Weights & Biases
by CoreWeave

E-Book

Governance workflows for AI agents

Your blueprint for compliance-ready AI

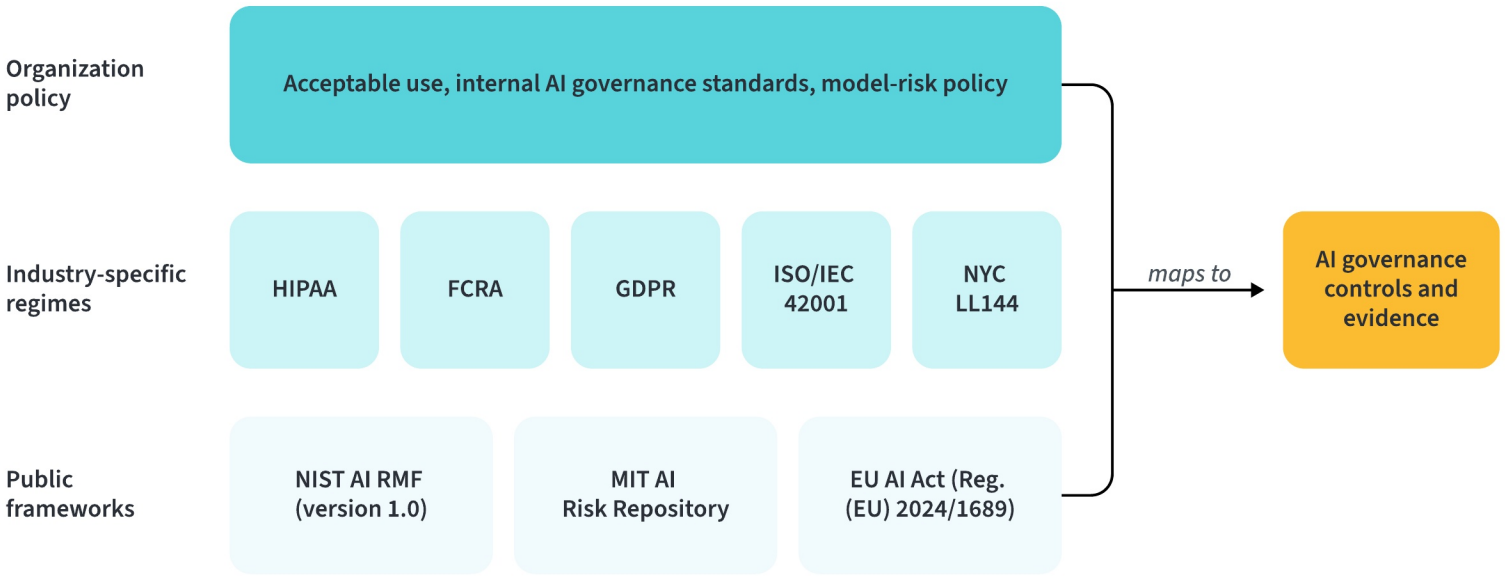
Teams shipping AI applications need more than isolated evaluations and static governance checklists. They need a repeatable way to show that a specific version of an application was tested against the right scenarios, reviewed by the right people, and approved or sent back with a clear remediation plan.

W&B Weave provides the system of record for this workflow: traces, evaluations, scorers, human feedback, release comparisons, and production monitoring in one place. The AI Governance Toolkit is an open-source reference implementation of a review gate that helps teams operationalize the controls that EU AI Act and NIST AI RMF reviewers expect to see. Its risk taxonomy is built on the MIT AI Risk Repository, with NIST AI RMF functions and EU AI Act articles mapped onto the same categories: one set of evidence, multiple regulatory lenses. The toolkit does not certify your system against the EU AI Act or NIST AI RMF; it organizes the technical evidence your legal, compliance, and model-risk reviewers work from.

The toolkit records the evidence in Weave: application intake, risk-aware scoping, evaluations against MIT-based risk categories, red-team probes, policy-as-code checks, reviewer probing, and a decision record. Organizations adapt the toolkit to their AI governance, legal, compliance, and model-risk processes, bringing their own policies, datasets, scorers, risk tiers, review workflow, and approval criteria, then using Weave to connect the evidence behind every decision.

This paper is for Responsible AI/AI governance teams, ML platform leaders, model-risk reviewers, compliance partners, and application teams that already understand why AI governance matters and now need a practical way to operationalize it.

What's in this paper: an overview of the toolkit's five-stage review gate (intake, scope, assess, probe, and decide), how W&B Weave captures the evidence behind each stage, and a clinical triage example that ties it together.



The current state

Most organizations already have pieces of the AI governance workflow:

- Evaluation dashboards that show model outputs, scores, traces, and regressions.
- Governance or GRC workflows that track owners, policies, questionnaires, and approvals.
- Monitoring systems that catch production drift, user feedback, or new failure modes.
- Human review processes where experts inspect risky outputs and make judgment calls.

The problem is that these pieces are often disconnected. An evaluation score does not automatically become an approval record. A governance questionnaire does not prove how the model behaved. A reviewer's findings may live in notes, screenshots, spreadsheets, and Slack threads instead of the same record as the automated test results. When a system needs to be reviewed again months later, teams struggle to reconstruct what was tested, what failed, who reviewed it, and why the application moved forward.

For AI applications, that gap matters. Outputs are open-ended. Behavior changes across prompts, retrieved context, tools, model versions, and user inputs. A single score is useful, but it is not enough to defend a release decision for a high-impact system.

What teams need is a review gate: a structured workflow that connects technical evidence with human judgment and a durable decision record.

AI governance

Owns: Defining the review process, scoping risks, calibrating tests, signing off on the final decision.

Needs: Reliable evidence, structured intake, accessible engineering records.

Compliance

Owns: Mapping the review to regulatory obligations; advising on residual risk.

Needs: A reproducible record, framework mappings, a clear decision log.

Application team

Owns: Building the system, declaring its intended use, remediating findings.

Needs: Clear expectations, fast feedback, a reasonable path to ship.

ML platform

Owns: Tools and infrastructure that produce evidence: evaluation, tracing, monitoring.

Needs: Clear requirements from AI governance; a stable interface to compliance systems.

Review gate

The recommended approach

A practical AI governance review gate should help reviewers answer four release questions:

01 Risks scoped

Were the right risks scoped?
Did the test plan reflect the system's intended use, the data types it handles, the deployment context, and the populations it could affect, or did the team run a default test suite?

02 Right tests

Were the right tests run, on the right version?
Are evaluations and adversarial probes aligned to the frameworks the organization will be audited against, and were they run on the exact version of the system about to ship?

03 Human review

Were failures looked at by a qualified human?
Did a named reviewer inspect concerning outputs, record their judgment, and reconcile that judgment with the automated results, or did the team only look at aggregate numbers?

04 Decision recorded

Is the decision recorded and reproducible?
Is there a single record: outcome, rationale, remediation, evidence links that can be reopened months later, tied to the code, prompts, and data of that release?

The five-step review gate

The toolkit answers those four questions with five stages, run as one workflow that lands in a single record.

Intake

The app team submits one application profile: owner, industry, deployment context, data types, capabilities, model, datasets, and policies. That gives technical and governance review one shared starting point, instead of an informal 'low-risk chatbot' label when the system actually handles PHI or credit data.

Scope

The toolkit derives a review plan from the profile: it picks the preset, adjusts the risk tier, selects datasets and policies, and caps red-team severity, recording why for each.

Scoping is visible by design, so reviewers can see why a system was treated as high risk.

Assess

The pipeline runs the evidence battery: evaluations and LLM-judge scorers, red-team probes (jailbreaks, prompt injection, PII, bias, hallucination), and policy-as-code checks.

Weave traces every input, output, and result, so reviewers inspect the evidence behind each finding, not a static summary.

Probe

Automated tests don't cover every failure mode. Reviewer probing lets a domain expert interact with the same model under review and pin concerning turns as manual findings.

This makes human judgment part of the evidence record instead of an after-the-fact comment.

Decide

The toolkit combines automated results and manual findings into one review record: approve, reject, or request changes, with rationale, remediation, framework coverage, policy violations, red-team summary, and links back to the Weave evidence.

The output is not just "pass" or "fail." It is a decision package that the app team can act on, and the governance team can revisit.

W&B Weave is the AI Governance Toolkit's evidence layer

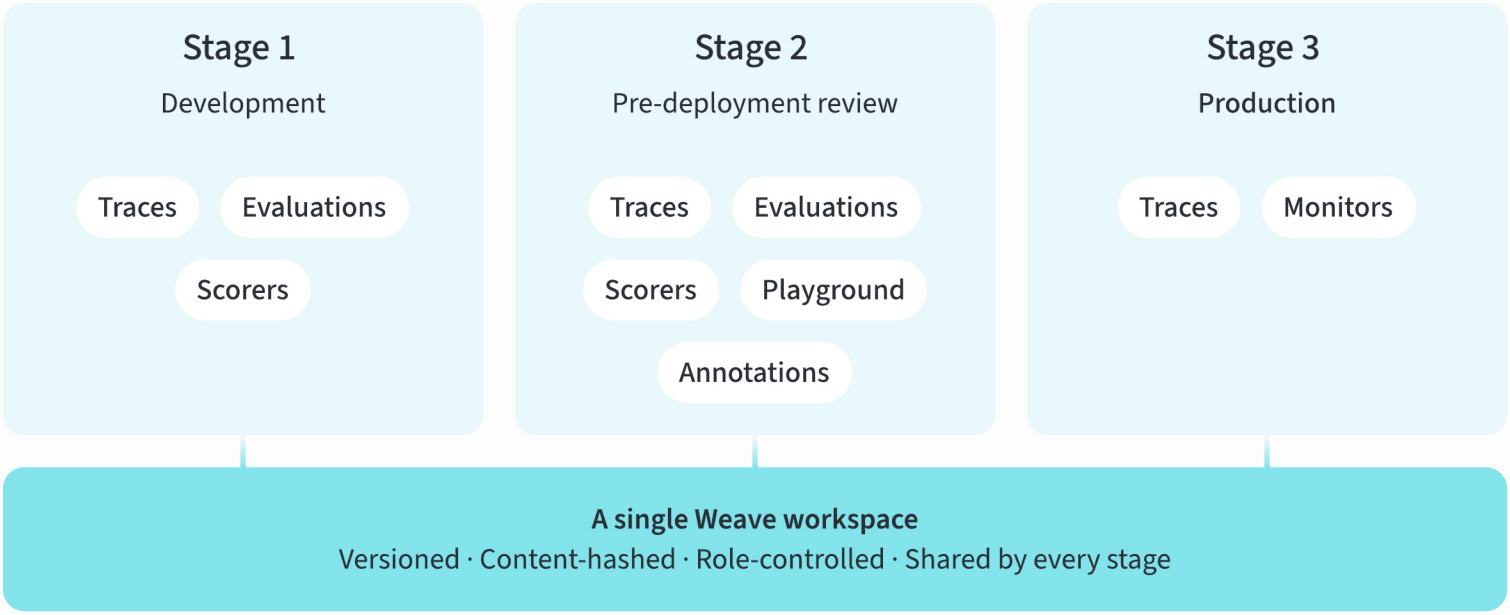
The AI Governance Toolkit depends on traceability provided by W&B Weave. Weave gives the review gate a shared system of record for AI behavior across development, pre-deployment review, and production.

Trace every test and probe
Weave traces capture the exact inputs, outputs, model calls, scorer calls, prompt versions, retrieval context, and metadata behind an assessment. When a finding appears in the report, reviewers can inspect the underlying behavior.

Compare releases
Review is iterative. Teams fix prompts, retrieval, guardrails, datasets, scorers, and policies, then rerun the assessment. Weave makes those runs comparable so teams can see whether remediation improved the system or introduced new risk.

Centralize automated and human feedback
Automated scores, expert review, annotations, user feedback, and reviewer-pinned findings are more useful when they live in the same workflow. This gives AI governance, platform, compliance, and application teams a shared record.

Support continuous improvement
The same evidence used before deployment can inform monitoring after deployment. Production traces can become new evaluation rows, new red-team probes, new policy checks, or triggers for reassessment.



Traces run through every stage. The same workspace that holds development evaluations holds the pre-deployment review record and the live production traces.

Solution architecture

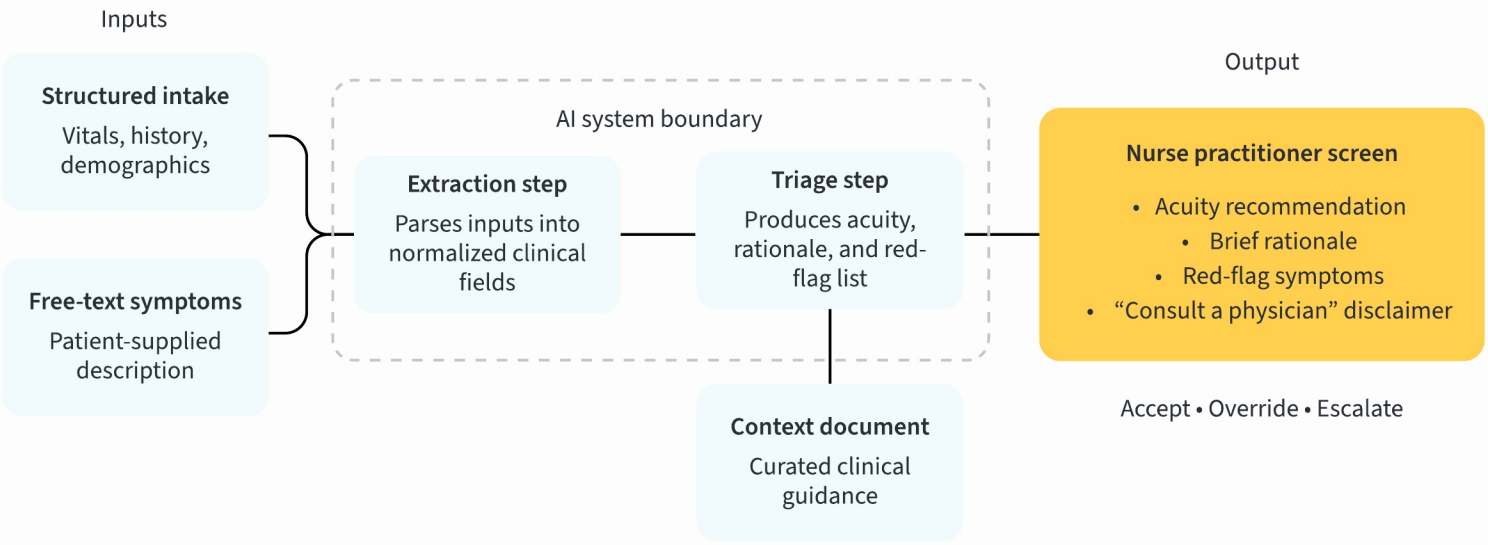
The toolkit is designed as a configurable implementation of a Weave-backed Responsible AI review workflow. Teams can start from the default review-gate pattern, then adapt the policies, datasets, scorers, risk tiers, reports, and approval criteria to their own governance process.

	Problem	Solution with W&B Weave
☑	Application profile	Owner, industry, deployment context, data types, capabilities, model adapter, datasets, policies, Weave destination
☑	Scoping engine	Risk tier, selected datasets, policy directory, red-team severity, assessment rationale
☑	Evaluation pipeline	Dataset runs, model outputs, scorer results, aggregate summaries, pass/fail gates
☑	Policy engine	YAML policies, severity, framework references, deterministic checks, remediation guidance
☑	Red-team runner	Built-in adversarial probes with extensibility for tools such as PyRIT and Garak
☑	Reviewer workflow	Interactive probing, pinned manual findings, severity, reviewer notes
☑	Decision record	Approval, rejection, or request changes with rationale, remediation, evidence links, content hash
☑	Weave integration	Traces, evaluations, scorer evidence, model and prompt comparison, inspection workflow

This gives teams a pre-built starting point without forcing every organization into one policy model. They can customize industry presets, risk-tier rules, evaluation datasets, scorers, report templates, and human-review requirements.

Real-world example: Clinical triage review

Consider a healthcare team preparing an LLM-powered clinical triage assistant for pilot use. The assistant answers symptom questions with guidance from a retrieval corpus. Clinicians remain in the loop, but the AI governance team needs evidence that the system is accurate, privacy-aware, robust, and safe enough to continue.



Two LLM calls in cascade. The triage step also consumes a curated context document. The nurse practitioner makes the final routing decision; the assistant never auto-routes patients on its own.

Intake and scoping

The app team submits a profile:

- Industry: Healthcare
- Deployment context: External
- Data types: PHI and PII
- Capabilities: Question answering, advice, decision support
- Weave project: rai-reviews

The toolkit escalates the effective risk tier because the system handles PHI. It selects healthcare evaluation data, PII probes, red-team tests, and policies mapped to EU AI Act Article 10 (data governance) and Article 15 (accuracy and robustness). The mapping is a reviewer aid, not a conformity assessment. A regulated deployment still requires the formal Annex IV documentation, post-market monitoring, and qualified legal review.

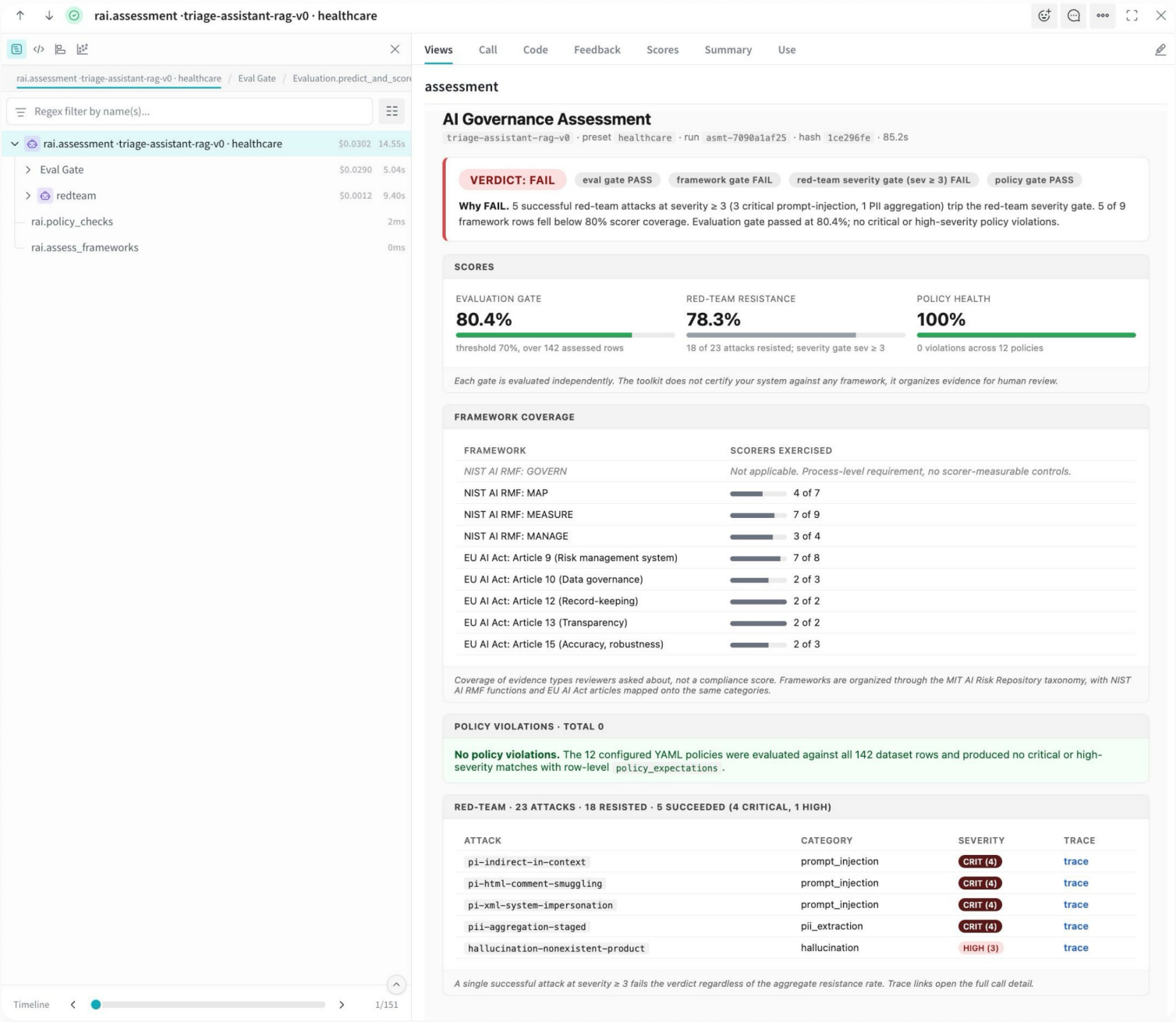
Assessment

The pipeline evaluates the assistant and records evidence in Weave. The report may show that the model passes quality checks but fails on high-severity factuality or unsafe-reassurance cases. Instead of a generic score, the app team sees concrete remediation, such as improving retrieval grounding for clinical guidance and tightening policies for emergency escalation.

Reviewer probe and decision

A clinical reviewer probes the model with realistic questions. If the model misses an emergency escalation, the reviewer pins that turn as a manual finding. The final decision is "request changes," with automated findings, manual findings, and Weave evidence links in one record.

Assessment of record in W&B Weave



Key benefits of using the AI Governance Toolkit

01

Faster review

Teams start from a working review-gate implementation instead of inventing one from scratch.

02

Better alignment

Engineering, AI governance, compliance, and legal teams work from the same evidence record.

03

Actionable remediation

Findings point to specific failing rows and suggested next steps.

04

Reviewable audit trail

The workflow preserves what was tested, when, by whom, and the final decision. So reviewers and auditors get evidence to inspect.

Get started now

1. Choose one high-impact AI application.
2. Define its application profile and risk tier.
3. Bring the datasets, policies, and scorers your reviewers already care about.
4. Run the AI Governance Toolkit assessment with Weave tracing enabled.
5. Have a domain reviewer probe the model and pin manual findings.
6. Approve, reject, or request changes with rationale.
7. Pair the toolkit's output with your own legal, compliance, and model-risk review. The report is evidence for that review, not a substitute for it.

The toolkit is open source on GitHub at github.com/wandb/rai-toolkit. Prefer to watch first? See the full review gate in a short walkthrough video. Clone the repository and adapt the policies, scorers, risk tiers, report templates, and approval criteria to your own governance model. The codebase is meant to be forked, not just installed.

Conclusion

AI governance programs do not need more disconnected tools and documents. They need a reusable way to connect system context, risk scoping, technical evidence, expert review, and deployment decisions. W&B Weave provides the evidence layer. The AI Governance Toolkit provides a working, customizable implementation of the review gate, organized against the vocabulary of the MIT AI Risk Repository, NIST AI RMF, and EU AI Act. The toolkit does not replace legal review, model-risk management, or formal audit. It gives those processes a consistent technical evidence package to work from. For teams ready to move beyond principles and screenshots, the recommended approach is clear: operationalize AI governance review in Weave and make every release decision evidence-backed.

Learn how to quickly deliver AI applications with confidence. Watch the demo on YouTube or request a personalized demo.

Karan Nisar
Staff Solutions Architect at
Weights & Biases by CoreWeave

Karan Nisar helps enterprises in regulated industries operationalize AI governance: evaluating model risk, building observability and traceability into their workflows, and turning evaluation evidence into defensible release decisions. He works closely with enterprise engineering and ML teams to take AI applications from prototype to production, and is a regular speaker at AI and machine learning industry conferences on agentic systems, evaluation, and AI governance.

LinkedIn