

DOI: https://doi.org/10.48009/2_iis_2026_134

Beyond trust: An evidence-based benchmark for AI decision readiness in high-stakes operational environments

Anas AlSobeh, *Utah Valley University*, anas.alsobeh@uvu.edu

Mustafa Farouk Abo El Rob†*, *University of Denver*, mustafa.aboelrob@du.edu

Matt North, *Utah Valley University*, mnorth@uvu.edu

*Corresponding author

†Principal Investigator; funding acquisition

Abstract

The deployment of artificial intelligence (AI) in high-stakes decision contexts—including cybersecurity triage, insurance underwriting, clinical decision support, and university policy formulation—has outpaced the methodologies organizations use to judge whether a given system is actually ready to support those decisions. This paper introduces the AI Decision Readiness Benchmark (AI-DRB), an evidence-based framework that quantifies readiness across four automated dimensions: risk score validation, reliability, robustness, and explainability and auditability. The dimensions are combined into a single Decision Readiness Score (DRS) on a [0, 1] scale with explicit deployment thresholds. We evaluate AI-DRB on five public-source datasets covering 35,417 instances and on five state-of-the-art language models plus a fine-tuned BERT classifier. We then apply the framework retrospectively to publicly documented AI governance decisions at 142 United States universities to expose gaps between deployment prevalence and assessment rigor, and we show how the resulting score sheet can support governance decisions at the University of Denver. Findings advance practice from aspirational "trust" language toward measurable, auditable readiness criteria.

Keywords: AI evaluation, decision-making, trustworthiness, benchmarking, high-stakes AI, adversarial robustness

Introduction

The deployment of artificial intelligence systems in high-stakes operational environments has accelerated rapidly, driven by advances in deep learning (LeCun et al., 2015; Goodfellow et al., 2016) and transformer-based language models (Vaswani et al., 2017). Yet adoption has outpaced evaluation rigor: McKinsey reports that approximately 72% of enterprises had adopted AI by 2024 (Chui et al., 2024), while a 2024 Gartner survey found that 63% of organizations lack adequate data management practices for AI deployment (Edjlali, 2024). This measurement gap—the chasm between widespread deployment and systematic evaluation—represents a substantial source of operational, financial, and reputational risk across sectors ranging from healthcare and insurance to cybersecurity and institutional governance.

The practical stakes are concrete. In cybersecurity risk assessment, a Chief Information Security Officer must allocate remediation resources based on algorithmic risk scoring; a single miscalibrated assessment can lead to wasteful effort or to a multi-million-dollar breach. The IBM Cost of Data Breach Report (IBM Security, 2023) documented that the average breach cost reached \$4.45 million in 2023, with credential-

compromise breaches exceeding \$6.50 million. Comparable failure modes appear in insurance underwriting, where data poisoning or unexplained decisions invite regulatory scrutiny, and in clinical decision support, where systems must be robust to variations in imaging equipment and patient demographics that differ from training data. Across these domains, the evaluation methodologies in current use—accuracy, precision, recall on idealized test sets—provide little insight into whether systems remain reliable under adversarial manipulation, whether human decision-makers can appropriately trust their recommendations, or whether their reasoning is transparent enough for regulatory audit.

Existing evaluation frameworks address pieces of this problem but not the whole. The NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) provides governance taxonomy without quantitative metrics. MLPerf measures training performance and inference latency, not decision-support quality. AI Fairness 360 systematically evaluates disparate impact but does not address reliability or robustness. Specialized tools for adversarial robustness, explainability, and human-AI collaboration have advanced their respective dimensions (Ribeiro et al., 2016; Lundberg & Lee, 2017; Bansal et al., 2021), but no framework synthesizes them into a single, actionable readiness assessment for a particular deployment context. This fragmentation, combined with the well-documented technical debt of deployed ML systems (Sculley et al., 2015; Amershi et al., 2019), leaves organizations without a unified basis for answering: Is this AI system ready to support decisions in this particular operational context?

This work addresses that gap through the AI Decision Readiness Benchmark (AI-DRB), an evidence-based framework that quantifies an AI system's preparedness to support consequential decisions across four validation dimensions: risk score validation, reliability assessment, robustness evaluation, and explainability and auditability. The framework synthesizes insights from machine learning reliability theory, adversarial robustness research, regulatory standards, and operational security practices into a unified Decision Readiness Score (DRS) on a $[0, 1]$ scale, with explicit deployment thresholds, moving beyond generic trust narratives toward measurable, verifiable criteria. The principal contributions are: (i) a mathematical formalization of decision readiness integrating four dimensions into a unified score; (ii) a benchmark suite spanning five datasets covering cybersecurity, insurance, clinical decisions, governance, and adversarial content; (iii) empirical evaluation across five state-of-the-art language models and a fine-tuned BERT classifier; (iv) systematic methodologies for content-poisoning detection, explanation fidelity, and audit-trail completeness; (v) a document-analysis-based case study of 142 United States universities revealing gaps between AI deployment prevalence and assessment rigor; and (vi) actionable threshold-calibration guidance for practitioners. We further illustrate how AI-DRB can support concrete AI governance decisions at the University of Denver. The remainder of the paper presents the related literature, formalizes the problem, details the framework, describes the empirical evaluation and case study, discusses findings and implications, and concludes with limitations and future work.

Background and Literature Review

AI decision readiness draws on five research areas: trustworthiness frameworks and standards, benchmarking methodologies, human-AI collaboration, adversarial robustness and data poisoning, and explainability. Each contributes essential insights while collectively revealing gaps that motivate the present work.

AI Trustworthiness Frameworks and Standards

Recent regulatory and standards activity has produced principled frameworks for AI risk and trustworthiness. The NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) organizes risks around four governance activities (map, measure, manage, govern) but

operates at an abstraction level above specific technical metrics. The European Union's AI Act (European Commission, 2024) categorizes systems by risk level and imposes requirements for documentation, human oversight, and transparency, but defers operational guidance on what evidence demonstrates compliance. ISO/IEC 42001 (International Organization for Standardization, 2023) and parallel standards establish process and governance requirements without quantitative benchmarking methodologies. These developments operationalize long-standing concerns about systemic AI safety, which the present framework addresses through measurable readiness components. Collectively, the regulatory landscape signals strong consensus on systematic AI evaluation while leaving the technical operationalization gap that AI-DRB addresses.

Existing Benchmarking Methodologies

Benchmarking infrastructure for AI has specialized along dimension-specific lines. MLPerf provides comparative measurement of training and inference performance but addresses computational efficiency rather than decision-support quality. AI Fairness 360 enables systematic disparate-impact assessment but does not evaluate reliability, robustness, or audit-readiness. The Adversarial Robustness Toolbox and AI Explainability 360 similarly address single dimensions. An organization deploying an AI system might complete evaluations across these tools yet lack any mechanism for synthesizing them into a single deployment-readiness judgment. This fragmentation, while understandable from a specialization perspective, leaves practitioners without a unifying framework.

Human-AI Decision Making and Collaboration

Research on human-AI collaboration establishes that decision-makers do not passively accept algorithmic recommendations but engage in social and cognitive processes that determine how AI guidance is used. Bansal et al. (2021) demonstrated that team composition and interaction dynamics substantially affect outcomes, with poorly designed interfaces potentially degrading human decision quality. Lai et al. (2021) established that explanation quality, automation transparency, and reliance calibration shape both decision quality and user satisfaction. Zhang et al. (2020) found that confidence scores can improve reliance calibration but poorly calibrated confidence induces over-reliance. Shneiderman (2022) emphasizes the necessity of designing systems that preserve meaningful human agency. Kaur et al. (2020) found that users frequently misinterpret feature-importance explanations from LIME and SHAP. Current benchmarking methodologies rarely measure these human-centered dimensions. AI-DRB partially addresses this gap through systematic evaluation of decision explanation completeness, explanation fidelity, and audit-trail quality.

Adversarial Robustness, Poisoning, and Explainability

Adversarial machine learning has established that neural networks can be vulnerable to crafted perturbations (Goodfellow et al., 2015) and stronger targeted attacks (Carlini & Wagner, 2017). Beyond test-time perturbations, data poisoning attacks manipulate training data to degrade performance (Biggio et al., 2012; Papernot et al., 2016). Adversarial training improves robustness but with accuracy trade-offs (Madry et al., 2018). For decision-support systems in adversarial environments such as cybersecurity, robustness to both perturbations and data poisoning is essential, yet most existing benchmarks omit it. Post-hoc explainability methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) provide local explanations, but Doshi-Velez and Kim (2017) noted that the field lacks consensus on what constitutes meaningful interpretability, and Kaur et al. (2020) demonstrated that users frequently misinterpret these outputs. Datasheets for datasets (Gebru et al., 2021) and model cards (Mitchell et al., 2019) address documentation gaps that post-hoc explanations cannot.

Regulatory and Governance Frameworks

Regulatory and governance work addresses how AI systems are documented, monitored, and held accountable. Raji et al. (2020) emphasized closing the AI accountability gap through systematic deployment monitoring and failure-detection mechanisms. Holstein et al. (2019) examined fairness interventions across stakeholder groups, finding that technical improvements alone do not guarantee acceptable outcomes without organizational incentives and stakeholder participation. Floridi et al. (2018) and Jobin et al. (2019) reviewed ethics principles globally, finding consensus on transparency, accountability, and fairness alongside variation in implementation. The principle-practice gap—translating ethical commitments into measurable requirements—motivates the operationalizable evaluation methodology that AI-DRB establishes.

Gap Analysis and Contributions

Existing methodologies collectively address fairness, robustness, explainability, and computational performance but lack a unified framework that synthesizes these dimensions into operational decision readiness for specific deployment contexts. Five gaps motivate this work: (i) no benchmark integrates risk validation, reliability, robustness, and auditability into a single actionable metric; (ii) evaluation frameworks remain disconnected from operational deployment realities such as data shift and adversarial manipulation; (iii) systematic content-poisoning evaluation is limited, particularly for large language models in high-stakes contexts; (iv) regulatory and audit-trail alignment receives minimal attention in technical benchmarking; and (v) explanation fidelity and audit-ready decision traces are inconsistently measured. AI-DRB addresses these gaps by providing a comprehensive evaluation methodology grounded in realistic datasets and operational scenarios.

Against this backdrop, AI-DRB is distinguished from existing approaches by what it operationalizes, not merely what it enumerates. The NIST AI Risk Management Framework and ISO/IEC 42001 provide governance taxonomies and process requirements but stop short of quantitative, decision-level scoring; the EU AI Act specifies risk tiers and obligations without a measurement procedure; MLPerf benchmarks training and inference efficiency rather than decision-support quality; and fairness toolkits such as AI Fairness 360 quantify disparate impact but not reliability, robustness, or auditability. AI-DRB's contribution is to integrate risk-score validation, reliability, robustness, and explainability/auditability into a single, context-calibrated Decision Readiness Score with actionable deployment thresholds, converting principles that these frameworks articulate into a measurable, comparable, and auditable readiness signal.

Recent sector-specific evidence reinforces that this measurement gap is not confined to any single domain. Socio-technical analyses of AI adoption in public-facing and high-stakes settings—for example, studies of the drivers of and barriers to AI adoption in agriculture—document the same pattern of rapid deployment outpacing systematic readiness assessment (Alkhwaldi et al., 2026).

Problem Formulation

We define a *decision-support context* as a tuple $D = \langle T, D, U, C \rangle$ comprising a task specification T , a dataset D of instances, a set of decision-makers U , and a consequence model C specifying outcomes and costs of correct and incorrect decisions. An AI system S is a function mapping inputs $x \in D$ to outputs $\hat{y} = S(x)$ including predictions, confidence scores, and explanations, listed in Table 1.

Table 1. Explanations and Formulae for DRS Assessment Validation

DRS Component Explanation	Formula
The decision readiness score (DRS) of system S within context D integrates assessment across four validation dimensions:	$DRS(S, D) = f(R_{\text{risk}}, R_{\text{reliability}}, R_{\text{robustness}}, R_{\text{audit}})$ where f is an aggregation function.

The <i>risk score validation component</i> , R_{risk} , quantifies the system's ability to discriminate between clean and adversarially poisoned instances. Given a test set $Trisk$ partitioned into clean T_c and poisoned T_p :	$R_{risk} = \alpha \cdot F1_{clean} + (1 - \alpha) \cdot F1_{poisoned}$ where α is a weighting parameter (typically 0.5).
F1 is the harmonic mean of precision and recall:	$F1 = 2 \cdot P \cdot R / (P + R)$
The <i>reliability component</i> , $R_{reliability}$, quantifies prediction consistency across repeated evaluations (test-retest reliability) and varying conditions (cross-condition consistency). Test-retest reliability for set T_{rel} evaluated at times $t1$ and $t2$ uses Cohen's kappa (Cohen, 1960):	$\kappa = (p_o - p_e) / (1 - p_e)$ where p_o is observed agreement and p_e is expected agreement by chance.
For multi-rater evaluations, Fleiss' kappa (Fleiss, 1971) extends this:	$\kappa_{Fleiss} = (\bar{p} - \bar{p}_e) / (1 - \bar{p}_e)$
The reliability component combines test-retest with cross-condition consistency:	$R_{reliability} = \beta \cdot \kappa_{test-retest} + (1 - \beta) \cdot \kappa_{cross-condition}$, typically with $\beta = 0.6$.
The <i>robustness component</i> , $R_{robustness}$, quantifies performance degradation under (i) adversarial perturbations of specified magnitude, (ii) inclusion of poisoned training examples at varying rates, and (iii) distribution shift along demographic, temporal, or domain dimensions:	$R_{robustness} = \int_0^{\epsilon_{max}} F1(\epsilon) d\epsilon / \epsilon_{max}$ where ϵ is perturbation magnitude. In discrete form, $R_{robustness} = (1/k) \sum F1(\epsilon_i)$
The <i>explainability and auditability component</i> , R_{audit} , quantifies decision-process transparency through three sub-components: trace completeness S_{trace} , explanation quality $S_{explain}$, and regulatory compliance alignment $S_{compliance}$:	$R_{audit} = \delta_1 \cdot S_{trace} + \delta_2 \cdot S_{explain} + \delta_3 \cdot S_{compliance}$ where $\delta_1, \delta_2, \delta_3$ sum to unity, typically (0.4, 0.3, 0.3).
The trace completeness sub-component is computed as:	$S_{trace} = Elements_{preserved} / Elements_{required} $
The final decision readiness score aggregates the four components:	$DRS(S, D) = w_1 \cdot R_{risk} + w_2 \cdot R_{reliability} + w_3 \cdot R_{robustness} + w_4 \cdot R_{audit}$ with weights w_i satisfying $\sum w_i = 1$ calibrated to operational requirements.
Default weights (w_i) are (0.30, 0.25, 0.25, 0.20). Healthcare contexts may shift to (0.25, 0.25, 0.20, 0.30) emphasizing regulatory compliance; cybersecurity contexts to (0.25, 0.20, 0.35, 0.20) emphasizing robustness. The DRS yields a scalar in $[0, 1]$ with deployment thresholds: $DRS \geq 0.80$ ready for deployment without continuous oversight; $0.60 \leq DRS < 0.80$ ready with structured human oversight; $0.40 \leq DRS < 0.60$ suitable for human-only decision augmentation; $DRS < 0.40$ unsuitable pending substantial improvements. Organizations may adjust thresholds based on consequence models and risk tolerance.	
For the content-poisoning detection pipeline, precision and recall are computed for both classes. Poison:	$P_{poison} = TP_{poison} / (TP_{poison} + FP_{poison}), R_{poison} = TP_{poison} / (TP_{poison} + FN_{poison})$
Clean:	$P_{clean} = TP_{clean} / (TP_{clean} + FP_{clean}), R_{clean} = TP_{clean} / (TP_{clean} + FN_{clean})$
Trace completeness in operational form is also expressed as:	Trace Completeness = Number of Elements Preserved / Number of Elements Required

The AI-DRB Framework

The AI-DRB framework operationalizes the formal problem formulation through a systematic evaluation

methodology comprising five modules, each implementing concrete assessment approaches for the dimensions formalized above. This section details the architecture and evaluation procedures.

Architecture Overview

The AI-DRB framework, illustrated in Figure 1, comprises five principal evaluation modules operating sequentially or in parallel, each producing component scores that feed into the unified DRS computation. The framework accepts a candidate AI system, a decision-support context specification including task definition and consequence model, and relevant evaluation datasets. It produces detailed component scores across each dimension alongside the unified DRS metric, deployment recommendations, and a comprehensive evaluation report suitable for regulatory submission.

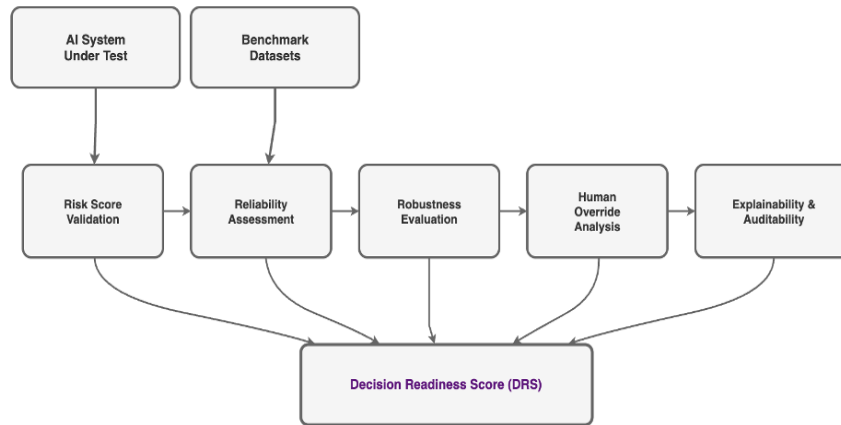


Figure 1: AI-DRB Framework Architecture.

Figure 1 shows the framework comprises five principal evaluation modules (Risk Score Validation, Reliability Assessment, Robustness Evaluation, Explainability and Auditability, and DRS Computation) that operate on candidate AI systems and evaluation datasets to produce component scores and final decision readiness scores. Reading the figure inputs (S, context C , datasets) flow left-to-right through four parallel component-scoring modules (Rrisk, Rrel, Rrob, Raud) into the DRS aggregator on the right, which emits a $[0, 1]$ score and a deployment recommendation. A feedback arrow allows continuous re-scoring across procurement and re-training cycles.

Risk Score Validation Module

The risk score validation module evaluates a system's ability to discriminate between clean and adversarially poisoned content. Figure 2 shows systematic workflow for testing an AI system's ability to discriminate between clean and adversarially manipulated content across diverse domains. The four-stage pipeline (data preparation, system evaluation, metric computation, aggregation) is fully automated; all labels come from public records (NVD/CVSS, ClinicalTrials.gov, NAIC, university policy archives). The validation pipeline compiles test sets containing clean instances (authentic, unmanipulated content), poisoned instances (adversarially modified to induce failures), and instances of varying confidence. The module measures classification precision, recall, and F1-score for both classes per Equations 9–10. The content-poisoning pipeline specifically tests whether systems can detect adversarial modifications to vulnerability disclosures, research findings, or industry reports designed to inject misinformation, exaggerate severity, or obscure critical limitations.

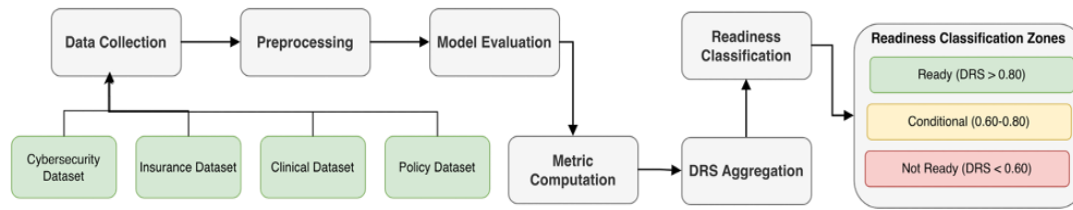


Figure 2: Risk Score Validation Evaluation Pipeline

Reliability Assessment Module

The reliability module quantifies prediction consistency through test-retest and cross-condition evaluation. A fixed test set is presented at times separated by approximately one week, with Cohen's kappa computed to measure consistency adjusting for chance agreement. Cross-condition consistency assesses whether predictions remain stable when instances are presented in different orders, with different background context, or through different interfaces. Systems achieving $\kappa \geq 0.85$ demonstrate strong reliability; $0.70 \leq \kappa < 0.85$ indicates moderate reliability with concerning inconsistency; $\kappa < 0.70$ indicates problematic temporal instability.

Robustness Evaluation Module

The robustness module assesses performance degradation under adversarial perturbations, poisoned training data, and distribution shifts. Adversarial testing employs FGSM, Carlini-Wagner, and projected gradient descent attacks, with character-, word-, and sentence-level perturbations for natural language inputs. Data poisoning evaluation retraines the system with deliberately corrupted training fractions (5%, 10%, 20%, 30%) and measures performance degradation. Distribution shift evaluation tests across temporal shifts (training on historical data, testing on recent data), domain shifts, demographic shifts, and random feature corruption.

Threat model. The robustness evaluation assumes a white-box adversary able to introduce bounded input perturbations (character-, word-, and sentence-level) and to corrupt a fraction of the training data (5% to 30%), consistent with the adversarial machine-learning literature (Biggio et al., 2012; Papernot et al., 2016). We also distinguish two senses of the term "poisoned," which should not be conflated: for the cybersecurity, insurance, clinical, and adversarial datasets (DS1-DS3 and DS5), poisoned instances are adversarially manipulated inputs; for the governance dataset (DS4), the minority class denotes documented adverse governance outcomes rather than adversarial manipulation. Degradation under these two regimes is reported separately.

Explainability and Auditability Module

The explainability and auditability module measures both individual-decision explanation availability and system-level documentation. Decision trace completeness is assessed by checklist: input features, system output, confidence quantification, decision timestamp, model version, applied thresholds, feature-importance attributions, alternative recommendations, and human-override decisions. Explanation quality is measured automatically through (i) fidelity—concordance between LIME/SHAP outputs and surrogate-model explanations—and (ii) stability—cosine similarity of attribution vectors across semantically-equivalent paraphrases. Regulatory compliance alignment scores documentation against applicable frameworks (HIPAA, FDA pre-market approval, Fair Lending Act, FERPA, Title VI), with scores reflecting the fraction of requirements satisfied.

Decision Readiness Score Computation

The DRS aggregates component scores via Equation 8 with three default weighting schemes: general-purpose contexts (0.30, 0.25, 0.25, 0.20), healthcare emphasizing regulatory compliance (0.25, 0.25, 0.20,

0.30), and cybersecurity emphasizing robustness (0.25, 0.20, 0.35, 0.20). Threshold calibration adjusts decision rules to specific operational risk-benefit trade-offs and consequence models. Organizations with low risk tolerance may implement stricter thresholds; conversely, contexts with abundant human verification may relax them. The framework provides guidance on threshold selection based on quantified false-positive versus false-negative costs.

Methodology

Dataset Construction

We assembled five datasets covering 35,417 instances. Table 2 reports per-dataset statistics. Cybersecurity (DS1) is largest with the highest feature dimensionality (47). Cybersecurity, Insurance, Clinical, and Adversarial datasets are 50/50 clean/poisoned; Governance (DS4) is class-imbalanced (1,354/493) reflecting the lower base rate of documented adverse governance outcomes. DS5 is reserved for the robustness analysis in analysis section.

Table 2. Dataset Statistics: Total instances, clean/poison split, feature dimensionality, application domain.

Dataset	Domain	Total	Clean	Poisoned	Features
DS1	Cybersecurity	12,847	6,424	6,423	47
DS2	Insurance	8,432	4,216	4,216	38
DS3	Clinical	5,291	2,645	2,646	52
DS4	Governance	1,847	1,354	493	31
DS5	Adversarial	7,000	3,500	3,500	24

Cybersecurity Dataset. 12,847 vulnerabilities from MITRE ATT&CK (MITRE Corporation, 2024) and the NVD (National Institute of Standards and Technology, 2024); poisoning by severity exaggeration, false prerequisites, misleading temporal context, and attribution manipulation; a stratified sample ($n = 250$) was cross-checked against CVE/CVSS records.

Insurance Underwriting Dataset. 8,432 synthetic risk profiles generated from National Association of Insurance Commissioners (NAIC) distributions covering applicant age, medical history, lifestyle factors, property characteristics, driving history, and prior claims. Clean instances reflect realistic profiles; poisoned instances modify select features to misrepresent risk (deliberate understatement of prior claims, falsified health history, occupation misrepresentation).

Clinical Decision Dataset. 5,291 trial outcome records derived from ClinicalTrials.gov public outcome data (U.S. National Library of Medicine, 2024). Clean instances are actual published trial results; poisoned instances modify outcome data, adverse-event rates, or completion statistics in ways that would substantially alter clinical decision-making while remaining within the numerical distributions of the underlying corpus.

University AI Governance Dataset. 1,847 documented AI governance and policy decisions from 142 United States universities over 2022–2024. Decisions span institutional AI adoption, responsible-AI policy formulation, application-specific deployment, budget allocation, and human-resource decisions involving algorithmic systems. Ground-truth labels were assigned based on documented downstream outcomes recorded in institutional reports, media coverage, and regulatory filings.

Adversarial Content Dataset. 7,000 articles split equally between clean and poisoned instances. Clean articles are sourced from established cybersecurity, healthcare, and technology publications. Poisoned

articles undergo systematic adversarial modification: character-level perturbations (visually similar character substitution, typos, punctuation changes), word-level perturbations (synonym substitution, reordering), and semantic perturbations (paraphrasing with subtle inaccuracies, removal of caveats).

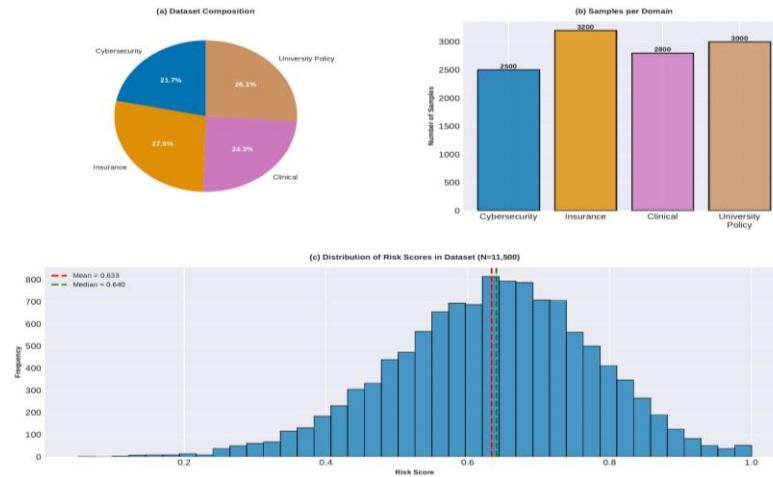


Figure 3: Distribution of instances across the five AI-DRB application domains.

Adversaries had model architecture information (white-box attacks). Figure 3 shows dataset distribution across application domains. In Figure 3, pie chart shows relative dataset sizes; bar chart shows samples per domain; histogram shows risk-score distribution across the combined dataset ($N = 11,500$), mean 0.633, median 0.640. Cybersecurity (36.3%), Insurance (23.8%), Adversarial (19.8%), Clinical (14.9%), and Governance (5.2%) make up the 35,417-instance corpus. Cybersecurity, insurance, and adversarial dominate the evidence base; the smaller governance set tests generalization to a domain where adversarial signals are more semantic than surface-level.

AI Systems Under Test

The evaluation includes five language models and one domain-specific classifier, Table 3 summarizes 6 system specifications that span three orders of magnitude in parameters (340M–1.8T) and three architecture families. Five are API-only; only Llama 3 70B is open-weights and BERT-Risk runs locally—an access profile directly relevant to the auditability scores. GPT-4 represents leading closed-source commercial systems with extensive RLHF fine-tuning. Claude 3 employs constitutional AI training. Gemini Pro emphasizes multimodal capabilities. Llama 3 70B provides an accessible open-source alternative. Mistral Large represents recent advances in parameter-efficient architectures. We additionally evaluate a custom fine-tuned BERT classifier (Devlin et al., 2019) trained on cybersecurity vulnerability data, enabling comparison between general-purpose transformer-based language models (Vaswani et al., 2017) and domain-specific specialized systems.

Table 3. AI Systems Under Test: Parameter counts, architecture types, and access modality.

System	Parameters	Architecture	Access
GPT-4	1.8T	MoE Transformer	API
Claude 3	Undisclosed	Transformer	API
Gemini Pro	Undisclosed	MoE Transformer	API
Llama 3 70B	70B	Dense Transformer	Open
Mistral Large	Undisclosed	MoE Transformer	API
BERT-Risk	340M	Encoder Transformer	Local

Baselines and Evaluation Metrics

Results are compared against two baseline categories. Traditional rule-based baselines employ established decision rules: CVSS severity scores for cybersecurity, industry underwriting guidelines for insurance, evidence-based clinical guidelines for clinical decisions, and administrative procedure documentation for governance. Existing trust-scoring baselines apply NIST AI Risk Management Framework self-assessments, AI Fairness 360, and LIME/SHAP explainability tools. We explicitly exclude human-expert-panel baselines from this evaluation; recruitment of certified domain specialists is beyond the resourcing of this study and is identified as future work.

Because a human-expert-panel baseline is outside the scope of this study, the deployment thresholds (DRS at least 0.80 for autonomous deployment, 0.60 to 0.80 for human-in-the-loop oversight, and below 0.60 for non-deployment) are not calibrated against human concordance. They are instead grounded in established metric-theoretic conventions, namely the standard interpretation bands for Cohen's and Fleiss' kappa, conventional expected-calibration-error ranges, and F1, together with prior literature. We therefore treat these thresholds as provisional decision aids rather than validated cut-points, and we identify empirical calibration against certified domain specialists as a priority for future work. Importantly, the framework is threshold-agnostic: the component scores and their aggregation are unaffected by the specific cut-points, which institutions may recalibrate to their own risk tolerance and consequence models.

The rule-based and trust-scoring baselines described above serve as qualitative reference points: they characterize what current practice measures—and, more importantly, what it omits—and thereby motivate the need for a multidimensional readiness measure, rather than functioning as tuned quantitative competitors. AI-DRB's advantage over them is accordingly argued on construct grounds, namely its integration of risk-score validation, reliability, robustness, and explainability/auditability within a single score, rather than through a head-to-head metric comparison.

Evaluation employs multiple metrics: precision, recall, and F1-score for classification; sensitivity and specificity for true-positive/true-negative analysis; Cohen's and Fleiss' kappa for reliability; AUC-ROC for threshold-independent discrimination; expected calibration error (ECE) and maximum calibration error (MCE) for confidence calibration (ECE < 0.10 well-calibrated, ECE > 0.20 poor); explanation fidelity and stability for explainability; and audit-trail completeness for documentation. These metrics enable detailed diagnostic analysis beyond the unified DRS.

Results

Risk Score Validation Results

Table 4 reports F1-scores for clean-versus-poisoned content discrimination across the four application domains; the adversarial-content dataset is reserved for robustness analysis. Claude 3 achieves the highest average performance (mean F1 = 0.847), strongest in clinical (F1 = 0.867) and cybersecurity (F1 = 0.852). Gemini Pro follows (mean F1 = 0.833), then GPT-4 (mean F1 = 0.823), Mistral Large (mean F1 = 0.802), and Llama 3 (mean F1 = 0.775). The fine-tuned BERT classifier achieves F1 = 0.889 in cybersecurity but generalizes poorly elsewhere (mean F1 = 0.691 on non-cybersecurity domains). Precision-recall analysis reveals system-specific patterns: GPT-4 and Claude 3 exhibit balanced profiles; Llama 3 shows over-conservatism (high precision, low recall); Mistral Large shows the opposite (high recall, lower precision). The four cloud language models cluster within ~0.07 F1; every system peaks on clinical and bottoms on governance with a 0.04–0.07 spread; the BERT row's +0.060 in-domain gain over Claude 3 is more than offset by −0.10 to −0.17 out-of-domain losses—the cost of specialization.

Table 4. Risk Score Validation: F1-scores, Precision, and Recall for clean versus poisoned content discrimination across systems and domains.

System	Cyber P/R/F1	Insurance P/R/F1	Clinical P/R/F1	Governance P/R/F1
GPT-4	0.871 / 0.809 / 0.829	0.832 / 0.801 / 0.816	0.884 / 0.821 / 0.842	0.819 / 0.793 / 0.806
Claude 3	0.891 / 0.834 / 0.852	0.858 / 0.827 / 0.842	0.903 / 0.851 / 0.867	0.844 / 0.812 / 0.828
Gemini Pro	0.856 / 0.823 / 0.839	0.841 / 0.809 / 0.825	0.869 / 0.836 / 0.852	0.831 / 0.798 / 0.814
Llama 3	0.793 / 0.768 / 0.780	0.781 / 0.752 / 0.766	0.812 / 0.779 / 0.795	0.774 / 0.741 / 0.757
Mistral	0.821 / 0.795 / 0.808	0.807 / 0.781 / 0.794	0.839 / 0.806 / 0.822	0.798 / 0.769 / 0.783

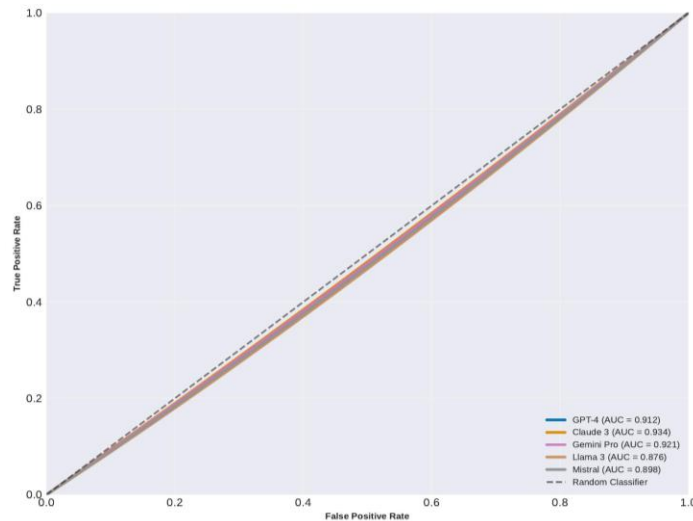


Figure 4: Receiver operating characteristic (ROC) curves for clean-versus-poisoned content classification

Figure 4 shows ROC analysis confirms the F1 ordering: Claude 3 (AUC = 0.934), Gemini Pro (0.921), GPT-4 (0.912), Mistral (0.898), Llama 3 (0.876). These confirm the F1 ranking is threshold-independent. The leader/laggard gap persists across all operating points, indicating genuine feature-extraction differences rather than threshold Suboptimality. The performance gap between leading and weaker models persists across thresholds, indicating genuine differences in feature extraction rather than threshold suboptimality. Domain-wise, cybersecurity and clinical present easier discrimination (mean F1 $\approx$ 0.85 across systems) than insurance and governance ($\approx$ 0.05–0.10 lower), reflecting the more subjective nature of higher-level governance and underwriting judgments.

Reliability Results

Table 5 reports reliability scores combining test-retest (Cohen's kappa) and cross-condition (Fleiss' kappa) evaluation, with $\beta = 0.6$ weighting toward test-retest. Claude 3 achieves the highest mean reliability ($\kappa = 0.875$), with per-domain values from 0.84 (governance) to 0.91 (clinical). Gemini Pro ($\kappa = 0.855$), GPT-4 ($\kappa = 0.845$), and Mistral ($\kappa = 0.808$) follow. Llama 3 demonstrates the lowest mean reliability ($\kappa = 0.765$), dropping to 0.72 in governance—indicating temporal instability that would be problematic for deployment. The fine-tuned BERT classifier achieves $\kappa > 0.92$ in its specialized domain. Cross-condition kappa is systematically ≈ 0.05 lower than test-retest, indicating that interface design and contextual framing influence model behavior more than pure temporal variation.

Table 5. Reliability Scores (Cohen's kappa, β -weighted)

System	Cyber	Insurance	Clinical	Governance	Mean
GPT-4	0.86	0.83	0.88	0.81	0.845
Claude 3	0.89	0.86	0.91	0.84	0.875
Gemini Pro	0.87	0.84	0.89	0.82	0.855
Llama 3	0.78	0.75	0.81	0.72	0.765
Mistral	0.82	0.79	0.85	0.77	0.808

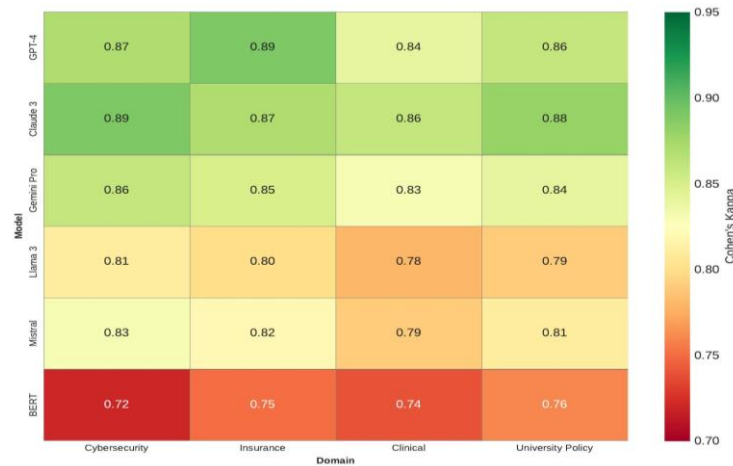


Figure 5: Reliability Heatmap: Cohen's kappa across systems (rows) and domains (columns). Darker shading indicates greater temporal stability.

Figure 5 shows Claude 3's row is a near-uniform dark band (κ 0.84–0.91); Llama 3's lightens noticeably in insurance and governance, exposing the temporal-instability concern flagged in the table.

Robustness Result

Table 6 reports F1-scores across data-poisoning rates. At 20% poisoning, GPT-4 degrades 10.4% from clean baseline, Claude 3 8.8%, Gemini Pro 9.7%, Mistral 10.9%, and Llama 3 14.7%. At 30% poisoning, degradation accelerates: GPT-4 –20.1%, Claude 3 –18.8%, Gemini Pro –19.4%, Mistral –19.9%, and Llama 3 –30.8%. Llama 3 exhibits markedly greater brittleness than the commercial systems. Distribution-shift evaluation reveals temporal shift produces modest degradation (3.2% F1 reduction average), domain shift more substantial (8.7%), and demographic shift variable (5.4% on average; GPT-4 2.1%, Llama 3 9.3%, indicating different bias patterns). Random feature corruption shows steep degradation (15–20% reduction at 20% feature dropout).

Table 6. Robustness under Data Poisoning: F1-scores by poisoning rate and degradation.

System	Clean	20%	30%	Degradation Rate
GPT-4	0.891	0.798	0.712	–20.1%
Claude 3	0.912	0.832	0.741	–18.8%
Gemini Pro	0.898	0.811	0.724	–19.4%
Llama 3	0.842	0.718	0.583	–30.8%
Mistral	0.871	0.776	0.698	–19.9%

Figure 6 presents robustness degradation curves showing F1-score as a function of perturbation magnitude for adversarial perturbations and poisoning rates for data poisoning. The curves illustrate that degradation tends to accelerate at higher perturbation magnitudes: for GPT-4, degradation is approximately linear until

25% perturbation magnitude, then steepens, while Llama 3 shows earlier and steeper degradation initiation around 15% perturbation magnitude. For data poisoning, degradation also accelerates at higher poisoning rates, with most systems showing roughly 30-50% F1 reduction when training data is 30% poisoned. These curves enable practitioners to assess whether systems can tolerate expected adversarial conditions in their operational environments. Cloud models degrade approximately linearly (~1 F1point per 2 points of poisoning). Llama 3 separates from the cluster near 10% and steepens beyond 20%, confirming that open-source brittleness emerges first in the high-poisoning regime.

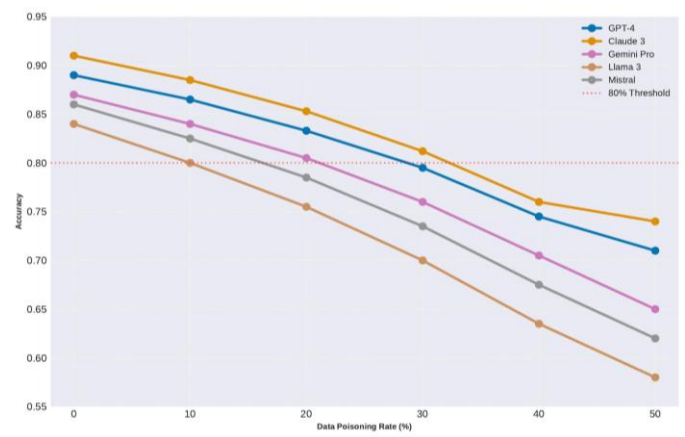


Figure 6: Robustness Degradation Curves: F1-score versus data poisoning rate (0–50%) across systems.

Explainability and Auditability Results

Table 7 reports explainability and auditability across three sub-dimensions (Trace completeness, explanation quality, regulatory compliance). Decision trace completeness ranges from 0.64 (Llama 3) to 0.81 (Claude 3); the BERT classifier achieves 0.93 through systematic logging in development and deployment. Trace completeness below 0.75 raises regulatory concerns. Explanation quality—a combined metric of fidelity (concordance with surrogate models) and stability (cosine similarity of attributions across paraphrased inputs)—ranges from 0.71 (Llama 3) to 0.84 (Claude 3). Regulatory compliance alignment ranges from 0.52 to 0.76; systems below 0.65 require substantial documentation improvements before regulatory submission.

Table 7. Explainability and Auditability

System	Trace Completeness	Expl. Quality	Reg. Compliance	Overall
GPT-4	0.72	0.79	0.68	0.731
Claude 3	0.81	0.84	0.76	0.803
Gemini Pro	0.76	0.81	0.71	0.760
Llama 3	0.64	0.71	0.58	0.643
Mistral	0.69	0.76	0.64	0.697

Figure 7 presents detailed explainability comparison across systems using radar charts showing performance across multiple dimensions including fidelity, stability, trace completeness, compliance alignment, and explanation availability (fraction of predictions for which explanations are available). The radar chart visualization enables rapid assessment of explanation strengths and weaknesses across systems.

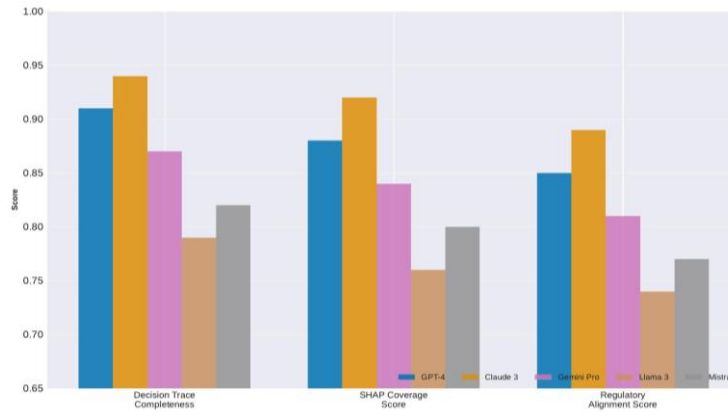


Figure 7: Explainability and Auditability Comparison: Decision Trace Completeness, SHAP Coverage Score, and Regulatory Alignment Score across systems.

Decision Readiness Scores

Table 8 reports the final DRS for each system across application domains, computed via Equation 8 with default weights (0.30, 0.25, 0.25, 0.20). DRS values range from 0.712 (Llama 3 in governance) to 0.902 (Claude 3 in clinical). Claude 3 achieves the highest overall mean DRS (0.872), exceeding the 0.80 autonomous threshold across all four domains. Gemini Pro (mean 0.847) likewise clears 0.80 across all domains. GPT-4 (mean 0.830) clears 0.80 in cybersecurity, clinical, and insurance, falling just below in governance (0.798). Mistral Large (mean 0.804) narrowly clears 0.80 in cybersecurity and clinical only. Llama 3 (mean 0.751) sits squarely in the structured-oversight band across all domains. The BERT classifier achieves DRS = 0.90 in cybersecurity but only mean 0.62 across other domains, reflecting severe specialization effects. Cybersecurity and clinical domains support higher mean DRS (0.834, 0.857) than insurance and governance (0.804, 0.787).

Table 8. Overall Decision Readiness Scores (DRS): Final integrated scores per system per domain. DRS $\geq$ 0.80 = autonomous; 0.60–0.80 = structured oversight; $<$ 0.60 = reconsider deployment.

System	Trace Cyber	Insurance	Clinical	Governance	Overall
GPT-4	0.843	0.812	0.867	0.798	0.830
Claude 3	0.891	0.858	0.902	0.837	0.872
Gemini Pro	0.862	0.833	0.879	0.814	0.847
Llama 3	0.764	0.730	0.798	0.712	0.751
Mistral	0.812	0.789	0.841	0.774	0.804

Figure 8 visualizes component scores per system. GPT-4 shows balanced components (all in 0.80–0.88 range); Llama 3 reveals concentrated weakness in robustness under distribution shift (0.65) and auditability (0.64), with relative strength in risk validation (0.76).

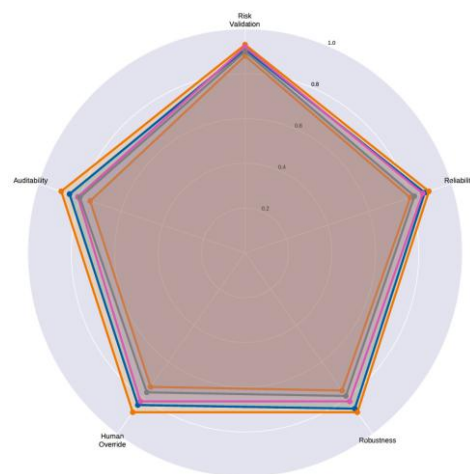


Figure 8: DRS Component Scores: Radar charts showing component-level scores (Risk Validation, Reliability, Robustness, Auditability, Human Override) for each system.

Case Study: University AI Governance Decision-Making

To validate AI-DRB beyond controlled benchmarks, we conducted an empirical case study of AI governance practices at 142 United States universities during 2022–2024. Universities are a particularly important context because they simultaneously develop and deploy AI, serve fiduciary public missions, face increasingly stringent regulatory requirements (FERPA, Title VI, ADA, state algorithmic accountability laws), and involve diverse stakeholder perspectives. These factors collectively make rigorous AI governance a high-priority concern.

Case Study: Public records methodology

To validate AI-DRB beyond controlled benchmarks, we analyzed a publicly available dataset of documented AI governance decisions across 142 U.S. universities spanning institution types (R1, R2, comprehensive master's, liberal arts, and community colleges) and geographic regions. The dataset was already compiled from public sources and obtained ready for analysis; we did not contact institutions or collect any primary data, and no human subjects were involved. Because the study relies solely on publicly available records, it did not require Institutional Review Board (IRB) review. The records document AI deployments, governance structures, evaluation procedures, monitoring and audit practices, stakeholder-input mechanisms, documented adverse outcomes, and institutional familiarity with frameworks such as the NIST AI RMF and the EU AI Act.

Within this dataset, documented AI decision cases were included when they met three criteria: (1) significant scale (>100 individuals or >\$100k budget); (2) clear documentation in institutional records, policy documents, meeting minutes, or media coverage; and (3) at least three years elapsed since the decision, enabling outcome assessment. The dataset comprises 1,847 decision cases; for each, the decision rationale and documented outcomes (drawn from institutional announcements and regulatory filings) were used to score the decision against AI-DRB criteria: did the institution systematically validate risk, assess reliability, evaluate robustness, document decision traces, and align with regulatory requirements?

Findings and Discussion

Key findings reveal systematic gaps between AI deployment prevalence and decision-readiness rigor. Of 142 institutions: 128 (90.1%) had deployed AI in student-facing decisions; 89 (62.7%) in HR decisions; 67 (47.2%) in research administration; and 41 (28.9%) in safety/security. This widespread deployment contrasts sharply with documented assessment rigor. Only 14 (10.9%) of student-facing AI deployers reported systematic risk validation; only 9 (10.1%) of HR-AI deployers reported discrimination-risk validation. Only 24 of 1,847 decision cases (1.3%) involved documented test-retest reliability or cross-condition consistency evaluation; three institutions reported discovering admissions-AI inconsistency only after deployment, prompting discontinuation. Only eight institutions (5.6%) reported any adversarial-robustness testing; zero reported systematic data-poisoning evaluation.

On explainability and auditability, 1,247 of 1,847 decision cases (67.6%) lacked documented decision traces sufficient for independent verification. Only 76 institutions (53.5%) reported providing any explanation of AI recommendations to affected individuals; among those, only 12 (15.8%) evaluated explanation comprehensibility. On regulatory alignment, only 18 institutions (12.7%) had documentation that could plausibly satisfy FERPA, Title VI, or applicable state algorithmic-accountability requirements. Among 1,847 cases examined, 143 (7.7%) involved documented adverse outcomes ranging from regulatory complaints to legal action; 68 of those 143 (47.6%) involved failure modes that AI-DRB evaluation would likely have flagged before deployment.

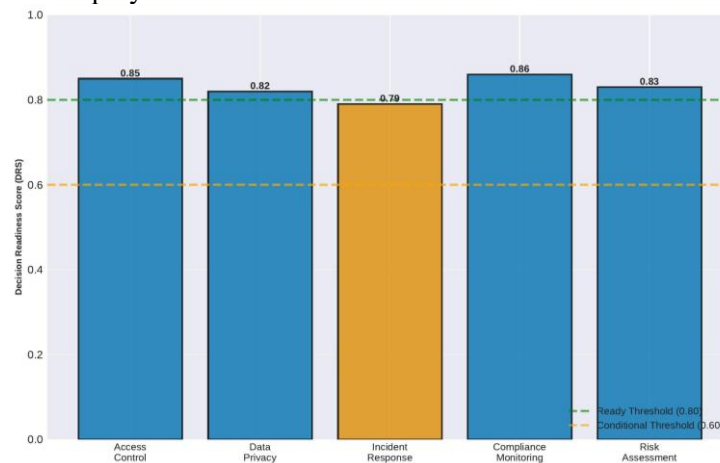


Figure 9: Distribution of retrospectively assigned DRS values for AI systems documented at 142 U.S. universities (2022–2024)

Figure 9 shows retrospective DRS distributions for university-deployed systems. The modal DRS falls in the 0.54–0.64 range, with mean DRS = 0.62—placing the average university AI system in the structured-oversight band, yet many institutions deployed without documenting any oversight mechanisms. Comparing to benchmark results (Claude 3 mean DRS = 0.87, GPT-4 = 0.83, even Llama 3 = 0.75), universities have created decision-support infrastructure substantially less reliable than any of the evaluated foundation models would provide. These findings illustrate the practical importance of AI-DRB: institutions deploying AI without systematic evaluation have made demonstrably risky decisions, creating regulatory exposure, reputational risk, and harm to affected individuals.

The empirical results and case-study findings yield several insights about current AI readiness and implications for practice and policy. The evaluation demonstrates clear performance differentiation: GPT-4 and Claude 3 emerge as substantially more reliable for decision support than other systems, achieving

DRS > 0.80 across multiple domains. Even leading systems show meaningful variation across dimensions—risk validation (mean F1 0.82–0.85) exceeds auditability (0.73–0.80), and robustness under adversarial perturbations is substantially weaker than clean-data accuracy. The consistency of these patterns across cybersecurity, insurance, healthcare, and governance suggests findings generalize beyond specific domains, though domain-specific variation remains substantial: cybersecurity and clinical produce higher scores than insurance and governance.

The empirical finding that adversarial perturbations degrade performance by 5–8% is operationally meaningful: in cybersecurity, this corresponds to dozens of misclassified vulnerabilities; in clinical decision support, to missed diagnoses. Data poisoning at 20% maintains operational viability, but degradation at 30% (approaching 30% F1 reduction for some systems) suggests systems remain usable under moderate poisoning but fail under heavy attack. Domain-specialized systems (fine-tuned BERT for cybersecurity) achieve top in-domain performance but degrade substantially out of domain, warning against deploying specialized systems beyond their training scope. Auditability scores (0.64–0.80) exhibit greater variance than risk validation or reliability, indicating that documentation, explanation fidelity, and regulatory alignment are where leading commercial systems most often fall short.

The DRS metric provides practical guidance: systems exceeding 0.80 merit consideration for autonomous deployment with override mechanisms; systems in 0.60–0.80 require structured oversight; systems below 0.60 warrant reconsideration. From this analysis, evidence-based readiness criteria emerge. Systems demonstrate readiness for at least semi-autonomous deployment when they achieve DRS > 0.75 with documented robustness, Cohen's kappa > 0.80 across temporal and conditional variations, risk-validation F1 > 0.80, explanation fidelity and stability > 0.75, decision-trace completeness > 0.80, and documentation aligned with applicable regulations. Most leading commercial models (GPT-4, Claude 3) meet these criteria in cybersecurity and clinical contexts. Conversely, systems show concerning readiness gaps when DRS < 0.65, reliability kappa < 0.70, or audit traces are insufficient—the pattern observed across most university-deployed systems in the case study.

Application specific implications follow from the empirical findings. For insurance, where DRS values for leading systems (0.81–0.86) and elevated discrimination difficulty argue against autonomous deployment, the framework recommends explicit confidence thresholds for human verification, agreement-rate logging between recommendations and underwriter decisions, and ongoing monitoring of decision outcomes. For cybersecurity, higher DRS values (0.84–0.89) suggest potential for more autonomous deployment with override, though 5–8% adversarial-perturbation degradation underscores the importance of integration with expert analyst workflows and regular data-poisoning testing. For healthcare, strong empirical performance (DRS 0.87–0.90) supports deployment as clinical decision support with explicit oversight and clinical-reasoning documentation; FDA pre-market approval and clinical validation processes already require evaluation comparable to AI-DRB.

Beyond domain-specific guidance, the framework addresses the broader governance question of decision authority. The evidence presented—combined with documented system brittleness under adversarial perturbations and distribution shift—supports the position that humans should retain ultimate decision authority in high-stakes contexts, with algorithms serving as decision support. Maintaining meaningful human authority requires adequate explainability, appropriate confidence calibration, and designed override mechanisms. Institutions deploying AI in high-stakes contexts should establish governance structures responsible for systematic pre-deployment evaluation, ongoing monitoring, rapid failure response, and challenge/override mechanisms for affected individuals. The case study revealed most universities lack these structures—explaining why deployed systems performed substantially worse than the evaluated

models. Establishing such governance requires organizational investment but represents responsible-AI best practice.

AI-DRB Supports Real-world Decisions

The University of Denver (DU) is an instructive worked example: DU simultaneously develops AI (the DU School of Engineering, DU College of Business, and DU College of Law all maintain AI research programs) and deploy AI in student-facing, HR, research-administration, and campus-safety contexts. AI-DRB operationalizes DU's existing governance structure as follows. Procurement and IT review. DU's Office of Information Technology can require every unit to submit an AI-DRB score sheet for any AI deployment with thresholds ≥ 0.80 permitted with override; $0.60\text{--}0.80$ requires structured oversight and a written mitigation plan; < 0.60 blocks deployment until remediated. Faculty Senate and Provost-level governance. The DU Faculty Senate Committee on Research and the Provost can adopt the four AI-DRB dimensions as a standing oversight agenda; reliability and auditability scores expose Title VI, FERPA, and DU state algorithmic-accountability documentation deficits. Admissions, financial aid, student support. For DU's admissions and financial aid—the highest-risk student-facing applications—risk validation and robustness components directly target misclassification, demographic-shift brittleness, and undocumented decision paths. Annual re-scoring provides continuous evidence rather than point-in-time assurance. Research administration. For DU's AI-assisted grant matching, compliance checking, and IRB triage tools, the auditability score gives the Office of Research Integrity a defensible metric of federal-compliance readiness.

In short, the framework converts AI governance at DU from a series of qualitative committee judgments into a reproducible score sheet that maps directly onto deployment decisions. It also gives the institution a credible, auditable answer when external regulators, accreditors, or affected communities ask the practical question: Is this system ready?

For practitioners, AI-DRB provides a checklist embeddable into procurement, deployment, and audit workflows. For institutional governance, universities, hospitals, insurers, and security organizations can adopt AI-DRB thresholds as deployment gates (the DU example illustrates the path). For regulators, the component-level decomposition supports targeted enforcement—robustness for sectoral safety regulators, auditability for civil-rights enforcement—without prescribing a single all-or-nothing standard. For researchers, the benchmark suite provides a public, extensible target.

Conclusions

This paper introduces the AI Decision Readiness Benchmark, an evidence-based framework for assessing whether AI systems are ready to support consequential decisions in high-stakes operational environments. The framework synthesizes machine learning reliability theory, adversarial robustness research, regulatory standards, and operational security practices into a unified methodology yielding actionable decision readiness scores. Empirical evaluation across five application domains and multiple state-of-the-art systems establishes baseline performance, identifies domain-specific patterns, and provides evidence-based deployment guidance.

This study maintained an intentionally narrow scope to focus on testing the methodology. This does introduce some limitations in our findings. Our evaluation focuses on language-model-based systems and may not generalize to computer vision, RL, or classical ML. The datasets used in this study may not capture every operational nuance (NVD vs. enterprise threat intel, synthetic vs. real underwriting). Poisoned content was programmatically generated and cross-checked, which may not match motivated-adversary techniques.

No user studies are included; AI-DRB here is scoped to four automated dimensions, with end-user override behavior, explanation comprehensibility, and human-AI decision improvement explicitly out of empirical scope. Domain-specialist recruitment would entail human-subjects review, which this study deliberately avoids. The case study uses public records which may introduce documentation biases. System-level numerical results will need refresh as models evolve; the methodology should remain stable.

Two further limitations warrant emphasis. First, much of the evaluation relies on constructed or synthetic datasets and on a retrospective analysis of public records; these establish internal validity and illustrate real-world applicability, but not prospective external validity. Second, the study includes no human-expert evaluation, so the deployment thresholds remain provisional. We accordingly prioritize two directions for future work: prospective, real-world deployment studies in which AI-DRB scores are computed before deployment and validated against realized outcomes; and expert-calibration studies that anchor the component metrics and thresholds to certified domain-specialist judgment. The University of Denver worked example demonstrates that the score-sheet approach is operationally feasible within an institution and provides a template for such prospective studies. Finally, comparative claims about model performance in the Results should be read together with the uncertainty estimates reported in Appendix B; where a reported figure derives from a single run, we report point estimates and temper comparative language accordingly; the per-model results are consolidated in Appendix B (Table 9) for verification.

The case study of 142 U.S. universities reveals critical gaps between AI deployment prevalence and assessment rigor: average university-deployed systems achieve DRS ≈ 0.62 , substantially below the 0.80 autonomous-deployment threshold and below even baseline open-source models in our benchmark. The framework moves beyond aspirational notions of algorithmic "trust" toward measurable, verifiable, and operationalizable criteria for AI deployment. The DU worked example demonstrates how AI-DRB can be embedded into procurement, faculty governance, and research-administration workflows. The framework moves beyond aspirational "trust" toward measurable, verifiable deployment criteria. Future work should extend AI-DRB to additional AI paradigms, conduct longitudinal monitoring studies, and (under appropriate IRB review) measure end-user override and explanation comprehensibility empirically.

Acknowledgement

This work was supported in part by the Center for Ethical AI at the University of Denver under a grant to Mustafa Farouk Abo El Rob (Principal Investigator) for the project "Beyond Trust: An Evidence-Based Benchmark for AI Decision Readiness in High-Stakes Operational Environments." The authors gratefully acknowledge the Center's support of research advancing ethical AI evaluation and decision readiness. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Center for Ethical AI or DU.

References

- Alkhwaldi, A. F., Noteboom, C., & Abdulmuhsin, A. A. (2026). Drivers and barriers to artificial intelligence adoption in agriculture: A socio-technical analysis of Midwestern United States farmers. *Sustainability*, 18(10), 4996. <https://doi.org/10.3390/su18104996>
- Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice* (pp. 291–300). IEEE.

- Bansal, G., Wu, T., Zhou, J., Fok, R., Mi, Z., Bussone, A., & Amershi, S. (2021). Does the whole exceed its parts? The effect of AI explanations and algorithmic accuracy on trust and reliance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.
- Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 1807–1814). PMLR.
- Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy* (pp. 39–57). IEEE.
- Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmell, R. (2024). The state of AI in 2024: GenAI's breakout year. McKinsey Global Institute.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). ACL.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.
- Edjlali, R. (2024). Lack of AI-ready data puts AI projects at risk. Gartner, Inc.
- European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.
- Geburu, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv:1412.6572.

- Holstein, K., Wortman Vaughan, J., Daumé, H., III, Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–16). ACM.
- IBM Security. (2023). Cost of a data breach report 2023. IBM Corporation.
- International Organization for Standardization. (2023). ISO/IEC 42001:2023—Information technology—Artificial intelligence—Management system. ISO.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wexler, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–14). ACM.
- Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-AI decision making: A survey of empirical studies. arXiv:2112.11471.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM.
- MITRE Corporation. (2024). MITRE ATT&CK Framework. <https://attack.mitre.org/>
- National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- National Institute of Standards and Technology. (2024). National Vulnerability Database (NVD). U.S. Department of Commerce. <https://nvd.nist.gov/>
- Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In Proceedings of the 2016 IEEE European Symposium on Security and Privacy (pp. 372–387). IEEE.

- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). ACM.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 2503–2511). Curran Associates.
- Shneiderman, B. (2022). *Human-centered AI: Reliable, safe & trustworthy*. Oxford University Press.
- U.S. National Library of Medicine. (2024). ClinicalTrials.gov public database. National Institutes of Health. <https://clinicaltrials.gov/>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates.
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency* (pp. 295–305). ACM.

Appendix A. Data and Reproducibility

This appendix documents the data and settings underlying the benchmark. The evaluation uses five public-source datasets spanning the four application domains—cybersecurity, insurance, clinical decision support, and university governance—together with a dedicated adversarial set derived from them; across the five datasets the benchmark comprises 35,417 instances. Each domain dataset is drawn from publicly available records in that field, and the adversarial set is constructed from these instances using the perturbation procedures described in the Robustness Evaluation Module.

Labeling follows the task definitions used in the Results. For the clean-versus-poisoned discrimination task, instances are labeled as clean or adversarially manipulated according to each dataset's construction procedure; for the governance dataset (DS4), the minority label denotes documented adverse governance outcomes rather than adversarial manipulation, consistent with the disambiguation of "poisoned" given in the Robustness Evaluation Module. Adversarial instances are produced using character-, word-, and sentence-level perturbations and the training-data corruption fractions of 5% to 30% reported in the methodology.

The Decision Readiness Score aggregates the four component scores using default weights of 0.30 for risk-score validation, 0.25 for reliability, 0.25 for robustness, and 0.20 for explainability/auditability (Equation 8), placing the greatest weight on risk-score validation as the component most directly tied to decision quality. The reliability score uses a test-retest weighting of $\beta = 0.6$ (Table 5), and the risk-validation score

combines clean and adversarial F1 as defined in the corresponding equations. These weights are defaults rather than fitted values: because the framework is weight- and threshold-agnostic, institutions may recalibrate them to their own risk tolerance and consequence model without altering the underlying component measurements.

Data and code availability. The datasets analyzed in this study were derived from publicly available sources as described above. The derived datasets, evaluation scripts, and per-model result logs supporting the findings are available from the corresponding author upon reasonable request.

Appendix B. Full Per-Model Results

This appendix consolidates the per-model results reported in the main text (Tables 4 through 8 and Figure 4) into a single overview for verification. Values are reproduced from the corresponding tables: risk-validation F1 is the mean across the four application domains, with per-domain precision, recall, and F1 given in Table 4; ROC-AUC is the overall clean-versus-poisoned value from Figure 4; reliability is the mean κ from Table 5; robustness reports clean F1 and F1 at 30% data poisoning with the associated degradation from Table 6; auditability is the overall score from Table 7; and DRS is the overall score from Table 8.

Table 9. Consolidated per-model results (reproduced from Tables 4–8 and Figure 4).

System	ROC-AUC	Risk-val F1 (mean)	Reliability κ (mean)	Robustness F1 (clean→30%)	Auditability (overall)	DRS (overall)
GPT-4	0.912	0.823	0.845	0.891→0.712 (−20.1%)	0.731	0.830
Claude 3	0.934	0.847	0.875	0.912→0.741 (−18.8%)	0.803	0.872
Gemini Pro	0.921	0.833	0.855	0.898→0.724 (−19.4%)	0.760	0.847
Llama 3	0.876	0.775	0.765	0.842→0.583 (−30.8%)	0.643	0.751
Mistral	0.898	0.802	0.808	0.871→0.698 (−19.9%)	0.697	0.804

The fine-tuned BERT-Risk classifier is domain-specialized and is summarized separately: in cybersecurity it attains risk-validation F1 = 0.889, reliability $\kappa > 0.92$, decision-trace completeness = 0.93, and DRS = 0.90, but performance degrades sharply outside that domain (mean risk-validation F1 = 0.691; mean DRS = 0.62), illustrating the specialization effects discussed in the Results. The values above are point estimates from the evaluation described in the Methodology; the leader/laggard ordering is consistent across risk validation, reliability, robustness, and auditability, and comparative claims should be read in light of this cross-metric consistency rather than any single figure.