in Table 4, the study used a dataset of fifteen scenarios ($C_1 - C_{15}$) stratified across high-impact domains ranging from Critical Infrastructure (C_8, C_{15}) to SMEs (C_{13}) to provide a basis for evaluating context adaptability. The dataset was further divided into two validation streams:

- **Historical Incident Recreation ($C_1 - C_{12}$):** A diverse set of widely documented real-world cybersecurity incidents spanning critical sectors, including Finance, Defense, Energy, and Healthcare. This stream combines high-impact ransomware events (e.g., Colonial Pipeline, C_8) with complex supply-chain attacks and logic flaws (e.g., SolarWinds, Optus) to compare model outputs against historically reported outcomes.
- **Comparative Context Analysis ($C_{13} - C_{15}$):** A controlled comparative study that applies a *real-world technical vulnerability* (unpatched RCE) across three distinct organizational contexts (SME vs. Large Enterprise vs. Critical Infrastructure). This stream evaluates how the framework adapts risk scores for the same technical flaw under varying contextual and operational conditions. To ensure methodological clarity, the technical

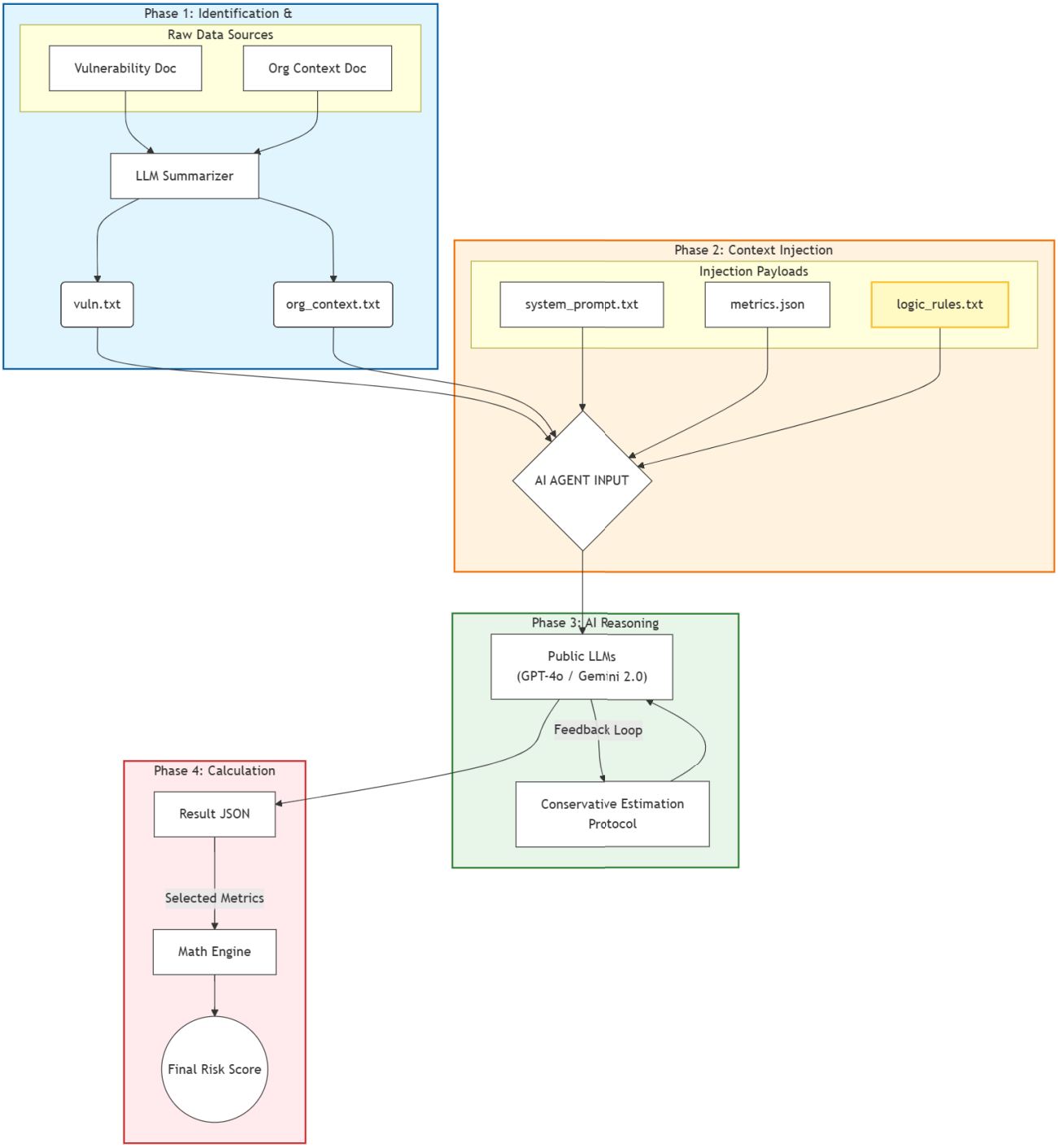

FIGURE 4. AI Agent Risk Assessment Workflow. The process involves pre-processing raw data into structured summaries, injecting them alongside metric definitions into the AI agent, and applying a conservative fallback protocol before external mathematical calculation.

vulnerability input remains constant across C_{13}–C_{15}, while only organizational attributes and operational constraints vary. Any variation in the final risk score can therefore be attributed to the framework's interpretation of business criticality and environmental reality. In C_{13} specifically, the observed reduction in risk score reflects both SME classification and the activation of the air-gap override rule. This scenario intentionally evaluates the dominance of physical isolation over raw technical severity, ensuring that deterministic governance constraints supersede context-agnostic vulnerability reports when exploitation is operationally infeasible.

TABLE 3. Risk calculator metrics box used in Phase 1 (Pilot) with survey-derived weights.

Main Parameters	Description	Subparameters with Scores	WF	(L/I)
Attack Surface and Exploitability				
Attack Vector	Remote, local, or physical access requirements	Physical (0.29), Local (0.56), Adjacent Network (0.78), Network (0.99)	0.1341	L
Privileges Required	User privilege level required	High (0.39), Low (0.91), None (1.00)	0.0665	L
User Interaction Required	Whether user interaction is needed	Required (0.65), None (1.00)	0.0437	L
Exploit Chain	Whether the vulnerability can be exploited in a chain	Controlled (0.53), Uncontrolled (0.91)	0.0121	L
Authentication Complexity	Extra authentication needed for exploitation	Multi-factor (0.17), Two-Factor (0.47), Single-factor (0.99), None (1.00)	0.0902	L
Patch Availability	Whether a patch is available for the vulnerability	Fully Patched (0.25), Partial Patching (0.88), No Patch (1.00)	0.1078	L
Exposure Duration	How long the vulnerability has been known/exploited	0–30 Days (0.63), 1–6 Months (0.80), 6+ Months (0.95)	0.0057	L
Exploit Maturity	Exploit Code Maturity	Feasible Exploit (0.72), Active Exploit (1.00)	0.0193	L
Threat Actor and Capabilities				
Skills Required	Attacker's technical skill level	Proficient (0.31), Adept (0.50), Operational (0.74), Minimal (0.91), None (1.00)	0.0503	L
Resources	Resources required for the attack	Individual (0.40), Team (0.69), Organization (0.89), Government (0.90)	0.1867	L
Objectives	Intended damage or data exfiltration level	Copy (0.69), Damage (0.78), Destroy/Deny (0.82), Take (0.85), All of the Above (0.98)	0.0378	L
Limits	Legal, technical, or ethical limitations of the attacker	Code of Conduct (0.22), Legal (0.50), Extra-Legal (0.91)	0.0028	L
Visibility	Likelihood of detection	Overt (0.39), Covert (0.87)	0.0088	L
Automation Level	Degree of automated attack execution	Manual (0.44), Scripted (0.87), Fully Automated (0.95)	0.0326	L
Motivation Type	Attacker's primary goal	Non-Hostile (0.18), Hacktivism (0.63), Insider Threat (0.75), Espionage (0.76), Financial (0.84)	0.0771	L
Supply Chain Attack	Use of supply chain or island-hopping techniques	No (0.59), Yes (0.87)	0.0278	L
Target Profile and Exposure				
Brand Value	Importance of the organization's reputation	Low (0.21), Medium (0.49), High (0.77), Critical (0.93)	0.0234	L
Employee Awareness	Security awareness level of employees	Advanced (0.24), Intermediate (0.57), Basic (0.94), None (1.00)	0.0156	L
Maturity Level	Organization's cybersecurity maturity level	Optimized (0.12), Managed (0.31), Defined (0.52), Developing (0.78), Initial (0.97)	0.0578	L
Industry Sensitivity	How critical the industry is to national/international security	Low (0.20), Medium (0.53), High (0.83), Critical (0.95)	0.1753	I
Incident Response Capability	Speed of detecting and mitigating threats	Advanced (0.20), Adequate (0.50), Slow (0.81), Nonexistent (0.97)	0.0425	I
Regulatory Impact	Level of compliance and regulatory oversight	None (0.10), Low (0.36), Medium (0.67), Strict (0.95)	0.1058	I
Damage and Impact				
Confidentiality Impact	Impact on data confidentiality	None (0.00), Low (0.28), Medium (0.67), High (0.95)	0.2586	I
Integrity Impact	Risk of data manipulation	None (0.00), Low (0.27), Medium (0.65), High (0.91)	0.0544	I
Availability Impact	Risk of system downtime	None (0.00), Low (0.33), Medium (0.67), High (0.92)	0.1336	I
Technical Impact	Damage to the technological infrastructure	No impact (0.00), Minimal (0.22), Moderate (0.51), Severe (0.78), Critical (0.93)	0.0321	I
Blast Radius	Scope of affected systems	Single (0.30), Multiple (0.71), Widespread (0.94)	0.0229	I
Financial Impact	Estimated monetary loss from the breach	None (0.00), Minor (0.32), Major (0.69), Catastrophic (0.89)	0.0683	I
Reputational Impact	Level of brand damage from breach disclosure	None (0.11), Low (0.40), High (0.80), Irreparable (0.94)	0.0850	I
Duration	Duration of the impact (Hours, Days, Months)	Volatile (Hours) (0.26), Volatile (Days) (0.52), Volatile (Months) (0.77), Persistence (0.96)	0.0069	I
Health and Safety Impact	Impact on human health and safety	None (0.00), Minor (0.27), Major (0.55), Extreme (0.74)	0.0145	I

1) AI AGENT CONFIGURATION AND REPRODUCIBILITY

Assessments were performed using two publicly available LLMs to examine model agnosticism: (i) GPT-4o (OpenAI) and (ii) Gemini 2.0 Flash (Google), selected to evaluate real-time processing capabilities. Both models were configured with a low temperature ($T = 0.1$) to promote stable (low-variance) outputs, using the text-based context-injection workflow described in Section IV. To support reproducibility and maintain strict experimental control, the input data was deliberately scoped to vulnerability reports and asset profiles. Results were compared against a control group of ten senior human experts ($N = 10$).

All LLM-based assessments were executed within a fixed evaluation window (2025-11-24 to 2025-11-29). For each provider, we logged the *provider-returned* model identifier/version string at run time (e.g., `gpt-4o-<version>` and `gemini-2.0-flash-<version>`) together with the run month (`2025-11`) to support traceability across model updates. Decoding parameters were held constant across all runs: temperature = 0.1, top_p = 1.0, max_tokens = 512, presence_penalty = 0, frequency_penalty = 0. No tool/function calling was enabled, and no external retrieval was used beyond the fixed text-based context injection described in Section IV. We fixed the system prompt, the metric schema (Table 3), and the conservative selection protocol for all cases; the full prompt template and JSON output schema are available from the authors upon reasonable request. All experiments used immutable system instructions (Appendix A) and standardized input contexts (Appendix B), which indicates a methodology intended to maintain comparability despite potential upstream model updates.

2) HUMAN EXPERT BASELINE PROTOCOL

A control group of ten senior cybersecurity analysts (CISSP/CISM-certified, average 12 years of experience) assessed the same 15 scenarios. The objective of this evaluation is not to claim an absolute ground truth for cyber risk, but to assess whether the proposed metric schema provides sufficient clarity to support consistent human judgment and whether LLM agents can operationalize this schema at scale. To support a baseline for algorithmic comparison, experts were not asked to provide intuitive scores. Instead, they were instructed to derive risk scores by applying the definitions in the proposed "Metric Box" schema (Table 3). This controlled protocol was intended to minimize subjective variance by providing a standardized analytical structure, where the high inter-rater reliability (Cronbach's $\alpha = 0.996$) suggests that the schema provides sufficient clarity to discipline expert judgment across complex risk parameters.

3) CONTROLLED EXPERIMENT DESIGN ($C_{13} - C_{15}$)

Unlike scenarios $C_1 - C_{12}$, which are retrospective analyses of historical events, scenarios $C_{13} - C_{15}$ were constructed as a controlled-variable test to isolate the effect of the "Organizational Context" parameter. The objective was to examine whether the framework distinguishes between *technical severity* (constant) and *business risk* (context-dependent).

For these three cases, the technical vulnerability input provided to the AI agent was fixed and identical:

> *"A critical unauthenticated Remote Code Execution (RCE) vulnerability (Log4Shell (CVE-2021-44228) (CVSS Base Score: 10.0)) has been detected on the external-facing web server 'SRV-WEB-01'. The exploit is publicly available."*

However, the organizational context documents injected into the prompt were varied as follows:

- C_{13} **(SME Context):** The asset is identified as an air-gapped workstation used for logistics management (Kitchen Menu Draft PC). This scenario is specifically designed to evaluate the framework's Logic Override mechanism (Rule 1) by introducing a deliberate conflict between technical vulnerability reports (e.g., an external RCE) and physical isolation. The vulnerability description is intentionally context-agnostic and represents a generic CVE feed entry rather than a live, externally reachable service within the SME environment. This controlled contradiction enables a precise evaluation of whether the framework correctly prioritizes organizational reality and network reachability over raw technical severity indicators.
- C_{14} **(Enterprise Context):** The asset is identified as a primary e-commerce portal for a large enterprise. It is integrated with backend payment gateways and stores confidential customer PII.
- C_{15} **(Critical Infrastructure Context):** The asset is flagged as an IT/OT gateway for a regional power distribution unit. Compromise of this server could allow adversaries to bridge into the SCADA network, potentially causing physical disruption to utility services.

This design isolates the context variable such that variation in the final risk score can be attributed to the agent's interpretation of asset criticality and impact context.

B. STABILITY UNDER REPEATED RUNS

To assess output stability, we re-ran each scenario $N = 10$ times per model under identical inputs (same context files, system prompt, and metric schema) and identical decoding parameters (temperature $T = 0.1$, top_p = 1.0). We report (i) *selection agreement* and (ii) *score variability*, measured by the standard deviation (std. dev.) of the final score.

Selection Agreement (SA) quantifies the consistency of the model's categorical metric selections across repeated iterations. For each metric j, the frequency of the most common selection (f_j) across N runs is identified. The SA is then defined as the average of these ratios across all M metrics

TABLE 4. Summary of real-world case studies used for validation ($C_1 - C_{15}$).

Case ID	Industry & Context	Vulnerability Profile	Asset Criticality
C_1	**Finance (Central Banking):** National Critical Infrastructure	Network Segmentation Failure, Lack of Logging & APT Malware [36]	Critical (Sovereign Funds & SWIFT Credentials)
C_2	**Defense Industry:** Strategic National Security Organization	Prompt Injection (OWASP LLM01) & Broken Access Control	Critical (Classified Design Schematics & R&D Data)
C_3	**Aerospace / Defense:** Global Tier-1 Defense Contractor [38]	Unpatched Vulnerability (Citrix Bleed) & Ransomware (LockBit 3.0) [37]	Critical (Design IP, Classified Data & Fleet Support)
C_4	**Healthcare / CNI:** SaaS Provider for Healthcare	Ransomware, Unpatched Vulnerability (ZeroLogon) & Lack of MFA [39]	High (Special Category Medical Data)
C_5	**E-Commerce (B2C):** Large Scale Retailer	XXE to Remote Code Execution (RCE) (CosmicSting) [40]	Critical (Customer PII, Payment Tokenization)
C_6	**Global SaaS / Cloud:** Publicly Traded Enterprise	Broken Authentication (Logic Flaw) & Privilege Escalation (ATO) [41]	Critical (Identity Management & Financial Data)
C_7	**Cybersecurity:** Crowdsourced Marketplace	Business Logic Flaw & IDOR (Reputation Manipulation) [42]	Critical (Reputation System Integrity & Platform Trust)
C_8	**Energy (Critical Infrastructure):** Fuel Pipeline Operator	Leaked VPN Credential (No MFA) & Ransomware (DarkSide) [43], [44]	Critical (Fuel Supply & Billing Continuity)
C_9	**Supply Chain (Software):** National Security Vendor	Compromised Build Pipeline (Sunburst Backdoor) [45], [46]	Critical (Govt. Networks & Enterprise Trust)
C_{10}	**Hospitality (Tourism):** Global Casino & Hotel Operator	Social Engineering (Vishing via Help Desk) & IdP Compromise [47], [48]	High (Guest Services & Casino Operations)
C_{11}	**Automotive (Manufacturing):** JIT Supply Chain	Third-Party Supplier Compromise (JIT Failure) [49], [50]	Critical (Production Line Continuity)
C_{12}	**Telecommunications:** National Carrier	Unauthenticated API Endpoint (BOLA/IDOR) [51], [52]	Critical (Customer PII - 9.8M Records)
Context Sensitivity Scenarios (Identical Technical Vulnerability: Log4Shell CVE-2021-44228, CVSS: 10.0)			
C_{13}	**SME (Logistics):** Kitchen Menu Draft PC	Asset is air-gapped and physically isolated; contains no sensitive PII or financial data.	Low (Operational overhead only)
C_{14}	**Large Enterprise:** Internal Corporate Network	Server hosts employee portals and integrated payment gateways with confidential PII.	High (Internal Operations & Confidentiality)
C_{15}	**Critical Infrastructure:** Life Support IoT Gateway	Aggregates telemetry from ventilators; data loss or DoS directly impacts patient safety.	Critical (Public Safety & National Security)

in the schema as follows:

$$\mathrm{SA} = \frac{1}{M} \sum_{j=1}^{M} \left(\frac{f_j}{N} \right) \times 100 \tag{7}$$

where $N = 10$ and M represents the total number of metrics in the schema. This granular approach ensures that stability is measured at the specific sub-parameter level, reflecting the agent's deterministic alignment across runs. For aggregate reporting, we compute selection agreement averaged across scenarios (C_1–C_{15}) and report the mean score std. dev. across scenarios, as summarized in Table 6.

TABLE 5. Profile of human experts participating in the validation study.

Expert ID	Role / Title	Exp. (Years)
E1	CISO / Information Security Manager	18
E2	CISO / Information Security Manager	15
E3	Risk Management / Compliance Specialist	12
E4	CISO / Information Security Manager	18
E5	Application Security Specialist	9
E6	SOC / Incident Response Analyst	7
E7	Risk Management / Compliance Specialist	10
E8	SOC / Incident Response Analyst	9
E9	Application Security Specialist	8
E10	Risk Management / Compliance Specialist	12

C. COMPARATIVE ANALYSIS RESULTS

1) RISK SCORING ALIGNMENT

The comparative analysis, visualized in Fig. 5, indicates that LLM-driven assessments are closely aligned with the human median, particularly in high-severity scenarios. Across the validation set (C_1 to C_{15}), both GPT-4o and Gemini 2.0 Flash exhibited a strong positive correlation with the control group of ten senior experts.

In critical-infrastructure scenarios, specifically Healthcare Ransomware (C_4), Energy Sector Pipeline (C_8), and SolarWinds Supply Chain Compromise (C_9), the AI agents produced median scores above 9.0. These results fall within the human interquartile range, suggesting that the "Metric Box" architecture can constrain model outputs in a manner consistent with established cybersecurity practice without requiring manual score overrides.

The results also indicate lower dispersion. In complex cases such as Defense Industry Prompt Injection (C_2) and JIT Supply Chain Failure (C_{11}), human assessors showed a wider spread of scores, reflected by larger box plots, suggesting differences in interpretation. In contrast, the AI agents exhibited narrower dispersion across runs. This pattern is consistent with reduced subjectivity relative to manual assessment under the same schema. Furthermore, the score drop observed in C_{13} relative to C_{14} provides initial evidence of context sensitivity, which is examined further in subsequent sections.

TABLE 6. Stability results over repeated runs (aggregated across C_1–C_{15}).

Model	Runs per scenario (N)	Selection agreement (%)	Score std. dev.
GPT-4o	10	97.4	0.089
Gemini 2.0 Flash	10	98.2	0.066

2) STATISTICAL CORRELATION AND RELIABILITY ANALYSIS

While the boxplot analysis provides visual evidence of alignment, we further quantified this relationship through statistical testing across the 15 validation scenarios ($N = 15$).

First, we evaluated the reliability of the human control group to establish a baseline. The expert cohort exhibited very high internal consistency, yielding a Cronbach's alpha of $\alpha = 0.996$. This high degree of agreement indicates that the structured definitions within the Metric Box effectively disciplined expert judgment, establishing a stable reference that limits the influence of subjective variability in the baseline data.

Against this median-based baseline, the GPT-4o agent exhibited a strong linear correlation ($r = 0.9717, p < 0.001$) and a strong monotonic relationship (Spearman $\rho = 0.9276$). Similarly, the Gemini 2.0 Flash agent demonstrated strong alignment ($r = 0.9390$, $p < 0.001$; $\rho = 0.8472$). These results, visualized in Fig. 6, indicate that the Metric Engine can translate LLM outputs into risk scores that are consistent with the expert baseline.

Beyond trend alignment, we quantified absolute agreement between the AI agents and human experts. To quantify the absolute scoring deviation from the human expert baseline, the MAE is computed. Following our validation methodology, the MAE is calculated by comparing the mean risk score of the AI agent's 10 runs ($\bar{R}_{\text{AI}}$) against the human expert consensus (median, $R_{\text{human\_med}}$) for each scenario i as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |R_{\text{human\_med},i} - \bar{R}_{\text{AI},i}| \tag{8}$$

where $n = 15$ represents the total number of primary validation scenarios. MAE across the 15 validation scenarios was 0.5991 for GPT-4o and 0.6613 for Gemini 2.0 Flash. On the 0 to 10 risk scale, these values correspond to an average deviation of less than 6.7%, suggesting that the agents not only follow the human trend but also produce scores with relatively low error. Additionally, the bias analysis indicates a positive shift for Gemini 2.0 (+0.5571), consistent with the framework's conservative "Safety-First" posture under ambiguity, whereas GPT-4o exhibited a near-neutral bias (−0.1454).

3) CONSISTENCY AND STABILITY

A primary concern regarding LLM-based systems is stochasticity. However, our extended stability analysis (see Table 6) indicates that variability is limited within the proposed architecture. Across 150 experimental iterations, GPT-4o achieved a Selection Agreement of 97.4%, while Gemini 2.0 Flash reached 98.2%. These results suggest that the agents selected the same metric parameters in most runs, consistent with the "Metric Box" acting as a guardrail against hallucination.

Consistent with this observation, Fig. 7 shows that the AI agents maintained low score variability, with most cases remaining below $\sigma \approx 0.15$. This represents improved stability relative to the human baseline, for which the standard deviation frequently exceeded $\sigma \approx 0.30$ and reached up to $\sigma \approx 0.45$ in complex scenarios.

This divergence is particularly evident in cases such as the SME Context (C_{13}) and the Defense Prompt Injection (C_2). In these scenarios, human experts exhibited greater disagreement regarding final severity, resulting in larger variance. In contrast, both GPT-4o and Gemini 2.0 Flash exhibited very low deviation across repeated runs. Overall, the framework supports reproducible risk scoring under high alert volumes and time-constrained assessment conditions.

4) OPERATIONAL EFFICIENCY

The human baseline for evaluating a single case was approximately 6 minutes (360 seconds). In contrast, the AI agents completed the same task in under 4 seconds on average (see Fig. 8), corresponding to an approximate 100× increase in throughput. Notably, the human experts were already familiar with the metric schema prior to evaluation; thus, the reported time primarily reflects analytical effort rather than learning overhead. This gain can support near-real-time risk assessment at scale by reducing the human bottleneck in the initial triage phase.

D. QUALITATIVE GAP ANALYSIS

Beyond numerical scoring, the AI Agent demonstrated an ability to capture contextual nuances that standard automated tools often miss. As detailed in Table 7, in Case C_1 (Bank Heist), the AI elevated the risk to "Critical" by recognizing the "Sovereign Funds" context, whereas standard technical audits typically rate comparable architectural flaws as Medium. These observations suggest that the RAG-style context injection helps bridge the gap between technical vulnerability scanning and business-centric risk management.

E. CONTEXT SENSITIVITY ANALYSIS

A critical requirement for modern risk assessment is the ability to distinguish between technical severity and business impact. To examine this capability, scenarios C_{13}, C_{14},

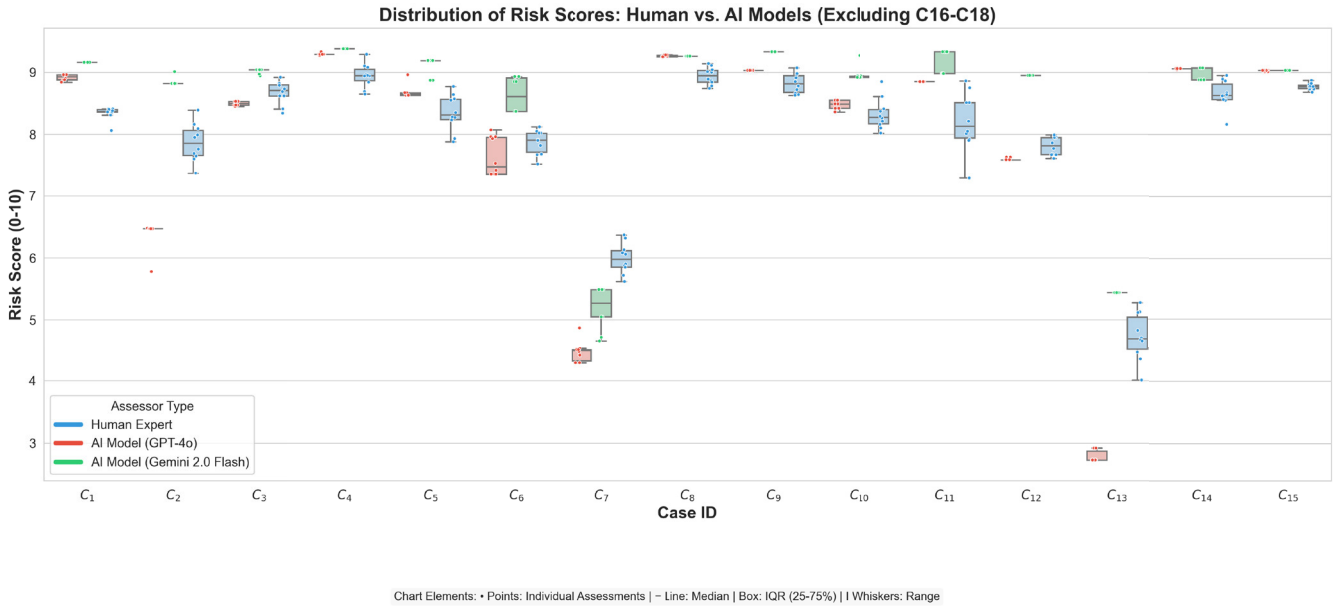

FIGURE 5. Distribution of Risk Scores across 15 validation cases ($C_1 - C_{15}$). The boxplots compare the spread of Human Expert assessments (Blue) against the AI Agents: GPT-4o (Red) and Gemini 2.0 Flash (Green). The AI models demonstrate high alignment with the human median in critical cases (e.g., C_4, C_9) while exhibiting lower variance in ambiguous scenarios (e.g., C_2, C_{11}), indicating superior consistency.

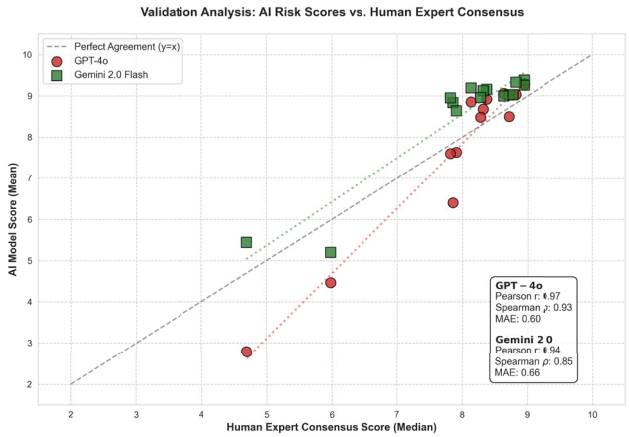

FIGURE 6. Correlation analysis between Human Median Scores (x-axis) and AI Agent Scores (y-axis). The regression analysis reveals a strong positive correlation for both GPT-4o ($r = 0.9717$) and Gemini 2.0 ($r = 0.9390$), confirming high alignment with the expert baseline.

and C_{15} used the same technical vulnerability (unpatched RCE) but injected distinct organizational contexts: an SME brochure website, a Large Enterprise payment portal, and a Critical Infrastructure gateway.

As shown in Fig. 9, the framework adapted risk scores based on the asset definition. In the SME context (C_{13}), the AI agents produced lower scores (GPT-4o: 2.8; Gemini 2.0 Flash: 5.4), reflecting the absence of critical data and lateral movement paths. When the same vulnerability was presented in an Enterprise context (C_{14}) and a Critical Infrastructure context (C_{15}), both models produced scores in the Critical range ($\approx$ 9.0), consistent with the higher asset criticality.

Overall, these results suggest that RAG-style context injection enables the AI agents to incorporate organizational context into scoring, differentiating risks based on business impact rather than relying solely on static technical severity ratings.

F. MITIGATION AND STRESS TESTING ANALYSIS

To assess the framework's logic-based reasoning beyond baseline vulnerability scoring, we conducted a stress test using the ShopFast Inc. Enterprise environment (C_{14}) as a reference case. We introduced three variants of the same high-severity Log4Shell vulnerability scenario ($C_{16} - C_{18}$), in which the technical finding remained constant but the operational context differed due to specific compensating controls.

1) BASELINE VS. MITIGATED SCENARIOS

As illustrated in the Sensitivity Stress Test (see Fig. 10), the AI models de-escalated risk scores in response to the presence and apparent effectiveness of compensating controls, rather than relying solely on the CVSS base score of 10.0.

- **Baseline Enterprise (C_{14}):** In the absence of mitigation, the AI agents assessed the risk in the Critical range (GPT-4o: 9.1, Gemini: 9.0), referencing the "Brand Trust" and "PCI DSS" implications of an exposed Payment Gateway.
- **False Positive (C_{16}):** When the context indicated that the scanner flagged a "non executable log file" and the vulnerable library was not loaded in memory, the AI

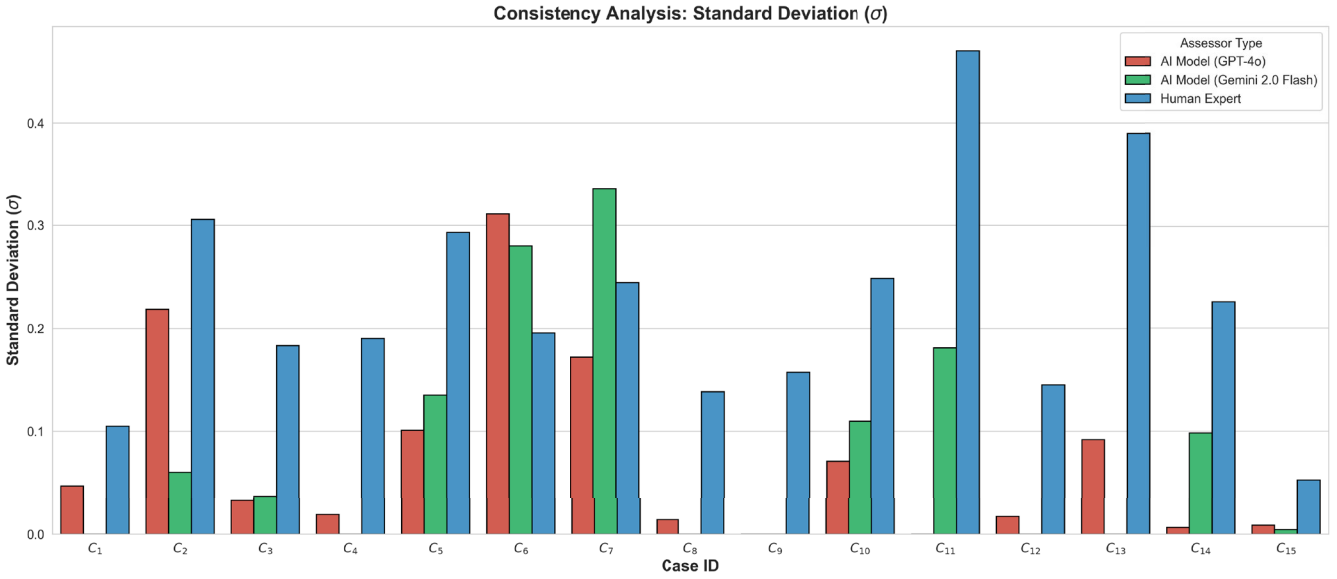

FIGURE 7. Consistency Analysis: Standard Deviation (σ) of scores. Lower values indicate higher consistency. The chart reveals that Human Experts (Blue) exhibit significantly higher variability across almost all cases, particularly in ambiguous contexts like C_{13}. The AI models (Red and Green) demonstrate superior stability, proving the reproducibility of the automated assessment.

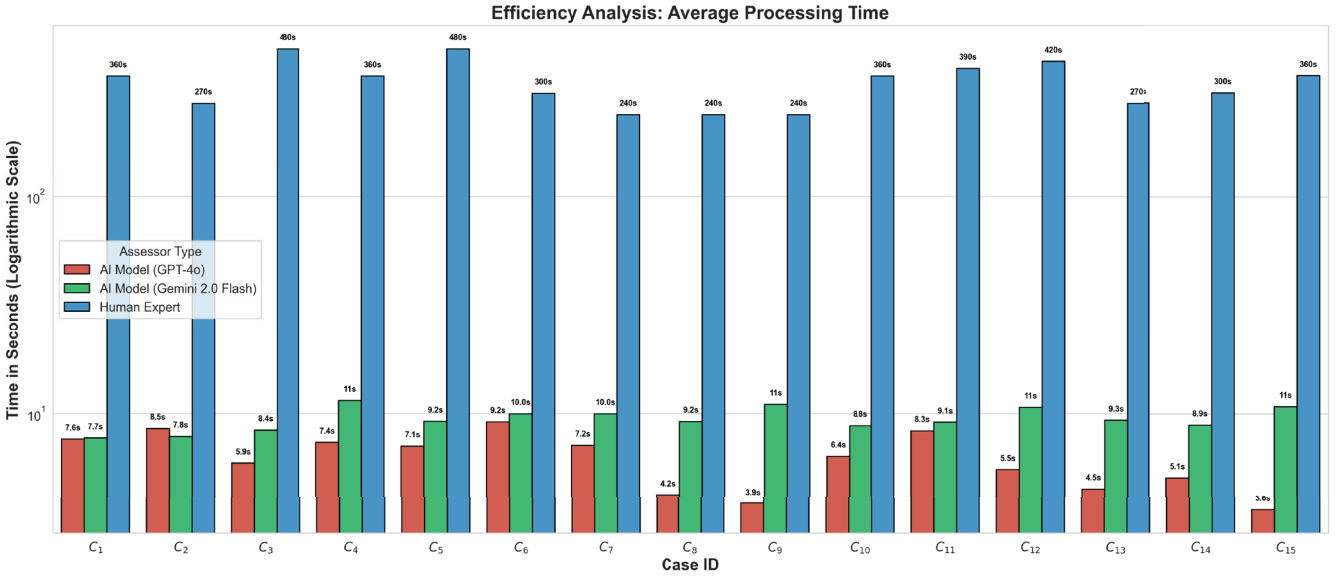

FIGURE 8. Efficiency Analysis: Average Processing Time (Logarithmic Scale). The transition from manual assessment (measured in minutes, Blue bars) to AI-driven assessment (measured in seconds, Red and Green bars) represents a $> 100x$ efficiency gain. GPT-4o demonstrates particularly high velocity in complex reasoning tasks (C_8, C_9).

models reduced the score to the Low range (GPT-4o: 3.8, Gemini: 4.4). Notably, the score did not drop to 0.00, reflecting the operational cost of triage and residual risk due to potential misdiagnosis, a nuance often overlooked by binary static scanners.

- **Air Gapped / Offline** (C_{17})**:** For the backup server physically disconnected in a secure vault, the score decreased to the 3.6–5.4 range. The AI outputs reflected that, although the software flaw remains present, the lack of connectivity limits the feasible attack vector, leaving residual risk primarily associated with physical or insider access.
- **WAF Mitigation** (C_{18})**:** In the scenario where a Cloud WAF was active in ''BLOCK'' mode, the risk was adjusted to the Medium/High range (GPT-4o: 6.1, Gemini: 6.3). Rather than treating this as a fully ''Fixed'' status, the outputs reflected that WAFs may be bypassed, categorizing the state as ''Mitigated'' rather

TABLE 7. Validation of AI agent scores against standard benchmarks and real-world ground truth.

Case ID	Scenario	Standard Benchmark *(Tech. Severity)*	Real-World Truth *(Impact Indicator)*	GPT-4o	Gemini	Insight (Gap Analysis)
C_1	Bank Heist	**N/A** (Arch. Flaw) *Standard audits: Med.*	**$81M Theft** *NCSC Study [36].*	**8.93**	**9.16**	The AI agent elevated risk by identifying the 'Sovereign Funds' context.
C_2	Defense	**N/A** (Logic Flaw) *Invisible to scanners.*	**Pentest: Critical** *Verified (Table 4).*	**6.47**	**8.82**	AI modeled the logic flaw consistent with internal risk appetite.
C_3	Aerospace	**Critical (9.4)** *CVE-2023-4966 [37].*	**Breach: Boeing** *CISA Alert [38].*	**8.50**	**9.04**	AI aligned with industry standards and verified major breach events.
C_4	Healthcare	**Critical (10.0)** *CVE-2020-1472.*	**£3.08m Fine** *ICO Notice [39].*	**9.29**	**9.38**	Validated by real-world disruption (NHS 111) and regulatory fines.
C_5	E-Commerce	**Critical (9.8)** *CVE-2024-34102.*	**Mass-Exploit** *4,000+ Stores [40].*	**8.65**	**9.19**	Consistent with destructive nature; stable global threat modeling.
C_6	SaaS	**Medium (4.8)** *Vendor Score.*	**$3,500 Bounty** *High Impact [41].*	**7.48**	**8.61**	**Gap:** High bounty confirmed impact where static scores failed.
C_7	Cybersecurity	**Medium (5.3)** *Vendor Score.*	**$2,500 Bounty** *Med. Impact [42].*	**4.49**	**5.27**	AI aligned with benchmark, identifying 'Platform Trust' context.
C_8	Energy / CNI	**High (7.0)** *Misses dependency.*	**Natl. Emergency** *CISA Alert [43].*	**9.26**	**9.26**	AI correctly deduced the **"Billing Blindness"** paradox.
C_9	Supply Chain	**Low / Trusted** *False Negative.*	**Global Espionage** *Sunburst [45].*	**9.03**	**9.33**	Identified **"Trusted Pipeline"** risk in signed binaries.
C_{10}	Hospitality	**N/A** (Process) *Help Desk gaps.*	**Op. Paralysis** *SEC Filing [47].*	**8.49**	**8.92**	Recognized **"Human-Centric Failure"** via Vishing (IdP bypass).
C_{11}	Automotive	**Low / External** *Ignored 3rd party.*	**Production Halt** *14 plants [49].*	**8.85**	**9.33**	Modeled **"JIT Fragility"**: 3rd party breach halts main line.
C_{12}	Telecom	**Medium** *Missed BOLA.*	**9.8M PII Leak** *OAIC Invest. [52].*	**7.59**	**8.95**	Identified **"Shadow API"** risk; overrode "Test" label.

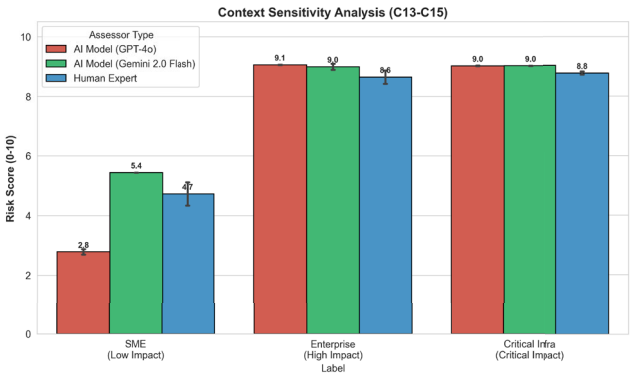

FIGURE 9. Context Sensitivity Analysis: Risk scores for the identical vulnerability across three contexts ($C_{13} - C_{15}$). The models correctly escalate risk from SME to Critical Infrastructure.

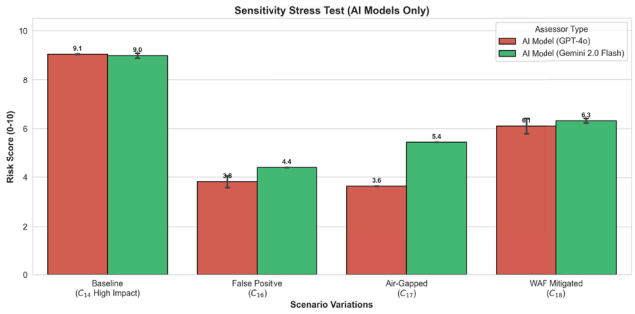

FIGURE 10. Sensitivity Stress Test (C_{14} vs. $C_{16} - C_{18}$). Using the Enterprise scenario (C_{14}) as a baseline (Score: ≈ 9.1), the chart illustrates how AI agents dynamically reduce risk scores based on compensating controls. Notice that False Positives (C_{16}) and Air Gapped assets (C_{17}) retain non zero "Residual Risk" scores (≈ 3.6 – 4.4), reflecting operational reality rather than binary classification.

than "Remediated" and maintaining a realistic vigilance level.

Overall, this differential scoring provides evidence that the framework operationalizes "Residual Risk" by distinguishing between theoretical vulnerability severity (CVSS) and context-dependent business risk.

VI. DISCUSSION

The results of this study challenge the common assumption that automated risk assessment necessarily requires complex, closed box deep learning models. By leveraging the semantic reasoning capabilities of publicly available LLMs anchored by expert-validated metrics, our findings suggest

that high-level cognitive tasks, such as contextualizing a vulnerability, can be automated with substantial gains in speed and precision. During development, however, we observed several recurring behavioral artifacts in LLMs, which were mitigated through targeted prompt engineering strategies.

A. ALGORITHMIC BIAS AND MITIGATION STRATEGIES

During iterative development of the framework, we identified and addressed two recurring behavioral patterns. These observations provide practical insight into the operational characteristics of LLMs in risk assessment.

1) ADDRESSING INSTRUCTION LEAKAGE VIA SEPARATION OF CONCERNS

In the early prototype phase, governance rules (e.g., ''If context is missing, choose High'') were embedded directly into the JSON Output Schema descriptions. We observed that this contributed to a phenomenon we term *Instruction Leakage*, in which the models produced spurious threat inferences to satisfy schema-embedded logic constraints. To mitigate this issue, we adopted a *Separation of Concerns* design in the final architecture (see Appendix A). The JSON Schema was restricted to data-structuring definitions (e.g., `"insufficient_information": false`, see Fig. 11), while reasoning logic was moved to the System Prompt and enforced via dynamic logic injection (see Appendix A). This decoupling reduced schema-induced hallucinations and provides a practical design pattern for improving the stability of LLM-based risk engines.

2) QUANTIFYING THE ''SAFETY-FIRST'' BIAS

Even with the optimized prompt, our analysis indicates a deliberate architectural *Safety-First Bias*, defined as a systematic positive deviation relative to the human median. As illustrated in Fig. 5, the AI agents produced higher risk scores in 11.5 scenarios on average (up to 14 for Gemini 2.0 Flash) out of 15, corresponding to an average shift of 2.059% (0.2059 points) above the human baseline. This behavior is evident in the False Positive scenario (C_{16}), where the technical finding was present but effectively neutralized (non-executable log file).

Rather than reducing the risk score to 0.0, both agents retained a ''Residual Risk'' score in the Moderate range (GPT-4o: 3.86, Gemini 2.0: 4.39). While this may appear as an overestimation relative to a theoretical zero, it is consistent with the cost asymmetry of cybersecurity risk: the impact of a false negative can be severe due to undetected compromise, whereas a false positive primarily incurs operational overhead for manual triage. This conservative posture increases the safety margin by prioritizing security integrity over noise reduction.

From an operational perspective, this systematic bias ensures that data ambiguity is treated as an inherent risk factor rather than arbitrary noise. By distinguishing between established technical risks and risks stemming from information uncertainty, the framework maintains a consistent prioritization hierarchy in large backlogs. This approach provides a safety margin for critical assets, ensuring they are not de-prioritized due to incomplete intelligence, thus optimizing the trade-off between precision and defensive resilience.

B. CONTEXTUAL SENSITIVITY AND THE MATHEMATICAL FLOOR

Our results indicate that a ''Zero Risk'' score (0.00) is not attainable under the operational bounds defined by our metric schema. As observed in the comparison between C_{14} (Baseline) and C_{16} (False Positive), the score decreased from 9.06 to 3.86 but did not reach zero. This residual score reflects the contribution of parameters with non-zero operational floors, such as ''Sector Sensitivity'' and ''Patch Status.''

Similarly, the transition from C_{14} to C_{17} (Air-Gapped) illustrates the model's use of *definition-based logic*. Rather than relying solely on the vulnerability report, the agent reclassified the Attack Vector from ''Network'' to ''Physical,'' consistent with the isolation context. The contrast between C_{14} (9.06) and C_{13} (SME context, 2.73) further suggests that the model incorporates organizational factors (e.g., ''Brand Value'' and ''Business Impact'') into scoring, acting not only as a calculator but also as a context-sensitive evaluator.

C. HIGH-FIDELITY REPRODUCIBILITY

A key finding of this study is that the framework can yield highly stable outputs from probabilistic models under controlled conditions. Although the framework architecture (Fig. 3) is designed to support multi-source data integration, the experimental focus in this study was deliberately narrowed to primary vulnerability and organizational inputs. This intentional scoping suggests that high-fidelity reproducibility is facilitated by prioritizing scientific rigor and reducing extraneous variables during evaluation. Critics of LLMs often cite non-determinism as a barrier to adoption in audit-compliant environments.

However, our empirical results (Table 6) indicate that this limitation is substantially mitigated within the proposed architecture. The high selection agreement ($> 97\%$) suggests that when unstructured inputs are mapped to a strict expert-validated schema (the Metric Box), publicly available LLMs can behave comparably to static algorithms while retaining semantic interpretation capabilities. This, in turn, suggests that instability is often associated with unconstrained prompting rather than being inherent to the models themselves.

D. INTERPRETATION OF FINDINGS: SPEED VS. ACCURACY

The observed $\sim 100\times$ reduction in assessment cycle time suggests that a key bottleneck in risk management is not data availability but *data synthesis*. While human analysts spent an average of six minutes evaluating the interplay among technical vulnerabilities, asset profiles, and organizational context, the LLM Agent performed this synthesis in under

TABLE 8. Operational management of system complexity.

Dimension	Traditional Approaches	Proposed Framework
Scalability	Manual effort bottleneck (avg. 6 mins/case) (Section I, Section II)	Automated throughput (>100x speedup) (Section V-C4)
Explainability	Closed box opacity or subjective variance (Section II)	Systemic Transparency (Decoupled Logic) (Section IV-D)
Maintainability	Static weights or high retraining costs (Section II)	Modular Logic Injection (No retraining) (Section VI-A1, Section VII)

four seconds on average. Importantly, this efficiency gain was not associated with a loss of alignment with the human baseline. The low variability (e.g., $\sigma < 0.1$) and strong correlation (e.g., $r \approx 0.97$) relative to the human median indicate that the ''Metric Box'' structure functions as an effective guardrail. Collectively, these results suggest that the agent follows human scoring trends while retaining the architectural Safety-First bias discussed earlier, operating at a more conservative threshold to account for risk asymmetry.

E. THE ROLE OF EXPERT VALIDATION

Critics may argue that reliance on commercial LLMs introduces opacity. We address this concern by decoupling the *reasoning* phase from the *scoring* phase. In end-to-end deep learning approaches, the rationale is often latent. In contrast, our framework requires the AI agent to justify its selections by mapping unstructured context to specific, human-readable sub-parameters. Consequently, while the semantic processing component is proprietary, the resulting risk score remains mathematically traceable to ROC weights derived from the expert cohort, supporting *process transparency*.

F. MANAGEMENT OF SYSTEM COMPLEXITY

To address the concerns regarding the integration of CTI, NLP/LLM, and expert-derived weights, Table 8 provides a comparison of how this complexity is managed relative to traditional approaches. As discussed in Section VI-A and Section VII, by utilizing a ''Separation of Concerns'' design, the framework maintains high scalability and explainability without the maintenance overhead typical of end-to-end closed box models.

G. LIMITATIONS

Despite these results, limitations remain. First, the framework depends in part on the LLM's underlying knowledge; while context injection provides current inputs, the model's training cut-off may limit its handling of newly emerging (e.g., zero-day) techniques. Second, the *Conservative Estimation Protocol* may yield slightly inflated scores in ambiguous cases, potentially requiring manual adjustment by the Risk Owner. Third, the ROC-derived weights reported in Table 3 should be interpreted as an expert-driven initial baseline rather than a universal constant. Although based on 101 professionals, these weights may reflect regional or sectoral perspectives. The framework therefore supports periodic re-calibration and allows organizations to customize weights to match their risk appetite, as detailed in Section III-E2 and supported by Eq. 2 and Eq. 3. Finally, longitudinal testing in a live SOC environment is needed to assess long-term stability.

H. ABLATION STUDY OF LOGIC OVERRIDES

To address the influence of deterministic rules on the probabilistic reasoning of LLMs, we conducted a comprehensive ablation study by disabling the *logic_rules.txt* across all primary scenarios (C_1–C_{18}). The results, summarized in Table 9 and Table 10, reveal a calculated trade-off between human-AI alignment and Logic Consistency within the framework.

TABLE 9. Ablation Analysis: Risk scores with and without logic overrides.

Case ID	Gemini		GPT	
	Rule OFF	Rule ON	Rule OFF	Rule ON
C_1	9.16	9.16	8.88	8.92
C_2	8.53	8.84	8.49	6.40
C_3	8.94	9.02	8.54	8.50
C_4	9.20	9.38	9.28	9.30
C_5	8.50	9.13	8.63	8.68
C_6	7.89	8.63	7.83	7.63
C_7	5.45	5.20	5.47	4.46
C_8	9.32	9.26	9.24	9.26
C_9	9.26	9.33	9.03	9.03
C_{10}	8.93	8.96	8.51	8.48
C_{11}	8.98	9.19	8.91	8.85
C_{12}	8.50	8.95	7.72	7.60
C_{13}	3.60	5.44	3.60	2.79
C_{14}	9.06	8.99	9.06	9.06
C_{15}	9.01	9.03	9.01	9.03
C_{16}	9.03	4.39	9.01	3.92
C_{17}	8.63	5.44	8.25	3.64
C_{18}	8.65	6.32	8.65	6.10

1) IMPACT ON STANDARD SCENARIOS (C_1–C_{15})

In scenarios where the LLM's semantic reasoning was sufficient to interpret the context, the introduction of deterministic overrides led to an increase in the MAE. For GPT-4o, the MAE rose from 0.402 (OFF) to 0.614 (ON), while Gemini 2.0 Flash exhibited a shift from 0.491 (OFF) to 0.654 (ON). This confirms that while the LLM's inherent reasoning is effective for standard cases, the rules introduce an ''architectural rigidity'' that deviates from the nuanced median of human experts to ensure systemic adherence to security policies in all conditions.

2) IMPACT ON EDGE CASES (C_{16}–C_{18})

The necessity of the logic overrides becomes evident in edge cases where technical vulnerability data conflicts with operational reality. Without rules (OFF), both models failed to de-escalate risk scores for Air-Gapped (C_{17}) or False

TABLE 10. Impact of logic overrides on scoring performance (C_1-C_{15} for MAE, C_{16}-C_{18} for consistency).

Model	MAE (OFF)	MAE (ON)	Logic Consistency
Gemini 2.0 Flash	0.491	0.654	Consistent
GPT-4o	0.402	0.614	Consistent

Positive (C_{16}) scenarios, maintaining scores in the critical range (≈ 8.63) solely based on technical severity. When rules were enabled (ON), the models dynamically de-escalated risk scores to align with the actual environmental context.

3) SAFETY-FIRST ENFORCEMENT ANALYSIS

Although MAE increased under override activation for both models, this outcome requires careful interpretation. MAE measures statistical proximity to the human median, whereas logic overrides introduce a deterministic governance layer that enforces predefined security constraints in edge conditions. These represent orthogonal evaluation axes: the former captures human-alignment, while the latter ensures policy-consistent risk boundaries.

Accordingly, the observed increase in MAE should not be interpreted as diminished model capability, but as evidence of intentional governance rigidity, where deterministic policy enforcement supersedes stochastic alignment with human scoring variance under predefined operational constraints. By acting as a final logical gatekeeper, the framework preserves strict adherence to organizational risk policies even when expert assessments may diverge.

4) HUMAN BASELINE AND GROUND TRUTH CLARIFICATION

It is important to emphasize that the human expert median serves as an empirical validation baseline rather than an absolute ground truth. While expert consensus provides a reliable reference point, human assessments may incorporate discretionary judgment and contextual variance. The deterministic layer of the proposed framework is explicitly designed to supersede such variance under predefined governance constraints, ensuring consistent enforcement of organizational risk policies even when individual expert interpretations diverge.

I. DATA PRIVACY AND ARCHITECTURAL SECURITY

Utilizing public LLM services involves inherent risks and threats to privacy, most notably the potential for unauthorized data disclosure. To mitigate these concerns, it is crucial to ensure that any information transmitted to external APIs is strictly anonymized, with careful attention to excluding sensitive or identifiable details from the prompts. Furthermore, for organizations with high security requirements, the adoption of on-premise LLMs is highly recommended. By deploying models on the institution's own internal resources in a closed-loop environment, cyber risk assessments can be performed without exposing sensitive data to external infrastructure.

SYSTEM IDENTITY & PURPOSE
ROLE: You are a Senior Cyber Risk Analyst.
GOAL: Perform a scientifically rigorous risk assessment based ONLY on the provided input data.

OUTPUT FORMAT
You must output a strictly valid JSON object. Do not include markdown fences or introductory text.

JSON STRUCTURE:

```
{
  "asset_info": {
    "name": "Extract from context",
    "id": "Extract from context"
  },
  "selected_metrics": {
    // Map EVERY parameter from Schema
    // to one of its exact Options.
    // Ex: "Attack Vector": "Network"
  },
  "parameter_explanations": {
    "Attack Vector": {
      "reason": "Brief justification...",
      "insufficient_information": false
    }
  }
}
```

FIGURE 11. System Prompt constraints enforcing JSON structure.

***** CRITICAL LOGIC OVERRIDES *****
You must analyze the `<organization_context>` and apply the following overrides STRICTLY. These rules serve to map compensating controls to their functional risk-equivalent states in the Metric Engine, superseding technical severity values.

— RULE 1: OFFLINE / AIR-GAPPED ASSETS —
IF the context mentions "Air-Gapped", "Offline", "Disconnected", "No Internet", or "Cold Storage":

1) SELECT "Physical" for "Attack Vector". (Reason: Remote attack is impossible)
2) SELECT "Fully Patched" for "Patch Availability". (Reason: Physical isolation functionally neutralizes the remote vulnerability, reaching the baseline residual risk score of 0.25)
3) SELECT "None" for "Confidentiality Impact".
4) SELECT "None" for "Integrity Impact".
5) SELECT "None" for "Availability Impact".
6) SELECT "No impact" for "Technical Impact".

— RULE 2: FALSE POSITIVE —
IF the context mentions "False Positive", "Invalid Alert", "Investigation Closed", or "No Risk":

1) SELECT "None" or "No impact" for ALL impact metrics.
2) SELECT "Fully Patched" for "Patch Availability". (Reason: Non-existent vulnerabilities are mapped to the baseline security state)
3) SELECT "Physical" for "Attack Vector".

— RULE 3: WAF / MITIGATED —
IF the context mentions "WAF", "Blocked", "Mitigated", or "Virtual Patch":

1) SELECT "Fully Patched" for "Patch Availability". (Reason: Virtual patching/mitigation provides operational risk neutralization equivalent to the baseline risk state)
2) SELECT "No impact" or "Minimal" for "Technical Impact".
3) SELECT "None" for "Confidentiality Impact".

END OF RULES

FIGURE 12. The `logic_rules.txt` input injected into the prompt to enforce safety-first overrides by mapping compensating controls to baseline risk equivalents.

J. ADVERSARIAL RISKS AND THREAT MODELING

As the framework processes external inputs such as CTI feeds and vulnerability reports, it is inherently susceptible to adversarial threats, including prompt injection and data poisoning. A malicious actor could attempt to manipulate the risk assessment by embedding misleading technical

```
{
  "Attack Surface and Exploitability": {
    "Attack Vector": {
      "WF": 0.1341,
      "Description": "Context of access required.",
      "Options": {
        "Network": {
          "Score": 0.99,
          "Description": "Bound to network stack."
        },
        "Local": { "Score": 0.56, "Desc": "..." },
        "Physical": { "Score": 0.29, "Desc": "..." }
      }
    },
    "Privileges Required": {"WF":0.0665,"Opts":"..."},
    "...": "Other metrics (e.g.,Patch Avail.) omitted."
  },
  "Damage and Impact": {
    "Confidentiality Impact": {
      "WF": 0.2586,
      "Description": "Data loss impact.",
      "Options": {
        "High": {
          "Score": 0.95,
          "Description":
            "Loss of critical PII/Secrets."
        },
        "Medium": { "Score": 0.67, "Desc": "..." },
        "None": { "Score": 0.0, "Desc": "No leak." }
      }
    },
    "...": "Other metrics (e.g. Financial) omitted."
  }
}
```


FIGURE 13. Abbreviated JSON Schema demonstrating the weight (W_F) and score structure used by the agent.

*** CYBER THREAT INTELLIGENCE REPORT ***
ID: CVE-2021-44228 (Log4Shell)
SEVERITY: CRITICAL (CVSS v3.1 Base Score: 10.0)
VECTOR: CVSS:3.1/AV:N/AC:L/PR:N/UI:N/S:C/C:H/I:H/A:H

DESCRIPTION:
Apache Log4j2 2.0-beta9 through 2.15.0 JNDI features... do not protect against attacker controlled LDAP... An attacker who can control log messages can execute arbitrary code.

TECHNICAL DETAILS:

- Exploitability: Extremely Easy (No auth required).
- Attack Vector: Network (Remote).
- Impact: Full System Control (RCE).

THREAT LANDSCAPE:
This vulnerability is currently being actively exploited in the wild by multiple threat actors. It allows complete takeover of the target server.

FIGURE 14. The shared 'vuln.txt' input injected into the prompt for all controlled scenarios.

context within a vulnerability report to artificially lower its score. To mitigate these threats, our framework relies on its decoupled architecture and the "Safety-First" residual risk floor as a multi layered defense. Since the scoring logic is governed by a deterministic Metric Engine with expert validated weights, it acts as a structural safeguard. Even if the LLM's semantic interpretation is partially compromised by an adversary, the "Safety-First" protocol ensures that for

*** LOW PRIORITY ASSET STATUS MEMO ***
Asset ID: Dev-Test-Server-04 (Kitchen Menu Draft PC)
Owner: Operations Manager, Ankara QuickFix Logistics

1. Company Profile:
Small logistics subcontractor (40 employees). No public-facing brand.

2. Compliance & Data Scope:
This server does NOT process regulated data (No PII/Financial). Used exclusively by kitchen staff to draft weekly lunch menus. A compromise poses NO risk of regulatory fines.

4. Asset Technical Context:

- **Network Status:** Air-gapped. Physically isolated.
- **Impact Analysis:** Zero criticality. If fails, we write the menu on a whiteboard.
- **Action:** Do not allocate budget for patching.

FIGURE 15. Context input for Scenario C_{13} (SME). The asset is isolated and has negligible business value.

*** CRITICAL INFRASTRUCTURE SAFETY REPORT ***
Asset ID: ICU-GW-LSS-01 (Life Support IoT Gateway)
Owner: Biomedical Eng. Dept, National University Hospital

1. Organizational Context:
Tier-1 Trauma Center. Reputation is "Critical" (National Trust).

2. Compliance & Regulatory Obligations:
Class III Medical Device. Strict adherence to MDR and Patient Safety standards.

4. Asset Technical Profile & Impact:

- **Network Zone:** Medical VLAN.
- **Safety Impact:** CATASTROPHIC. Aggregates telemetry from ventilators. Data loss or DoS could lead to failure to alarm during cardiac arrest (Loss of Life).
- **Tolerance:** Zero. 99.999% availability required.

FIGURE 16. Context input for Scenario C_{15} (Critical Infrastructure). The asset is a life-critical medical gateway.

critical assets or ambiguous inputs, the final score cannot drop below a predefined conservative baseline. This separation ensures that the decision making logic remains immutable to prompt-level manipulations while maintaining a necessary safety margin.

VII. CONCLUSION AND FUTURE WORK

By combining static vulnerability data with organizational context in near real time, this work proposes a framework that operationalizes expert knowledge and supports a shift from static tabular lookups toward context-aware reasoning in cyber risk management.

The main contributions and findings of this study are summarized as follows:

- **Expert-Validated Methodology:** We introduced a methodology for weight elicitation, using the ROC method to convert ordinal rankings from 101 cybersecurity professionals into objective parameter weights.
- **Architectural Innovation (Separation of Concerns):** We identified a phenomenon termed *"Instruction Leakage,"* in which embedding logic rules directly into the output schema contributed to model hallucinations. We mitigated this issue by adopting a separation-of-concerns design that decouples the *Logic Layer* (System Prompt) from the *Definition Layer* (JSON Schema).
- **Behavioral Analysis (Safety-First Bias):** Our analysis indicated an architectural *Safety-First Bias*, defined as

a systematic positive deviation relative to the human median. Quantitative results showed this pattern in an average of 11.5 out of 15 scenarios, with an average shift of 2.059% (0.2059 points) above the human baseline. This shift is consistent with a risk-management posture that prioritizes reducing false negatives over minimizing the operational burden of false positives. Using definition-based logic, the framework captured residual risk in stress tests (C_{16}), supporting a balance between numerical scoring and operational safety considerations.

- **Performance & Validation:** The framework exhibited strong agreement with the human median (Pearson r ranging from 0.9390 to 0.9717) while reducing the assessment cycle time by more than 100×. Comparative evaluation on real-world incidents (e.g., Colonial Pipeline, Optus) further indicated that the system can differentiate between technical severity and business impact.

Future extensions of this research will focus on two directions:

1) **Privacy-Preserving On-Premise Models:** Transitioning from general-purpose public LLMs to fine-tuned Small Language Models (SLMs) (e.g., Llama-3 8B) deployed locally to address data-sovereignty requirements in regulated sectors.
2) **Automated Risk Treatment (SOAR Integration):** Expanding the framework to generate machine-readable mitigation playbooks (e.g., Ansible/Terraform scripts) based on the identified risk profile, enabling increased automation of cyber defense workflows.

By pursuing these extensions, the framework is intended to mature toward an enterprise-ready solution for real-time, AI-assisted, and legally defensible risk governance.

DECLARATION OF AI-BASED LANGUAGE ASSISTANCE

The authors confirm that the intellectual content, methodology, and experimental analyses presented in this study are the original work of the human authors. Generative AI tools were used exclusively for linguistic refinement, grammatical corrections, and enhancing the overall readability of the manuscript. The authors maintain full responsibility for the technical accuracy and integrity of the final work, in accordance with IEEE guidelines on AI-generated content.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study (including validation scenarios, raw scoring logs, and the full JSON schema) are available from the corresponding author upon reasonable request.

APPENDIX A
REPRODUCIBILITY AND EXPERIMENTAL ARTEFACTS

To support transparency and reproducibility of the proposed deterministic risk scoring methodology, we provide the core system instructions, output schema, and a real-world case study used in the experiments.

A. AGENT INSTRUCTION SET (SYSTEM PROMPT)

The following instruction set provides the core directive for the LLM agent and enforces the JSON output format defined in Fig. 11.

B. LOGIC RULES INJECTION

To support the ''Safety-First'' design and reduce hallucination risk in ambiguous contexts, the following rule set is dynamically injected into the user prompt at runtime. These overrides function as a deterministic guardrail for specific edge cases (e.g., air-gapped assets), constraining the LLM's probabilistic outputs.

C. JSON OUTPUT SCHEMA (METRIC BOX)

To facilitate automated parsing and statistical analysis, the agent is constrained to a fixed JSON structure corresponding to the ''Metric Box'' defined in Table 3. Due to space constraints, Fig. 13 presents an abbreviated schema that illustrates the hierarchy of weights (W_F) and scoring options.

APPENDIX B
CONTROLLED EXPERIMENT DATA: CONTEXT SENSITIVITY

We present the input data used in the ''Context Sensitivity Analysis'' (C_{13} vs. C_{15}) to illustrate how the framework differentiates risk scores for the same technical vulnerability as a function of organizational context.

A. SHARED VULNERABILITY INPUT (VULN.TXT)

Across all controlled scenarios, the AI agent received the same vulnerability report corresponding to the Log4Shell exploit (CVE-2021-44228), as shown in Fig. 14.

B. ORGANIZATIONAL CONTEXT INPUTS (C_{IN})

While the vulnerability input remained constant, the organizational context was varied. Fig. 15 and Fig. 16 display the raw context files for the SME (C_{13}) and Critical Infrastructure (C_{15}) scenarios, respectively.

REFERENCES

[1] J. Yu, A. V. Shvetsov, and S. Hamood Alsamhi, ''Leveraging machine learning for cybersecurity resilience in industry 4.0: Challenges and future directions,'' *IEEE Access*, vol. 12, pp. 159579–159596, 2024, doi: 10.1109/ACCESS.2024.3482987.
[2] W. He, H. Li, and J. Li, ''Unknown vulnerability risk assessment based on directed graph models: A survey,'' *IEEE Access*, vol. 7, pp. 168201–168225, 2019, doi: 10.1109/ACCESS.2019.2954092.
[3] S. Drissi, M. Chergui, and Z. Khatar, ''A systematic literature review on risk assessment in cloud computing: Recent research advancements,'' *IEEE Access*, vol. 13, pp. 76289–76307, 2025, doi: 10.1109/ACCESS.2025.3561123.
[4] Forum of Incident Response and Security Teams (FIRST). (2023). *Common Vulnerability Scoring System (CVSS) V4.0: Specification Document*. [Online]. Available: https://www.first.org/cvss/v4.0/specification-document
[5] J. A. Freund and J. Jones, *Measuring and Managing Information Risk: A FAIR Approach*. London, U.K.: Butterworth, 2014.

[6] MITRE Corporation. (2011). *Common Weakness Scoring System (CWSS)*. [Online]. Available: https://cwe.mitre.org/cwss/

[7] *Guide for Conducting Risk Assessments*, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2012.

[8] *Information Security, Cybersecurity and Privacy Protection—Guidance on Managing Information Security Risks*, Standard ISO/IEC 27005, International Organization for Standardization, 2022.

[9] *Risk Management—Guidelines*, Standard ISO 31000, International Organization for Standardization, 2018.

[10] C. Alberts and A. Dorofee, *Managing Information Security Risks: The OCTAVE Approach*. Reading, MA, USA: Addison-Wesley, 2002.

[11] P. Cheimonidis and K. Rantos, "A proactive and time-sensitive cyber risk assessment model integrating Markov chains and Bayesian networks," *IEEE Access*, vol. 13, pp. 96911–96932, 2025, doi: 10.1109/ACCESS.2025.3575070.

[12] B. Al-Sada, A. Sadighian, and G. Oligeri, "Analysis and characterization of cyber threats leveraging the MITRE ATT&CK database," *IEEE Access*, vol. 12, pp. 1217–1234, 2024, doi: 10.1109/ACCESS.2023.3344680.

[13] L. Wang, Y. Ali, S. Nazir, and M. Niazi, "ISA evaluation framework for security of Internet of Health Things system using AHP-TOPSIS methods," *IEEE Access*, vol. 8, pp. 152316–152332, 2020, doi: 10.1109/ACCESS.2020.3017221.

[14] W. Xie, X. Yu, Y. Zhang, and H. Wang, "An improved Shapley value benefit distribution mechanism in cooperative game of cyber threat intelligence sharing," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Toronto, ON, Canada, 2020, pp. 810–815, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162739.

[15] A. G. Femi and M. Madu, "Enhancing adaptive cybersecurity risk management through AI-driven threat detection," *Int. J. Trendy Res. Eng. Technol.*, vol. 9, no. 2, pp. 1–14, Apr. 2025, doi: 10.54473/IJTRET.2025.9210.

[16] L. K. Jamili, H. Rawat, V. Garg, Bhanuvardhan, D. Semlani, and O. Goel, "Artificial intelligence for adaptive risk assessment in cloud-based security frameworks," in *Proc. Int. Conf. Netw. Cryptol. (NETCRYPT)*, May 2025, pp. 1583–1587, doi: 10.1109/netcrypt65877.2025.11102751.

[17] S. Islam, N. Basheer, S. Papastergiou, M. Ciampi, and S. Silvestri, "Intelligent dynamic cybersecurity risk management framework with explainability and interpretability of AI models for enhancing security and resilience of digital infrastructure," *J. Reliable Intell. Environments*, vol. 11, no. 3, Sep. 2025, Art. no. 12, doi: 10.1007/s40860-025-00253-3.

[18] A. Malik, K. Arshid, N. Noonari, and R. Munir, "Artificial intelligence-driven cybersecurity framework using machine learning for advanced threat detection and prevention," *Scholars J. Eng. Technol.*, vol. 13, no. 6, pp. 401–423, Jun. 2025, doi: 10.36347/sjet.2025.v13i06.005.

[19] J. M. Camacho, A. Couce-Vieira, D. Arroyo, and D. R. Insua, "A cybersecurity risk analysis framework for systems with artificial intelligence components," *Int. Trans. Oper. Res.*, vol. 33, no. 2, pp. 798–825, 2025, doi: 10.1111/itor.70049.

[20] Y. Hmimou, M. Tabaa, A. Khiat, and Z. Hidila, "A multi-agent system for cybersecurity threat detection and correlation using large language models," *IEEE Access*, vol. 13, pp. 150199–150215, 2025, doi: 10.1109/ACCESS.2025.3602681.

[21] R. Marinho and R. Holanda, "Automated emerging cyber threat identification and profiling based on natural language processing," *IEEE Access*, vol. 11, pp. 58915–58936, 2023, doi: 10.1109/ACCESS.2023.3260020.

[22] Y. Zhang, Z. Wang, Y. Wang, K. Lin, T. Li, H. Liu, C. Li, and B. Wang, "A risk assessment model for similar attack scenarios in industrial control system," *J. Supercomput.*, vol. 79, no. 14, pp. 15955–15979, Sep. 2023, doi: 10.1007/s11227-023-05269-1.

[23] N. M. Unal and B. Celiktas, "A metric-driven IT risk scoring framework: Incorporating contextual and organizational factors," in *Proc. Int. Conf. Artif. Intell., Comput., Data Sci. Appl. (ACDSA)*, Aug. 2025, pp. 1–7, doi: 10.1109/acdsa65407.2025.11166074.

[24] Y. Kawanishi, H. Nishihara, H. Yoshida, H. Yamamoto, and H. Inoue, "A study on threat analysis and risk assessment based on the 'Asset Containe," method and CWSS," *IEEE Access*, vol. 11, pp. 18148–18156, 2023, doi: 10.1109/ACCESS.2023.3246497.

[25] W. Wang, F. Shi, M. Zhang, C. Xu, and J. Zheng, "A vulnerability risk assessment method based on heterogeneous information network," *IEEE Access*, vol. 8, pp. 148315–148330, 2020, doi: 10.1109/ACCESS.2020.3015551.

[26] T. Wang, Q. Lv, B. Hu, and D. Sun, "CVSS-based multi-factor dynamic risk assessment model for network system," in *Proc. IEEE 10th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2020, pp. 289–294, doi: 10.1109/ICEIEC49280.2020.9152340.

[27] M. Ahmed, S. Panda, C. Xenakis, and E. Panaousis, "MITRE ATT&CK-driven cyber risk assessment," in *Proc. 17th Int. Conf. Availability, Rel. Secur.*, New York, NY, USA, Aug. 2022, pp. 1–10, doi: 10.1145/3538969.3544420.

[28] H. Yang, H. Yuan, and L. Zhang, "Risk assessment method of IoT host based on attack graph," *Mobile Netw. Appl.*, vol. 29, no. 5, pp. 1504–1513, Oct. 2024, doi: 10.1007/s11036-023-02198-4.

[29] M. U. Aksu, M. H. Dilek, E. I. Tatli, K. Bicakci, H. I. Dirik, M. U. Demirezen, and T. Aykir, "A quantitative CVSS-based cyber security risk assessment methodology for IT systems," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2017, pp. 1–8, doi: 10.1109/CCST.2017.8167819.

[30] V. C. W. Younang and A. Sen, "Security risk assessment using Bayesian attack graphs and complex probabilities for large scale IoT applications," *IEEE Trans. Dependable Secure Comput.*, vol. 22, no. 6, pp. 7360–7371, Nov. 2025, doi: 10.1109/TDSC.2025.3597186.

[31] U. Bansal, G. Sikka, L. K. Awasthi, and B. Bhargava, "Quantitative evaluation of extensive vulnerability set using cost benefit analysis," *IEEE Trans. Dependable Secure Comput.*, vol. 21, no. 1, pp. 298–308, Jan. 2024, doi: 10.1109/TDSC.2023.3253121.

[32] G. Abbas, M. Ali, M. Ahmad, and A. Khan, "CIRA-cyber intelligent risk assessment methodology for industrial Internet of Things based on machine learning," *IEEE Access*, vol. 13, pp. 77001–77016, 2025, doi: 10.1109/ACCESS.2025.3559617.

[33] F. R. Moreira, D. A. Da Silva Filho, G. D. A. Nze, R. T. de Sousa Junior, and R. R. Nunes, "Evaluating the performance of NIST's framework cybersecurity controls through a constructivist multicriteria methodology," *IEEE Access*, vol. 9, pp. 129605–129618, 2021, doi: 10.1109/ACCESS.2021.3113178.

[34] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, "MITRE ATT&CK: Design and Philosophy," The MITRE Corporation, McLean, LA, USA, Tech. Rep. MP180360R1, Mar. 2020. [Online]. Available: https://www.mitre.org/publications/technical-papers/mitre-attack-design-and-philosophy

[35] F. H. Barron and B. E. Barrett, "Decision quality using ranked attribute weights," *Manage. Sci.*, vol. 42, no. 11, pp. 1515–1523, Nov. 1996, doi: 10.1287/mnsc.42.11.1515.

[36] U.K. National Cyber Security Centre. *Nation-State Hackers Case Study: Bangladesh Bank Heist*. Accessed: Nov. 29, 2025. [Online]. Available: https://cyber.uk/areas-of-cyber-security/cyber-security-threat-groups-2/nation-state-hackers-case-study-bangladesh-bank-heist/

[37] Citrix Systems Inc. *Citrix Bleed Vulnerability and Advanced Data Theft: CVE-2023-4966*. Accessed: Nov. 29, 2025. [Online]. Available: https://nvd.nist.gov/vuln/detail/cve-2023-4966

[38] Cybersecurity and Infrastructure Security Agency (CISA). *#StopRansomware: LockBit 3.0 Ransomware Affiliates Exploit CVE 2023-4966 Citrix Bleed Vulnerability*. Accessed: Nov. 29, 2025. [Online]. Available: https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-325a

[39] Information Commissioner's Office (ICO). *Advanced Penalty Notice–Advanced Computer Software Group, 26 March 2025*. Accessed: Nov. 29, 2025. [Online]. Available: https://ico.org.uk/media2/gdlfddgc/advanced-penalty-notice-20250327.pdf

[40] Sansec Research Team. *CosmicSting Attack & Defense Overview*. Accessed: Nov. 29, 2025. [Online]. Available: https://sansec.io/research/cosmicsting

[41] *Unauthorized Partner Access Via Invitation Process–Report 2885269*. Accessed: Nov. 29, 2025. [Online]. Available: https://hackerone.com/reports/2885269

[42] *Business Logic Flaw Allowing Self-Creation of Testimonials for Reputation Manipulation–Report 2490953*. Accessed: Nov. 29, 2025. [Online]. Available: https://hackerone.com/reports/2490953

[43] (May 2021). *DarkSide Ransomware: Best Practices for Preventing Bus. Disruption From Ransomware Attacks*. Accessed: Nov. 29, 2025. [Online]. Available: https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-131a

[44] J. Blount, "Cyber Threats in the pipeline: Lessons from the federal response to the colonial pipeline ransomware attack," U.S. House Representatives, Committee Homeland Security, Washington, DC, USA, Tech. Rep. Serial No. 117-18, Jun. 2021. Accessed: Mar. 26, 2026. [Online]. Available: https://www.congress.gov/117/chrg/CHRG-117hhrg45310/CHRG-117hhrg45310.pdf

[45] Cybersecurity and Infrastructure Security Agency (CISA). *Emergency Directive 21-01: Mitigate SolarWinds Orion Code Compromise*. Accessed: Nov. 29, 2025. [Online]. Available: https://www.cisa.gov/news-events/directives/ed-21-01-mitigate-solarwinds-orion-code-compromise
[46] Microsoft Security Response Center. *Deep Dive Into the Solorigate 2nd Stage Activation: From SUNBURST To TEARDROP*. Accessed: Nov. 29, 2025. [Online]. Available: https://www.microsoft.com/en-us/security/blog/2020/12/18/analyzing-solorigate-the-compromised-dll-file-that-started-a-sophisticated-cyberattack-and-how-microsoft-defender-helps-protect/
[47] MGM Resorts International. (Oct. 2023). *Form 8-K: Report of Unscheduled Material Events or Corporate Event*. [Online]. Available: https://www.sec.gov/ix?doc=/Archives/edgar/data/789570/000119312523251667/d461062d8k.htm
[48] (Sep. 2023). *The MGM Resorts Attack: Analysis of the Identity-Based Compromise*. [Online]. Available: https://www.cyberark.com/resources/blog/the-mgm-resorts-attack-initial-analysis
[49] (Feb. 2022). *Toyota To Suspend All Japan Factory Operations March 1 After Cyberattack*. [Online]. Available: https://asia.nikkei.com/Business/Automobiles/Toyota-to-suspend-all-Japan-factory-operations-March-1-after-cyberattack
[50] C. Cimpanu. (Feb. 2022). *Toyota Halts Production After Suspected Cyberattack at Supplier*. [Online]. Available: https://therecord.media/toyota-halts-production-after-suspected-cyberattack-at-supplier
[51] (Sep. 2022). *Optus Notifies Customers of Cyberattack Compromising Customer Information*. [Online]. Available: https://www.optus.com.au/about/media-centre/media-releases/2022/09/optus-notifies-customers-of-cyberattack
[52] Office of the Australian Information Commissioner (OAIC). (Oct. 11, 2022). *OAIC Opens Investigation Into Optus Over Data Breach*. [Online]. Available: https://www.oaic.gov.au/news/media-centre/oaic-opens-investigation-into-optus-over-data-breach

NEZIH MAHMUT UNAL received the B.S. degree in electronic engineering from Bursa Uludağ University, in 2006. He is currently pursuing the M.S. degree in cybersecurity with Isik University. He worked for five years in network, system, and database administration roles before transitioning to cybersecurity, where he has accumulated over 14 years of professional experience. He is also a Cybersecurity Consultant, providing advisory services and professional training on information security management and organizational security practices. His expertise includes ISO 27001 implementation, cybersecurity risk management, and information security governance. His research interests include cybersecurity maturity, risk assessment methodologies, and threat analysis.

BARIS CELIKTAS received the B.S. degree in systems engineering from National Defense University, in 2008, the M.S. degree in applied informatics from Istanbul Technical University, in 2018, and the Ph.D. degree in cybersecurity engineering and cryptography from the Institute of Informatics, Istanbul Technical University, in 2022. He is currently an Assistant Professor with the Computer Engineering Department and the Director of the Cybersecurity Graduate Program, Isik University. In addition, he works as a Cybersecurity Consultant and an Architect, specializing in enterprise cybersecurity and cryptography solutions, cloud security, risk management, and governance. He holds numerous industry-recognized certifications, including CISSP, CCSP, CISM, CISA, CRISC, AAIA, SSCP, CCNP, Security+, CySA+, CIEH, and ISO/IEC 27001, 22301, 20000, 27701, and 42001 Lead Auditor/Lead Implementer credentials, as well as GDPR DPO and NIST cybersecurity consulting credentials. His research interests include cybersecurity, network security, cloud computing, cryptography, malware analysis, risk management, and security applications.