

The AI Deployment Audit Playbook: An Operational, Lifecycle-Oriented Framework for the Global South

César Giovanni Cárdenas Olivares, Itan Gabriel Ramirez Miramontes,
Francisco Segura Valencia
Universidad de Guadalajara, Centro Universitario de Tonalá (CUTonalá), Mexico

Track: AI Governance — **Sub-track:** System Auditing and Accountability
Code and Data: N/A (see Code and Data section)

Abstract

The accelerated deployment of artificial intelligence is occurring within a pronounced regulatory and institutional vacuum, where legal compliance has become an unreliable proxy for algorithmic safety amid shifting statutory deadlines and politically driven rollbacks. This paper presents the AI Deployment Audit Playbook, a voluntary, operational, and lifecycle-oriented workflow developed during the Global South AI Safety Hackathon (AI Governance track, System Auditing and Accountability sub-track). Rather than proposing new ethical principles, the playbook operationalizes existing instruments — including ISO/IEC 42001, ISO/IEC 23894, the NIST AI Risk Management Framework, the UNESCO Readiness Assessment Methodology, and OECD AI principles — into a ten-phase audit lifecycle organized around a three-block evaluation architecture and a double-perspective approval formula. Grounded in the volatile Mexican regulatory context, the framework is illustrated through two scenario walkthroughs rather than empirical field validation. We discuss its limitations, uncalibrated thresholds, and dual-use risks, including audit washing and superficial compliance.

1 Introduction

The governance of artificial intelligence (AI) currently suffers from a profound mismatch between the pace of technological adoption and the pace of institutional maturity. Highly capable, automated decision-making systems are being deployed at scale into legal and administrative environments characterized by outdated statutes, politically vulnerable regulatory agencies, and technical standards that frequently overlook the localized realities of the Global South.

This structural risk can be conceptualized as an adoption-maturity gap, $G = A - M$, where A denotes an index of AI adoption within a jurisdiction and M an index of institutional and regulatory maturity; we treat this as a theoretical heuristic rather than a calibrated metric (Section 5). The gap is visible in Mexico, where enterprise AI adoption is estimated above 80% (DevsData/IDC, 2026), even as the country's data protection authority, the Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI), was formally dissolved on 9 May 2025. Its functions transferred to the Secretaría Anticorrupción y de Buen Gobierno (SABG), an executive-branch body that absorbed roughly 80% of INAI's functions while retaining only about 35% of its organizational structure (Directorio Legislativo, 2025; Infobae, 2025).

Statutory compliance elsewhere has proven similarly unreliable as a safety metric. Colorado's SB 24-205, the first comprehensive U.S. state algorithmic-discrimination law, had its effective date delayed twice, was judicially paused in April 2026, and was ultimately repealed

and replaced in May 2026 by a narrower automated-decision-making statute, SB 26-189, before its risk-management and duty-of-care provisions ever took effect (Akin Gump, 2025; Crowell & Moring, 2026). California's AB 2013 training-data transparency law took effect on 1 January 2026 but remains under active legal challenge (Crowell & Moring, 2025; Norton Rose Fulbright, 2026). Internationally, the EU AI Act's Digital Omnibus deferred high-risk obligations by sixteen months in a political agreement reached on 7 May 2026 (Hogan Lovells, 2026; White & Case, 2026). Organizations cannot rely on the statutory calendar alone to guarantee algorithmic safety.

The problem this paper addresses is operational, not theoretical: if an organization wants to deploy an AI system, what should be audited, in what order, with what evidence, under what criteria, and with what outputs — before, during, and after deployment? We address this through the AI Deployment Audit Playbook, submitted to the AI Governance track (System Auditing and Accountability sub-track) of the Global South AI Safety Hackathon. The playbook does not claim to invent AI auditing or risk management. This submission makes three concrete contributions, all developed during the hackathon:

- A **three-block asymmetric evaluation architecture** that separates mandatory baseline compliance (Block I) from government-specific (Block II) and private-sector-specific (Block III) assessments, ensuring that no deployment can be approved without satisfying universal requirements regardless of sector.
- A **ten-phase, iterative audit lifecycle** that operationalizes existing governance principles into a concrete sequence of evidence activities, decision gates, and re-audit triggers — executable by an independent audit team under legally volatile conditions.
- A **double-perspective approval formula** that requires both a user-protection score (S_u) and an organizational-compliance score (S_o) to clear independent thresholds before a deployment can be approved, preventing either dimension from masking deficiencies in the other.

2 Related Work

The playbook synthesizes globally recognized governance instruments into functional audit checkpoints rather than summarizing them at length. ISO/IEC 42001:2023 specifies requirements for an organizational AI management system and informs the playbook's procedural-governance checks (Phase 4); ISO/IEC 23894:2023 provides risk-management guidance that informs the system-mapping and continuous-monitoring phases (Phases 3 and 9) (ISO/IEC, 2023a, 2023b). The NIST AI Risk Management Framework's map-measure-manage structure informs the requirements-gathering phase (Phase 2) (NIST, 2023). UNESCO's Readiness Assessment Methodology, developed under the 2021 Recommendation on the Ethics of Artificial Intelligence, and the OECD Framework for the Classification of AI Systems inform the sociotechnical impact assessment (Phase 6) by supplying legal, social, and technical readiness dimensions and a shared vocabulary for system context, particularly regarding their capacity to assess impacts on vulnerable populations amid the unequal access conditions and linguistic diversity characteristic of the Global South (UNESCO, 2021; OECD, 2022).

Jurisdictional instruments anchor the legal-review phase (Phase 5): the EU AI Act's fundamental-rights provisions (European Union, 2024), the algorithmic-discrimination parameters originally codified in Colorado's SB 24-205 and now superseded by SB 26-189 (Colorado General Assembly, 2024, 2026), California's AB 2013 training-data transparency mandate (California State Legislature, 2024), and Mexico's Ley Federal de Protección de Datos Personales en Posesión de los Particulares (LFPDPPP), now supervised by the SABG following INAI's dissolution (Pandectes, 2025).

A separate input, external to this paper, is VIGÍA, a sociotechnical risk taxonomy developed by Max Pinelo and Pilar Moncada in collaboration with AI Safety Mexico. VIGÍA is not authored by this paper’s authors; the playbook draws on its risk categories — including structural discrimination, concentration of power, and state surveillance — strictly as an input to Phase 6, in the same way it draws on UNESCO and OECD instruments (Pinelo & Moncada, 2025).

Table 1: Comparison of the proposed playbook with existing AI governance frameworks.

Framework	Core Focus	Risk Assessment	Audit Workflow	Decision Logic
ISO/IEC 42001	Organizational management systems	General	No	No
ISO/IEC 23894	Risk management guidance	Yes	No	No
NIST AI RMF	Map-Measure-Manage structure	Yes	No	No
UNESCO RAM	Ethical and readiness diagnostics	Yes	No	No
OECD Framework	System context classification	Yes	No	No
Proposed Playbook	Operational integration layer	Yes (Integrated)	Yes (10-Phase)	Yes (Formulaic)

None of these instruments independently provides a step-by-step, execution-focused workflow for an audit team operating under legally volatile conditions. ISO/IEC 42001 defines management-system requirements without prescribing the ordered sequence of evidence activities or decision gates that a third-party auditor should follow. ISO/IEC 23894 provides risk-guidance vocabulary but no deployment-decision formula. The NIST AI RMF’s map-measure-manage structure is deliberately high-level and sector-agnostic, leaving operationalization to the implementing organization. UNESCO’s Readiness Assessment Methodology and the OECD classification framework supply diagnostic readiness dimensions but do not constitute audit execution protocols. The playbook’s contribution is thus an integration layer: it chains these instruments’ principles into a phase-ordered, gate-driven procedure — with an explicit conjunctive approval formula — designed for audit teams operating under the institutional uncertainty characteristic of Global South regulatory environments (see Table 1).

3 Methods

The methodology replaces a traditional linear compliance checklist with an iterative, ten-phase audit lifecycle, executed by an independent audit team working alongside the deploying organization. The playbook organizes evaluation criteria into a Three-Block Asymmetric Evaluation Architecture. Block I comprises mandatory baseline compliance checks common to every deployment, regardless of sector. Block II adds requirements specific to government deployments, such as legal-basis review for cross-agency data convergence. Block III adds requirements specific to private-sector deployments, such as consumer-protection and terms-of-service review. A deployment cannot be approved unless it clears Block I in addition to its applicable sector-specific block (see Figure 1).

Table 2 summarizes the ten phases, condensing the objective, governing framework, and primary output and decision gate of each. The lifecycle is iterative rather than strictly linear: a re-audit trigger detected in Phase 10 returns the engagement to Phase 1.

To ground the framework in Global South realities, Phase 6 (Sociotechnical Impact Assessment) requires auditors to explicitly evaluate a deployment’s accessibility against regional socioeconomic inequalities and varying digital-literacy levels. This phase operationalizes inclusion by mandating concrete checks for cultural appropriateness, linguistic accessibility—including support for indigenous-language speakers—and the system’s potential to inadvertently exclude

Table 2: Ten-phase audit lifecycle: objective, governing framework, and output/decision gate by phase.

Phase	Core Objective	Governing Framework(s)	Key Output / Decision Gate
1. Engagement Initiation	Define audit scope, legal parameters, and Block II/III classification.	Internal scoping; no external framework	Output: signed NDA, Audit Plan. Gate: sufficient independent access?
2. Requirements Gathering	Establish technical, legal, and operational baseline.	NIST AI RMF (Map function)	Output: Consolidated Base Documentation. Gate: requirements sufficiently defined?
3. System and Data Mapping	Trace data lineage, actor responsibilities, and "Silent AI" components.	ISO/IEC 23894; NIST AI RMF	Output: Data Flow Diagram. Gate: can the organization map its own system?
4. Technical Assessment	Empirically verify access controls, robustness, and security claims.	ISO/IEC 42001; ISO/IEC 23894	Output: Technical Audit Report. Gate: does technical reality match documented claims?
5. Legal and Regulatory Review	Evaluate the deployment against binding jurisdictional law.	EU AI Act; Colorado SB 26-189; California AB 2013; Mexico's LFPDPPP	Output: Legal Compliance Checklist. Gate: statutory violations that prohibit deployment?
6. Sociotechnical Impact Assessment	Assess equity and human-rights impacts on marginalized users.	UNESCO RAM; OECD AI Classification; VIGÍA taxonomy (external input)	Output: Sociotechnical Impact Evaluation. Gate: does the system disenfranchise specific groups?
7. Risk Scoring and Prioritization	Quantify residual risk via the Double-Perspective Approval Formula.	Synthesizes Phases 4–6	Output: Prioritized Risk Matrix. Gate: do scores satisfy the conjunction formula?
8. Recommendations & Decision	Define remediation and issue a go/no-go decision.	Synthesizes all prior phases	Output: Final Audit Report, Improvement Plan. Gate: go/no-go on deployment?
9. Continuous Monitoring	Maintain post-deployment oversight for drift and degradation.	ISO/IEC 23894	Output: Monitoring Logs. Gate: operating within accepted risk parameters?
10. Re-Audit	Define triggers that mandate repeating the lifecycle.	Internal trigger logic	Output: Re-Audit Trigger Notification. Gate: does a detected change warrant a new Phase 1 engagement?

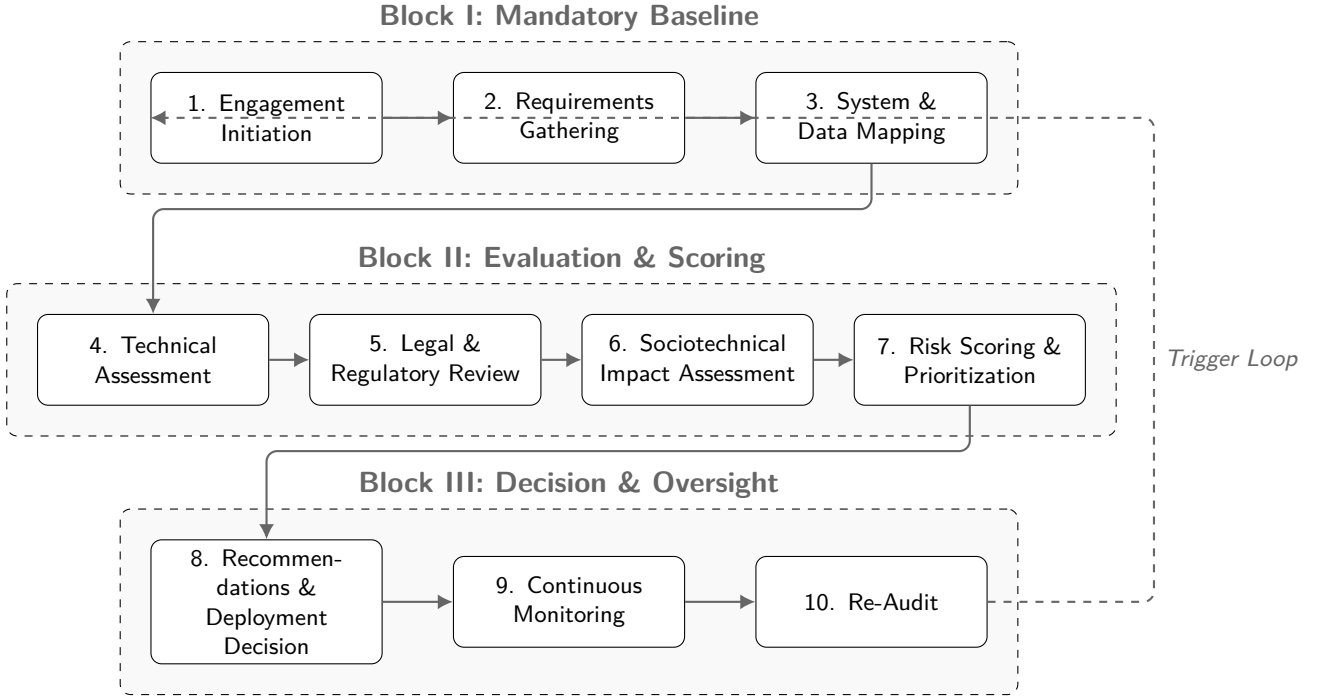


Figure 1: Overview of the AI Deployment Audit Playbook. The framework is organized into three operational blocks comprising ten audit phases. Block I establishes mandatory governance foundations. Block II executes risk and evidence assessments. Block III determines approval and ensures continuous oversight through monitoring and re-audits.

rural or underserved populations.

Following this impact evaluation, Phase 7 applies the playbook’s Double-Perspective Approval Formula, which requires a deployment to satisfy Block I and clear two independently scored thresholds before approval:

$$\text{APPROVE} \iff (\text{Block I}) \wedge (S_u \geq \theta_u) \wedge (S_o \geq \theta_o) \quad (1)$$

where S_u is an aggregate user-protection score drawn from Phases 5 and 6, S_o is an aggregate organizational-compliance score drawn from Phase 4, and θ_u, θ_o are minimum passing thresholds. The logical conjunction prevents strong organizational-management metrics from masking weak user protections, or vice versa.

We distinguish explicitly between what is operational and what remains theoretical. The ten phases, their evidence requirements, and their decision gates are operational: they were designed to be executed by an audit team and exercised against two illustrative scenarios (Section 4). The approval formula’s logical structure is operational, but its threshold values (θ_u, θ_o) are uncalibrated placeholders requiring empirical baseline data from real-world audits before the formula can be applied in practice. The adoption-maturity gap equation introduced in Section 1 is likewise a theoretical heuristic, not a validated quantitative model.

4 Results

Because field-testing the playbook against a live organizational deployment was not feasible within the hackathon timeframe, we do not report empirical validation results. Instead, Table 3 presents two illustrative walkthroughs tracing the playbook’s mechanics against publicly documented regulatory scenarios discussed in Section 1.

The walkthroughs are designed to demonstrate three properties: (i) internal consistency — the ten phases execute in sequence and produce a non-contradictory decision; (ii) phase traceability — each finding maps to a specific phase and its governing framework; and (iii) operational applicability — the decision gates function as designed, halting the engagement when Block I criteria are unmet (Scenario A) and sustaining user protections independently of statutory rollback (Scenario B). These walkthroughs do not constitute field validation, do not establish that the playbook detects real-world harms more reliably than existing audit practice, and do not validate the approval formula’s scoring thresholds.

Table 3: Illustrative audit walkthroughs (not empirical validation).

Scenario	Context	Audit Finding (Phases 3, 5, 6)	Outcome
A: Government biometric ID rollout	A government mandates centralized biometric data collection under a weakened privacy regulator.	Phase 3 surfaces unverified transborder data-transfer nodes; Phase 5 finds the regulator lacks proportional oversight capacity; Phase 6 flags extreme state-surveillance risk.	Fails Block I; audit halts at Phase 8 before mandatory rollout.
B: Private-sector hiring tool under regulatory rollback	An enterprise deploys an AI hiring system as the relevant state algorithmic-discrimination law is delayed by political pressure.	Phase 6 demands empirical non-discrimination evidence independent of the (subsequently repealed) statute.	Phase 9 monitoring sustains user protections despite the legislative rollback.

Both walkthroughs are reconstructions built from publicly reported facts about the underlying regulatory episodes; they are not derived from an organization that was actually audited using this playbook. We present them as proof-of-concept demonstrations of the workflow’s mechanics, not as evidence of effectiveness.

5 Discussion and Limitations

Limitations. Four limitations constrain the playbook’s present form. First, the adoption-maturity gap equation ($G = A - M$) remains an unvalidated heuristic; A and M have no agreed quantitative operationalization. Second, the playbook is an organizational and governance audit workflow; it deliberately excludes technical AI safety vectors such as mechanistic interpretability, adversarial robustness testing, and data-poisoning detection, limiting its applicability to laboratory-level foundation-model evaluation. Third, the Double-Perspective Approval Formula’s thresholds are uncalibrated placeholders that require empirical baseline data from real audits before the formula can be computed in practice. Fourth, Phase 6’s evaluation of consent “comprehensibility” lacks a validated psychometric instrument; what counts as comprehensible consent is asserted rather than measured. Similarly, the playbook does not yet provide validated quantitative metrics for evaluating cultural appropriateness, linguistic accessibility, or digital inclusion, meaning future iterations must develop standardized assessment instruments to robustly measure these dimensions.

Dual-use considerations. An audit framework can be misused as readily as it can be used well. We flag five specific risks. *Audit washing*: an organization could commission a playbook-based audit and cite its completion as evidence of due diligence without implementing any of its

recommendations. *Superficial compliance*: because Phases 1–8 generate substantial documentation, an organization could optimize for producing the required artifacts (NDAs, checklists, reports) rather than for reducing the underlying risks those artifacts are meant to evidence. *Legitimizing harmful systems*: a passing audit score does not guarantee a system is safe to deploy at scale, particularly for systems whose risks emerge only post-deployment (e.g., model drift, adversarial misuse); presenting a Phase 8 approval as a permanent safety guarantee would itself be a misuse of the framework. *Optimizing paperwork over risk reduction*: reliance on aggregate scores creates an incentive to game component scores rather than address the underlying harms they represent. *Selective disclosure*: because the playbook produces a Final Audit Report rather than mandating public disclosure of all phase-level findings, an organization could publicize a favorable summary while withholding adverse phase-level evidence — a dynamic structurally analogous to selective reporting in voluntary environmental and financial audits.

Mitigation. These risks are mitigated, though not eliminated, by three design choices already present in the playbook: the requirement that the audit team be independent of the deploying organization (Phase 1); the Phase 9 continuous-monitoring obligation, which prevents a one-time Phase 8 approval from substituting for ongoing oversight; and the conjunctive structure of the approval formula, which requires both scores to clear their thresholds rather than allowing either to compensate for the other. Future versions should consider mandatory minimum disclosure requirements for Phase 8 reports and third-party verification of Phase 4 evidence, both of which remain outside the scope of the present hackathon submission.

6 Conclusion

AI governance cannot rely exclusively on static, frequently revised legal frameworks. The AI Deployment Audit Playbook responds to the adoption-maturity gap by translating existing governance instruments — ISO/IEC 42001, ISO/IEC 23894, the NIST AI RMF, UNESCO’s Readiness Assessment Methodology, and OECD AI principles — into an operational, ten-phase audit lifecycle organized around a three-block evaluation architecture and a double-perspective approval formula. The contribution of this hackathon submission is the workflow itself: an explicit sequence of evidence activities and decision gates, illustrated through two scenario walkthroughs rather than empirical validation. The framework’s thresholds remain uncalibrated, and its dual-use risks — audit washing, superficial compliance, and selective disclosure chief among them — require active mitigation rather than passive trust in the audit process. We see the playbook as a starting point for field-tested, Global-South-grounded AI governance tooling, not as a finished or validated instrument.

Code and Data

No code or dataset was produced for this submission. The playbook is a methodological and procedural framework; its phases, evidence requirements, and decision gates are documented in full in Section 3. Code and Data: N/A.

Author Contributions

All three authors (C.G.C.O., I.G.R.M., F.S.V.) jointly conceived, designed, and wrote this submission during the Global South AI Safety Hackathon. A detailed, role-specific contribution breakdown (e.g., following the CRediT taxonomy) was not specified by the authors at the time of submission;

we note this honestly rather than retrofitting a division of labor that was not formally tracked during the event.

References

- Akin Gump Strauss Hauer & Feld LLP. (2025, September 22). Colorado postpones implementation of Colorado AI Act, SB 24-205. *AI Law and Regulation Tracker*. <https://www.akingump.com/en/insights/ai-law-and-regulation-tracker/colorado-postpones-implementation-of-colorado-ai-act>
- California State Legislature. (2024). AB 2013: Generative artificial intelligence — training data transparency. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013
- Colorado General Assembly. (2024). SB24-205: Consumer protections for artificial intelligence. <https://leg.colorado.gov/bills/sb24-205>
- Colorado General Assembly. (2026). SB26-189: Automated decision-making technology. <https://leg.colorado.gov/bills/sb26-189>
- Crowell & Moring LLP. (2025, December 8). California's AB 2013 requires generative AI data disclosure by January 1, 2026. <https://www.crowell.com/en/insights/client-alerts/californias-ab-2013-requires-generative-ai-data-disclosure-by-january-1-2026>
- Crowell & Moring LLP. (2026, May). Colorado hits reset on AI regulation: SB 26-189 repeals and reenacts the Colorado AI Act. <https://www.crowell.com/en/insights/client-alerts/colorado-hits-reset-on-ai-regulation-sb-26-189-repeals-and-reenacts-the-colorado-ai-act>
- DevsData. (2026). Mexican AI industry: AI in Mexico overview (citing IDC enterprise-adoption estimates). <https://devsdata.com/mexican-ai-industry-overview/>
- Directorio Legislativo. (2025). Following INAI's closure: Changes in transparency and access to information in Mexico. <https://directoriolegislativo.org/en/following-inais-closure-changes-in-transparency-and-access-to-information-in-mexico>
- European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- Hogan Lovells. (2026, May 7). EU legislators agree to delay for high-risk AI rules. <https://www.hoganlovells.com/en/publications/eu-legislators-agree-to-delay-for-high-risk-ai-rules>
- Infobae. (2025, May 10). INAI desaparece; Secretaría Anticorrupción y Buen Gobierno asume funciones. <https://www.infobae.com/mexico/2025/05/10/inai-desaparece-secretaria-anticorrupcion-y-buen-gobierno-asume-funciones>
- International Organization for Standardization / International Electrotechnical Commission. (2023a). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system.
- International Organization for Standardization / International Electrotechnical Commission. (2023b). ISO/IEC 23894:2023 — Information technology — Artificial intelligence — Guidance on risk management.
- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1).
- Norton Rose Fulbright. (2026, March 31). California District Court upholds transparency requirements for generative AI training data. <https://www.nortonrosefulbright.com/en-us/knowledge/publications/c1df8419/california-district-court-upholds-transparency-requirements-for-generative-ai-training-data>

- Organisation for Economic Co-operation and Development. (2022). OECD Framework for the Classification of AI Systems.
- Pandectes. (2025, June 2). Mexico implements new data protection framework: Impacts on businesses. <https://pandectes.io/blog/mexico-implements-new-data-protection-framework/>
- Pinelo, M., & Moncada, P. (2025). VIGÍA: Government AI Impact Assessment in Mexican Regulation. AI Safety Mexico. Retrieved June 2026, from <https://vigia-observatorio.github.io/index-en.html>
- United Nations Educational, Scientific and Cultural Organization. (2021). Recommendation on the Ethics of Artificial Intelligence; Readiness Assessment Methodology.
- White & Case LLP. (2026, May 14). EU agrees Digital Omnibus deal to simplify AI rules. <https://www.whitecase.com/insight-alert/eu-agrees-digital-omnibus-deal-simplify-ai-rules>

LLM Usage Statement

This submission was prepared with the assistance of large language model tools, including Claude, Gemini, Cursor, and ChatGPT, used for research exploration, drafting, and editing assistance — including reformatting the manuscript to the hackathon template and verifying time-sensitive regulatory facts (e.g., the 2026 status of Colorado SB 24-205/SB 26-189 and the EU AI Act Digital Omnibus) through web search. The authors reviewed all LLM-assisted output and remain fully responsible for the accuracy of every claim, the integrity of the methodology, and the final content of this paper.