

Standards Body · Foundational source, public edition · Released July 17, 2026

Canonical record: <https://standardsbody.ai/library/foundational-source/foundations-appendix/>

Standards Body is an independent research and institutional-design project. It is not currently a regulator, accreditation body, certification body, or governmental authority. This document is research; it is not an adopted standard.

FOUNDATIONS_APPENDIX.md

Foundations Appendix

Project: Standards Body

Primary domain: standardsbody.ai

Core line: Foundations for Frontier AI

Document type: Canonical cross-foundation integration, dependency, and implementation appendix

Version: 1.0

Status: Approved foundational source

Document owner: Standards Body

Applies to: The eight-foundation series, future standards development, institutional design, evaluation programs, research agendas, templates, scorecards, website summaries, contributor work, partnerships, and document-control decisions

Source documents: FOUNDATION_01_DYNAMIC_EVALUATION_PROTOCOLS.md through FOUNDATION_08_GLOBAL_INTEROPERABILITY.md

Related canonical sources: PROJECT_IDENTITY.md, TERMINOLOGY.md, FOUNDATIONS.md

Review cycle: Annual review, with event-triggered revision after any material change to a foundation, project identity, terminology, evidence standard, governance framework, standards-development process, or institutional role

Document Purpose

This appendix integrates the eight Standards Body foundations into one coherent system.

The individual foundation papers are intentionally deep and specialized. Each addresses a distinct institutional problem:

1. Keeping evaluation valid as AI systems change
2. Preserving evidentiary value through held-out testing
3. Matching evaluation rigor to high-consequence capability
4. Establishing credible independent expert review
5. Building a competent third-party assurance ecosystem
6. Progressing from research to stronger standards and requirements
7. Aligning incentives and prestige with integrity
8. Making evidence portable across institutions and borders

This appendix explains how those parts depend on one another.

It is the authoritative source for:

- The combined architecture of the eight foundations
- The order in which the foundations interact
- Shared assumptions and nonnegotiable principles
- Cross-foundation definitions and interfaces
- Dependency and responsibility mapping
- Consolidated governance, evidence, security, and versioning requirements
- The integrated evaluation and assurance lifecycle
- Common maturity levels
- Consolidated failure modes
- Shared evidence gaps and research priorities
- A unified implementation pathway
- The complete index of operational templates and scorecards
- Rules for maintaining coherence across the foundation library

This document is not a replacement for the individual foundation papers.

It does not shorten or override their specialized requirements.

It provides the system-level view needed to use them together.

Executive Summary

The eight foundations form one institutional chain.

A frontier AI evaluation system begins with a question:

What consequential claim or decision requires evidence?

Foundation 1 turns that question into a dynamic, versioned evaluation protocol.

Foundation 2 determines which tasks, methods, or evidence must remain held out to preserve measurement integrity.

Foundation 3 determines whether the capability or decision is high stakes and therefore requires stronger rigor, domain expertise, uncertainty analysis, thresholds, and safeguards.

Foundation 4 determines when independent expert review is required and what access, competence, mandate, conflicts, dissent, and publication rights are necessary.

Foundation 5 turns isolated reviews into a scalable ecosystem by defining evaluator competence, quality systems, proficiency testing, accreditation concepts, surveillance, registries, and mutual recognition.

Foundation 6 determines how a validated practice may progress from research into guidance, voluntary frameworks, standards, assurance schemes, procurement requirements, recognized codes, and binding obligations.

Foundation 7 examines the incentives surrounding every stage so that organizations are rewarded for valid evidence, disclosure, correction, maintenance, and public contribution rather than gaming, superficial compliance, or prestige accumulation.

Foundation 8 makes the resulting evidence, protocols, qualifications, incidents, and recognition decisions understandable across institutions, languages, sectors, and jurisdictions.

The foundations therefore form a loop rather than a linear sequence.

Evidence from evaluations, independent reviews, audits, incidents, deployments, and international use should feed back into:

- Protocol changes
- holdout renewal
- capability models
- threshold decisions
- reviewer requirements
- evaluator scopes
- standards revisions
- incentive redesign
- interoperability mappings

The system rests on several common positions.

1. The evaluated object must be identified precisely

A model name is not enough.

Every consequential result should identify the relevant model, system configuration, tools, prompts, retrieval, safeguards, access conditions, deployment context, protocol, evaluator, and date.

2. Evaluation is decision-linked

Evaluation is not an abstract score-producing activity.

A protocol should state:

- The decision it informs
- The claim it tests
- The cost of error
- The relevant uncertainty
- The conditions under which the result expires

3. Capability is not risk

Capability evidence is one input to risk judgment.

Risk also depends on:

- Propensity
- access
- exposure
- vulnerability
- actor
- safeguards
- consequence
- context
- uncertainty

4. Passing is not proof of safety

A result can support a bounded claim.

It cannot establish that an AI system is universally safe, secure, harmless, aligned, compliant, or incapable of untested behavior.

5. Protected content and transparent governance can coexist

Exact tasks, solutions, attack libraries, or sensitive methods may remain confidential.

The purpose, construct, governance, evaluator role, uncertainty, limitations, and result status should remain legible.

6. External is not the same as independent

Independent review depends on:

- Competence
- access
- funding
- conflicts
- method freedom
- publication rights
- accountability
- security
- absence of controlling influence

7. Assurance is an ecosystem

Testing, evaluation, inspection, audit, certification, and accreditation are distinct functions.

They should be connected without being collapsed into one vague service.

8. Standards should mature through evidence

A research method should not become a binding requirement merely because the issue is important.

Formality should increase as evidence, consequence, implementation capacity, assurance readiness, and legitimacy increase.

9. Incentives are part of technical design

Metrics, rankings, access, awards, procurement, funding, certification, and penalties change behavior.

Every major protocol or standard should examine the incentives it creates.

10. Interoperability does not require uniformity

Institutions can share terminology, metadata, protocols, result structures, and evaluator evidence while retaining different legal systems, thresholds, and public-policy judgments.

The combined architecture is designed to produce evidence that is:

- Current
- protected where necessary
- proportionate to consequence
- independently challengeable
- institutionally scalable
- progressively formalized
- incentive-aware
- internationally interpretable
- correctable
- retireable

The foundations are strongest when implemented together.

A dynamic protocol without protected tasks may become easy to game.

A held-out test without transparent governance may become unaccountable.

A high-stakes threshold without independent review may become self-serving.

Independent review without evaluator infrastructure may remain scarce and inconsistent.

A certification scheme without mature standards may create false assurance.

A standard without incentive analysis may reward paperwork rather than outcomes.

A national system without interoperability may create duplication and fragmentation.

Interoperability without local authority and capacity may export one institution's assumptions globally.

The purpose of this appendix is to prevent those failures by treating the eight foundations as one integrated institutional system.

1. Role of the Foundations Appendix

1.1 Integration

This appendix combines the foundations at the level of:

- Concepts
- roles
- evidence
- process
- governance
- security
- incentives
- standards
- international use

1.2 Dependency Control

It identifies which foundation supplies prerequisites for another.

Example:

A third-party evaluator cannot be meaningfully accredited without:

- A defined protocol
- protected evidence controls
- high-stakes competence requirements
- independent review standards
- a mature conformity-assessment scheme

1.3 Contradiction Control

It establishes cross-foundation rules that should remain consistent.

1.4 Implementation Control

It provides a sequence for moving from theory to pilots, standards, assurance, and recognition.

1.5 Document Navigation

It directs users to the correct foundation, template, or scorecard for a given task.

1.6 Change Impact

It identifies which other files should be reviewed when one foundation changes.

1.7 Authority Boundary

It preserves the distinction between:

- Standards Body's current research and institutional-design role
 - A proposed future standard
 - A voluntary framework
 - A certification or accreditation system
 - A legal requirement
 - A regulatory decision
-

2. Foundation Library Inventory

2.1 Foundation 1

File: FOUNDATION_01_DYNAMIC_EVALUATION_PROTOCOLS.md

Subject: Maintaining valid, decision-relevant evaluation as models, systems, tasks, threats, and evidence change

Primary output: A versioned evaluation protocol with defined stable and dynamic layers

Central question: How can evaluation remain meaningful over time?

2.2 Foundation 2

File: FOUNDATION_02_HELD_OUT_EVALUATIONS.md

Subject: Protecting tasks, methods, environments, scoring, and other materials when exposure would reduce validity or create harm

Primary output: A governed held-out evaluation plan and custody system

Central question: What should remain protected, from whom, for how long, and under which controls?

2.3 Foundation 3

File: FOUNDATION_03_HIGH_STAKES_CAPABILITY_EVALUATION.md

Subject: Evaluating capabilities relevant to severe, strategic, systemic, or difficult-to-reverse consequences

Primary output: A multidimensional capability evidence case linked to defined thresholds and decisions

Central question: When should evaluation become more rigorous, and what evidence is required?

2.4 Foundation 4

File: FOUNDATION_04_INDEPENDENT_EXPERT_REVIEW.md

Subject: Establishing review that combines expertise, access, independence, accountability, security, and dissent

Primary output: A documented independent review mandate and findings process

Central question: Who should challenge consequential evidence, with what access and authority?

2.5 Foundation 5

File: FOUNDATION_05_THIRD_PARTY_AUDITOR_ECOSYSTEM.md

Subject: Scaling competent and impartial testing, evaluation, inspection, audit, certification, and accreditation

Primary output: A quality-controlled evaluator and assurance ecosystem

Central question: How can independent assurance become repeatable, credible, and available beyond isolated reviews?

2.6 Foundation 6

File: FOUNDATION_06_PROGRESSIVE_STANDARDS_AND_REQUIREMENTS.md

Subject: Progressing practices from research and voluntary use toward standards, procurement, recognition, and binding requirements

Primary output: A maturity and transition-gate system

Central question: When should an emerging practice become more formal or enforceable?

2.7 Foundation 7

File: FOUNDATION_07_INCENTIVES_AND_PRESTIGE.md

Subject: Aligning financial, professional, access-based, reputational, governance, and enforcement incentives

Primary output: Incentive programs that reward integrity, public goods, correction, and contribution

Central question: What behavior does the evaluation and standards system actually reward?

2.8 Foundation 8

File: FOUNDATION_08_GLOBAL_INTEROPERABILITY.md

Subject: Making evidence, protocols, evaluator competence, incidents, standards, and recognition portable across institutions and borders

Primary output: Shared terminology, metadata, profiles, crosswalks, and recognition records

Central question: How can plural institutions understand and use one another's evidence without forced uniformity?

3. Combined System Proposition

The eight foundations support one combined proposition:

Consequential frontier AI claims should be evaluated through current and protected protocols, interpreted through proportionate high-stakes analysis, challenged through competent independent review, supported by accountable third-party assurance, formalized only as evidence and institutions mature, reinforced by aligned incentives, and made portable through global interoperability.

This proposition contains eight requirements.

3.1 Current

Evidence must remain connected to changing systems, tasks, threats, and decisions.

3.2 Protected

Evaluation integrity must be preserved against contamination, leakage, gaming, and harmful disclosure.

3.3 Proportionate

The rigor of evaluation should rise with consequence and uncertainty.

3.4 Independently Challengeable

Consequential claims should not depend solely on the organization making them.

3.5 Institutionally Scalable

Evaluation should be supported by competent organizations, quality systems, surveillance, and accountability.

3.6 Progressively Formalized

Practices should become standards or requirements only when readiness is demonstrated.

3.7 Incentive-Compatible

The system should make accurate evidence and responsible correction more rewarding than manipulation or superficial conformity.

3.8 Interoperable

Evidence should be understandable and reusable across relevant institutional boundaries.

4. The Integrated Architecture

The architecture contains nine connected layers.

Layer 1: Decision and Claim

Define:

- The claim
- decision
- affected parties
- consequence of error
- authority
- evidence standard

Primary foundations:

- Foundation 3
- Foundation 6

Layer 2: Evaluated Object

Identify:

- Model
- system
- version
- configuration
- tools
- access
- deployment context

Primary foundations:

- Foundation 1
- Foundation 3
- Foundation 8

Layer 3: Protocol

Define:

- Construct
- task universe
- administration
- elicitation
- scoring
- uncertainty
- change control
- retirement

Primary foundation:

- Foundation 1

Layer 4: Integrity and Protection

Define:

- Held-out materials
- access
- provenance
- custody
- security
- compromise response
- rotation

Primary foundation:

- Foundation 2

Layer 5: High-Stakes Interpretation

Define:

- Capability dimensions
- threat actors
- uplift
- propensity
- safeguards
- consequence
- thresholds
- decision rules

Primary foundation:

- Foundation 3

Layer 6: Independent Challenge

Define:

- Review mandate
- competence
- independence
- access
- conflict
- dissent
- publication
- appeal

Primary foundation:

- Foundation 4

Layer 7: Assurance Ecosystem

Define:

- Evaluator organization
- quality system
- scope
- proficiency
- audit or certification scheme
- accreditation
- surveillance
- registry

Primary foundation:

- Foundation 5

Layer 8: Institutional Progression and Incentives

Define:

- Maturity
- voluntary adoption
- standards
- procurement
- recognition
- legal requirement

- rewards
- sanctions
- prestige
- public goods

Primary foundations:

- Foundation 6
- Foundation 7

Layer 9: Cross-Border Use and Feedback

Define:

- Terminology
- metadata
- protocol mapping
- recognition
- incidents
- secure exchange
- localization
- capacity

Primary foundation:

- Foundation 8

Every layer feeds the next.

Every later layer may also reveal defects in earlier layers.

5. End-to-End Institutional Flow

Stage 1: Need Identification

A credible concern, opportunity, claim, incident, policy question, procurement decision, or standards gap is identified.

Required questions:

- What decision needs evidence?
- Who is affected?
- What could happen if the decision is wrong?
- Is the issue technical, institutional, legal, or mixed?

Primary foundations:

- Foundation 3
- Foundation 6

Stage 2: Construct and Scope

Define:

- The capability, behavior, safeguard, process, or institutional property
- The evaluated object
- The intended interpretation
- Relevant exclusions

Primary foundations:

- Foundation 1
- Foundation 3
- TERMINOLOGY.md

Stage 3: Protocol Design

Create a complete evaluation protocol.

Primary requirements:

- Decision link
- task universe
- elicitation
- administration
- scoring
- baselines
- uncertainty
- versioning
- reporting
- retirement

Primary foundation:

- Foundation 1

Stage 4: Holdout and Security Design

Determine which components should remain protected.

Primary requirements:

- Threat model

- access classes
- provenance
- chain of custody
- least privilege
- separation of duties
- compromise response
- rotation
- release or retirement

Primary foundation:

- Foundation 2

Stage 5: High-Stakes Classification

Determine whether the capability or decision warrants enhanced rigor.

Primary considerations:

- Consequence
- access
- exposure
- autonomy
- reliability
- uplift
- safeguards
- uncertainty
- false-positive and false-negative cost

Primary foundation:

- Foundation 3

Stage 6: Validation and Pilot

Test:

- Construct validity
- task quality
- scoring
- reliability
- security
- operational feasibility
- burden
- fairness
- decision usefulness

Primary foundations:

- Foundations 1, 2, and 3

Stage 7: Independent Review

Appoint qualified reviewers under a documented mandate.

Primary requirements:

- Independence profile
- sufficient access
- conflict disclosure
- factual correction process
- dissent
- publication rights
- result limitations

Primary foundation:

- Foundation 4

Stage 8: Evaluator Qualification and Assurance

Where repeated third-party work is expected, establish:

- Evaluator scope
- personnel competence
- quality management
- proficiency testing
- surveillance
- complaints
- recognition
- accreditation or certification arrangements where mature

Primary foundation:

- Foundation 5

Stage 9: Operational Use

Apply evidence to:

- Internal decisions
- deployment decisions
- procurement
- insurance
- standards

- public reporting
- regulatory analysis
- research

Primary foundations:

- Foundations 3, 5, 6, and 7

Stage 10: Progression Decision

Determine whether the practice should remain:

- Research
- guidance
- voluntary framework
- reporting convention
- standard
- assurance scheme
- procurement requirement
- recognized code
- mandatory requirement

Primary foundation:

- Foundation 6

Stage 11: Incentive Integration

Examine:

- What adoption rewards exist?
- What gaming will occur?
- Are maintainers funded?
- Are incidents disclosed?
- Are evaluators rewarded for favorable results?
- Does prestige track contribution?
- Are small actors excluded?

Primary foundation:

- Foundation 7

Stage 12: Interoperability and Recognition

Create:

- Common metadata
- terminology mappings

- evaluator profiles
- protocol crosswalks
- recognition decisions
- incident exchange
- local modules

Primary foundation:

- Foundation 8

Stage 13: Monitoring and Feedback

Use:

- New model performance
- incidents
- evaluator findings
- deployment evidence
- market behavior
- international experience
- complaints
- appeals

to update the system.

All eight foundations apply.

Stage 14: Revision, Suspension, or Retirement

A protocol, result, evaluator scope, certificate, requirement, incentive program, crosswalk, or recognition may be:

- Revised
- corrected
- suspended
- superseded
- withdrawn
- retired

The reason and transition should be documented.

6. Foundation Profiles and Interfaces

6.1 Foundation 1: Dynamic Evaluation Protocols

Core Thesis

The protocol, not the dataset alone, is the appropriate unit of evaluation governance.

Institutional Role

Foundation 1 establishes how evidence remains meaningful as:

- Models improve
- systems change
- tools are added
- tasks leak
- benchmarks saturate
- threats evolve
- decisions change

Required Inputs

- Decision question
- construct
- evaluated object
- task universe
- evidence review
- baseline
- threat and deployment context

Primary Outputs

- Protocol identifier and version
- stable core
- dynamic layer
- administration rules
- elicitation policy
- scoring
- uncertainty
- change triggers
- transition plan
- expiration
- retirement

Interfaces

Foundation 1 supplies:

- Protocol structure to Foundation 2
- capability measurement to Foundation 3
- review object to Foundation 4
- technical scope to Foundation 5
- maturity evidence to Foundation 6
- metric and maintenance incentives to Foundation 7
- metadata and version lineage to Foundation 8

Failure if Absent

Without Foundation 1:

- Benchmarks become stale
- historical comparisons become misleading
- model configuration disappears from results
- protocol changes become arbitrary
- results lack expiration
- evaluation gaming becomes easier

6.2 Foundation 2: Held-Out Evaluations

Core Thesis

Some evaluation material should remain unavailable before testing when exposure would materially reduce validity or create harm.

Institutional Role

Foundation 2 establishes the integrity boundary around protected tasks, environments, scoring, methods, and evidence.

Required Inputs

- Evaluation protocol
- contamination threat model
- sensitivity assessment
- evaluated-party access
- security and governance capacity

Primary Outputs

- Holdout classification
- access matrix
- provenance record
- chain of custody

- task rotation plan
- compromise process
- disclosure policy
- retirement or release decision

Interfaces

Foundation 2 supplies:

- Protected evidence to Foundation 3
- access and security conditions to Foundation 4
- custody and quality requirements to Foundation 5
- transparency limits to Foundation 6
- disclosure incentives to Foundation 7
- restricted exchange rules to Foundation 8

Failure if Absent

Without Foundation 2:

- Training contamination weakens evidence
- developers optimize against visible tasks
- sensitive attack methods spread
- result confidence exceeds integrity
- secrecy may become arbitrary because no governed alternative exists

6.3 Foundation 3: High-Stakes Capability Evaluation

Core Thesis

Evaluation rigor should increase with the potential consequence of error, but capability should not be confused with risk.

Institutional Role

Foundation 3 links evaluation evidence to high-consequence decisions.

Required Inputs

- Protocol results
- system identity
- threat model
- access and deployment context
- human baselines
- safeguard evidence
- uncertainty

Primary Outputs

- Capability evidence case
- multidimensional profile
- alert and critical thresholds
- false-positive and false-negative analysis
- domain-expert findings
- decision implications
- re-evaluation triggers

Interfaces

Foundation 3 supplies:

- Consequence and rigor requirements to Foundation 4
- evaluator competence needs to Foundation 5
- transition triggers to Foundation 6
- priority and incentive targets to Foundation 7
- threshold and metadata requirements to Foundation 8

Failure if Absent

Without Foundation 3:

- Every capability receives the same burden
- knowledge tests are mistaken for operational capacity
- thresholds become unexplained labels
- capability and safeguard evidence are collapsed
- high-consequence decisions rely on single scores

6.4 Foundation 4: Independent Expert Review

Core Thesis

No organization should be the sole authority over the evaluation of its own most consequential systems.

Institutional Role

Foundation 4 creates credible challenge without confusing externality, independence, and competence.

Required Inputs

- Review question
- evidence package
- system access
- protocol
- high-stakes classification

- conflict information
- security constraints

Primary Outputs

- Review mandate
- reviewer selection
- independence profile
- conflict disclosures
- findings
- limitations
- minority report
- publication and correction record

Interfaces

Foundation 4 supplies:

- Review evidence to Foundation 5
- independent validation to Foundation 6
- professional incentives to Foundation 7
- reviewer profiles and recognition data to Foundation 8

Failure if Absent

Without Foundation 4:

- developer claims remain self-authored
- external reviewers may lack access
- conflicts remain hidden
- reviewed parties can negotiate conclusions
- dissent disappears
- advisory review is marketed as certification

6.5 Foundation 5: Third-Party Auditor Ecosystem

Core Thesis

Third-party assurance is an ecosystem of distinct activities, roles, competence systems, and recognition arrangements.

Institutional Role

Foundation 5 scales independent work beyond one-off expert panels.

Required Inputs

- Mature protocols
- defined requirements

- reviewer competence models
- quality criteria
- security and evidence-preservation needs
- assurance purpose

Primary Outputs

- Evaluator scopes
- organizational profiles
- quality management
- proficiency testing
- accreditation concepts
- certification-scheme roles
- surveillance
- complaints
- registries
- mutual-recognition conditions

Interfaces

Foundation 5 supplies:

- Assurance capacity to Foundation 6
- evaluator incentives and market structure to Foundation 7
- competence and recognition profiles to Foundation 8

Failure if Absent

Without Foundation 5:

- the word audit covers incompatible activities
- competence remains unverified
- evaluators overclaim broad scope
- evaluator shopping grows
- certificates become difficult to interpret
- markets concentrate without accountability

6.6 Foundation 6: Progressive Standards and Requirements

Core Thesis

Institutional force should increase only as consequence, evidence, implementation capacity, assurance readiness, and legitimacy justify it.

Institutional Role

Foundation 6 governs the path from emerging practice to stronger institutional expectation.

Required Inputs

- Validated method
- evidence of usefulness
- implementation data
- independent review
- evaluator capacity
- burden and competition analysis
- public-interest need

Primary Outputs

- Maturity classification
- transition-gate decision
- voluntary framework
- standard
- assurance scheme
- procurement clause
- recognized code
- mandatory requirement proposal
- sunset and review

Interfaces

Foundation 6 uses all earlier foundations.

It supplies:

- adoption pathways to Foundation 7
- recognized standards and evidence structures to Foundation 8

Failure if Absent

Without Foundation 6:

- immature tests become rigid rules
- voluntary practice persists despite severe externalities
- requirements lack phase-in
- standards become difficult to retire
- compliance formality outruns institutional capacity

6.7 Foundation 7: Incentives and Prestige

Core Thesis

Evaluation infrastructure will be no stronger than the incentives surrounding it.

Institutional Role

Foundation 7 examines what the ecosystem rewards, discourages, hides, and neglects.

Required Inputs

- Actor map
- desired behavior
- evidence measures
- market structure
- access dependencies
- public-goods gaps
- gaming threat model

Primary Outputs

- Incentive program
- recognition system
- grants
- prizes
- bounties
- procurement preferences
- corrective credit
- sanctions
- anti-gaming controls
- impact review

Interfaces

Foundation 7 affects every foundation.

It determines whether:

- Protocols are maintained
- holdouts remain protected
- high-stakes evidence is disclosed
- reviewers remain independent
- evaluators prioritize accuracy
- standards are adopted
- interoperability receives investment

Failure if Absent

Without Foundation 7:

- visible scores outrank valid methods
- evaluators depend on favorable clients
- incidents remain hidden
- public goods remain underfunded

- prestige substitutes for competence
- compliance metrics become targets

6.8 Foundation 8: Global Interoperability

Core Thesis

Frontier AI evidence should be portable across boundaries without requiring identical laws, protocols, thresholds, or policy judgments.

Institutional Role

Foundation 8 creates shared meaning and recognition across plural institutions.

Required Inputs

- Canonical terminology
- system identity
- protocol metadata
- evaluator scope
- result status
- incident taxonomy
- security classification
- legal and local context

Primary Outputs

- Interoperability profile
- protocol crosswalk
- bridge study
- evaluator recognition decision
- incident exchange record
- translation validation
- registry record
- local extension

Interfaces

Foundation 8 makes outputs from Foundations 1 through 7 usable across:

- Organizations
- sectors
- standards systems
- countries
- languages
- open and closed ecosystems

Failure if Absent

Without Foundation 8:

- evidence is repeatedly recreated
- numeric scores travel without context
- evaluator qualifications remain local and opaque
- incidents fail to cross borders
- one dominant framework may be mistaken for global consensus
- smaller jurisdictions become dependent rather than capable

7. Cross-Foundation Dependency Matrix

The following matrix identifies the primary dependency relationships.

Foundation	Depends most directly on	Supplies most directly to
1. Dynamic Evaluation Protocols	Decision question, terminology, construct definition, system identity	Held-out design, high-stakes evidence, review object, evaluator scope, interoperability metadata
2. Held-Out Evaluations	Protocol design, security threat model, task provenance	Protected evidence, integrity status, access conditions, compromise records
3. High-Stakes Capability Evaluation	Dynamic protocol, held-out evidence, system identity, domain expertise	Rigor level, threshold logic, decision implications, reviewer requirements
4. Independent Expert Review	Evidence package, review mandate, access, competence, conflicts	Independent findings, dissent, validation evidence, public limitations
5. Third-Party Auditor Ecosystem	Mature methods, defined criteria, review principles, quality requirements	Scalable assurance, competence recognition, surveillance, registries
6. Progressive Standards and Requirements	Evidence maturity, independent validation, assurance capacity, public-interest need	Standards pathway, procurement conditions, recognized codes, legal proposals
7. Incentives and Prestige	Actor behavior, evidence measures, market and professional context	Adoption, maintenance, disclosure, correction, public-goods support
8. Global Interoperability	Shared terminology, identity, metadata, protocol and evaluator evidence	Cross-border comparability, recognition, incident exchange, capacity coordination

7.1 Hard Dependencies

A hard dependency is a condition without which a later function should not proceed.

Examples:

- Certification requires specified requirements and a valid scheme.
- Accreditation requires defined evaluator scope and competence criteria.
- High-stakes threshold decisions require identified systems and protocols.

- Mutual recognition requires clear scope and status.
- Held-out result claims require task provenance and integrity status.
- Independent findings require sufficient access and a review mandate.

7.2 Soft Dependencies

A soft dependency improves quality but may not be required in every low-consequence case.

Examples:

- International crosswalks
- full external assurance
- public recognition programs
- continuous monitoring
- formal accreditation
- multilingual validation

7.3 Bidirectional Dependencies

Several relationships are reciprocal.

Protocol and Holdout

Protocol design determines what may be protected.

Compromise evidence determines when the protocol must change.

Capability and Safeguards

Capability evidence informs safeguard requirements.

Safeguard evidence changes practical risk and deployment conditions.

Review and Evaluator Ecosystem

Review principles define evaluator expectations.

Evaluator experience reveals where review principles need revision.

Standards and Incentives

Standards create incentives.

Incentive analysis determines whether standards produce meaningful behavior.

Interoperability and Local Practice

Shared structures enable cross-border use.

Local implementation reveals translation, capacity, and equivalence problems.

8. Cross-Foundation Invariants

The following positions should remain stable across all eight foundations unless formally revised.

8.1 Exact Object Identity

Every consequential result should identify the evaluated object precisely.

At minimum:

- Developer or source
- model family
- model or artifact identifier
- model version
- system version
- relevant prompts or configuration status
- tools and external systems
- safeguards
- access tier
- date
- evaluator
- protocol version

8.2 Decision Relevance

Every major evaluation, review, audit, certification, or standard should identify the decision or claim it supports.

8.3 Scope Limitation

No result should be interpreted beyond:

- The evaluated object
- construct
- tasks
- administration conditions
- threat model
- period
- requirement
- jurisdiction
- assurance level

8.4 Versioning

Protocols, task banks, models, systems, standards, evaluator scopes, certificates, recognition decisions, and registries should be versioned.

8.5 Expiration

High-consequence evidence should not remain valid indefinitely.

8.6 Correction

Material errors should produce visible correction, withdrawal, or supersession.

8.7 Independence

Consequential claims should receive scrutiny beyond the organization making them.

8.8 Competence

Independence does not compensate for lack of expertise.

8.9 Access

Competence and independence do not support broad conclusions without sufficient access.

8.10 Uncertainty

Uncertainty should be reported rather than hidden by a single score or categorical label.

8.11 Security

Sensitive information should be protected proportionately.

8.12 Transparency

Protected content should not become an excuse for opaque governance.

8.13 Dissent

Reasoned disagreement should be preserved when material.

8.14 Appeals and Complaints

Consequential decisions should have correction and challenge pathways.

8.15 Anti-Capture

Financial, organizational, ideological, professional, political, and geopolitical conflicts should be identified and managed.

8.16 Small-Actor Access

Requirements should include realistic pathways for smaller evaluators, researchers, developers, and open communities.

8.17 Public-Interest Purpose

The system should improve real decisions rather than produce institutional appearance.

8.18 Interoperability

Shared evidence structures are preferable to unnecessary duplication or forced uniformity.

8.19 No Universal Safety Claim

Passing a protocol, review, audit, or certification is not proof of universal safety.

8.20 Evaluate the Evaluator and the Evaluation

The protocol, reviewer, evaluator organization, standard, incentive program, and recognition arrangement should themselves be evaluated.

9. Shared Evaluated-Object Architecture

The foundations rely on a common representation of what is being assessed.

9.1 Model Layer

Record:

- Base model
- checkpoint
- weights
- tokenizer
- post-training
- fine-tuning
- quantization
- model date
- model hash where available

9.2 System Layer

Record:

- System prompt
- policies
- tools
- retrieval
- memory
- scaffolds
- orchestration
- interfaces
- monitoring
- safeguards

9.3 Access Layer

Record:

- Public or restricted access
- user verification
- rate limits
- permissions
- model-weight availability
- tool permissions
- fine-tuning access
- geographic restrictions

9.4 Deployment Layer

Record:

- Users
- sector
- scale
- autonomy
- critical integrations
- decision authority
- operating environment
- human oversight

9.5 Evaluation Layer

Record:

- Protocol

- tasks
- environment
- elicitation
- retries
- tools
- evaluator assistance
- scoring
- uncertainty
- validity status

9.6 Assurance Layer

Record:

- Reviewer or evaluator
- independence
- scope
- accreditation or qualification
- evidence access
- review level
- result status
- expiry

9.7 Change Layer

Record:

- Material system changes
- protocol changes
- safeguard changes
- deployment changes
- ownership changes
- evaluator changes
- incidents
- status transitions

9.8 Identity Rule

A result should not be inherited by a materially different system without a documented inheritance analysis.

10. Shared Evidence Architecture

10.1 Evidence Classes

The combined foundations recognize several evidence classes.

Direct Technical Evidence

Examples:

- Task outputs
- logs
- trajectories
- environmental states
- test results
- safeguard bypasses
- security artifacts

Process Evidence

Examples:

- Protocol records
- access logs
- custody records
- review records
- quality procedures
- change controls

Organizational Evidence

Examples:

- Governance
- staffing
- competence
- training
- funding
- conflicts
- incident response
- quality management

Operational Evidence

Examples:

- Deployment monitoring
- user incidents
- failure rates

- access patterns
- post-deployment changes

Expert Evidence

Examples:

- Domain judgment
- review findings
- dissent
- threat assessment
- legal or policy analysis

Comparative Evidence

Examples:

- Human baselines
- model baselines
- bridge studies
- cross-evaluator replication
- proficiency tests

Institutional Evidence

Examples:

- Accreditation
- certification
- recognition decisions
- procurement acceptance
- regulatory findings

10.2 Evidence Quality Dimensions

Every material evidence package should be evaluated for:

- Relevance
- validity
- reliability
- provenance
- completeness
- independence
- recency
- security
- reproducibility
- uncertainty
- decision utility

- representativeness
- conflict exposure

10.3 Evidence Portfolio

High-stakes decisions should normally use a portfolio rather than one test.

A portfolio may combine:

- Public benchmarks
- held-out tasks
- dynamic environments
- red-team findings
- domain-expert evaluation
- human-uplift studies
- safeguard testing
- deployment monitoring
- incident evidence
- independent review

10.4 Evidence Case

A consequential claim should be represented as an evidence case containing:

- Claim
- scope
- system identity
- protocol
- direct evidence
- supporting evidence
- contrary evidence
- assumptions
- uncertainty
- reviewer findings
- decision implications
- status
- expiration

10.5 Evidence Inheritance

Evidence may be reused only when:

- The system is materially equivalent
- The protocol remains applicable
- The requirement has not changed
- The evaluator scope remains valid

- The evidence has not expired
- No incident or compromise changes interpretation

10.6 Evidence Conflict

When credible evidence conflicts:

- Preserve both records
- examine protocol differences
- examine system identity
- examine elicitation
- examine task populations
- examine conflicts and access
- conduct replication or bridge study
- avoid forced averaging

10.7 Negative Evidence

Failure to demonstrate a capability is weaker than evidence that the capability is absent.

The distinction should remain visible.

10.8 Confidential Evidence

Confidential evidence may support a public conclusion only when:

- Qualified reviewers accessed it
- provenance is documented
- conflicts are controlled
- the public claim remains bounded
- limitations are stated
- security rules are followed

11. Shared Governance Architecture

11.1 Core Governance Functions

The foundations collectively require governance for:

- Protocol ownership
- task custody
- threshold decisions
- independent review
- evaluator qualification
- certification schemes

- standards development
- incentive programs
- interoperability and recognition
- complaints and appeals
- security incidents
- version control

11.2 Core Roles

Sponsor or Requesting Authority

Defines the decision need and funds or authorizes work.

Protocol Owner

Maintains the evaluation protocol.

Task or Evidence Custodian

Protects held-out material and chain of custody.

Evaluation Administrator

Executes the protocol.

Scorer or Judge

Applies scoring criteria.

Domain Expert

Provides subject-matter competence.

Independent Reviewer

Challenges evidence and conclusions.

Evaluator Organization

Operates quality-controlled evaluation activities.

Scheme Owner

Maintains certification or assurance scheme rules.

Accreditation Body

Recognizes conformity-assessment competence within scope.

Standards Development Body

Maintains the process for standards creation and revision.

Security Officer

Oversees information protection and incident response.

Appeals or Complaints Body

Reviews procedural or substantive challenges.

Registry Operator

Maintains current status records.

Public-Interest Participant

Contributes affected-party and public-consequence knowledge.

Legal or Regulatory Authority

Makes binding legal decisions where authorized.

11.3 Separation of Roles

Incompatible roles should not be combined without safeguards.

Examples:

- A task author should not unilaterally approve task validity.
- A developer should not be the sole evaluator of its high-stakes claim.
- An evaluator should not accredit itself.
- A scheme owner should not hide evaluator failures.
- An accreditation body should not depend financially on favorable recognition outcomes.
- An award sponsor should not select winners without conflict controls.
- A technical mapping body should not claim legal equivalence.

11.4 Decision Records

Material decisions should record:

- Decision question
- authority
- evidence
- reviewers
- conflicts
- dissent
- conditions

- effective date
- expiration
- appeal
- change triggers

11.5 Public and Protected Records

Governance should distinguish:

- Public rationale
- public status
- protected technical evidence
- confidential personal or commercial information
- restricted security information

11.6 Emergency Decisions

Emergency processes should be:

- Narrow
- time-limited
- documented
- reviewed afterward
- subject to correction
- prevented from becoming permanent by default

11.7 Governance Review

The governance system should be reviewed for:

- Capture
 - concentration
 - consistency
 - timeliness
 - participation
 - security
 - correction
 - burden
 - real-world effectiveness
-

12. Shared Independence and Competence Architecture

12.1 Independence Dimensions

The foundations recognize:

- Organizational independence
- financial independence
- methodological independence
- informational independence
- operational independence
- publication independence
- intellectual independence
- political independence
- security independence

12.2 Competence Dimensions

Competence should be scoped by:

- Activity
- domain
- method
- model or system type
- threat class
- assurance level
- security level
- jurisdiction where relevant

12.3 Access Dimensions

Access may include:

- Model endpoint
- weights
- system prompt
- safety policies
- tools
- logs
- training or evaluation records
- incidents
- internal thresholds
- personnel
- raw evidence

12.4 Independence Profile

Every consequential review should state:

- Who selected the reviewers
- Who paid them
- Whether future work depends on the reviewed party
- What evidence they could access
- Whether they selected methods
- Whether the reviewed party could delay or prevent publication
- What conflicts existed
- What limitations remained

12.5 Competence Evidence

Competence may be shown through:

- Education
- experience
- work samples
- proficiency tests
- peer review
- supervised practice
- domain credentials
- security qualification
- ongoing training

12.6 No Prestige Substitution

Institutional name, celebrity, government affiliation, or model access should not substitute for demonstrated scope-specific competence.

12.7 No Access Substitution

Access does not compensate for weak methods or conflicts.

12.8 Review of Reviewers

Reviewers and evaluators should be subject to:

- Complaints
- correction
- surveillance
- proficiency testing
- scope limitation
- suspension

- withdrawal
-

13. Shared Security and Transparency Architecture

13.1 Core Principle

Transparent governance and protected content should coexist.

13.2 Information Classes

The combined system should support at least:

- Public
- controlled
- confidential
- restricted
- highly restricted

13.3 Public Minimum

Even when technical content is protected, the public record should normally identify:

- Purpose
- construct
- broad methodology
- evaluated system
- evaluator role
- assurance level
- result summary
- uncertainty
- limitations
- status
- expiration
- governance
- correction history

13.4 Protected Components

Protection may apply to:

- Active tasks
- solutions
- scoring details
- attack libraries

- vulnerability information
- model weights
- system prompts
- personal data
- national-security information
- proprietary evidence

13.5 Protection Requirements

Protected information should have:

- Owner
- purpose
- access rule
- classification
- retention
- logging
- onward-disclosure rule
- incident response
- review date
- release or destruction path

13.6 Security Is Not Validity

A secure evaluation can still be invalid.

13.7 Transparency Is Not Validity

A fully open evaluation can still be invalid.

13.8 Compromise Response

A material compromise should trigger consideration of:

- Task suspension
- result status change
- notification
- investigation
- rotation
- re-evaluation
- public correction
- recognition impact

13.9 Responsible Disclosure

Disclosure processes should support:

- Safe reporting
 - acknowledgment
 - triage
 - remediation
 - publication timing
 - researcher protection
 - corrective learning
-

14. Shared Status and Version Architecture

14.1 Objects Requiring Status

Status should be maintained for:

- Protocols
- task banks
- evaluation results
- reviewer findings
- evaluator scopes
- certificates
- accreditation
- standards
- requirements
- incentive programs
- crosswalks
- recognition decisions
- registry records

14.2 Common Status Vocabulary

Use:

- Draft
- proposed
- approved
- active
- current
- expired
- suspended
- corrected
- superseded

- withdrawn
- deprecated
- retired
- archived

14.3 Change Triggers

Common triggers include:

- Model update
- system configuration change
- protocol revision
- task compromise
- incident
- threshold change
- evaluator personnel change
- loss of accreditation
- legal change
- new evidence
- security breach
- scheduled review

14.4 Materiality

A change is material when it could alter:

- Performance
- interpretation
- risk
- decision
- comparability
- assurance
- scope
- public claim

14.5 Version Inheritance

A new version should not automatically inherit status from a prior version.

14.6 Public Change Record

Material changes should identify:

- What changed
- why
- evidence

- compatibility
 - transition
 - affected records
 - effective date
-

15. Shared Decision Architecture

15.1 Decision Types

The foundations support decisions concerning:

- Protocol approval
- task protection
- high-stakes classification
- threshold escalation
- review finding
- evaluator recognition
- certification
- standards progression
- procurement eligibility
- disclosure
- incident response
- interoperability recognition

15.2 Decision Inputs

A consequential decision should consider:

- Claim
- evidence
- uncertainty
- consequence of error
- affected parties
- alternatives
- conflicts
- security
- burden
- implementation capacity
- appeal
- expiration

15.3 Decision Outputs

The output may be:

- Approve
- approve with conditions
- pilot
- defer
- request additional evidence
- limit scope
- suspend
- withdraw
- retire
- recognize
- conditionally recognize
- decline recognition

15.4 Thresholds as Process Triggers

A threshold should normally trigger a defined process.

It should not operate as an unexplained label.

15.5 False-Positive and False-Negative Costs

Both should be considered.

A false positive may:

- Restrict beneficial development
- create unnecessary cost
- misclassify capability
- distort public perception

A false negative may:

- permit unsafe access
- delay safeguards
- understate risk
- weaken preparedness

15.6 Reversibility

Where evidence is uncertain, prefer decisions that preserve correction and re-evaluation.

15.7 Binding Authority

Technical evidence may inform binding decisions.

Binding authority must come from the institution legally or contractually authorized to decide.

16. Shared Maturity Model

The eight foundation-specific maturity models can be integrated into six system levels.

Level 0: Fragmented and Informal

Characteristics:

- Static public benchmarks
- weak system identity
- no holdout governance
- developer-only interpretation
- informal external review
- no evaluator quality system
- ambiguous standards claims
- accidental incentives
- incompatible evidence

Primary risk:

False confidence from visible activity without institutional integrity.

Level 1: Defined and Documented

Characteristics:

- Protocol and version
- construct definition
- system identity
- basic protected-task controls
- documented high-stakes questions
- reviewer mandate
- evaluator role definition
- voluntary guidance
- explicit recognition criteria
- core terminology

Primary achievement:

The work becomes legible and reviewable.

Level 2: Validated and Independently Challenged

Characteristics:

- Protocol validation
- held-out integrity controls
- multidimensional capability evidence
- independent expert review
- conflict disclosure
- evidence-backed reporting
- pilot implementation
- early crosswalks

Primary achievement:

The evidence becomes more credible.

Level 3: Quality-Controlled and Scalable

Characteristics:

- Evaluator quality systems
- proficiency testing
- surveillance
- incident processes
- repeatable standards-development process
- procurement use
- public registries
- anti-gaming incentive programs
- conditional recognition

Primary achievement:

The work becomes institutionally repeatable.

Level 4: Formally Assured and Interoperable

Characteristics:

- Mature certification schemes where appropriate
- accreditation or equivalent competence recognition
- versioned standards
- cross-border evaluator recognition
- secure incident exchange
- multilingual validation
- result portability
- periodic impact review

Primary achievement:

Evidence and assurance can travel across organizations.

Level 5: Adaptive and Accountable Institutional Regime

Characteristics:

- Continuous evaluation
- dynamic standards maintenance
- integrated incident learning
- progressive legal or procurement requirements
- distributed international governance
- mature appeals and correction
- capacity inclusion
- retirement of obsolete systems
- evaluation of institutional outcomes

Primary achievement:

The entire regime learns and corrects itself.

Maturity Rule

No organization should claim a higher system maturity level because it possesses one advanced component.

Maturity depends on the weakest material layer relevant to the claim.

17. Consolidated Failure Register

The foundation library identifies recurring failure modes across technical, institutional, market, and international layers.

17.1 Measurement Failure

Static Benchmark Dependence

A fixed public benchmark becomes the primary evidence for changing frontier systems.

Controls:

- Dynamic protocol
- versioning
- held-out components
- new task forms

- result expiration

Construct Failure

The evaluation measures a proxy rather than the claimed capability, safeguard, or risk.

Controls:

- Construct definition
- domain review
- multi-method evidence
- criterion validation
- public limitations

Artificial Difficulty

Tasks become harder without becoming more valid or decision-relevant.

Controls:

- Task-universe definition
- content validity
- real-world linkage
- human baselines
- reviewer challenge

Score Collapse

A single aggregate score hides reliability, autonomy, cost, task family, or safeguard tradeoffs.

Controls:

- Multidimensional profiles
- decomposable scores
- uncertainty
- domain reporting
- threshold rationale

Historical Comparability Failure

Protocol versions are ranked together without demonstrated equivalence.

Controls:

- Anchor tasks
- bridge studies
- explicit discontinuity
- new baselines
- version-specific reporting

17.2 Integrity Failure

Training Contamination

Evaluation material or close derivatives enter training or development data.

Controls:

- Provenance
- held-out tasks
- contamination review
- rotation
- generated-at-test-time components

Leakage

Protected content is disclosed intentionally or accidentally.

Controls:

- Least privilege
- access logging
- custody
- security classification
- compromise response

Evaluator Preparation Leakage

The evaluated party receives enough information to optimize narrowly against the active task set.

Controls:

- Fair notice at construct level
- protected exact content
- alternate forms
- limited feedback
- exposure budgets

Hidden Compromise

A compromised task bank continues to support public claims.

Controls:

- Integrity status
- incident reporting
- result correction
- suspension
- re-evaluation

17.3 Elicitation Failure

Under-Elicitation

The system appears incapable because prompts, tools, time, retries, or integration are weak.

Controls:

- Elicitation matrix
- best-effort conditions
- developer input
- external elicitation
- documented budget

Overfitted Elicitation

Extensive task-specific optimization produces performance unavailable in practical use.

Controls:

- Separate capability targets
- standardized and practical conditions
- resource reporting
- generalization tests

Scaffold Confusion

Performance is attributed to the model when it depends materially on tools or orchestration.

Controls:

- System-level identity
- scaffold documentation
- component and end-to-end testing

17.4 High-Stakes Interpretation Failure

Capability-Risk Collapse

Capability is treated as sufficient proof of risk.

Controls:

- Risk decomposition
- propensity
- access
- exposure
- safeguards
- consequence
- uncertainty

Knowledge-Action Confusion

Correct answers are treated as proof of reliable operational performance.

Controls:

- End-to-end tasks
- environments
- autonomy
- reliability
- resource requirements
- human uplift

Threshold Theater

A threshold appears precise but lacks a valid construct or consequence model.

Controls:

- Threshold charter
- uncertainty
- trigger definition
- alert thresholds
- review
- change process

High-Stakes Inflation

Too many capabilities receive the highest evaluation burden.

Controls:

- Consequence criteria
- tiering
- proportionality
- review
- sunset

High-Stakes Minimization

Severe capability is treated as ordinary because evidence is uncertain.

Controls:

- Precautionary review
- alert thresholds
- interim safeguards
- continuous evaluation

17.5 Review Failure

Externality Without Independence

An outside organization is treated as independent despite client dependence or restricted judgment.

Controls:

- Independence profile
- funding disclosure
- publication rights
- conflict management

Prestige Without Competence

High-status reviewers are appointed outside their actual domain.

Controls:

- Scope-specific competence
- work samples
- proficiency
- diverse panels

Competence Without Access

Reviewers cannot inspect the evidence needed for the public claim.

Controls:

- Access disclosure
- mandate limitation
- confidential review mechanisms
- narrowed conclusion

Developer Veto

The reviewed party controls methods, reviewers, findings, timing, or publication.

Controls:

- Mandate
- independent selection
- factual correction only
- publication rights
- escalation

Consensus Theater

Material dissent is removed to create a unified public message.

Controls:

- Minority reports
- dissent records
- confidence bands
- separate findings

17.6 Evaluator-Ecosystem Failure

Scope Inflation

An evaluator claims broad competence from limited experience.

Controls:

- Granular scope
- public registry
- surveillance
- proficiency testing

Evaluator Shopping

A client seeks the most favorable evaluator or conclusion.

Controls:

- Engagement disclosure
- reason for replacement
- registry
- second review
- scheme rules

Client Capture

Repeat business weakens evaluator independence.

Controls:

- Client concentration limits
- fixed fees
- rotation
- pooled funding
- oversight

Certificate Inflation

Certificates proliferate without clear requirements or assurance.

Controls:

- Scheme governance
- claim control
- accreditation
- expiry
- surveillance

Accreditation Monopoly

One body controls evaluator legitimacy without accountability.

Controls:

- Peer review
- plural recognition
- appeals
- international comparison
- scope transparency

Auditor Bottleneck

Formal requirements activate before competent evaluators exist.

Controls:

- Capacity gate
- phase-in
- shared facilities
- training
- limited initial scope

17.7 Standards Failure

Premature Formalization

Immature methods become binding.

Controls:

- Readiness gates
- pilots
- independent evidence
- sunset

Permanent Voluntarism

Severe externalities persist because no minimum requirement emerges.

Controls:

- Nonadoption review
- incident triggers
- public authority
- progressive escalation

Prescriptive Lock-In

One method becomes mandatory and blocks superior alternatives.

Controls:

- Performance requirements
- equivalent methods
- experimental annexes
- review

Standards Capture

Dominant organizations shape requirements around their own systems.

Controls:

- Balanced participation
- conflict disclosure
- small-actor funding
- competition analysis
- minority reports

Regulatory Ratchet

Requirements strengthen but cannot be reduced after evidence changes.

Controls:

- Symmetric review
- sunset
- retirement
- impact analysis

Compliance Theater

Documents and checklists replace outcome evidence.

Controls:

- Operational testing
- incident evidence
- performance requirements
- independent review

17.8 Incentive Failure

Goodhart Effects

Participants optimize the metric rather than the intended objective.

Controls:

- Multiple measures
- held-out components
- counter-metrics
- outcome review
- metric rotation

Prestige Capture

Existing status determines awards, access, funding, and authority.

Controls:

- Evidence-based recognition
- open nominations
- distribution audit
- rotation
- scoped prestige

Corrective Credit Abuse

Organizations create or repeat preventable failures while seeking praise for correction.

Controls:

- Negligence analysis
- recurrence tracking
- remaining consequences
- independent verification

Bounty Distortion

Programs reward finding quantity, duplication, or unsafe disclosure.

Controls:

- Severity
- novelty
- triage
- safe harbor
- responsible disclosure

Crowding Out

External rewards weaken intrinsic motivation, cooperation, or public service.

Controls:

- Fair compensation
- autonomy
- mission
- community review
- nonfinancial recognition

17.9 Interoperability Failure

False Equivalence

Different protocols or certificates are accepted as interchangeable without evidence.

Controls:

- Purpose-bounded mapping
- bridge studies
- local modules
- conditional recognition

Uniformity Disguised as Interoperability

One institution's framework becomes the global default.

Controls:

- Common minimum
- plural governance
- local extensions
- multiple trust anchors

Lowest Common Denominator

International alignment weakens high-stakes requirements.

Controls:

- Baseline plus advanced profiles
- local safeguards
- conditional recognition

Registry Monopoly

One registry controls global legitimacy.

Controls:

- Open schemas
- federated records
- signatures
- mirrored status

Capacity Exclusion

Institutions are expected to implement systems they lack the resources to use.

Controls:

- Training
- regional facilities
- grants
- translation
- shared infrastructure
- local governance

Translation Error

Literal translation changes construct meaning.

Controls:

- Domain review
- cultural adaptation
- bridge studies
- local human baselines

17.10 Meta-Institutional Failure

Authority Inflation

A research project appears to exercise certification, regulatory, or official standards authority.

Controls:

- PROJECT_IDENTITY.md
- authority disclaimers
- status labels
- public-claims review

Institutional Theater

The appearance of maturity outruns competence, governance, and evidence.

Controls:

- Stage boundaries
- pilots
- independent critique
- transparent limitations

Founder or Leadership Capture

One person becomes the unchallengeable source of institutional identity and decisions.

Controls:

- Governance
- decision records
- contributors
- appeals
- succession

Mission Drift

The project expands into unrelated commentary, promotion, or consulting.

Controls:

- Mission test
- canonical architecture
- portfolio review

18. Cross-Foundation Tensions

The foundations contain deliberate tensions that should be managed rather than erased.

18.1 Transparency Versus Security

Foundation 2 protects active evaluation material.

Foundations 4, 6, and 8 require accountability and public legibility.

Resolution:

- Publish purpose, construct, governance, evaluator role, uncertainty, and status.
- Protect exact material only when disclosure would reduce validity or increase harm.
- Use qualified confidential review.
- Review restrictions periodically.

18.2 Stability Versus Adaptation

Foundation 1 requires change.

Measurement and interoperability require comparability.

Resolution:

- Maintain a stable construct and metadata core.
- Change only what evidence requires.
- use anchors and bridge studies.
- declare discontinuity when continuity cannot be defended.

18.3 Developer Cooperation Versus Independent Judgment

Developers possess essential system knowledge.

Reviewers must remain free to disagree.

Resolution:

- Permit technical input and factual correction.
- Preserve method selection, interpretation, dissent, and publication independence.

18.4 Speed Versus Legitimacy

Frontier systems change quickly.

Standards and public institutions require review and participation.

Resolution:

- Use staged outputs.
- issue research protocols and temporary guidance before mature standards.
- use emergency processes with expiration.
- avoid presenting speed as authority.

18.5 Precaution Versus False Positives

Severe risks can justify early action.

Weak evidence can restrict beneficial work.

Resolution:

- Use alert thresholds before critical restrictions.
- choose reversible interim safeguards.
- report uncertainty.
- review both false-positive and false-negative costs.

18.6 Standardization Versus Innovation

Standards improve consistency.

Prescriptive rules can freeze immature methods.

Resolution:

- Prefer performance-based requirements where outcomes can be verified.
- permit equivalent methods.
- use experimental annexes.
- revise and retire.

18.7 Market Scale Versus Independence

Commercial evaluator markets can increase capacity.

Client funding can weaken judgment.

Resolution:

- fixed or pooled payment
- no outcome-dependent fees
- client concentration monitoring
- surveillance
- public scopes
- complaints

18.8 International Alignment Versus Sovereignty

Shared evidence improves coordination.

Legal and cultural decisions remain local.

Resolution:

- recognize evidence separately from legal effect.
- use common cores and local modules.
- preserve national and institutional decision authority.

18.9 Open Participation Versus Security

Open communities improve scrutiny and access.

Some evidence is sensitive.

Resolution:

- create tiered participation.

- use secure access pathways.
- recognize open tools and community expertise.
- avoid equating openness with unrestricted disclosure.

18.10 Prestige Versus Accountability

Recognition can motivate contribution.

Prestige can shield error.

Resolution:

- scope recognition.
 - require evidence.
 - allow correction and withdrawal.
 - separate awards from authority.
-

19. Consolidated Evidence Gaps

The following evidence gaps recur across the foundation series.

19.1 Protocol Validity Over Time

How quickly do evaluation protocols lose predictive or decision value?

19.2 Comparability Across Versions

Which bridge-study methods work for dynamic, agentic, and held-out evaluations?

19.3 Contamination Detection

How reliably can direct, indirect, synthetic, and derivative exposure be detected?

19.4 Elicitation Sufficiency

How should evaluators estimate plausible maximum capability without unlimited resources?

19.5 Agent Evaluation

How should long-horizon reliability, environmental change, and trajectory quality be measured?

19.6 High-Stakes Predictive Validity

Which evaluation results predict real-world capability, misuse, incidents, or safeguard failure?

19.7 Threshold Governance

How should uncertain capability evidence connect to alert, critical, or policy thresholds?

19.8 Independent Reviewer Access

Which access arrangements support credible review while protecting security and intellectual property?

19.9 Reviewer Independence

How do funding, repeat clients, model access, and publication rights affect conclusions?

19.10 Evaluator Proficiency

Which proficiency-testing systems predict actual evaluation quality?

19.11 Accreditation Adaptation

How should existing accreditation concepts adapt to fast-changing AI methods?

19.12 Certification Validity

Which claims are mature enough for certification, and how do users interpret certificates?

19.13 Standards Impact

Do AI standards improve system behavior, or mainly organizational documentation?

19.14 Procurement Effects

Do evaluation and assurance requirements improve safety without entrenching incumbents?

19.15 Incentive Effects

Which rewards increase disclosure, maintenance, replication, and correction?

19.16 Crowding Out

When do external rewards reduce intrinsic safety and public-interest motivation?

19.17 Incident Reporting

Which classification and protection systems improve cross-organizational learning?

19.18 International Recognition

Which forms of evaluator and result recognition preserve rigor across jurisdictions?

19.19 Translation Validity

How can capability and risk constructs remain valid across languages and cultures?

19.20 Capacity Building

Which models create durable local evaluation institutions rather than dependency?

19.21 Registry Governance

How can status records remain current, distributed, secure, and legitimate?

19.22 Open-Weight Systems

How should evidence, responsibility, and assurance travel across model forks and decentralized deployments?

19.23 Institutional Outcomes

How should the effectiveness of the full evaluation and standards regime be measured?

20. Consolidated Research Agenda

Standards Body should organize cross-foundation research into the following programs.

Program 1: Dynamic Measurement

Develop:

- Protocol versioning
- anchor methods
- bridge studies
- result expiration
- dynamic task quality
- protocol-performance metrics

Program 2: Evaluation Integrity

Develop:

- Contamination detection
- task provenance
- custody systems
- generated-at-test-time methods
- compromise response
- exposure budgets

Program 3: Capability Elicitation

Develop:

- Elicitation budgets
- tool and scaffold standards
- external elicitation
- capability ceiling estimates
- practical capability profiles
- agent evaluation

Program 4: High-Stakes Evidence

Develop:

- Capability graphs
- threshold governance
- false-positive and false-negative analysis
- human uplift
- safeguard integration
- decision-linked evidence cases

Program 5: Independent Review

Develop:

- Review mandates
- access models
- independence metrics
- dissent procedures
- reviewer accountability
- secure publication

Program 6: Evaluator Quality

Develop:

- Scope taxonomies
- competency frameworks
- proficiency testing
- evaluator security
- quality-management requirements
- surveillance

Program 7: Assurance and Accreditation

Develop:

- Testing and audit distinctions
- certification readiness
- scheme governance
- accreditation pathways
- registry models
- mutual recognition

Program 8: Progressive Standards

Develop:

- Readiness gates
- voluntary framework evaluation
- procurement pilots
- legal recognition models
- safe harbors
- sunset and retirement

Program 9: Incentive Alignment

Develop:

- Public-goods funding
- corrective credit
- bounty programs
- maintenance grants
- recognition systems
- anti-Goodhart mechanisms

Program 10: Incident Learning

Develop:

- Incident taxonomy
- near-miss reporting
- root-cause analysis
- restricted exchange
- protocol-update mechanisms
- public summaries

Program 11: Global Interoperability

Develop:

- Terminology mappings
- common metadata
- model identity
- protocol crosswalks
- recognition levels
- federated registries

Program 12: Multilingual and Regional Evaluation

Develop:

- Translation validation
- cultural adaptation
- regional task development
- local human baselines
- capacity programs
- regional evaluator networks

Program 13: Open Ecosystems

Develop:

- Model lineage
- community evaluation
- decentralized incident reporting
- maintainership funding
- open-tool assurance
- functional compliance pathways

Program 14: Institutional Evaluation

Develop methods for evaluating:

- Standards bodies
- evaluator markets
- accreditation bodies
- incentive programs
- recognition networks
- Standards Body itself

Program 15: Public Claims and Comprehension

Study how developers, purchasers, governments, journalists, and the public interpret:

- Scores
- thresholds
- review findings
- certificates
- accreditation
- uncertainty
- status labels
- safety claims

21. Consolidated Role Map

Role	Primary responsibilities	Primary foundation
Decision owner	Defines the consequential decision and evidence need	3 and 6
Protocol owner	Maintains construct, method, version, and retirement	1
Task author	Creates valid task material	1 and 2
Task custodian	Protects held-out content and custody	2
Evaluation administrator	Executes the protocol under controlled conditions	1 and 2
Elicitation specialist	Configures prompts, tools, scaffolds, and resources	1 and 3
Scorer or judge	Applies scoring and records uncertainty	1
Domain expert	Assesses content and real-world validity	3 and 4
Security specialist	Protects models, tasks, evidence, and exchange	2, 4, 5, and 8
Independent reviewer	Challenges evidence and conclusions	4
Review chair	Preserves mandate, process, dissent, and records	4
Evaluator organization	Operates quality-controlled evaluation services	5

Role	Primary responsibilities	Primary foundation
Auditor	Conducts criteria-based audits within scope	5
Certification body	Issues certification under a defined scheme	5
Accreditation body	Recognizes conformity-assessment competence	5
Scheme owner	Maintains assurance or certification rules	5 and 6
Standards committee	Develops and maintains standards	6
Purchaser	Uses evidence and requirements in procurement	6 and 7
Insurer	Uses evidence in underwriting and loss prevention	7
Incentive-program owner	Designs grants, prizes, recognition, or bounties	7
Registry operator	Maintains current public or controlled records	5 and 8
Recognition authority	Accepts evidence, competence, or process for a purpose	8
Legal authority	Makes binding legal decisions within jurisdiction	6 and 8
Public-interest participant	Contributes affected-party and social-context knowledge	4, 6, 7, and 8
Contributor	Provides documented research, review, infrastructure, or implementation work	All

22. Consolidated Metrics

The combined system should be evaluated across seven dimensions.

22.1 Measurement Quality

- Construct validity
- task quality
- reliability
- discrimination
- uncertainty
- comparability
- real-world relevance

22.2 Integrity

- Contamination rate
- leakage incidents
- provenance completeness
- access anomalies
- compromise response
- task rotation

22.3 Independence and Competence

- Reviewer conflicts
- access sufficiency
- proficiency results
- correction rate
- client concentration
- dissent handling

22.4 Operational Performance

- Cost
- time
- evaluator availability
- repeatability
- evidence completeness
- incident response
- re-evaluation latency

22.5 Institutional Quality

- Governance
- participation
- appeals
- transparency
- security
- capture risk
- competition
- small-actor access

22.6 Decision Utility

- Decisions improved
- safeguards changed
- incidents prevented
- procurement value
- legal or policy relevance
- reduced uncertainty
- reduced duplication

22.7 Adaptation and Interoperability

- Revision speed
- obsolete-method retirement
- bridge-study quality

- cross-border recognition
 - translation validity
 - registry freshness
 - local capacity
-

23. Consolidated Template Index

The eight foundations contain an integrated operational library.

Foundation 1 Templates

Protocol Template

Use to define a complete evaluation protocol.

Change Request Template

Use to propose and review a protocol change.

Protocol Scorecard

Use to evaluate protocol quality, integrity, governance, and decision value.

Foundation 2 Templates

Held-Out Evaluation Plan Template

Use to define protected content, access, custody, administration, rotation, and disclosure.

Access Request Template

Use to authorize and record access to protected material.

Compromise Incident Template

Use after suspected or confirmed leakage, contamination, or security failure.

Holdout Scorecard

Use to evaluate validity, protection, fairness, governance, and lifecycle.

Foundation 3 Templates

High-Stakes Evaluation Plan Template

Use to define domain, capability, consequence, protocol, experts, safeguards, and decisions.

Capability Evidence Case Template

Use to assemble multidimensional evidence for a high-stakes capability claim.

Threshold Change Request Template

Use to propose a new or revised alert or critical threshold.

High-Stakes Evaluation Scorecard

Use to evaluate consequence linkage, capability evidence, uncertainty, safeguards, and governance.

Foundation 4 Templates

Independent Review Mandate Template

Use to define review question, authority, access, methods, outputs, and publication.

Independence Profile Template

Use to disclose organizational, financial, methodological, and publication independence.

Reviewer Conflict Disclosure Template

Use to identify and manage reviewer conflicts.

Review Finding Template

Use to present evidence, findings, confidence, limitations, and required action.

Minority Report Template

Use to preserve material dissent.

Independent Review Scorecard

Use to evaluate competence, independence, access, security, accountability, and utility.

Foundation 5 Templates

Evaluator Organizational Profile Template

Use to document evaluator governance, personnel, methods, quality, security, and client structure.

Accreditation Scope Template

Use to define the exact activities and limits proposed for competence recognition.

Evaluator Engagement Template

Use to govern client, scope, payment, evidence, access, reporting, and conflicts.

Surveillance Report Template

Use for ongoing review of evaluator status and performance.

Evaluator Complaint Template

Use to submit and resolve complaints.

Ecosystem Scorecard

Use to evaluate evaluator-market competence, impartiality, capacity, recognition, and concentration.

Foundation 6 Templates

Requirement Maturity Assessment Template

Use to determine whether a practice is ready for a more formal stage.

Transition-Gate Decision Template

Use to approve, defer, condition, or reject institutional progression.

Progressive Requirement Specification Template

Use to define scope, trigger, evidence, assurance, implementation, enforcement, and sunset.

Procurement Clause Template

Use to incorporate evaluation and assurance into purchasing.

Requirement Impact Review Template

Use to assess effectiveness, burden, competition, and unintended effects.

Progressive Standards Scorecard

Use to evaluate readiness and legitimacy of a standard or requirement.

Foundation 7 Templates

Incentive Program Design Template

Use to define actor, behavior, mechanism, reward, gaming, distribution, and evaluation.

Contribution Recognition Template

Use to recognize evidence, infrastructure, safety, institutional, or public-interest contribution.

Responsible Disclosure Reward Template

Use to handle bounty or disclosure recognition.

Corrective Credit Template

Use to evaluate whether an organization should receive credit for disclosure and remediation.

Recognition and Prestige Scorecard

Use to test whether status and rewards track genuine contribution.

Foundation 8 Templates

Interoperability Profile Template

Use to describe supported terminology, protocols, metadata, security, localization, and recognition.

Protocol Crosswalk Template

Use to compare two protocols and determine compatibility.

Recognition Decision Template

Use to accept, conditionally accept, or decline external evidence or competence.

International Incident Exchange Template

Use to share incident evidence across institutions under controlled conditions.

Translation and Localization Validation Template

Use to validate language and regional adaptations.

Registry Record Template

Use to publish current status, scope, evidence, and recognition.

Global Interoperability Scorecard

Use to assess semantic, protocol, measurement, assurance, legal, security, and capacity interoperability.

24. Scorecard Coordination

The eight scorecards should be used in sequence.

24.1 Protocol Scorecard

Question:

Is the evaluation method valid, current, versioned, and decision-relevant?

24.2 Holdout Scorecard

Question:

Is protected evidence governed, secure, proportionate, fair, and maintainable?

24.3 High-Stakes Evaluation Scorecard

Question:

Does the evidence support the high-consequence capability interpretation and decision?

24.4 Independent Review Scorecard

Question:

Was the evidence challenged by competent reviewers with sufficient independence and access?

24.5 Ecosystem Scorecard

Question:

Can third-party assurance be delivered consistently and accountably at scale?

24.6 Progressive Standards Scorecard

Question:

Is the practice mature and legitimate enough for a stronger institutional stage?

24.7 Recognition and Prestige Scorecard

Question:

Do the incentives reward the underlying objective rather than superficial measures?

24.8 Global Interoperability Scorecard

Question:

Can the evidence and assurance be understood and used across boundaries?

24.9 No Averaged Master Score

Standards Body should not combine all eight scorecards into one universal numerical rating.

A failed critical dimension should remain visible rather than being offset by unrelated strengths.

25. Unified Implementation Pathway

The foundations should be implemented as a staged program rather than eight disconnected projects.

Phase 1: Canonical Control Layer

Complete and approve:

- PROJECT_IDENTITY.md
- TERMINOLOGY.md
- FOUNDATIONS_APPENDIX.md
- EVIDENCE_STANDARDS.md
- RESEARCH_METHODODOLOGY.md
- TAXONOMY.md
- EVALUATION_PHILOSOPHY.md

Purpose:

- Establish identity
- stabilize language
- define evidence
- prevent contradiction
- create research discipline
- establish evaluation principles

Exit criteria:

- Canonical source hierarchy exists
- material terms are defined
- claims can be classified
- revision and status rules are clear
- all eight foundations are cross-referenced

Phase 2: Institutional Design Layer

Complete and approve:

- INSTITUTION_DESIGN.md
- GOVERNANCE_FRAMEWORK.md

- STANDARDS_DEVELOPMENT_PROCESS.md
- EVALUATOR_ACCREDITATION_FRAMEWORK.md
- CONTRIBUTOR_FRAMEWORK.md
- TRANSPARENCY_FRAMEWORK.md
- PARTNERSHIP_PRINCIPLES.md
- LONG_TERM_ROADMAP.md

Purpose:

- Define roles
- establish decision rights
- control conflicts
- create contribution pathways
- distinguish research from authority
- prepare legitimate future standards work

Exit criteria:

- Governance roles and decisions are mapped
- standards progression has a documented process
- conflicts and recusals are governed
- contributor credit and conduct are defined
- transparency and security levels are defined
- partnerships cannot imply unsupported endorsement

Phase 3: Living Knowledge Infrastructure

Complete and maintain:

- WEBSITE_SOURCE_OF_TRUTH.md
- SOURCES.md
- PERSON_PROFILES.md
- TIMELINE.md
- CASE_STUDIES.md
- FAILURE_DATABASE.md
- EARLY_SIGNALS.md
- FUTURE_WORK.md
- VERSION_HISTORY.md

Purpose:

- Preserve source quality
- monitor change
- learn from real cases
- track institutional and technical failures
- keep public materials consistent
- maintain current versions

Exit criteria:

- Every public claim has an identifiable source
- every canonical file has current status
- major cases and incidents are recorded
- website copy matches approved identity
- superseded work remains discoverable

Phase 4: Pilot Protocol

Select one bounded domain.

Recommended characteristics:

- High enough consequence to test rigor
- Measurable tasks
- available domain experts
- controllable environment
- meaningful human baselines
- realistic security
- multiple possible evaluators
- international relevance

The foundations repeatedly identify autonomous cyber capability as a strong candidate.

Build:

- Dynamic protocol
- held-out task bank
- high-stakes evidence case
- independent review
- evaluator qualification pilot
- requirement-maturity assessment
- incentive program
- interoperability profile

Exit criteria:

- All eight scorecards have been used
- critical findings are documented
- a public summary and protected evidence package exist
- no unsupported certification or approval claim is made

Phase 5: Multi-Evaluator Replication

Add independent evaluator organizations.

Test:

- Protocol portability
- task custody
- scoring consistency
- proficiency
- review disagreement
- secure access
- cost
- burden
- correction

Exit criteria:

- Evaluator scopes are defined
- major variance is explained
- complaints and correction work
- no evaluator is treated as universally competent

Phase 6: Voluntary Framework

Convert pilot learning into a voluntary framework.

Include:

- Minimum protocol requirements
- reporting
- independent review triggers
- evaluator criteria
- result status
- incident reporting
- interoperability metadata

Exit criteria:

- Multiple organizations can implement it
- burden is understood
- small-actor pathway exists
- public claims are bounded
- revision governance exists

Phase 7: Standards Development

Use STANDARDS_DEVELOPMENT_PROCESS.md.

Develop:

- Technical specification

- public review
- pilot implementation
- dissent record
- version governance
- retirement process

Exit criteria:

- Construct and methods are sufficiently mature
- implementation evidence exists
- affected actors participated
- assurance readiness is tested
- competition effects are reviewed

Phase 8: Assurance Scheme Pilot

Where justified, define:

- Scheme owner
- certification or assurance object
- requirements
- evaluator competence
- evidence
- surveillance
- complaints
- claims
- expiry

Exit criteria:

- The scheme does not imply universal safety
- sufficient evaluators exist
- accreditation or competence-recognition path is credible
- certificate interpretation has been tested
- withdrawal can propagate publicly

Phase 9: Procurement and Market Use

Pilot:

- Procurement clauses
- insurer evidence
- investor due diligence
- platform access conditions
- professional expectations

Measure:

- Adoption
- risk reduction
- cost
- concentration
- gaming
- innovation
- small-actor access

Phase 10: International Interoperability

Develop:

- Shared metadata
- protocol crosswalks
- evaluator profiles
- recognition decisions
- multilingual adaptations
- incident exchange

Exit criteria:

- Noncomparability can be stated
- local authority is preserved
- sensitive evidence remains protected
- underrepresented regions participate meaningfully

Phase 11: Formality Review

Determine whether the practice should:

- Remain research
- remain voluntary
- become a consensus standard
- become a procurement expectation
- receive legal recognition
- become a mandatory requirement
- be narrowed or retired

This decision belongs to the authorized institution, not automatically to Standards Body.

26. Integrated Standards Body Pilot

26.1 Pilot Name

Frontier Evaluation Integrity and Assurance Pilot

26.2 Purpose

Demonstrate the complete eight-foundation architecture in one bounded evaluation domain.

26.3 Candidate Domain

Autonomous cyber capability within controlled environments.

26.4 Pilot Claim

The pilot should not ask:

Is this AI system safe?

It should ask a bounded question such as:

Under the identified system configuration and protocol, what level of autonomous cyber task capability was demonstrated, how reliable was that performance, which safeguards affected practical access, and how much confidence should a defined decision-maker place in the evidence?

26.5 Foundation 1 Workstream

Create:

- Protocol charter
- task universe
- stable construct
- dynamic task layer
- elicitation conditions
- environment versions
- scoring
- uncertainty
- expiration

26.6 Foundation 2 Workstream

Create:

- Held-out task bank

- provenance records
- access matrix
- custody system
- compromise process
- rotation schedule
- delayed-release rules

26.7 Foundation 3 Workstream

Create:

- Capability graph
- autonomy and reliability dimensions
- human baseline
- human-uplift study
- safeguards profile
- alert thresholds
- evidence case
- decision matrix

26.8 Foundation 4 Workstream

Create:

- Review mandate
- reviewer panel
- independence profiles
- access agreement
- conflict register
- review findings
- minority report if needed
- public limitations

26.9 Foundation 5 Workstream

Create:

- Evaluator organizational profiles
- competence scopes
- quality requirements
- proficiency exercise
- surveillance plan
- complaint channel
- pilot registry

26.10 Foundation 6 Workstream

Create:

- Maturity assessment
- voluntary framework
- transition-gate analysis
- sample procurement clause
- impact review
- sunset plan

26.11 Foundation 7 Workstream

Create:

- Maintainer funding
- evaluator payment rules
- responsible-disclosure bounty
- contribution recognition
- corrective-credit policy
- anti-gaming review
- distribution report

26.12 Foundation 8 Workstream

Create:

- Common metadata
- system manifest
- protocol profile
- evaluator profile
- crosswalk with a second protocol
- international incident simulation
- recognition exercise
- registry record

26.13 Required Participants

The pilot should include:

- Protocol designers
- cyber domain experts
- model evaluation specialists
- security personnel
- independent reviewers
- at least two evaluator organizations

- open-source or open-weight expertise
- public-interest participant
- international or regional participant
- standards and assurance expertise

26.14 Public Outputs

Publish:

- Pilot purpose
- protocol summary
- system identity summary
- evaluator identities and scopes
- review mandate
- result profile
- uncertainty
- limitations
- integrity status
- decision implications
- correction history
- public lessons

26.15 Protected Outputs

Protect as necessary:

- Active tasks
- solutions
- attack methods
- vulnerable configurations
- sensitive model information
- personal data
- restricted incident evidence

26.16 Success Criteria

The pilot succeeds if:

- Evidence remains valid after adversarial review
- held-out integrity is demonstrated
- high-stakes interpretation is multidimensional
- reviewers have sufficient access
- evaluators produce materially consistent results or explain differences
- no broad safety claim is made
- incentive gaming is identified and corrected
- another institution can interpret the result

- status and expiration are clear
- the process produces reusable standards infrastructure

26.17 Failure Is Informative

The pilot may still succeed institutionally if it reveals that:

- The construct is immature
- tasks are invalid
- evaluator capacity is insufficient
- recognition is premature
- burden is excessive
- the standard should not progress

A credible decision not to formalize is a valid output.

27. Source-of-Truth and Document Hierarchy

The foundation library should operate within the following hierarchy.

27.1 Identity and Mission

1. PROJECT_IDENTITY.md
2. PROJECT_MANIFESTO.md
3. BRAND_POSITIONING.md

27.2 Foundational System

1. FOUNDATIONS.md
2. FOUNDATIONS_APPENDIX.md
3. The eight foundation papers

27.3 Terminology and Evidence

1. TERMINOLOGY.md
2. EVIDENCE_STANDARDS.md
3. RESEARCH_METHODODOLOGY.md
4. TAXONOMY.md
5. EVALUATION_PHILOSOPHY.md

27.4 Institutional Process

1. INSTITUTION_DESIGN.md
2. GOVERNANCE_FRAMEWORK.md

3. STANDARDS_DEVELOPMENT_PROCESS.md
4. EVALUATOR_ACCREDITATION_FRAMEWORK.md
5. TRANSPARENCY_FRAMEWORK.md
6. CONTRIBUTOR_FRAMEWORK.md
7. PARTNERSHIP_PRINCIPLES.md

27.5 Public and Living Records

1. WEBSITE_SOURCE_OF_TRUTH.md
2. SOURCES.md
3. VERSION_HISTORY.md
4. Other living knowledge-base files

27.6 Conflict Rule

When two files conflict:

- The specialized approved canonical source governs within its domain.
 - A specialized source may not silently change project identity or authority.
 - Terminology changes require update to TERMINOLOGY.md.
 - Foundation changes require review of this appendix.
 - Public copy changes require consistency with PROJECT_IDENTITY.md.
-

28. Cross-Foundation Change Control

28.1 Material Foundation Change

A material change includes:

- Core thesis change
- definition change
- new institutional role
- changed threshold philosophy
- new assurance claim
- changed security rule
- changed standards progression
- changed recognition model
- changed authority boundary

28.2 Impact Review

A material change to one foundation should trigger review of all connected foundations.

Foundation 1 Change

Review:

- Foundation 2 task protection
- Foundation 3 thresholds
- Foundation 4 review scope
- Foundation 5 evaluator methods
- Foundation 8 comparability

Foundation 2 Change

Review:

- Foundation 1 integrity
- Foundation 4 access
- Foundation 5 security
- Foundation 8 evidence exchange

Foundation 3 Change

Review:

- Foundation 4 review depth
- Foundation 5 competence scope
- Foundation 6 trigger logic
- Foundation 8 threshold mapping

Foundation 4 Change

Review:

- Foundation 5 evaluator governance
- Foundation 6 independent-validation gate
- Foundation 7 reviewer incentives
- Foundation 8 recognition

Foundation 5 Change

Review:

- Foundation 6 assurance readiness
- Foundation 7 evaluator market incentives
- Foundation 8 mutual recognition

Foundation 6 Change

Review:

- Foundation 5 schemes
- Foundation 7 adoption incentives

- Foundation 8 legal and standards crosswalks

Foundation 7 Change

Review all foundations because incentives affect every component.

Foundation 8 Change

Review:

- Terminology
- protocol metadata
- evaluator scopes
- recognition and status records
- international claims

28.3 Required Change Record

Record:

- Foundation
- version
- proposed change
- reason
- evidence
- affected files
- compatibility
- public effect
- implementation
- transition
- approval
- effective date

28.4 Emergency Change

Emergency corrections may be issued for:

- Security compromise
- materially invalid evidence
- harmful public claim
- incorrect authority statement
- legal or institutional status error

Full review should follow.

29. How to Use the Foundation Library

29.1 Designing an Evaluation

Begin with:

1. Foundation 1
2. Foundation 2
3. Foundation 3

Then use Foundation 4 if the claim is consequential.

29.2 Commissioning Independent Review

Begin with:

1. Foundation 4
2. Foundation 3
3. Foundation 1
4. Foundation 2

Use Foundation 5 for repeated or formal third-party work.

29.3 Building an Evaluator Organization

Begin with:

1. Foundation 5
2. Foundation 4
3. Foundation 1
4. Foundation 2
5. Foundation 3

29.4 Developing a Standard

Begin with:

1. Foundation 6
2. Foundation 1
3. Foundation 3
4. Foundation 4
5. Foundation 5

Use Foundation 7 for adoption and gaming.

Use Foundation 8 for international use.

29.5 Designing a Certification Scheme

Begin with:

1. Foundation 5
2. Foundation 6
3. Foundation 4
4. Foundation 1
5. Foundation 3

Do not proceed without defined requirements and claim boundaries.

29.6 Creating a Reporting Framework

Use:

- Foundation 3 for evidence
- Foundation 4 for review
- Foundation 6 for institutional stage
- Foundation 7 for incentive effects
- Foundation 8 for metadata and portability

29.7 Designing an International Partnership

Use:

- Foundation 8
- Foundation 4
- Foundation 5
- Foundation 2
- Foundation 7

29.8 Responding to an Incident

Use:

- Foundation 2 for integrity and custody
 - Foundation 3 for capability implications
 - Foundation 4 for independent review
 - Foundation 5 for evaluator consequences
 - Foundation 6 for requirement changes
 - Foundation 7 for disclosure and corrective credit
 - Foundation 8 for cross-border exchange
-

30. Integrated Decision Rules

30.1 Protocol Rule

Do not rely on a benchmark alone when:

- The decision is consequential
- tasks are public or saturated
- system configuration matters
- threats or deployment conditions change
- historical comparability is uncertain

30.2 Holdout Rule

Protect evaluation content only when:

- Exposure would materially reduce validity or increase harm
- protection is proportionate
- governance and review remain transparent
- access and duration are defined

30.3 High-Stakes Rule

Increase evaluation rigor when:

- Consequence is severe
- uncertainty is high
- access is broad
- autonomy is meaningful
- uplift is substantial
- safeguards are unproven
- decisions are difficult to reverse

30.4 Independent Review Rule

Require independent review when evidence influences:

- High-stakes deployment
- critical capability classification
- public safety claims
- standards thresholds
- certification or procurement
- binding requirements

30.5 Evaluator-Ecosystem Rule

Do not create a formal assurance scheme until:

- Requirements are defined
- methods are assessable
- evaluators can demonstrate competence
- surveillance and complaints exist
- claim boundaries are clear

30.6 Progression Rule

Do not move a practice to a stronger institutional stage merely because:

- Concern is high
- one company already complies
- a standards body seeks relevance
- a certificate market seeks demand
- another jurisdiction adopted a similar rule

30.7 Incentive Rule

Do not attach a high-value reward to a metric without:

- Gaming analysis
- counter-metrics
- verification
- correction
- distribution analysis
- sunset

30.8 Interoperability Rule

Do not declare two protocols, evaluators, certificates, or legal requirements equivalent without:

- Defined purpose
- scope
- evidence
- mapping
- conditions
- version
- review

30.9 Public-Claim Rule

A public claim should state:

- What
- which system
- under which conditions
- by whom
- with what evidence
- with what uncertainty
- for how long
- within which scope

30.10 Retirement Rule

Retire or suspend a protocol, result, evaluator scope, certificate, standard, incentive program, or recognition when its authority exceeds its current evidence.

31. Canonical Cross-Foundation Positions

Standards Body adopts the following consolidated positions.

1. Frontier AI evaluation should be governed at the protocol level, not the dataset level alone.
2. Evaluation protocols should be dynamic, versioned, traceable, and retireable.
3. Some evaluation material should remain held out when exposure would reduce validity or increase harm.
4. Protected content does not justify opaque governance.
5. Capability and risk are distinct.
6. High-stakes evaluation should be multidimensional.
7. Knowledge tests are not sufficient proof of operational capability.
8. The deployed system should be evaluated, not only the base model.
9. Elicitation conditions are part of the result.
10. Failure to demonstrate a capability is not proof of inability.
11. Passing an evaluation is not proof of safety.
12. Results should identify the exact system, protocol, evaluator, date, and limitations.
13. Results should expire after material change or defined time.
14. Developer evaluation is necessary but insufficient for the most consequential claims.

15. External review is not automatically independent.
16. Independence, competence, and access are jointly necessary.
17. Reviewed parties may correct facts but should not control conclusions.
18. Material dissent should remain visible.
19. Testing, evaluation, inspection, audit, certification, and accreditation are distinct.
20. Accreditation recognizes competence within scope, not universal reliability.
21. Certification applies to defined requirements, systems, scopes, and periods.
22. Broad safe-AI certification claims should be avoided.
23. Evaluator competence should be scope-specific.
24. Evaluators should face proficiency testing, surveillance, complaints, and correction.
25. Result-dependent evaluator compensation should be prohibited.
26. Standards should mature through evidence-based stages.
27. Not every voluntary practice should become mandatory.
28. Severe externalities and persistent nonadoption may justify stronger requirements.
29. Standards should normally be common floors rather than ceilings.
30. Requirements should include review, appeal, correction, and retirement.
31. Formality should not outrun evaluator and implementation capacity.
32. Small-actor and open-community participation is part of legitimacy.
33. Incentive design is part of technical and institutional design.
34. Public goods such as maintenance, replication, and incident databases require deliberate support.
35. Recognition should follow evidence and contribution rather than existing prestige.
36. Responsible disclosure and correction should receive protection and credit.
37. Corrective credit should not excuse negligence or concealment.
38. Rankings should not be used when construct validity and comparability are weak.
39. Interoperability should preserve legitimate difference.
40. Shared evidence structure should generally precede shared policy outcomes.
41. Numeric scores should not travel without measurement context.

42. Noncomparability is a legitimate conclusion.
 43. Recognition of evidence, competence, process, and legal effect should remain distinct.
 44. International cooperation should preserve local and national decision authority.
 45. Capacity building is part of interoperability.
 46. English-language evaluation should not define global capability by default.
 47. Open-weight model lineage and configuration should be recorded precisely.
 48. Registries should be current, portable, signed where appropriate, and independently governable.
 49. Every component of the regime should be subject to correction and retirement.
 50. The complete evaluation and standards system should itself be continuously evaluated.
-

32. Completeness Criteria for the Foundation Series

The foundation series is complete enough for early institutional use when:

- All eight complete foundation papers exist
- PROJECT_IDENTITY.md governs present status
- TERMINOLOGY.md governs language
- this appendix governs integration
- evidence and research methods are documented
- institutional governance is documented
- public copy is controlled
- versions are maintained
- templates can be used operationally
- no file implies unsupported authority

The series is not complete enough for formal certification, accreditation, regulatory recognition, or binding standards merely because the documents exist.

Those roles require:

- Tested processes
- competent people
- operational quality systems
- public legitimacy
- secure infrastructure
- independent oversight
- authority
- implementation evidence
- continuing accountability

33. Final Perspective

The eight foundations are not eight separate ideas.

They are eight controls on the same institutional problem.

Frontier AI systems can change faster than the evidence used to judge them.

Evaluation content can lose value through exposure.

Capability can be misinterpreted as risk.

Developers can become the sole judges of their own claims.

Independent reviewers can lack access or competence.

Auditor markets can create certificates before standards are mature.

Voluntary frameworks can remain weak, while premature regulation can freeze poor methods.

Metrics and prestige can reward the appearance of safety instead of valid evidence.

National systems can produce incompatible conclusions about the same technology.

The foundation architecture answers these problems through a connected design.

Dynamic protocols keep evidence current.

Held-out evaluations protect integrity.

High-stakes evaluation matches rigor to consequence.

Independent review creates challenge.

Third-party assurance creates scale.

Progressive standards connect learning to institutional force.

Incentive design aligns behavior with integrity.

Global interoperability makes evidence portable without erasing plural authority.

None of these elements is sufficient alone.

Together they create a path from a technical question to a credible institutional decision.

The governing idea of the foundation library is:

Evidence should be current, protected, proportionate, independently challengeable, institutionally scalable, progressively formalized, incentive-compatible, and internationally interpretable.

That is the combined foundation for Standards Body.

Revision Record

Version 1.0

Date: July 16, 2026

Change type: Complete foundational edition

Summary: Establishes the canonical integration of the eight Standards Body foundations. Defines the combined architecture, foundation inventory, dependency matrix, cross-foundation invariants, evaluated-object and evidence architecture, governance, independence, security, transparency, status, decision systems, shared maturity model, consolidated failure register, tensions, evidence gaps, research agenda, role map, metrics, complete template and scorecard indexes, unified implementation pathway, integrated pilot, document hierarchy, change control, usage guidance, decision rules, canonical positions, and completeness criteria.

Status: Approved foundational source.