DEPLOY-5: Audit-Ready Deployment Gates for Agentic AI in Legal Research and Decision Workflows

Khazretgali Sapenov
Decision Grade Labs
San Jose, California, USA
ksapen@decision-grade.com

Abstract

Agentic AI systems—LLM-based legal research assistants, document analysis agents, and workflow planners—are rapidly entering legal practice, where outputs influence discovery strategy, contract negotiation, compliance decisions, and litigation risk assessments. Yet AI governance is often aspirational: principles are declared while concrete deployment decisions lack auditable evidence, deterministic go/no-go criteria, and clear responsibility allocation. We introduce DEPLOY-5, a deployment-time governance execution layer that operationalizes oversight through five sequential, fail-closed gates requiring verifiable artifacts: (1) context and stakes definition, (2) evidence and evaluation sufficiency, (3) safety and control mechanisms, (4) operational readiness and monitoring, and (5) accountability and residual-risk acceptance. DEPLOY-5 is designed for socio-technical settings where model internals are partially opaque and where risk arises from tool composition and workflow embedding rather than accuracy alone. We contribute an adoptable artifact package—binary checklist, scoring rubric, evidence log template, and walkthrough protocol—and demonstrate how the gates surface failure modes that generic benchmarks miss (e.g., unverifiable citations, silent tool degradation, and non-reconstructable decision chains). We discuss implications for AI-enabled legal processes: defensible audit trails, liability-aware sign-off, and governance that is executable rather than performative.

Keywords

agentic AI, AI governance, deployment gates, legal AI, auditability, fail-closed semantics, e-discovery, legal workflows, accountability

1 Introduction

1.1 The Agentic Turn in Legal Practice

Artificial intelligence has moved from passive legal-research tools—case-retrieval systems, citation checkers, contract keyword scanners—to agentic systems that autonomously plan multi-step workflows, invoke external databases and APIs, and produce outputs that directly inform legal decisions [2, 3]. Modern agentic legal assistants search case law across jurisdictions, generate privilege-review recommendations on millions of documents, draft contract redlines, and flag regulatory non-compliance—all with minimal human intervention per document.

Importantly, large-scale responsiveness review and privilege review should not be conflated. Technology-assisted review for identifying responsive documents has a substantial empirical and doctrinal foundation, including work demonstrating that TAR can match or exceed the effectiveness of exhaustive manual review [8] and judicial decisions approving computer-assisted review in appropriate cases [5, 10]. Privilege and work-product determinations, by contrast, often require contextual information outside the four corners of an individual document: the role of the communicant, whether counsel was acting in a legal or business capacity [16], the purpose of the communication, waiver risk, and matter-specific facts. DEPLOY-5 therefore does not assume that privilege review at scale is a solved technical problem. It treats privilege review as a high-stakes, context-sensitive deployment setting in which evidence thresholds, escalation rules, reconstruction logs, and attorney sign-off must be specified before operational use.

These systems act as semi-autonomous decision participants whose outputs shape litigation strategy, discovery scope, contract terms, and compliance posture.

From an AI-and-law perspective, this shift matters because deployment is no longer a technical endpoint—it is an organizational and institutional process with decision rights, accountability structures, and audit requirements imposed by professional-conduct rules, evidence law, and emerging AI regulation. The governance challenge is not primarily about model accuracy; it is about the socio-technical infrastructure that determines *when*, *how*, and under *what conditions* an agentic system should enter operational legal use.

1.2 The Governance Execution Gap in Legal AI

Many law firms, legal-tech vendors, and in-house legal departments have adopted responsible-AI principles, published ethics guidelines, and established AI review committees. Yet a persistent gap exists between governance intent and governance execution. Frameworks emphasize transparency, fairness, and accountability as aspirational objectives but rarely specify what must be true—in terms of evidence, artifacts, and decision records—before a system influences a legal outcome. The result is a characteristic failure pattern: agentic systems are deployed because benchmark metrics appear acceptable, but no audit trail exists for why the deployment was considered safe, appropriate, and proportionate to the professional-risk context.

This gap is especially consequential in legal practice, where the correctness of agent outputs directly affects clients' rights and interests. A privilege-review agent that misclassifies documents, a contract-analysis agent that fabricates clause interpretations, or a litigation-support agent that omits adverse authority can misdirect strategy, expose firms to sanctions, and compromise clients—outcomes that post-hoc auditing cannot reverse. Professional responsibility rules (e.g., ABA Model Rules 1.1, 5.3) require competent

supervision of work regardless of the tool used [1]; demonstrating that supervision occurred requires audit evidence that current governance practices do not systematically produce.

1.3 Contributions

This paper introduces DEPLOY-5, a five-gate governance execution layer designed for deployment-time oversight of agentic AI systems, with particular application to legal workflows. Our contributions are:

- **C1 (Framework).** DEPLOY-5 reconceptualizes AI deployment governance as a socio-technical execution problem. It defines five sequential, fail-closed gates that require verifiable evidence artifacts before an agentic system enters operational use, shifting the governance locus from algorithmic properties to organizational decision processes.
- **C2 (Artifact Package).** DEPLOY-5 delivers a deployable artifact package—binary compliance checklist, scoring rubric, evidence log template, and walkthrough protocol—that legal organizations can adopt within existing governance architectures without structural redesign.
- **C3 (Legal Application).** We apply all five gates to an e-discovery privilege-review agent, demonstrating that DEPLOY-5 surfaces failure modes invisible to generic benchmarks (unverifiable citation chains, silent privilege-model degradation, non-reconstructable review decisions) and yields liability-aware, defensible deployment decisions.
- **C4 (Legal Relevance).** We connect DEPLOY-5's evidence and accountability requirements to established legal standards—attorney competence rules, model-rule supervision obligations, and emerging AI-Act compliance requirements—providing a governance vocabulary aligned with legal professionals.

2 Background and Positioning

2.1 Agentic Systems in Legal Workflows

Agentic AI systems are composite architectures combining language models with tool access, memory, planning modules, and external data sources [13, 18]. In legal contexts, these composite systems are embedded in high-stakes socio-technical infrastructures: their outputs feed into human decision-making under professional-responsibility obligations, their errors propagate through case strategy and client advice, and their governance requires institutional structures—decision rights, accountability assignments, audit trails—that extend well beyond model evaluation.

Four legal workflow categories are particularly affected:

1. **E-discovery and document review.** Privilege-detection agents process millions of documents; misclassification produces inadvertent waiver or over-designation with spoliation risk.
2. **Legal research and case analysis.** Citation-generation agents that hallucinate authority or omit adverse precedent mislead legal strategy and may constitute misrepresentation to courts [17].
3. **Contract review and due diligence.** Clause-extraction and risk-flagging agents that miss jurisdiction-specific enforceability conditions expose clients to contract risk.
4. **Regulatory compliance monitoring.** Workflow agents that generate compliance assessments must be auditable to satisfy regulatory-authority inquiry; unverifiable reasoning chains are inadmissible as compliance evidence.

2.2 Limitations of Existing Governance Approaches

Three patterns characterize current legal AI governance that DEPLOY-5 addresses.

Over-reliance on model transparency. Governance frameworks assume that understanding model internals is the primary path to trust. For agentic systems composed of opaque models, third-party APIs, and emergent tool compositions, this assumption is impractical [12]. DEPLOY-5 requires reconstructability of actions and evidence, not interpretability of internals.

Principles without decision gates. Ethics guidelines state that AI should be fair, transparent, and accountable but do not specify the decision procedures, evidence thresholds, or artifact requirements that would make these principles actionable at deployment time [7, 11, 15].

Monitoring as afterthought. Deployment is treated as a one-time approval event. Post-deployment monitoring focuses on system availability rather than governance-relevant signals: privilege-classification drift, citation-hallucination rate, or reconstruction-gap emergence.

2.3 Positioning DEPLOY-5

DEPLOY-5 operates as a deployment-time execution layer complementing existing frameworks: NIST AI RMF [15], EU AI Act [7], ISO/IEC 42001 [11], and legal-specific standards (ABA Model Rules, state bar AI guidance). It specifies what must be true before a deployment decision, what evidence must exist, and who must accept residual risk—filling the gap between high-level governance intent and operational legal-deployment practice.

3 The DEPLOY-5 Framework

3.1 Design Objectives

DEPLOY-5 is guided by four structural objectives reflecting the reality of governing partially opaque, socio-technically embedded agentic systems in professional settings:

Auditability. Every deployment decision must be supported by verifiable evidence artifacts stored in a structured evidence log that is available for professional-responsibility review and regulatory inquiry.

Deterministic Decision Semantics. Deployment readiness is evaluated through explicit pass/fail gates. There is no partial credit; each criterion is satisfied or not.

Fail-Closed Default. Absence of sufficient evidence blocks deployment. The burden of proof rests on the deployer, not on an assessor to demonstrate non-readiness.

Context Proportionality. Required evidence scales with the stakes of the deployment context, as defined by the deployer in

Gate 1, including professional-responsibility risk and client-impact severity.

3.2 The Five Gates

Gate 1: Context and Stakes Definition. Requires explicit definition of deployment scope, stakeholder identification (including client-impact scope), harm taxonomy proportionate to professional risk, impact bounds with enforcement mechanisms, and decision authority boundaries. In legal settings: which matters or matter-types does the agent touch? What is the worst-case professional-responsibility or client-harm outcome? Who has deployment authority? *Sub-gates*: (1A) Scope and Stakeholder Definition; (1B) Harm and Impact Modeling.

Gate 2: Evidence and Evaluation Sufficiency. Requires decision-grade evidence demonstrating acceptable performance within the defined legal context. Evidence must include versioned system documentation, context-relevant evaluation (not benchmark-only), robustness and edge-case analysis across adversarial legal documents, known limitations, a staged rollout plan, and an operational feedback loop. Aggregate accuracy metrics alone do not satisfy this gate—the evaluation must address specific legal-workflow failure modes (e.g., privilege-classification rates by document type, citation-hallucination rate, clause-extraction precision across jurisdictions). *Sub-gates*: (2A) System Baseline Documentation; (2B) Contextual Evaluation and Robustness.

Gate 3: Safety and Control Mechanisms. Requires defined containment and mitigation structures: explicit behavioral constraints (enforced, not advisory), safe fallback and rollback procedures, escalation triggers with attorney ownership, fail-closed execution semantics, and security and abuse resistance. In legal contexts: automatic escalation when confidence falls below threshold; attorney sign-off required for any output that directly enters a pleading, disclosure, or client advice. *Sub-gates*: (3A) Bounded Behavior and Fail-Closed Design; (3B) Security and Abuse Resistance.

Gate 4: Operational Readiness and Monitoring. Deployment is an ongoing operational state, not a static approval event. This gate requires monitoring signals aligned with the risk profile from Gate 1, alert thresholds with attorney ownership, incident documentation procedures, drift-detection mechanisms, and logging architecture supporting causal reconstruction of material decisions—prerequisite for retrospective professional-responsibility review. *Sub-gates*: (4A) Monitoring and Telemetry; (4B) Incident Response and Rollback.

Gate 5: Accountability and Residual Risk Acceptance. No deployment is risk-free. This gate formalizes responsibility: explicit assignment of decision rights (supervising attorney of record), documented deployment approval, a residual risk statement signed by an accountable authority, and an exception register with expiry conditions. This gate converts governance from implicit assumption into institutional accountability, directly satisfying Model Rule 5.3 supervision requirements. *Sub-gate*: (5A) Human Oversight and Residual Risk Acceptance.

3.3 Structural Properties

The five gates exhibit four systemic properties critical for legal deployment governance. *Sequential integrity*: later gates cannot compensate for failures in earlier ones—an organization cannot substitute monitoring (Gate 4) for absent evaluation (Gate 2). *Evidence traceability*: each gate produces artifacts stored in a structured evidence log, available for bar-association review or regulatory inquiry. *Escalation control*: failures trigger defined remediation paths, not workarounds. *Portability*: the model applies across legal workflow types and organizational structures.

4 Legal Relevance and Professional Standards Alignment

4.1 Competence and Supervision Obligations

ABA Model Rule 1.1 (Competence) requires that attorneys understand the benefits and risks of relevant technology [1]. Rule 5.3 (Supervision of Nonlawyers) requires that supervising attorneys ensure work meets professional obligations regardless of who or what performs it. DEPLOY-5's evidence artifacts—particularly the Gate 4 evidence log and Gate 5 residual-risk acceptance—provide a structured basis for demonstrating that required supervision occurred: who reviewed what, on what evidence, under what conditions, and with what escalation triggers.

4.2 Emerging Regulatory Requirements

The EU AI Act [7] classifies legal-process AI systems as high-risk, imposing requirements for conformity assessment, technical documentation, human oversight, and logging. DEPLOY-5's gate structure maps directly onto these requirements: Gate 1 (context and scope) fulfills intended-purpose documentation; Gate 2 (evaluation sufficiency) fulfills accuracy and robustness testing; Gate 3 (safety controls) fulfills human-oversight and corrective-action requirements; Gate 4 (monitoring) fulfills post-market surveillance; Gate 5 (accountability) fulfills conformity-declaration and responsibility assignment.

4.3 Audit-Trail Defensibility

When AI-generated outputs are challenged—as in *Mata v. Avianca* [17], where court-sanctioned attorneys had submitted LLM-hallucinated citations—the ability to produce an evidence log demonstrating what the system was authorized to do, how its outputs were verified, and who accepted residual risk becomes legally significant. DEPLOY-5's structured evidence log is designed precisely to produce this documentation.

Hallucination risk extends well beyond the *Mata* sanctions order. General LLM surveys identify hallucination as a persistent reliability failure [9]; legal-domain empirical studies find that both general-purpose and legal-specific systems generate false, unsupported, or misleading legal propositions [6, 14]; and a growing court-level database documents repeated instances of AI-generated false citations entering legal filings across multiple jurisdictions [4].

These findings reinforce DEPLOY-5's emphasis on citation verification (Gate 2), reconstruction logs (Gate 4), and attorney-owned escalation (Gate 3) rather than reliance on model-level assurances.

5 Artifact Package

5.1 Artifacts Provided

DEPLOY-5 delivers a complete artifact package designed for immediate organizational adoption:

Binary Compliance Checklist. 25 criteria across five gates, each with a verification method and pass/fail/N/A assessment. All criteria within a gate must pass for the gate to pass; all five gates must pass for deployment.

Scoring Rubric. 0–2 scoring per criterion (Not Met / Partial / Fully Met) providing diagnostic granularity for governance-maturity tracking. The rubric complements but does not replace the binary checklist.

Evidence Log Template. Structured, append-only record of all artifacts produced during gate evaluation, serving as the primary audit trail and basis for independent reconstruction—and, in legal contexts, for professional-responsibility defense.

Case Walkthrough Protocol. Structured protocol for applying DEPLOY-5 to a specific legal deployment, with gate-by-gate evidence assessment and remediation-path generation.

The complete 25-criterion checklist, scoring rubric, evidence log template, and walkthrough protocol are provided in Appendix A and as a supplement to this submission.

5.2 Illustrative Application: E-Discovery Privilege-Review Agent

To demonstrate actionable, discriminating results in a legal context, we apply all five gates to an e-discovery privilege-review agent: an LLM-based system that processes raw document productions, classifies documents as privileged/non-privileged/redact, and writes determinations to the matter management system. Agent outputs directly inform disclosure decisions with discovery sanctions, waiver, and client-confidentiality consequences.

Gate 1 passes: the organization defined the agent's scope as "first-pass privilege review on corporate M&A matters only," established that a supervising attorney reviews all privilege calls above a confidence threshold, and documented client-impact scope (thousands of documents per matter, potential waiver on $50M+ transactions).

Gate 2 fails: evaluation evidence covers aggregate precision/recall on a generic privilege dataset but not on M&A-specific document types (board minutes, deal-counsel emails, foreign-law instruments). No degraded-dependency testing (e.g., behavior when document OCR quality is poor) is documented.

Gate 3 partially fails: behavioral constraints exist (output routed to attorney review queue) but the fallback procedure if the review queue is backlogged is undefined— creating a path where unreviewed determinations could enter the disclosure set.

Gate 4 fails: infrastructure monitoring (API availability, latency) is in place, but governance-relevant signals are absent: no alert threshold for privilege-call reversal rate (which would indicate model drift), no logging of which model version produced which determination (precluding causal reconstruction for sanctions defense).

Gate 5 is conditional: general partner approval is documented, but the approval lacks a structured residual-risk statement ("we accept the risk that X% of borderline calls may require correction") and no exception register exists.

Overall verdict: Not Deployable. Three specific, remediable failures block deployment; Gate 1 and partial Gate 5 indicate the organization has a functioning governance intent—it has simply failed to execute it. DEPLOY-5 produces a prioritized remediation list: (1) M&A-specific evaluation corpus, (2) degraded-OCR robustness tests, (3) defined backlog fallback with automatic hold, (4) governance monitoring signals with attorney-owner alerts, (5) structured residual-risk statement and exception register.

6 Evaluation Approach

6.1 Expert Structured Walkthrough

DEPLOY-5 is evaluated through a structured expert walkthrough protocol. Participants (target: 5–8) are recruited from legal technology, legal-operations, and AI-governance practitioner communities. Each participant independently applies the DEPLOY-5 checklist and walkthrough protocol to the privilege-review agent scenario during a 45–60 minute structured session.

Assessment measures include: clarity of gate definitions and criteria, completeness of evidence requirements relative to legal-workflow failure modes, feasibility of assessment within legal-operations constraints, adoption intent (Likert 1–5), estimated time-to-complete per gate, and friction points and suggested improvements. Optional: inter-rater agreement on binary gate pass/fail for a standardized scenario.

6.2 Threats to Validity

The walkthrough is applied to a single scenario; generalizability across legal workflow types requires additional case applications. The expert walkthrough evaluates perceived quality and feasibility, not longitudinal adoption. The framework addresses deployment-time governance; it does not address model-training governance or post-deployment organizational learning in depth.

7 Discussion and Research Agenda

7.1 Governance Theater vs. Governance Execution

DEPLOY-5 is designed to detect "governance theater"—organizations that formally adopt responsible-AI artifacts while failing to execute them substantively. The privilege-review walkthrough illustrates this: the organization had executive approval (Gate 5 intent) and scope constraints (Gate 1) but had not produced the evaluation evidence (Gate 2) or monitoring infrastructure (Gate 4) that would make governance operational. Gate-based governance with fail-closed semantics prevents approval of incompletely evidenced deployments, regardless of how strong the governance narrative is at higher levels.

7.2 Propositions for Empirical Research

P1 (Governance Determinism). Gate-based governance with fail-closed semantics reduces governance theater compared to principle-based governance in legal organizations.

P2 (Evidence Traceability). Organizations maintaining structured DEPLOY-5 evidence logs achieve faster and more reliable audit completion for AI deployment reviews under bar-association or regulatory inquiry.

P3 (Liability-Aware Deployment). Legal organizations using DEPLOY-5 Gate 5 accountability artifacts demonstrate measurable improvement in professional-responsibility defensibility when AI-assisted outputs are challenged.

7.3 Implications for Legal AI Practice

For legal practitioners, DEPLOY-5 provides three immediate adoption paths. *Procurement*: include DEPLOY-5 gate criteria in RFP requirements for legal-AI vendors. *Internal governance*: use the checklist as deployment go/no-go criteria within existing matter-management and technology-approval workflows. *Professional responsibility defense*: use the evidence log as the primary artifact for demonstrating Model Rule 5.3 compliance when AI-assisted outputs are challenged.

8 Conclusion

DEPLOY-5 makes AI deployment governance in legal practice executable, inspectable, and auditable. By specifying five sequential gates with pass/fail criteria, required evidence artifacts, and fail-closed semantics, it translates governance intent into deployment-time enforcement aligned with professional-responsibility obligations and emerging AI regulation.

The artifact package—checklist, rubric, evidence log, and walk-through protocol—lowers adoption barriers by integrating within existing legal-operations governance architectures. The illustrative e-discovery walkthrough demonstrates that DEPLOY-5 produces actionable, discriminating results: an organization with documented executive approval and defined scope controls still fails three of five gates on specific, remediable criteria.

The governance of agentic AI in legal practice is too consequential to remain aspirational. DEPLOY-5 provides the execution layer that makes it enforceable.

References

[1] American Bar Association. 2023. Model Rules of Professional Conduct and Formal Opinion 512 on Generative AI. ABA Standing Committee on Ethics and Professional Responsibility.

[2] Kevin D. Ashley. 2017. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age.* Cambridge University Press.

[3] Trevor Bench-Capon and Henry Prakken. 2012. Using Argument Schemes for Hypothetical Reasoning in Law. *Artificial Intelligence and Law* 18, 2 (2012), 153–174.

[4] Damien Charlotin. 2024. AI Hallucination Cases Database. https://www.damiencharlotin.com/hallucinations/. Database of court decisions involving AI-generated false citations, false quotes, and misrepresented content; accessed May 2026.

[5] Da Silva Moore v. Publicis Groupe. 2012. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012). First judicial approval of technology-assisted review (computer-assisted review) in U.S. e-discovery; United States District Court, S.D.N.Y. (Judge Andrew J. Peck).

[6] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (2024), 64–93. doi:10.1093/jla/laae003

[7] European Parliament and Council. 2024. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (AI Act). Official Journal of the European Union.

[8] Maura R. Grossman and Gordon V. Cormack. 2011. Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review. *Richmond Journal of Law and Technology* 17, 3 (2011), 11. https://scholarship.richmond.edu/jolt/vol17/iss3/5/ Article 11.

[9] Lei Huang et al. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *Comput. Surveys* (2025). doi:10.1145/3703155 DOI: 10.1145/3703155.

[10] In re Broiler Chicken Antitrust Litigation. 2018. In re Broiler Chicken Antitrust Litigation, No. 1:16-cv-08637 (N.D. Ill. Jan. 3, 2018). TAR/search methodology validation protocol approved by the court; United States District Court, N.D. Illinois.

[11] International Organization for Standardization. 2023. ISO/IEC 42001:2023 Information Technology — Artificial Intelligence — Management System.

[12] Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. The Flash Crash: High-Frequency Trading in an Electronic Market. *Journal of Finance* 72, 3 (2017), 967–998.

[13] LangChain. 2024. LangChain Documentation: Agents. https://docs.langchain.com/.

[14] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies* (2025). doi:10.1111/jels.12413

[15] National Institute of Standards and Technology (NIST). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

[16] Supreme Court of the United States. 1981. Upjohn Co. v. United States, 449 U.S. 383 (1981). Establishing scope of attorney-client privilege, including the role of the communicant and whether counsel acted in a legal or business capacity.

[17] United States District Court, S.D.N.Y. 2023. Mata v. Avianca, Inc., No. 22-cv-1461 (S.D.N.Y. June 22, 2023). Sanctions order addressing LLM-generated hallucinated citations.

[18] Shuyan Zhou et al. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. *arXiv preprint arXiv:2307.13854* (2024).

A DEPLOY-5 Criteria and Artifact Package

DEPLOY-5 operationalizes governance through 25 verifiable criteria across five sequential gates. Table 1 lists every criterion and its primary verification artifact. *Fail-closed default*: if evidence is absent, insufficient, or ambiguous, the criterion fails. All criteria within a gate must pass for the gate to pass; all five gates must pass for deployment. The complete artifact package—formatted binary checklist, per-criterion scoring rubric (0–2 scale), evidence log template, and walkthrough protocol—is provided as a supplement to this submission.

Table 1: DEPLOY-5: 25-Criterion Binary Compliance Checklist with Verification Artifacts

ID	Gate	Criterion	Verification Artifact
Gate 1: Context & Stakes Definition			
1.1	G1	Deployment scope and intended use explicitly documented	Review scope document for completeness and specificity
1.2	G1	Stakeholders identified with roles, exposure, and decision authority	Verify stakeholder register covers affected parties
1.3	G1	Harm taxonomy defined, proportionate to deployment risk level	Review harm taxonomy; verify proportionality rationale
1.4	G1	Impact bounds defined with enforcement mechanism (not advisory)	Attempt to exceed declared bounds; verify enforcement
1.5	G1	Decision authority boundaries documented (who can approve, escalate, halt)	Trace escalation path from operator to accountable executive
Gate 2: Evidence & Evaluation Sufficiency			
2.1	G2	Versioned system baseline documentation exists (model, tools, configs, dependencies)	Review documentation; verify version control
2.2	G2	Context-relevant evaluation completed (not benchmark-only)	Confirm test suite maps to deployment context, not generic benchmarks
2.3	G2	Robustness and edge-case analysis documented (degraded dependencies, adversarial input)	Review stress test results; verify failure-mode coverage
2.4	G2	Known limitations and uncertainty explicitly characterized	Review limitations statement; verify it is specific, not boilerplate
2.5	G2	Staged rollout plan with monitoring gates (time/volume-boxed)	Verify rollout plan with explicit go/no-go criteria per stage
2.6	G2	Feedback loop with triage cadence and gate-relevant tagging operational	Trace a sample issue through triage to resolution
Gate 3: Safety & Control Mechanisms			
3.1	G3	Explicit behavioral constraints defined and enforced (hard limits, not advisory)	Attempt to violate each constraint; verify enforcement
3.2	G3	Safe fallback and rollback procedures defined and tested	Execute rollback drill; verify recovery time and completeness
3.3	G3	Escalation triggers defined with ownership and response SLA	Verify trigger conditions, escalation path, and response timeline
3.4	G3	Fail-closed execution semantics implemented (ambiguity or missing evidence blocks action)	Remove required input; verify system halts rather than proceeds
3.5	G3	Security and abuse resistance assessed (threat model, adversarial testing)	Review threat model; verify adversarial test coverage
Gate 4: Operational Readiness & Monitoring			
4.1	G4	Monitoring signals defined and aligned with risk profile from Gate 1	Map each risk to at least one monitoring signal
4.2	G4	Alert thresholds defined with ownership and escalation path	Verify thresholds are specific (not TBD) and owners are assigned
4.3	G4	Incident documentation and response procedures defined	Review IR runbook; verify roles, communication plan, evidence preservation
4.4	G4	Drift detection mechanisms operational (model, data, tooling, environment)	Verify drift signals are active and reviewed on declared cadence
4.5	G4	Logging architecture supports causal reconstruction of material actions	Select 5 material actions; verify full causal chain is reconstructable from logs
Gate 5: Accountability & Residual Risk Acceptance			
5.1	G5	Decision rights explicitly assigned (no orphaned responsibilities)	Trace each gate to an accountable individual; verify no gaps
5.2	G5	Deployment approval documented with approver identity and date	Review approval record
5.3	G5	Residual risk statement signed by accountable authority	Review statement; verify it names specific accepted risks, not generic boilerplate
5.4	G5	Exception register maintained with expiry conditions and review schedule	Review register; verify exceptions have expiry dates and review owners