Upstream Governance Framework

A Structural Standard for AI Governance at the Training, Fine-Tuning, and Deployment Layers

Kindred Labs Foundation

Version 1.0

www.kindredlabsfoundation.org

Table of Contents

Executive Summary

The Problem

AI governance programs are typically scoped to the outputs of systems that were already built, by processes governance had no role in, encoding assumptions governance was never positioned to examine.

The decisions that determine what an AI system learns, what it is trained to recognize, what it is optimized to produce, and whose existence its training data was built to see clearly: these decisions are made at the training and fine-tuning layers, before the model exists. By the time a governance program is in place, those decisions are infrastructure. They cannot be undone by auditing outputs or documenting deployment decisions.

This scope gap — governance defined as the management of consequences rather than the authorization of decisions — is what the Upstream Governance Framework addresses.

The Three-Layer Model

The framework organizes upstream governance across three layers where consequential AI decisions are made:

- The training layer is where a model acquires its understanding of the world. Dataset curation, category design, optimization objectives, and representational review are governance decisions made here. For organizations deploying vendor models, this layer is controlled by the vendor. The framework defines what organizations can require vendors to disclose and be accountable for.
- The fine-tuning layer encompasses every decision an organization makes about how a vendor base model is configured, adapted, or extended for organizational use. It includes proprietary data selection, behavioral configuration, retrieval setup, and any mechanism that shapes model behavior. These are governance decisions.
- The deployment layer is where the organization decides which AI system to deploy, for what purposes, and under what conditions. Vendor selection, use case authorization, and deployment scoping are governance decisions made here. For organizations that do not train or fine-tune their own models, this is where upstream governance begins.

The framework does not replace deployment-layer governance programs or existing standards. It governs the layer those programs assume someone else already handled.

What the Framework Contains

The framework consists of four components:

1. The Three-Layer Domain Model defines the governance domains present at each layer, the specific decisions that constitute governance decisions within each domain, the structural controls required, and the indicators that governance is absent.
2. The Vendor Transparency Standards define what organizations have the right to require of vendors at selection and throughout the vendor relationship. They operationalize the governance obligation that exists at the deployment layer for organizations that cannot directly access vendor training decisions. They do not require vendors to provide direct access to training data. They require vendors to disclose what they can disclose and accept accountability for what they cannot.
3. The Fine-Tuning Governance Protocol defines what organizations must govern when using proprietary data for fine-tuning. It covers data provenance requirements, consent documentation standards, category review process, and ongoing audit requirements.
4. The Maturity Assessment Rubric is the diagnostic instrument that translates the framework into an organizational assessment. It scores governance maturity across all twelve domains in the three-layer model, produces a layer-by-layer maturity profile, and identifies the highest-priority governance gaps.

How This Framework Relates to Existing Standards

The NIST AI Risk Management Framework operates primarily at the deployment and output layers. The upstream governance framework operates at the layer prior to where NIST AI RMF engages. The two frameworks are compatible and complementary.

The EU AI Act's technical documentation requirements for high-risk systems include training data governance and bias assessment requirements at the training and fine-tuning layers. The upstream governance framework provides the governance architecture from which compliant documentation is produced.

ISO 42001 establishes AI management system requirements at the organizational level. The upstream governance framework specifies the operational content that an ISO 42001 management system must cover at the upstream layer.

Licensing

Framework Introduction

Scope

The Upstream Governance Framework defines governance as the structural authorization of decisions before they are made, at the point where they can still be changed.

This framework governs the decisions that precede model deployment. It does not govern incident response, model monitoring, or output auditing. This framework is designed to operate prior to those functions, not instead of them.

This framework does not require organizations to access vendor training data. The vendor transparency standards defined in this framework require vendors to disclose what they can disclose and to accept accountability for what they cannot. The governance obligation the organization carries is to require that accountability, not to perform the vendor's governance for them.

This framework does not replace NIST AI RMF, the EU AI Act, or ISO 42001. It operates at the layer those frameworks assume has already been governed.

How to Use This Framework

The framework is organized in four parts. The Three-Layer Domain Model defines the governance domains, governance decisions, required structural controls, and absence indicators at each of the three layers. This is the specification against which an organization's governance program is measured.

The Vendor Transparency Standards define what to require of vendors at selection and throughout the relationship. The Fine-Tuning Governance Protocol defines what to govern when using proprietary data for fine-tuning.

The Maturity Assessment Rubric scores governance maturity across all twelve domains and identifies priority governance gaps.

The Standards Mapping section defines the standards mapping, licensing, and versioning terms

.

Supporting documents are provided in three appendices: Appendix A contains field guidance for the Data Provenance Record, Appendix B contains field guidance for the Consent Documentation Record, and Appendix C contains relevant definitions. The Data Provenance

Record, Consent Documentation Record, and Residual Risk Acceptance Record forms are available as standalone documents at kindredlabsfoundation.org.

Conventions

The following terms carry defined meanings throughout this document:
Must denotes a binding requirement. An organization implementing this framework is required to meet every "must" statement.
Should denotes a recommended practice. Departure from a "should" statement is permitted where the organization documents its reasoning.
May denotes a permitted option. No preference or requirement is implied.

Three-Layer Domain Model

The domain model defines what governance means at each of the three layers where consequential AI decisions are made. All subsequent components are built against the specifications established here.

For each of the three governance layers, this part defines the governance domains present at that layer, the specific decisions that constitute governance decisions within each domain, the structural controls required, and the indicators that governance is absent.

A doctrinal governance program expresses values and principles without creating mechanisms to enforce them. A structural governance program creates named authority, formal process, documentation requirements, and defined consequences for non-compliance. This framework requires structural governance at every layer.

Training

Definition

The training layer encompasses every decision made about what data a model learns from, what it is trained to recognize or predict, and how its learning process is structured. Governance at this layer precedes the model's existence and cannot be retroactively applied.

Governance Domains

Domain 1.1: Dataset Curation

The decisions governing what data enters the training corpus. Includes source selection, inclusion and exclusion criteria, data cleaning and filtering decisions, and the handling of gaps in representation.

Governance decisions at this domain:

- What sources are eligible for inclusion and on what basis
- What criteria determine whether a data source is representative
- Who reviews inclusion and exclusion decisions and with what authority
- What consent basis is required for data about identifiable populations
- What representational gaps are acceptable and who decides acceptability
- How data cleaning decisions are documented and reviewed

Domain 1.2: Category Design

The decisions governing what the system is trained to recognize, classify, or predict. Category design decisions determine what distinctions the model learns to make, what outcomes it learns to predict, and what boundaries it treats as fixed. These are governance decisions with consequences for every population the system subsequently affects.

Governance decisions at this domain:

- What categories the training process treats as fixed versus contextual
- Who has authority to define or challenge category boundaries
- What review process exists before categories are encoded
- How contested or politically significant categories are handled
- What happens when a category boundary produces discriminatory outcomes
- How category decisions are documented for future audit

Domain 1.3: Training Architecture

The decisions governing how the system learns: what it optimizes for, what it is penalized for, and what trade-offs are built into the learning process. Governance at this domain requires that optimization objectives are stated, justified, and reviewed before training begins.

Governance decisions at this domain:

- What the model is optimized to produce and who authorized that objective
- What trade-offs between accuracy, fairness, and performance are acceptable
- Who reviews the optimization objective before training begins
- How changes to training objectives are authorized and documented
- What evaluation criteria determine whether training is complete
- Who has authority to halt or restart training based on evaluation results

Domain 1.4: Representational Review

The structured process by which the training corpus and category design are evaluated for representational adequacy before training proceeds. Representational review requires named authority to assess, before training begins, which populations the training data represents adequately and which it does not.

Governance decisions at this domain:

- Who conducts representational review and what authority they hold
- What populations or communities require explicit representation assessment
- What standard determines adequate representation
- What remediation is required when representation is inadequate
- Whether review findings can halt training and under what conditions
- How review findings are documented and retained

Required Structural Controls

For governance at this layer to be structural rather than doctrinal, the following controls must exist as formal processes with named authority, documentation requirements, and defined consequences for non-compliance:

- A pre-training review gate: no training proceeds without documented sign-off across all four domains

- Named authority for each domain: a specific role or body with defined scope and enforcement power
- A dataset provenance record: documentation of source, consent basis, and representational assessment for every dataset
- A category decision record: documentation of every category definition decision, rationale, and approving authority
- A training objective authorization: formal sign-off on optimization objectives before training begins, accompanied by a trade-off analysis documenting what was deprioritized between competing objectives and the governance basis for that decision
- A representational review report: findings, the specific standard or criteria applied to determine adequacy, remediation taken, and residual gaps acknowledged before training proceeds

Absence of Governance Indicators

An organization is ungoverned at this layer when:

- Training data is selected by the technical team without a formal review process
- Category boundaries are defined by model architects without governance involvement
- Optimization objectives are set by product or technical leadership without documented review
- No named authority exists for representational adequacy
- Dataset provenance is tracked for legal compliance only, not governance
- No pre-training gate exists: training proceeds without documented governance sign-off

Fine-Tuning

Definition

The fine-tuning layer encompasses every decision an organization makes about how a vendor base model is configured, adapted, or extended for organizational use. It includes decisions about what proprietary data the model learns from, what behavioral changes the organization is introducing, how retrieval is configured, and how any mechanism that shapes model behavior is designed and authorized. These are governance decisions.

Fine-tuning governance has two distinct upstream moments: what assumptions the vendor's base model already carries into the fine-tuning process, and what the organization's own configuration decisions are adding to or reinforcing in those assumptions. Both require governance.

The fine-tuning layer as defined here extends beyond model weight modification. Prompt engineering applied systematically at scale, agent tool configuration, memory system design, and policy classifiers each shape model behavior in ways that carry the same governance

obligations as direct fine-tuning. Any mechanism through which an organization configures how the model behaves falls within this layer's scope.

Governance Domains

Domain 2.1: Proprietary Data Selection

The decisions governing what organizational data enters the fine-tuning process. This is the domain where the organization's own data about its employees, customers, users, or operations becomes training material. Data selection decisions determine what the model is trained on, whose data is included, and what consent basis covers that use.

Governance decisions at this domain:

- What categories of proprietary data are eligible for fine-tuning use
- What consent basis is required for each data category and how it is documented
- Who reviews data selection decisions and with what authority
- What data about employees, customers, or users requires elevated review
- How data that reflects historical organizational bias is identified and handled
- What documentation is required before proprietary data enters fine-tuning

Domain 2.2: Fine-Tuning Objectives

The decisions governing what behavioral changes the organization is introducing through fine-tuning. Fine-tuning objectives define what the model will do differently after training than it did before. These objectives encode organizational priorities, assumptions about users, and judgments about what constitutes a correct output.

Governance decisions at this domain:

- What the fine-tuning is designed to change in the model's behavior and who authorized that objective
- What assumptions about users or use cases are built into the fine-tuning objective
- Who reviews fine-tuning objectives before training begins
- How conflicts between fine-tuning objectives and base model behavior are identified and resolved
- What evaluation criteria determine whether fine-tuning has achieved its objective
- How fine-tuning objective decisions are documented for future audit

Domain 2.3: Customization Scope

The decisions governing how far the organization's customization extends: what behaviors the organization is modifying, what constraints it is adding or removing, and what the boundaries of authorized customization are. Customization scope defines the difference between what the vendor's model does and what the organization's version of it does. That difference is entirely the organization's governance responsibility.

Governance decisions at this domain:

- What behaviors the organization is authorized to modify and who sets that authorization
- What constraints from the base model the organization is permitted to relax
- What new constraints the organization is adding and on what basis
- How customization decisions are reviewed before deployment
- What happens when customization produces unexpected behavior in production
- How the boundary between vendor responsibility and organizational responsibility is documented

Domain 2.4: RAG Configuration

The decisions governing what content a model can retrieve and use at inference time. RAG configuration determines what sources are authorized for retrieval, what the model treats as authoritative, and how retrieved content is weighted against trained knowledge. These decisions shape every output the model produces in a RAG-enabled deployment.

Governance decisions at this domain:

- What sources are authorized for retrieval and on what basis
- Who reviews the retrieval corpus and how often
- How outdated, incorrect, or biased content in the retrieval corpus is identified and remediated
- What authority exists to add or remove sources from the retrieval configuration
- How retrieval configuration changes are authorized and documented
- What evaluation process confirms that retrieval is producing appropriate outputs
- What constitutes a material change to the retrieval corpus requiring governance review, as distinct from routine content updates that do not alter source categories, retrieval weighting, or authorization scope

Required Structural Controls

For governance at this layer to be structural rather than doctrinal, the following controls must exist as formal processes with named authority, documentation requirements, and defined consequences for non-compliance:

- A pre-fine-tuning review gate: no fine-tuning begins without documented review across all four domains
- Named authority for fine-tuning governance with power to approve, pause, or halt
- A fine-tuning data record: documentation of every dataset, including category, consent basis, representational limitations, and approval authority
- A fine-tuning objective authorization: formal sign-off before training begins
- A customization scope boundary document: explicit definition of what the organization has and has not modified, maintained as a living document
- A RAG corpus governance log: documentation of corpus contents, authorization, and review date. A material change — defined as any addition or removal of source

categories, change to retrieval weighting logic, or modification of authorization scope — requires a new governance review before the change is deployed
- A base model assumption register: documented assessment of vendor model assumptions prior to fine-tuning

Absence of Governance Indicators

An organization is ungoverned at this layer when:

- Fine-tuning data is selected by the technical or product team without formal review
- Fine-tuning objectives are not subject to governance review before fine-tuning begins
- No inventory exists of what proprietary data has been used in fine-tuning
- The organization cannot answer what consent basis covers its fine-tuning corpus
- RAG configuration is treated as an infrastructure decision with no governance review
- Customization decisions are not documented separately from general system architecture
- The governance function has no visibility into fine-tuning decisions or activities

Deployment

Definition

The deployment layer encompasses the decisions an organization makes about which AI system to deploy, for what purposes, and under what conditions. It includes vendor selection, use case authorization, deployment scoping, and accountability boundary definition. For organizations that do not train or fine-tune their own models, this is where upstream governance begins.

Governance Domains

Domain 3.1: Vendor Selection

The decisions governing which AI system the organization adopts. Vendor selection determines what assumptions, training data, category design, and optimization objectives arrive pre-encoded in the system the organization will deploy. Governance evaluation of vendors includes what the system was built to optimize for, what its training data was designed to represent, and what structural accountability the vendor accepts for its upstream decisions.

Governance decisions at this domain:

- What governance criteria are included in vendor evaluation alongside capability and cost
- What vendor disclosures are required before selection proceeds
- Who in the governance function has a formal role in vendor selection decisions

- What published documentation is required: model cards, system cards, bias evaluations, third-party audits
- What the vendor's position is on downstream failures that originate in training decisions
- How vendor selection decisions are documented including governance criteria applied and findings

Domain 3.2: Use Case Authorization

The decisions governing what problems a system is authorized to solve, for whom, and under what conditions. Use case authorization defines what the system is and is not permitted to do, what populations it is authorized to affect, and what process governs the expansion of authorized use over time.

Governance decisions at this domain:

- What process exists to formally authorize each use case before deployment
- Who has authority to authorize a use case and what criteria they apply
- What populations are affected by each use case and what elevated review that triggers
- What use cases are explicitly prohibited and how that prohibition is enforced
- How use case creep is monitored and brought back into the authorization process
- How use case authorization decisions are documented and reviewed

Domain 3.3: Deployment Scoping

The decisions governing how the system is deployed: what it can access, what actions it is authorized to take, what human oversight is required, and where the boundaries of its authority are. Deployment scoping defines the operational envelope of the system. Governance at this domain requires that the envelope is explicitly defined and reviewed on a regular cycle.

Governance decisions at this domain:

- What data and systems the deployed model is authorized to access
- What actions the model is authorized to take versus recommend
- Where human review is required before the model's output is acted upon
- What the escalation path is when the model produces an unexpected or high-stakes output
- How deployment scope changes are authorized and documented
- What monitoring is in place and who reviews monitoring outputs with what authority

Domain 3.4: Accountability Boundaries

The decisions governing who is responsible for what when an AI system produces a harmful or unexpected outcome. Accountability boundaries define the line between vendor responsibility and organizational responsibility, between the governance function and the technical function, and between the organization and the people its AI system affects.

Governance decisions at this domain:

- What the organization accepts responsibility for that the vendor does not cover
- What the governance function is accountable for versus the technical function
- What remediation process exists for people harmed by deployment decisions
- How failures are investigated and at what layer root cause is assessed
- What triggers an escalation from deployment governance to upstream review
- How accountability decisions are documented and communicated internally

Required Structural Controls

For governance at this layer to be structural rather than doctrinal, the following controls must exist as formal processes with named authority, documentation requirements, and defined consequences for non-compliance:

- A vendor governance scorecard: formal evaluation criteria including upstream governance requirements
- A use case authorization register: maintained inventory of every authorized use case, approving authority, populations affected, and review schedule
- A deployment scope document: explicit definition of access, action authority, and human oversight requirements, reviewed on a defined cycle
- Named accountability assignments for each deployment governance domain
- A failure escalation protocol: defined triggers and process for escalating a deployment failure to upstream review
- A vendor disclosure file: maintained record of vendor disclosures, updated at each contract renewal
- A Residual Risk Acceptance Record: required when a vendor declines one or more Vendor Transparency Standards disclosure or accountability requirements, documenting the specific gap, the governance risk accepted, and the named authority approving the decision to proceed

Absence of Governance Indicators

An organization is ungoverned at this layer when:

- Vendor selection is made by procurement and product without formal governance involvement
- No use case authorization process exists: systems are deployed without formal governance review
- The governance function cannot produce an inventory of deployed AI systems and their purposes
- Accountability for AI failures is assigned after the fact rather than defined before deployment
- Deployment scope is not explicitly defined or subject to governance review

- Vendor disclosures are not reviewed by the governance function at selection or renewal
- Failures are investigated at the output layer without any mechanism to escalate to upstream review

Layer Integration

Purpose

The three layers do not operate independently. Decisions made at the training layer create conditions that fine-tuning governance must account for. Decisions made at fine-tuning create conditions that deployment governance must account for. Vendor decisions made outside the organization's direct control create governance obligations the organization must fulfill anyway. This section defines the handoff points between layers, the vendor decision surface, and the temporal structure that governs when each layer must be governed relative to the others.

Layer Handoff Points

Training to Fine-Tuning

What the training layer encodes arrives at the fine-tuning layer as the base model's assumptions. This has two governance implications.

First, the fine-tuning governance program must account for what it is inheriting. The base model assumption register defined in Layer 2 is the mechanism for this. Before fine-tuning begins, the organization must document what it knows and does not know about the assumptions already encoded in the model it is adapting.

Second, fine-tuning can reinforce, correct, or compound training layer assumptions. A fine-tuning objective designed without assessment of base model assumptions may deepen a representational gap the organization did not introduce. The pre-fine-tuning review gate must include explicit assessment of base model assumptions relative to the proposed fine-tuning objective.

Handoff governance requirement: the base model assumption register must be completed and reviewed before any fine-tuning objective is authorized.

Fine-Tuning to Deployment

What the fine-tuning layer produces arrives at the deployment layer as the customized model the organization will deploy. The deployment governance program must have documented knowledge of what fine-tuning has been done and what it was designed to change before use case authorization proceeds.

The customization scope boundary document defined in Layer 2 is the primary handoff artifact. It must be available to the deployment governance function before use case authorization proceeds. The use case authorization process must include explicit assessment of whether the proposed use case is within the appropriate scope of the system's customized behavior.

Handoff governance requirement: the customization scope boundary document must be current and reviewed by the deployment governance function before use case authorization for any fine-tuned system.

Deployment Back to Upstream

The handoff does not only run forward. Deployment failures carry diagnostic information about upstream decisions. A system that produces discriminatory outputs in production may reflect assumptions encoded at the training or fine-tuning layer. The framework requires a defined escalation path from deployment failure back to upstream review.

The failure escalation protocol defined in Layer 3 is the mechanism. It must define specific triggers that move investigation from deployment governance to fine-tuning governance or training governance.

Handoff governance requirement: every deployment failure investigation must include an explicit determination of whether the failure signals a fine-tuning or training layer assumption. That determination must be documented regardless of whether escalation follows.

Vendor Decision Surface

Organizations deploying vendor models do not control the training layer directly. This creates a governance obligation that operates differently from the internal governance obligations defined in Layers 1 and 2.

The vendor decision surface is the set of upstream decisions the vendor has made that shape the behavior of the system the organization deploys. It includes what the model was trained on, what categories it was trained to recognize, what it was optimized to produce, and what representational gaps exist in its training data.

The organization cannot govern these decisions directly. It can require the vendor to be accountable for them. That accountability operates through three mechanisms:

- **Disclosure requirements at selection**: the vendor must provide documentation of training data sources, known limitations, bias evaluation findings, and optimization objectives as a condition of selection. This is the vendor governance scorecard defined in Layer 3.
- **Contractual accountability**: the vendor must accept defined responsibility for failures that originate in training decisions, including a defined process for investigating whether a deployment failure has a training layer cause.
- **Ongoing transparency obligations**: the vendor must disclose material changes to the model that affect the assumptions the organization relied on at selection. Model updates that alter category behavior, representational characteristics, or optimization objectives are governance events requiring organizational review.

Where a vendor declines to meet one or more of the disclosure or accountability requirements defined the Vendor Transparency Standards, the organization must not proceed as though the requirement does not apply. The governance function must produce a Residual Risk Acceptance Record documenting which requirements the vendor declined, what governance risk that creates, and the named authority's affirmative decision to proceed despite that risk. Proceeding without a completed Residual Risk Acceptance Record is a governance failure at Domain 3.1.

The Vendor Transparency Standards operationalize these three mechanisms into contract language and procurement requirements.

The Residual Risk Acceptance Record is available as a standalone form at kindredlabsfoundation.org. The record must document: the specific Vendor Transparency Standards requirement the vendor declined, the governance risk the gap creates, any compensating controls the organization is implementing, and the named authority's affirmative sign-off on the decision to proceed. The record must be completed before deployment proceeds and retained as a governance record.

Temporal Structure

Upstream governance is defined by its temporal logic. The decisions that require governance must be governed before the next layer begins.

The required temporal sequence is:

- Training layer governance must be complete before training begins. Dataset curation reviewed, categories approved, optimization objectives authorized, representational review completed and documented. If any of these are incomplete, training does not proceed.
- Fine-tuning layer governance must be complete before fine-tuning begins. Base model assumptions registered, proprietary data reviewed, fine-tuning objectives authorized, customization scope defined. The customization scope boundary document must reflect the current state of the model before fine-tuning adds to it.
- Deployment layer governance must be complete before deployment begins. Vendor governance scorecard completed, use cases authorized, deployment scope defined, accountability assignments documented, failure escalation protocol in place.

Each layer's governance completion is a formal gate with authority to stop what follows if requirements are not met. Post-deployment documentation does not satisfy the governance requirement.

The temporal structure also defines when governance re-engages after initial deployment. Fine-tuning changes require the fine-tuning governance process to run again before the updated model is deployed. Use case expansions require new use case authorization before the expansion goes live. Vendor model updates that materially change base model behavior require the base model assumption register to be updated before the updated model is deployed in production.

Vendor Transparency Standards

Purpose

The Vendor Transparency Standards define what organizations deploying vendor AI systems have the right to require, what vendors are obligated to disclose, and what accountability structures must exist as conditions of the vendor relationship. These standards operationalize the governance obligation that exists at the deployment layer for organizations that do not control the training layer directly.

The standards are organized into four components: disclosure requirements at selection, ongoing transparency obligations, contractual accountability structures, and vendor evaluation criteria.

These standards do not require vendors to provide direct access to training data. They require vendors to disclose what they can disclose and accept accountability for what they cannot.

Disclosure Requirements at Selection

Before an organization selects a vendor AI system for deployment, the vendor must provide documentation sufficient for the organization's governance function to assess what upstream decisions have been made and what their governance implications are. The organization cannot authorize use cases for a system whose upstream decisions it cannot characterize.

1.1 Training Data Documentation

The vendor must provide a written description of the training corpus including: the categories of data sources used, the date range of training data, the geographic and demographic scope of the data, and the criteria used to include or exclude data sources. The description need not identify proprietary sources but must be sufficient for the organization to assess representational scope.

1.2 Known Limitations and Bias Findings

The vendor must disclose all known limitations of the system including published and internal bias evaluation findings, populations or use cases for which the system has demonstrated reduced accuracy or reliability, and any use cases the vendor explicitly does not recommend. This disclosure must include findings from third-party audits where they exist.

1.3 Optimization Objectives

The vendor must describe what the system was optimized to produce, what trade-offs were made between competing objectives during training, and what the system treats as a correct output. This description must be sufficient for the organization to evaluate whether the optimization objective is appropriate for its intended use cases.

1.4 Category Architecture

The vendor must disclose what categories the system is trained to recognize, classify, or predict, and must identify any categories that were treated as fixed during training. Where categories involve human characteristics, behaviors, or identities, the vendor must describe the basis on which those categories were defined.

1.5 Governance Documentation

The vendor must provide documentation of its own internal governance process for the system being deployed, including:

- what review process governed training data selection,
- what authority existed to challenge or halt training,
- and whether any independent review of the system's training was conducted.

Ongoing Transparency Obligations

The assumptions the organization relies on when it authorizes use cases are based on the system as it existed at selection. When the vendor changes the system materially, the organization's governance assessments may no longer be valid. Ongoing transparency obligations ensure the organization is informed when material changes occur.

2.1 Material Model Updates

The vendor must notify the organization in advance of any model update that materially alters the system's behavior, category architecture, optimization objectives, or known limitations. Notification must occur before the update is deployed to the organization's instance and must include sufficient description for the organization to assess whether its existing use case authorizations remain valid.

2.2 New Bias or Limitation Findings

The vendor must disclose any new bias evaluation findings, limitation discoveries, or third-party audit results that affect the system the organization is deploying, within a defined period of the vendor becoming aware of them. The organization's governance function must be notified directly, not through general product release notes.

2.3 Training Data Changes

Where the vendor retrains or significantly updates the model's training corpus, the vendor must disclose the nature of the change and its governance implications. The organization must be given sufficient time and information to update its base model assumption register before the retrained model is deployed.

2.4 Incident Disclosure

The vendor must disclose any material incident involving the system that the organization is deploying, including incidents at other client organizations where the failure mode is relevant to the organization's use cases. Incident disclosure must include the vendor's assessment of whether the incident originated in a training layer decision.

Contractual Accountability Structures

The following contractual provisions give the disclosure requirements structural force. They define consequences for non-disclosure, assign responsibility for training layer failures, and create an escalation path from deployment failures to vendor accountability.

3.1 Training Layer Responsibility

The vendor must accept contractual responsibility for failures that originate in training decisions the vendor made. This responsibility does not transfer to the organization simply because the organization deployed the system. The contract must define a process for determining whether a deployment failure has a training layer cause and must assign the vendor responsibility for remediation where training layer causation is established.

3.2 Disclosure Breach Remedy

The contract must define a specific remedy for the vendor's failure to make required disclosures, including failure to disclose known limitations at selection, failure to notify of material model updates, and failure to disclose new bias findings. The remedy must be sufficient to incentivize compliance, not merely acknowledge the breach.

3.3 Audit Rights

The organization must retain the right to commission an independent third-party audit of the vendor system at defined intervals or upon defined triggers. The vendor must cooperate with such audits within defined scope and timeframe. Audit rights must survive contract renewal and must not be conditioned on additional fees beyond reasonable scope costs.

3.4 Upstream Escalation Process

The contract must define a formal escalation process by which the organization can require the vendor to investigate whether a deployment failure originates in a training layer decision. The vendor must respond within a defined timeframe, must conduct the investigation at its own cost where training layer causation is plausible, and must provide the organization with findings sufficient to update its governance documentation.

3.5 Governance Representation

The vendor must represent that the disclosures made at selection are accurate and complete to the best of the vendor's knowledge, that no material information relevant to the organization's governance assessment has been withheld, and that the vendor's internal

governance process for the system is as described. Misrepresentation of these matters must constitute a material breach.

Vendor Evaluation Criteria

The vendor governance scorecard is the instrument organizations use to apply these standards at the selection stage. It translates the disclosure requirements and accountability structures into an evaluation framework the governance function applies during vendor selection.

The scorecard evaluates vendors across four dimensions:

4.1 Disclosure Quality

Does the vendor's documentation for each required disclosure provide sufficient information for a governance assessment, or does it satisfy the form of disclosure without the substance?

Evaluation criteria:

- Completeness of training data description
- Specificity of known limitations disclosure
- Adequacy of optimization objective description
- Granularity of category architecture disclosure
- Evidence of internal governance process

4.2 Governance Process Evidence

Does the vendor provide credible evidence that upstream governance decisions were made through a structured process?

Evaluation criteria:

- Evidence of pre-training review process
- Existence of third-party audit findings
- Structured bias evaluation methodology
- Named governance authority for training decisions

4.3 Accountability Acceptance

Is the vendor willing to accept the contractual accountability structures required by these standards?

Evaluation criteria:

- Willingness to accept training layer responsibility provision
- Willingness to define disclosure breach remedy
- Acceptance of audit rights
- Willingness to define upstream escalation process

4.4 Ongoing Transparency Track Record

For vendors with existing deployments, what is the vendor's track record on ongoing transparency obligations?

Evaluation criteria:

- Documented history of proactive disclosure
- Responsiveness to governance function inquiries
- Quality of incident disclosure where incidents have occurred

Fine-Tuning Governance Protocol

Purpose

The Fine-Tuning Governance Protocol defines what organizations must govern when they use proprietary data to fine-tune a vendor base model. It covers every organization that goes beyond deploying a vendor model as-is: those fine-tuning with employee data, customer data, behavioral data, operational data, or any other proprietary data source. It also covers organizations using retrieval-augmented generation at scale, where the retrieval corpus functions as a continuous fine-tuning mechanism.

The underlying principle governing this protocol is that no data may be used to shape model behavior without documented provenance, consent basis, and governance review.

The protocol is organized into four components: data provenance requirements, consent documentation standards, category review process, and ongoing audit requirements.

Data Provenance Requirements

Data provenance governance establishes where data came from, what it contains, and what it was originally collected for, before the organization determines whether it is appropriate for fine-tuning use. Provenance documentation is required for every dataset that enters the fine-tuning process. The governance question is not only whether the organization has the right to use the data, but whether using this data for this purpose is within the scope of authorized fine-tuning objectives.

See Appendix A: Data Provenance Record Template for the implementation form for this component.

1.1 Source Identification

The origin of the dataset must be documented: what system, process, or activity generated it, what time period it covers, and what population it reflects. For data generated by organizational processes, the documentation must identify whether the process that generated the data was itself subject to bias or differential treatment that would be reproduced in the fine-tuning corpus.

1.2 Original Collection Purpose

The purpose for which the data was originally collected must be documented and assessed for compatibility with fine-tuning use. The provenance record must assess whether the individuals whose data this is had awareness that it might be used for AI model training.

1.3 Representational Scope

The provenance record must document what populations, behaviors, or conditions the data reflects and what it does not. A dataset generated by organizational processes reflects the organization's history, including historical patterns of access, opportunity, and treatment. Those patterns must be assessed as part of the provenance review.

1.4 Known Data Quality Issues

Any known quality issues must be documented: gaps, anomalies, periods of unreliable data collection, and conditions under which the data may not accurately reflect the population it purports to represent. The governance assessment must include whether known limitations would produce systematic misrepresentation when encoded through fine-tuning.

1.5 Prior Use History

Where the dataset has been used in prior fine-tuning or model development, that history must be documented. Cumulative fine-tuning on the same dataset can compound representational assumptions across model generations. The provenance record must reflect prior use history so the governance review can assess cumulative effect.

Consent Documentation Standards

Consent documentation establishes the basis on which the organization is authorized to use data about identifiable people for fine-tuning purposes. This is distinct from the legal question of whether the organization has terms of service that permit broad data use. The governance question is whether the people whose data is being used for fine-tuning were meaningfully informed that their data would be used for this purpose and whether they had a meaningful opportunity to object. Both the legal basis and the governance basis must be documented.

See Appendix B: Consent Documentation Template for the implementation form for this component.

2.1 Consent Basis Classification

Every dataset entering fine-tuning must be classified by consent basis. Four classifications apply:

Explicit consent: the individuals whose data is included were specifically informed that their data would be used for AI training or fine-tuning and affirmatively agreed.

Informed general consent: the individuals whose data is included were informed that their data might be used for purposes including AI development, in terms sufficiently specific that AI fine-tuning was a reasonably foreseeable use.

Implied organizational consent: the data was generated within an employment or service relationship where AI development use was disclosed as an organizational practice at the time the relationship was established.

No documented consent basis: the data is being used without a documented basis for the individuals' awareness of this use. This classification does not automatically prohibit use. It

triggers elevated review. The governance function must make an affirmative decision that the use is appropriate despite the absence of a documented consent basis, and that decision must be recorded.

2.2 Consent Basis Documentation

For each dataset, the documentation that establishes the consent basis must be identified and retained:

- the specific terms of service language,
- employment agreement provision,
- privacy notice,
- or other instrument that establishes the basis.

Where no such document exists, the absence must be recorded and the elevated review process initiated.

2.3 Consent Scope Assessment

The documented consent basis must be assessed against the specific fine-tuning use. The assessment must determine whether a reasonable person who provided this consent would understand it to cover this specific use.

2.4 Withdrawal and Deletion Obligations

The consent documentation must address what happens when an individual withdraws consent or requests deletion of their data. The protocol must define:

- whether withdrawal from the original data source requires removal from the fine-tuning corpus,
- what the process is for removing data from a fine-tuning corpus that has already been used,
- and what the organization's obligation is to individuals whose data has already shaped a deployed model.

Category Review Process

Category review examines what the fine-tuning is teaching the model to recognize, distinguish, or predict. Fine-tuning objectives encode assumptions about what distinctions matter and how to make them. Category review is the governance process that ensures those assumptions are explicitly identified and authorized before fine-tuning begins.

3.1 Category Inventory

Before fine-tuning begins, the governance function must produce an inventory of the categories the fine-tuning objective will reinforce or introduce. A category in this context is any distinction the model is being trained to make:

- between high-performing and low-performing employees,
- between high-risk and low-risk customers,
- between qualified and unqualified applicants.

The inventory must name each category explicitly. Unnamed categories cannot be reviewed.

3.2 Category Origin Assessment

For each category in the inventory, the review must assess where the category came from. Three origins require different governance treatment:

- Vendor-inherited categories: distinctions the base model already makes that the fine-tuning will reinforce. These require assessment of whether the vendor's category design is appropriate for the organization's use case and population.
- Data-derived categories: distinctions that emerge from patterns in the fine-tuning data rather than from explicit design. These require careful review because they may encode historical organizational patterns that were not part of the stated fine-tuning objective.
- Explicitly designed categories: distinctions the organization has deliberately chosen to build into the fine-tuning objective. These require the clearest governance rationale and the most explicit authorization.

3.3 Contested Category Protocol

Any category that involves human characteristics, behaviors, identities, or assessments of human value or capability is a contested category and requires elevated review. The elevated review must include:

- A named authority who approves the category's inclusion
- A documented rationale for why the distinction is appropriate for the use case
- An assessment of differential outcomes across protected or vulnerable populations
- A defined review schedule for reassessment after deployment

3.4 Category Authorization

No fine-tuning may proceed until every category in the inventory has been reviewed and authorized by named governance authority. The authorization must be recorded: who authorized each category, on what basis, and what conditions or review requirements attach to the authorization. Category authorization is not permanent. It must be renewed when the fine-tuning objective changes, when the data corpus changes materially, or on the defined review schedule.

Ongoing Audit Requirements

Purpose

Fine-tuning governance does not end when fine-tuning is complete. The model the fine-tuning produced continues to operate on the basis of what it learned. Ongoing audit requirements ensure the organization maintains visibility into what its fine-tuning decisions produced and retains the ability to identify when a correction is needed.

4.1 Fine-Tuning Inventory

The organization must maintain a current inventory of all fine-tuning that has been applied to models in active deployment. The inventory must include: what fine-tuning was applied, when, using what data, for what objective, under what authorization, and what version of the base model it was applied to. This inventory is the organizational record of what proprietary data has shaped what deployed models.

4.2 Audit Cycle

The fine-tuning data and category authorizations for each deployed model must be reviewed on a defined cycle, not to exceed twelve months. The review must assess: whether the data corpus remains representative of the population the model is currently serving, whether any new quality issues or bias findings have emerged, whether the consent basis for any data in the corpus has changed, and whether the category authorizations remain appropriate given how the model has performed in deployment.

4.3 Out-of-Cycle Review Triggers

The following events must trigger an out-of-cycle review regardless of where the organization is in its audit schedule:

- A deployment failure routed to the fine-tuning layer
- A material change in the population the model is serving
- A significant change in the organizational processes that generated the fine-tuning data
- A legal or regulatory development affecting the consent basis for data in the corpus
- A vendor finding regarding the base model that may interact with the organization's fine-tuning decisions

4.4 Correction Protocol

Where an audit identifies a fine-tuning governance failure, the correction protocol must define: what immediate steps are required to limit harm from the deployed model while correction is underway, what the process is for correcting the fine-tuning corpus or objective, what authority must authorize the corrected fine-tuning before it is deployed, and what documentation is required to close the governance finding.

4.5 Sunset and Retirement

Fine-tuning applied to a model that is retired from deployment does not automatically cease to be a governance concern. The organization must document:

- what happened to the fine-tuned model at retirement,
- whether the fine-tuning corpus or any derivatives remain in organizational systems,
- and whether any subsequently deployed model inherited assumptions from the retired model's fine-tuning.

The governance record for retired models must be retained for a defined period.

Maturity Assessment Rubric

Purpose

The Maturity Assessment Rubric is the diagnostic instrument that translates the framework into an organizational assessment. It scores governance maturity across every domain defined in the Three-Layer Domain Model, identifies specific gaps, and produces a layer-by-layer maturity profile.

The rubric is built against the absence of governance indicators defined in each layer, the structural controls required at each layer, and the handoff requirements defined in the Layer Integration section.

The rubric is designed for two uses:

- self-assessment with facilitated guidance,
- and governance architecture engagement.

The questions are written to be legible to governance and compliance functions without requiring technical expertise to answer.

Maturity Levels

Level 0: Absent

No governance structure exists at this domain. The decisions within this domain are being made without any formal review, documentation, or named authority.

Level 1: Nascent

Awareness of the governance need exists but no formal process or documentation has been established. Governance at this domain happens informally or inconsistently, dependent on individual judgment rather than organizational structure.

Level 2: Developing

A formal process exists but is incomplete, inconsistently applied, or lacks one or more required structural elements. Documentation exists but may be incomplete. Named authority exists but may lack enforcement power. The governance structure is present in outline but not in full operation.

Level 3: Structural

Governance is embedded in decision-making processes with named authority, complete documentation requirements, defined consequences for non-compliance, and a functioning

review cycle. The governance structure operates independently of the individuals currently filling governance roles. It would survive personnel change.

Scoring Architecture

Each domain is scored 0 through 3 against the maturity levels above. The score reflects the weakest structural element present, not an average. A domain with a defined process but no named authority scores 1, not 2, because the absence of named authority means the process cannot be enforced.

Domains are grouped by layer. Layer scores are the average of domain scores within that layer, rounded down. The overall maturity profile is not a single aggregate score. It is a layer-by-layer profile that shows where governance is strongest and where the gaps are most consequential.

Weighting note: Layer 1 (Training) and Layer 2 (Fine-Tuning) gaps carry greater governance consequence than Layer 3 (Deployment) gaps because they are less recoverable. A deployment decision can be reversed. A training decision that encoded a structural assumption into a deployed model cannot be undone without retraining. The rubric does not numerically weight by layer but the diagnostic narrative must reflect this asymmetry.

Assessment Instrument

The instrument is organized by layer and domain. For each domain, the assessor asks the following questions in order. The first question answered "no" determines the domain score. If all questions are answered "yes," the domain scores Level 3.

Training

Organizations that do not train their own models score Layer 1 domains against what they know or can require their vendor to disclose. The base model assumption register in Layer 2 is the mechanism for documenting vendor-dependent domain assessments.

Domain 1.1: Dataset Curation

Level 0 check:

Does any formal review process exist for training data selection?
No = Level 0

Level 1 check:

Is there documented awareness within the governance function that dataset curation decisions are governance decisions?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal review process exist for dataset curation that includes governance function involvement?
No = Level 1

Level 3 check:

Does the formal process include all of the following:

- *named review authority,*
- *documented inclusion and exclusion criteria,*
- *consent basis documentation for each data source,*
- *representational scope assessment,*
- *and a pre-training gate that can halt training if review is incomplete?*

No = Level 2 | Yes = Level 3

Domain 1.2: Category Design

Level 0 check:

Does the governance function have any involvement in category boundary decisions before they are encoded?
No = Level 0

Level 1 check:

Does the governance function have awareness that category design is a governance decision, even without a formal process for influencing it?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal review process exist for category design decisions that includes governance function involvement before categories are encoded?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a category inventory produced before training begins,*
- *named authority for approving contested categories,*
- *documented rationale for each category decision,*
- *and a defined process for challenging category boundaries after deployment?*

No = Level 2 | Yes = Level 3

Domain 1.3: Training Architecture

Level 0 check:

Does any documented governance review exist for optimization objectives before training begins?
No = Level 0

Level 1 check:

Does the governance function have visibility into training architecture decisions even without formal authority over them?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal review process exist for optimization objectives that includes governance sign-off before training begins?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a training objective authorization document,*
- *named authority for sign-off,*
- *defined evaluation criteria for training completion,*
- *and authority to halt or restart training based on governance findings?*

No = Level 2 | Yes = Level 3

Domain 1.4: Representational Review

Level 0 check:

Does a named authority or formal process exist for assessing representational adequacy before training?
No = Level 0

Level 1 check:

Is representational adequacy considered informally during training data selection, even without a structured process?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a structured representational review process exist that assesses the training corpus against defined population coverage criteria before training proceeds?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a named review authority with power to halt training,*
- *defined standards for adequate representation,*

- *a documented remediation process for identified gaps,*
- *and a representational review report retained as a governance record?*

No = Level 2 | Yes = Level 3

Fine-Tuning

Organizations that do not fine-tune their own models but use RAG at scale must assess Domain 2.4 against their RAG configuration governance and score remaining domains as Not Applicable with documentation of why.

Domain 2.1: Proprietary Data Selection

Level 0 check:

Does the governance function have any involvement in fine-tuning data selection decisions?
No = Level 0

Level 1 check:

Does the governance function have awareness that proprietary data selection for fine-tuning is a governance decision?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal review process exist for fine-tuning data selection that includes governance function involvement?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a fine-tuning data record for each dataset,*
- *documented consent basis for each data category,*
- *named review authority,*
- *elevated review for employee or customer data,*
- *and a pre-fine-tuning gate?*

No = Level 2 | Yes = Level 3

Domain 2.2: Fine-Tuning Objectives

Level 0 check:

Are fine-tuning objectives subject to any governance function involvement or review before fine-tuning begins?
No = Level 0

Level 1 check:

Does the governance function have visibility into fine-tuning objectives even without formal authority over them?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal review process exist for fine-tuning objectives that requires governance sign-off before fine-tuning begins?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a fine-tuning objective authorization document,*
- *named sign-off authority,*
- *documented assessment of assumptions encoded in the objective,*
- *defined evaluation criteria,*
- *and a process for documenting objective decisions for future audit?*

No = Level 2 | Yes = Level 3

Domain 2.3: Customization Scope

Level 0 check:

Does any governance-specific record exist for customization decisions, separate from general system architecture documentation?
No = Level 0

Level 1 check:

Does the governance function have awareness that customization scope decisions are governance decisions, even without a formal review process?

No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal process exist for reviewing and documenting customization decisions before deployment?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a customization scope boundary document maintained as a living record,*
- *named authority for authorizing customization decisions,*
- *a defined process for reviewing customizations that relax base model constraints,*
- *and handoff of the boundary document to the deployment governance function?*

No = Level 2 | Yes = Level 3

Domain 2.4: RAG Configuration

Level 0 check:

Is RAG configuration subject to any governance review?
No = Level 0

Level 1 check:

Does the governance function have awareness that RAG configuration is a governance decision?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal review process exist for RAG corpus authorization and configuration changes?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a RAG corpus governance log, named authority for source authorization,*
- *a defined review cycle for corpus content,*
- *a process for identifying and removing outdated or biased content,*
- *and an evaluation process for retrieval output quality?*

No = Level 2 | Yes = Level 3

Deployment

Domain 3.1: Vendor Selection

Level 0 check:

Does the governance function have any formal involvement in vendor selection decisions?
No = Level 0

Level 1 check:

Does the governance function have informal input into vendor selection even without a formal process or defined criteria?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal governance review process exist for vendor selection that includes governance function involvement and defined governance criteria?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a vendor governance scorecard applied at selection,*
- *required disclosures as a condition of selection,*
- *governance function sign-off before selection is finalized,*
- *a vendor disclosure file maintained as a governance record,*
- *and defined review at contract renewal?*

No = Level 2 | Yes = Level 3

Domain 3.2: Use Case Authorization

Level 0 check:

Does a formal process exist for authorizing use cases before deployment?
No = Level 0

Level 1 check:

Does informal consideration of use case appropriateness occur before deployment even without a structured authorization process?
No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal use case authorization process exist that requires governance review before a use case goes into production?
No = Level 1

Level 3 check:

Does the process include all of the following:

- *a use case authorization register,*
- *named authority for authorization decisions,*
- *defined criteria including population impact assessment,*
- *a list of explicitly prohibited use cases with enforcement mechanism,*
- *and a process for identifying and reviewing use case creep?*

No = Level 2 | Yes = Level 3

Domain 3.3: Deployment Scoping

Level 0 check:

Does a formal governance review of deployment scope exist that covers access, action authority, and human oversight requirements?
No = Level 0

Level 1 check:

Does informal consideration of deployment scope occur even without a formal process or documented scope definition?

No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a formal process exist for defining and reviewing deployment scope before a system goes into production?

No = Level 1

Level 3 check:

Does the process include all of the following:

- *a deployment scope document,*
- *named authority for scope decisions,*
- *defined human review requirements for high-stakes outputs,*
- *a process for authorizing scope changes,*
- *and a monitoring review process with defined authority?*

No = Level 2 | Yes = Level 3

Domain 3.4: Accountability Boundaries

Level 0 check:

Is there a pre-defined accountability structure in place to handle AI failures before they occur?

No = Level 0

Level 1 check:

Does informal understanding of accountability exist within the governance function even without documented assignments?

No = Level 0 | Yes = Level 1 minimum

Level 2 check:

Does a documented accountability structure exist that defines governance function versus technical function responsibility before a failure occurs?

No = Level 1

Level 3 check:

Does the structure include all of the following:

- *named accountability assignments for each deployment governance domain,*
- *a documented vendor versus organizational responsibility boundary,*
- *a defined remediation process for affected individuals,*
- *a failure escalation protocol with upstream review triggers,*

- *and a process for documenting and learning from accountability decisions?*

No = Level 2 | Yes = Level 3

Scoring and Output Format

After completing the instrument, the assessor calculates domain scores and produces the maturity profile in the following format:

- Layer 1 Training: Domain scores for 1.1, 1.2, 1.3, 1.4. Layer score (average, rounded down).
- Layer 2 Fine-Tuning: Domain scores for 2.1, 2.2, 2.3, 2.4. Layer score (average, rounded down). Note any domains scored Not Applicable.
- Layer 3 Deployment: Domain scores for 3.1, 3.2, 3.3, 3.4. Layer score (average, rounded down).

The profile is presented as a layer-by-layer summary rather than an aggregate score.

Gap narrative: The output must include a written gap narrative that identifies the two or three highest-priority governance gaps, the layer at which each gap exists, what decisions are being made without governance at each gap, and what the first governance action at each gap would be.

Rubric Revision Protocol

This rubric is Version 1.0, built against the domain model defined in the Three-Layer Domain Model. It will be revised as implementation experience produces data on rubric performance in practice.

Revision triggers:

- A scoring question that consistently produces ambiguous answers in practice
- A governance domain that real organizations consistently describe in terms the rubric does not capture
- A maturity level distinction that does not hold up when applied to real governance programs
- Any systematic gap between rubric scores and observed governance quality in the gap narrative

The revision process: document the trigger, propose the specific question or level description change, apply the proposed revision to the prior engagement data to assess whether it produces a more accurate result, and update the rubric with a version increment and change record.

Standards Mapping, Licensing, and Versioning

Standards Mapping

Purpose

The standards mapping defines how the Upstream Governance Framework relates to existing AI governance standards. It identifies where the frameworks are complementary, where they overlap, and where they address distinct governance problems.

NIST AI Risk Management Framework

The NIST AI RMF organizes AI governance around four functions: Govern, Map, Measure, and Manage. It is a risk management framework applied to AI. Its Govern function addresses organizational roles and responsibilities. Its Map function addresses AI system context and risk identification. Its Measure function addresses risk analysis and evaluation. Its Manage function addresses risk response and recovery.

Relationship to the upstream governance framework: the NIST AI RMF operates primarily at the deployment layer and at the output layer. Its Map function touches the upstream layer in identifying AI system context but does not define governance requirements for the decisions that created that context. The upstream governance framework operates at the layer prior to where NIST AI RMF engages.

Compatibility: the upstream governance framework is designed to be compatible with NIST AI RMF. The governance records and documentation requirements defined in the upstream framework provide input to NIST AI RMF's Map and Measure functions. Organizations that are NIST AI RMF aligned can adopt the upstream governance framework without replacing or conflicting with their existing alignment.

EU AI Act

The EU AI Act establishes risk-based requirements for AI systems deployed in the European Union, with the most stringent requirements applying to high-risk AI systems. It requires conformity assessments, technical documentation, human oversight, transparency, and post-market monitoring for high-risk systems.

Relationship to the upstream governance framework: the EU AI Act's technical documentation requirements for high-risk systems include requirements for training data governance, data quality, and bias assessment. These requirements operate at the training layer and the fine-tuning layer. The upstream governance framework provides a structured governance architecture for meeting these requirements. Organizations subject to the EU AI Act that implement the upstream governance framework will have governance documentation that supports their Act compliance obligations.

Compatibility: the upstream governance framework's documentation requirements for dataset provenance, category review, and training objective authorization directly support the EU AI

Act's technical documentation requirements for high-risk systems. The framework does not replace Act compliance but provides the governance structure from which compliant documentation is produced.

ISO 42001

ISO 42001 is the international standard for AI management systems. It follows the ISO management system structure familiar from ISO 9001 and ISO 27001: establishing context, leadership, planning, support, operation, performance evaluation, and improvement. It addresses AI governance at the organizational management system level.

Relationship to the upstream governance framework: ISO 42001 establishes management system requirements for responsible AI development and use. It addresses governance at the policy and management system level rather than specifying governance requirements for particular upstream decisions. The upstream governance framework specifies what the governance decisions at each layer are and what structural controls are required. It operates at the decision and process level where ISO 42001 operates at the management system level.

Compatibility: the upstream governance framework is designed to operate within an ISO 42001 management system. The governance domains, structural controls, and documentation requirements defined in the upstream framework are the operational content that an ISO 42001 AI management system governs. Organizations implementing ISO 42001 can use the upstream governance framework to specify what their management system must cover at the upstream layer.

Licensing and Attribution

License

Required Attribution

All use of the framework, including adaptation, incorporation into other documents, and organizational adoption, must include the following attribution:

> *This work is based on the Upstream Governance Framework developed by Kindred Labs Foundation, published under CC BY 4.0. The original framework is available at kindredlabsfoundation.org.*

For adaptations that materially alter the framework's content, the attribution must additionally state:

> *This is an adaptation of the original framework. The original has not been reviewed or endorsed by Kindred Labs Foundation.*

What Attribution Does Not Permit

Attribution does not constitute endorsement. Organizations that adopt or adapt the framework may not represent that Kindred Labs Foundation has assessed, certified, or endorsed their governance program. Adoption of the framework does not constitute certification against it. Certification is a separate function distinct from adoption of an open standard.

Conformance and Adoption Language

Organizations implementing this framework may use the following defined terms to describe their adoption status. Use of these terms carries the meaning defined here. No other conformance claims are authorized by this framework.

Self-Attested Alignment: The organization has reviewed the framework, applied the Maturity Assessment Rubric, and is implementing governance structures against the framework's domain model. The organization makes this determination internally without external review.

Independent Maturity Assessment: The organization has engaged an independent assessor to administer the Maturity Assessment Rubric and produce a maturity profile. The assessment was conducted against Version [X.X] of the framework. An independent maturity assessment does not constitute certification.

Certified Conformance: Reserved. No certification program exists under this framework as of Version 1.0. Organizations may not claim certified conformance. A certification program may be developed in a future version.

Organizations may not represent that Kindred Labs Foundation has reviewed, approved, or endorsed their governance program on the basis of any adoption status. Adoption of the framework and certification against it are distinct functions.

Document Versioning

The published framework is Version 1.0. The version number appears on the cover, in the header, and in the attribution language. All subsequent versions follow semantic versioning: major version increments for structural changes to the domain model or maturity levels, minor version increments for additions to protocols or rubric revisions, patch increments for editorial corrections.

A version changelog is maintained as a public document at kindredlabsfoundation.org alongside the framework. Each version entry records what changed, why, and what prompted the change.

Publication Location

The framework is published at kindredlabsfoundation.org as a freely downloadable document. The executive summary is available separately. The version changelog is maintained as a separate public document. The framework is not paywalled, gated, or restricted to registered users.

About Kindred Labs Foundation

Kindred Labs Foundation is an AI research and governance organization. The Foundation develops governance architecture, conducts original research, and publishes open standards for AI governance.

Appendix A: Data Provenance Record

This appendix provides field guidance for the Data Provenance Record. The form is available as a standalone document at kindredlabsfoundation.org. A completed record must be on file for every dataset that enters the fine-tuning process before fine-tuning may begin. Complete one record per dataset.

Record header fields

Record ID: a unique identifier assigned to this record at creation, including the date of creation. Record IDs must be retained in the fine-tuning inventory defined in Fine-Tuning Governance Protocol, Component 4.
Model: the name and version of the model this dataset is associated with.
Prepared by: the name and role of the person completing this record.
Reviewed by: the name and role of the governance authority reviewing and approving this record.

Section 1: Source Identification

Data source name: the name of the dataset or data source as it is identified within organizational systems.
Generating system or process: the system, application, or organizational process that produced this data. Sufficient detail must be provided for a reviewer unfamiliar with the system to understand what it is and what it produces.
Date range: the start and end dates of the data included in this dataset.
Population reflected: a description of who or what this data describes — the people, behaviors, transactions, or conditions it captures.
Known bias in generating process: an assessment of whether the process that generated this data was subject to differential treatment, historical bias, or conditions that would systematically over- or underrepresent any population. If yes or unknown, the nature of the bias or uncertainty must be described. This field requires an affirmative assessment, not a default assumption that no bias exists.

Section 2: Original Collection Purpose

What was this data originally collected for: the stated purpose at the time of collection, as documented in the system of record, privacy notice, or data governance documentation.
Is the original collection purpose compatible with fine-tuning use: an assessment of whether using this data for fine-tuning is within the scope of the purpose for which it was collected. Where compatibility is uncertain, the record must be routed to elevated review. The assessment must include a written explanation of the basis for the compatibility determination.
Individual awareness: an assessment of whether individuals whose data is included had awareness at the time of collection that their data might be used for AI model training. Yes, No, or Unknown. If No or Unknown, this must be noted in the consent documentation record for this dataset.

Section 3: Representational Scope

Populations represented: the populations, communities, or groups reflected in this dataset. This field requires specificity — a description sufficient for a reviewer to assess whether the dataset is appropriate for the intended fine-tuning use.
Populations underrepresented: the populations, communities, or groups that are absent or underrepresented relative to the population the model will subsequently affect. If none are identified, the basis for that determination must be documented.
Historical patterns: an assessment of whether this data reflects historical organizational patterns of access, opportunity, or differential treatment that would be reproduced through fine-tuning. If yes or unknown, the patterns must be described.
Representational adequacy assessment: a determination of whether the dataset's representational scope is adequate for the intended fine-tuning use. Three determinations are available:

- Adequate,
- Requires Remediation,
- or Requires Elevated Review.

A determination of Requires Remediation must be accompanied by a description of the remediation applied. A determination of Requires Elevated Review routes the record to the Elevated Review section.

Section 4: Data Quality

Known gaps or anomalies: documentation of any known gaps in data coverage, anomalies in data collection, or periods during which data collection was unreliable. If none are known, that determination must be stated affirmatively.
Conditions of unreliability: documentation of any conditions under which this data may not accurately reflect the population it purports to represent — seasonal variation, system outages, changes in organizational practice, or other factors affecting data integrity.
Quality assessment: a determination of whether data quality is sufficient for governance purposes. Three determinations are available:

- Adequate for Governance Purposes,
- Requires Disclosure in Fine-Tuning Record,
- or Requires Remediation.

A quality determination of Requires Disclosure means the limitation must be documented in the fine-tuning data record and assessed as part of the fine-tuning objective authorization.

Section 5: Prior Use History

Prior fine-tuning use: a determination of whether this dataset has been used in prior fine-tuning or model development. This field requires an affirmative determination — "not to our knowledge" is not sufficient if no record has been checked.
Prior use description: if prior use exists, the models the dataset was used with, the dates of that use, and the fine-tuning objectives it supported must be documented.
Cumulative effect assessment: an assessment of whether prior uses create compounding representational assumptions that require governance attention in the current use. If yes or unknown, the nature of the compounding effect must be described.

Elevated Review

The Elevated Review section is triggered when Section 2 (Original Collection Purpose) or Section 3 (Representational Scope) produces a determination of Requires Elevated Review. Elevated review must be conducted by a named governance authority above the standard review function.

Elevated review authority: the name and role of the governance authority conducting the review.

Routed from: identification of which section triggered elevated review.

Decision:

- Approved,
- Approved with Conditions,
- or Not Approved.

A decision of Not Approved means this dataset may not enter the fine-tuning process. A decision of Approved with Conditions must include a description of the conditions that attach to the approval.

Rationale: the basis for the elevated review decision, sufficient for a future auditor to understand why the decision was appropriate given the triggering finding.

Approval

Conditions on approval: any conditions, limitations, or ongoing review requirements attached to the approval of this record. If none, that must be stated.

Next review date: the date by which this record must be reviewed again. Review is required at the interval defined in the fine-tuning audit cycle or earlier if an out-of-cycle trigger is met.

Governance authority sign-off: signature, name, role, and date of the approving governance authority.

Appendix B: Consent Documentation Template

This appendix provides field guidance for the Consent Documentation Record. The form is available as a standalone document at kindredlabsfoundation.org. A completed record must be on file for every dataset that enters the fine-tuning process before fine-tuning may begin. Complete one record per dataset. A separate record is required for each distinct consent basis within a dataset where the dataset combines data from multiple consent contexts.

Record Information fields

Record ID: a unique identifier assigned to this record at creation, including the date of creation.
Data Provenance Record ID: the Record ID of the corresponding Data Provenance Record for this dataset. These records must be cross-referenced and retained together.
Model: the name and version of the model this dataset is associated with.
Prepared by: the name and role of the person completing this record.
Reviewed by: the name and role of the governance authority reviewing this record.
Review date: the date on which the governance authority reviewed and approved this record.

Section 1: Consent Basis Classification

Consent basis: the classification that applies to this dataset. Four classifications are defined in Fine-Tuning Governance Protocol, Component 2:

- Explicit consent — individuals were specifically informed their data would be used for AI training or fine-tuning and affirmatively agreed.
- Informed general consent — individuals were informed their data might be used for purposes including AI development, in terms sufficiently specific that AI fine-tuning was a reasonably foreseeable use.
- Implied organizational consent — the data was generated within an employment or service relationship where AI development use was disclosed as an organizational practice at the time the relationship was established.
- No documented consent basis — the data is being used without a documented basis for individual awareness of this use. This classification does not prohibit use but triggers elevated review.

Classification rationale: the basis for the classification applied, sufficient for a reviewer to assess whether the classification is appropriate.

Section 2: Consent Basis Documentation

Instrument establishing consent: the specific document that establishes the consent basis — the terms of service provision, employment agreement clause, privacy notice, or other instrument. The document must be identified by name and version or date.
Document location: where the instrument is retained within organizational systems.

Relevant provision: the specific language within the instrument that establishes the consent basis. The provision must be quoted or described with sufficient specificity for a reviewer to assess its scope.
Document absent: if no instrument exists, the absence must be recorded here and the elevated review section completed. An absent document does not default to any consent basis classification — it requires an affirmative governance decision.

Section 3: Consent Scope Assessment

Specific fine-tuning use: a precise description of what this data will be used to train the model to do. Vague descriptions of fine-tuning objectives are not sufficient for a consent scope assessment.
Scope assessment: a determination of whether a reasonable person who provided the documented consent would understand it to cover this specific use. Three determinations are available: Yes, No, or Uncertain. A determination of No or Uncertain routes the record to the Elevated Review section.
Scope assessment rationale: the basis for the scope determination, including any factors that support or complicate the assessment.
Scope gap: if the consent does not clearly cover this use, the nature of the gap must be described. This field is required when the scope assessment is No or Uncertain.

Section 4: Elevated Review

The Elevated Review section is triggered when Section 1 produces a classification of No Documented Consent Basis, or when Section 3 produces a scope assessment of No or Uncertain.
Elevated review authority: the name and role of the governance authority conducting the review.
Elevated review date: the date on which elevated review was conducted.
Routed from: identification of which section triggered elevated review — Section 1 (Consent Basis Classification) or Section 3 (Consent Scope Assessment).
Decision: Approved for Use, Approved with Conditions, or Not Approved. A decision of Not Approved means this dataset may not enter the fine-tuning process. A decision of Approved with Conditions must include a description of the conditions that attach to the approval.
Decision rationale: the basis for the elevated review decision, including why use is appropriate despite the consent gap or why it is not.
Conditions on approval: any conditions, limitations, or safeguards required as a condition of approval. Required when decision is Approved with Conditions.

Section 5: Withdrawal and Deletion Obligations

Withdrawal process: the defined process for handling individual consent withdrawal after data has been included in fine-tuning. This process must be defined before the dataset enters fine-tuning, not after a withdrawal request is received.

Removal from corpus: whether individual data can be removed from the fine-tuning corpus and the process for doing so. If removal is not technically feasible, the constraints and their governance implications must be documented.
Obligation for already-trained models: the organization's defined obligation to individuals whose data has already shaped a deployed model at the time a withdrawal or deletion request is received.
Deletion request process: the process by which individual deletion requests for data in the fine-tuning corpus are received, assessed, and acted upon.

Approval

Conditions on approval: any conditions, limitations, or ongoing obligations attached to the approval of this record. If none, that must be stated.
Next review date: the date by which this record must be reviewed again.
Governance authority sign-off: signature, name, role, and date of the approving governance authority.

Appendix C: Definitions

Base model. A foundation AI model produced by a vendor through large-scale training on external data. The base model arrives in an organization's environment with assumptions, category designs, and optimization objectives already encoded. Organizations that do not train their own models inherit these decisions at the point of vendor selection.

Base model assumption register. A governance record documenting the assumptions, known limitations, and representational characteristics of a vendor's base model as assessed prior to fine-tuning or deployment. Required under this framework before fine-tuning begins. The register ensures that the organization has formally accounted for what it is building on before it begins building.

Consent basis. The legal and ethical foundation on which data about identifiable individuals is collected, retained, and used. Under this framework, a documented consent basis is required for any data entering the training or fine-tuning process. Consent basis classification includes explicit consent, informed general consent, implied organizational consent, and no documented consent basis. See Appendix B for the Consent Documentation Template.

Contested category. A classification, label, or predictive target that encodes a definition of human identity, behavior, or condition that is not universally agreed upon, is subject to political or social dispute, or has documented potential for discriminatory application. Contested categories require named authority for approval and documented rationale under Domain 1.2 (Category Design).

Deployment layer. The governance layer at which an organization decides which AI system to deploy, for what purposes, and under what conditions. Encompasses vendor selection, use case authorization, deployment scoping, and accountability boundary definition. For organizations that do not train or fine-tune their own models, this is where upstream governance begins.

Doctrinal governance. A governance program that operates through the expression of values, principles, and policies without creating the structural mechanisms required to enforce them. A doctrinal program may produce documentation, convene review boards, and publish commitments. It does not create named authority, formal gates, or defined consequences for non-compliance. This framework requires structural governance, not doctrinal governance, at every layer.

Elevated review. A heightened governance review required when a dataset, use case, or deployment decision involves populations with protected characteristics, when consent basis is absent or uncertain, or when potential for harm is materially higher than baseline. Elevated review requires a named authority above the standard review function and documented rationale for the decision reached.

Fine-tuning. The process by which an organization adapts a vendor base model using proprietary data or configuration to produce different behavior than the base model exhibits. As defined in this framework, fine-tuning encompasses weight modification, prompt engineering applied systematically at scale, agent tool configuration, memory system design,

RAG configuration, and policy classifiers. Any mechanism through which an organization configures how the model behaves falls within this definition.

Fine-tuning layer. The governance layer encompassing every decision an organization makes about how a vendor base model is configured, adapted, or extended for organizational use. The fine-tuning layer is fully within the organization's control. See Layer 2 definition in the Three-Layer Domain Model.

Governance decision. A decision that determines what an AI system learns, what it is trained to recognize, what it is optimized to produce, or what it is authorized to do. Governance decisions require named authority, documented rationale, and formal review before they are made.

Governance gate. A formal checkpoint at which no further action may proceed without documented sign-off from named governance authority. This framework requires a pre-training gate at Layer 1, a pre-fine-tuning gate at Layer 2, and a use case authorization gate at Layer 3. A governance gate is structural: it creates a hard stop.

Governance record. A document produced as a required output of a governance process, retained as evidence that a governance decision was made with named authority and documented rationale. Governance records under this framework include dataset provenance records, consent documentation, training objective authorizations, customization scope boundary documents, vendor disclosure files, and use case authorization registers.

Material model update. A change to a deployed AI system substantial enough to require re-execution of one or more governance gates. Includes changes to training data, fine-tuning corpus, optimization objectives, retrieval corpus, customization scope, or deployment configuration that could alter model behavior in ways the original governance review did not assess. Organizations must define in advance what constitutes a material model update and what governance review it triggers.

Named authority. A specific role, individual, or governance body with formally assigned responsibility for a governance domain and the power to approve, pause, or halt decisions within that domain. The absence of named authority is an indicator of doctrinal rather than structural governance.

Optimization objective. The formal specification of what an AI system is trained to maximize, minimize, or produce. Optimization objectives encode organizational judgments about what constitutes a correct output and what trade-offs between accuracy, fairness, and performance are acceptable. Under this framework, optimization objectives are governance decisions requiring formal authorization before training begins.

Policy classifier. A model or rule system applied at inference time to evaluate, filter, or modify model outputs against defined behavioral policies. Policy classifiers shape model behavior and fall within the fine-tuning layer as defined in this framework, carrying the same governance obligations as direct model training.

Representational adequacy. The standard by which the training corpus or fine-tuning data is determined to include sufficient, accurate, and non-distorting representation of the populations the model will subsequently encounter. Representational adequacy is a governance determination, not a statistical one. It requires a named review authority with defined standards and the power to halt training when those standards are not met.

Residual Risk Acceptance Record. A governance record produced when a vendor declines to meet one or more disclosure or accountability requirements defined in the Vendor Transparency Standards. The record documents which requirements were declined, the governance risk created by the gap, and the affirmative decision of named authority to proceed despite that risk. A Residual Risk Acceptance Record is required before deployment may proceed with a non-compliant vendor.

Retrieval-augmented generation (RAG). A deployment architecture in which a model's outputs are shaped at inference time by content retrieved from an external corpus. RAG configuration determines what sources the model treats as authoritative and how retrieved content is weighted. RAG configuration decisions are governance decisions under Domain 2.4 of this framework.

Structural governance. A governance program that operates through named authority, formal process, documentation requirements, and defined consequences for non-compliance. Structural governance creates mechanisms that enforce governance requirements independent of the individuals currently filling governance roles. It would survive personnel change. This framework requires structural governance at every layer.

Training layer. The governance layer at which a model learns from data — what it is trained to recognize or predict, and how its learning process is structured. Encompasses dataset curation, category design, training architecture, and representational review. Governance at this layer precedes the model's existence and cannot be retroactively applied. See Layer 1 definition in the Three-Layer Domain Model.

Upstream governance. Governance of the decisions that determine what an AI system learns, what it is trained to recognize, and what it is optimized to produce, exercised before those decisions become embedded in a deployed system. Upstream governance operates at the training, fine-tuning, and deployment layers.

Upstream review. A formal investigation triggered when a deployment failure, audit finding, or governance escalation indicates that the root cause of a problem may originate at the training or fine-tuning layer rather than the deployment layer. Upstream review requires named authority to conduct the investigation and a documented protocol for what constitutes a finding and what remediation is required.

Use case creep. The expansion of an AI system's application beyond the boundaries of its original authorized use case, without formal re-authorization. Use case creep is a governance failure at Domain 3.2 (Use Case Authorization). This framework requires a defined process for monitoring use case expansion and returning unauthorized uses to the authorization process.