Standards Body · Foundation paper, public edition · Released July 17, 2026
Canonical record: https://standardsbody.ai/library/foundation-paper/third-party-auditor-ecosystem/

FOUNDATION_05_THIRD_PARTY_AUDITOR_ECOSYSTEM.md

Foundation 5: Third-Party Auditor Ecosystem

Series: Foundations for Frontier AI Evaluation Infrastructure
Version: 1.0
Status: Canonical working white paper
Project: Standards Body
Primary domain: standardsbody.ai
Core line: Foundations for Frontier AI
Research basis reviewed through: July 16, 2026
Document owner: Standards Body
Review cycle: Annual review, with event-triggered revision after material changes in accreditation practice, frontier AI auditing, evaluator markets, international recognition, or relevant standards

---

Document Purpose

This paper defines the Standards Body position on the development of a trustworthy third-party evaluator and auditor ecosystem for frontier artificial intelligence.

It is intended to serve as:

- A first-principles explanation of why independent evaluation must become an ecosystem rather than remain a collection of one-off expert engagements
- A framework for distinguishing testing, evaluation, inspection, audit, certification, validation, verification, and accreditation
- A design guide for evaluator competence, impartiality, quality management, security, market structure, and international recognition
- A bridge between independent expert review and scalable institutional assurance
- A reference for future accreditation requirements, standards, public registries, procurement systems, and evaluator partnerships
- A durable source document from which shorter articles, technical specifications, and operational manuals can be developed

This paper is not itself an accreditation standard.

It does not authorize Standards Body to accredit, certify, inspect, or regulate any organization.

It does not claim that the existing conformity-assessment system can be copied directly into frontier AI.

It identifies the institutional architecture that could allow third-party AI evaluation to become more competent, consistent, legible, competitive, secure, and internationally trusted.

Executive Summary

Independent expert review can improve frontier AI evaluation.

It does not automatically scale.

A small number of respected researchers may be able to evaluate one model, review one protocol, or advise one developer. That approach becomes insufficient when many organizations, governments, insurers, purchasers, standards bodies, and members of the public need evidence about:

- Model capabilities
- High-stakes risks
- Safeguard effectiveness
- Security practices
- Compliance with frontier safety frameworks
- Evaluation quality
- Deployment controls
- Incident response
- Organizational governance
- Claims of conformity with technical or institutional standards

Scaling this work requires more than additional experts.

It requires an ecosystem.

A trustworthy third-party auditor ecosystem should make it possible to answer:

- Who is competent to perform which evaluation?
- What exactly is within an evaluator's approved scope?
- Which methods have been validated?
- How independent is the evaluator?
- How are conflicts and client dependence managed?
- How is sensitive model access protected?
- Can another qualified evaluator reproduce the result?
- How are evaluators monitored after initial approval?
- What happens after serious error or misconduct?
- How can smaller evaluators enter the market?
- How are results recognized across borders?
- Who evaluates the evaluators?
- How can buyers and the public verify legitimate claims?

- How are audit, evaluation, certification, and accreditation kept conceptually separate?

These questions are not solved by branding an organization an "AI auditor."

The term can refer to radically different activities.

One organization may run capability benchmarks.

Another may inspect governance records.

Another may test cybersecurity controls.

Another may review compliance with a voluntary frontier safety framework.

Another may issue a certification mark.

Another may accredit the organization issuing the certificate.

If these roles are not distinguished, the public may incorrectly assume that:

- A benchmark score is an audit
- A review is a certification
- A certification proves safety
- A private commercial evaluator is independent
- Accreditation means approval of every result
- One successful assessment authorizes work in every domain
- International recognition exists where none has been established

A serious ecosystem should therefore separate at least six functions:

1. **Standards and scheme development**
   Defines requirements, methods, decision rules, and governance.

2. **Testing and evaluation**
   Produces technical evidence about a model or system.

3. **Inspection and audit**
   Examines systems, records, processes, controls, and claims against defined criteria.

4. **Validation and verification**
   Assesses whether claims, methods, statements, or evidence meet specified requirements.

5. **Certification or attestation**
   Issues a formal statement that defined requirements have been fulfilled within a stated scope.

6. **Accreditation**
   Provides independent recognition that a conformity-assessment body is competent and impartial to perform specified activities.

These functions can interact.

They should not collapse into one institution by default.

The ecosystem also needs multiple evaluator types.

No single organization is likely to possess world-class competence across:

- Cybersecurity
- biology
- autonomous agents
- persuasion
- critical infrastructure
- statistics
- model elicitation
- system security
- organizational governance
- safety cases
- international standards
- open-weight systems
- sociotechnical impact

Specialization should therefore be expected.

A mature system may include:

- Technical evaluation laboratories
- Domain-specialist evaluator teams
- Inspection bodies
- management-system auditors
- certification bodies
- accreditation bodies
- government AI institutes
- nonprofit public-interest evaluators
- academic consortia
- commercial assurance providers
- open-source evaluation communities
- proficiency-testing providers
- reference-material custodians
- standards organizations
- scheme owners
- regulators and public purchasers

Diversity can increase resilience and innovation.

It can also create fragmentation, inconsistent quality, duplicated burden, evaluator shopping, and public confusion.

Standards Body adopts the following core position:

Frontier AI evaluation should develop into a plural, accountable, and internationally interoperable third-party assurance ecosystem in which evaluator competence, impartiality, scope, methods, security, and performance are independently assessed and continuously monitored.

The ecosystem should be designed around several principles:

- **Competence before branding**
- **Scope-specific recognition**
- **Independence with disclosed limitations**
- **Validated methods**
- **Secure access**
- **Separation of incompatible roles**
- **Proficiency testing and inter-evaluator comparison**
- **Continuous surveillance**
- **Transparent registries**
- **Complaint and appeal mechanisms**
- **Proportionate liability**
- **Small-evaluator participation**
- **International mutual recognition**
- **No permanent monopoly**
- **No claim beyond evidence**

Existing conformity-assessment infrastructure provides important lessons.

ISO and IEC standards distinguish requirements for:

- Testing and calibration laboratories
- Inspection bodies
- certification bodies
- management-system audit and certification bodies
- validation and verification bodies
- accreditation bodies

The International Accreditation Forum and International Laboratory Accreditation Cooperation have built multilateral recognition arrangements intended to support international acceptance of accredited results.[^iaf-mla][^ilac-mra]

These systems demonstrate how competence, impartiality, consistent operation, peer evaluation, scopes of accreditation, surveillance, and recognition can support trust.

Frontier AI differs in important ways:

- Models change rapidly
- Evaluation constructs remain immature
- Methods can become obsolete quickly
- Access is concentrated in a small number of developers
- Systems may behave strategically
- Evidence may be security-sensitive

- Evaluation itself can create dual-use risk
- There may be no agreed threshold or standard
- A model can be updated after assessment
- Full reproducibility may be impossible
- The assessed object may include dynamic tools, prompts, users, and deployment infrastructure

The goal is therefore not to copy an existing laboratory or certification regime mechanically.

The goal is to adapt its strongest institutional principles to a field whose science and objects are changing.

A credible third-party ecosystem should eventually allow a decision-maker to distinguish:

- An unreviewed evaluator claim
- A reviewed evaluator method
- An accredited evaluator operating within scope
- A one-time evaluation
- A system audit
- A certification against defined requirements
- A continuous assurance arrangement
- A result recognized across jurisdictions

The fifth foundation of Standards Body is the infrastructure that makes those distinctions real.

---

1. Foundational Proposition

1.1 Core Thesis

> **Independent evaluation will not become reliable at scale unless the institutions conducting it are themselves evaluated, monitored, and held accountable.**

1.2 Ecosystem Thesis

> **No single evaluator should be expected to provide every form of frontier AI assurance. A plural ecosystem with clear roles and interoperable requirements is more resilient than a permanent monopoly.**

1.3 Accreditation Thesis

> **Recognition should be scope-specific and based on demonstrated competence, impartiality, consistent operation, security, and performance, not reputation alone.**

1.4 Market Thesis

Competition can improve evaluator innovation and capacity, but unmanaged commercial incentives can produce client capture, evaluator shopping, weak methods, and assurance theater.

1.5 International Thesis

Frontier AI assurance should move toward cross-border recognition of competent evaluators and credible results without requiring one global evaluator or one universal protocol.

1.6 Dynamic Thesis

Evaluator competence and accreditation should expire, narrow, or change when methods, systems, personnel, or evidence change materially.

1.7 Public-Interest Thesis

The evaluator ecosystem should be treated as part of critical public-interest infrastructure, even when many participants are private organizations.

---

2. Scope and Boundaries

2.1 What This Foundation Covers

This paper covers the ecosystem of organizations and processes that may provide:

- Model testing
- Capability evaluation
- Safeguard evaluation
- Security assessment
- Inspection
- organizational audit
- safety-framework compliance review
- validation
- verification
- certification
- evaluator accreditation
- proficiency testing
- interlaboratory comparison
- public registries
- international recognition
- surveillance and enforcement of evaluator competence

2.2 What This Foundation Does Not Establish

This paper does not establish:

- A legal right to audit frontier laboratories
- A binding certification scheme
- A specific accreditation body
- A universal audit checklist
- A single permitted business model
- A global regulator
- A certification mark
- Mandatory conformity assessment
- Liability rules
- Government recognition

2.3 Relationship to Foundation 4

`FOUNDATION_04_INDEPENDENT_EXPERT_REVIEW.md` defines the principles of meaningful independent review.

Foundation 5 asks how those principles can be institutionalized across many organizations and repeated engagements.

2.4 Relationship to the Accreditation Framework

This paper provides the philosophical and ecosystem foundation.

`EVALUATOR_ACCREDITATION_FRAMEWORK.md` should later define detailed operational requirements.

2.5 Relationship to Standards Development

An evaluator can only assess conformity when the criteria are sufficiently defined.

Where no accepted standard exists, the activity may be:

- Exploratory evaluation
- independent review
- assurance engagement
- benchmark testing
- method validation

It should not be marketed as certification against a standard that does not exist.

---

3. Canonical Definitions

3.1 First-Party Assessment

Assessment performed by the organization responsible for the system or claim.

Examples:

- Developer internal evaluation
- Internal audit
- Self-attestation

3.2 Second-Party Assessment

Assessment performed by a party with a user, purchaser, contractual, or direct stakeholder interest.

Examples:

- Customer audit
- Government procurement review
- Insurer assessment
- Platform assessment of a supplier

3.3 Third-Party Assessment

Assessment performed by a body sufficiently independent of the provider and immediate user interests to support an impartial judgment.

3.4 Testing

Determination of one or more characteristics of an object according to a specified procedure.

In AI, testing may include:

- Running tasks
- collecting outputs
- measuring performance
- testing safeguards
- evaluating system behavior

3.5 Evaluation

Structured production and interpretation of evidence about a model, system, process, or claim.

3.6 Inspection

Examination of a product, process, service, installation, or design and determination of conformity with specified or professional requirements.

ISO/IEC 17020 specifies competence, impartiality, and consistent-operation requirements for inspection bodies.[^iso-17020]

3.7 Audit

A systematic, independent, and documented process for obtaining and evaluating evidence against defined criteria.

In frontier AI, audit can cover:

- Safety-framework compliance
- security practices
- evaluation processes
- governance
- incident response
- deployment controls
- claims

3.8 Validation

Confirmation that specified requirements are adequate for an intended future use or result.

3.9 Verification

Confirmation that specified requirements have been fulfilled based on objective evidence.

3.10 Certification

Third-party attestation related to products, processes, services, persons, or management systems.

Certification is stronger and more formal than general review language.

3.11 Attestation

Issue of a statement based on a decision following review that fulfillment of specified requirements has been demonstrated.

3.12 Conformity Assessment

Demonstration that specified requirements relating to a product, process, system, person, or body are fulfilled.

3.13 Conformity-Assessment Body

An organization performing conformity-assessment activities.

Commonly abbreviated as CAB.

3.14 Accreditation

Independent recognition that a conformity-assessment body is competent and impartial to perform specified activities.

ISO/IEC 17011 specifies requirements for accreditation bodies assessing and accrediting conformity-assessment bodies.[^iso-17011]

3.15 Accreditation Body

An authoritative body that performs accreditation.

3.16 Scheme

A system of rules, procedures, criteria, governance, and responsibilities for a defined conformity-assessment activity.

3.17 Scheme Owner

The organization responsible for developing and maintaining a conformity-assessment scheme.

3.18 Scope of Accreditation

The specific activities, methods, domains, systems, and limits for which an evaluator has been recognized as competent.

3.19 Surveillance

Ongoing monitoring of an accredited or recognized body after initial assessment.

3.20 Reassessment

Periodic comprehensive review to determine whether recognition should continue.

3.21 Proficiency Testing

Evaluation of participant performance against pre-established criteria through comparison exercises.

3.22 Interlaboratory Comparison

Organization, performance, and evaluation of measurements or tests on the same or similar objects by two or more laboratories.

3.23 Reference Material

Material sufficiently homogeneous and stable with respect to specified properties for use in measurement, validation, quality control, or proficiency testing.

In AI, analogous reference assets may include:

- Frozen models
- calibrated task suites
- known-output systems
- controlled environments
- validated incident cases
- benchmark artifacts

3.24 Impartiality

Presence of objectivity and management of conflicts so that assessment judgments are not improperly influenced.

3.25 Evaluator Shopping

Selecting among evaluators based on the likelihood of receiving a favorable result rather than competence or fit.

3.26 Scope Creep

An evaluator making claims beyond its recognized competence or approved activity.

3.27 Assurance Level

A defined degree of review depth, access, rigor, continuity, and confidence.

3.28 Mutual Recognition

Acceptance by one body or jurisdiction of results issued under another recognized system.

3.29 Peer Evaluation

Assessment of an accreditation or recognition body by peers to determine whether it meets shared requirements.

3.30 Continuous Assurance

Ongoing or repeated evidence collection, assessment, and review rather than a one-time engagement.

---

4. Why an Ecosystem Is Necessary

4.1 Volume

The number of:

- Models
- systems
- deployments
- domains
- jurisdictions
- safeguards
- standards
- incidents

will exceed the capacity of a small expert community.

4.2 Specialization

Cyber evaluation and biological evaluation require different expertise, infrastructure, security, and methods.

4.3 Independence

Reliance on a few evaluator organizations can create:

- Access dependence
- market concentration
- conflicts
- intellectual monoculture
- single points of failure

4.4 Reproducibility

Multiple qualified evaluators can challenge and reproduce results.

4.5 International Reach

National institutes cannot perform every assessment worldwide.

4.6 Market Legibility

Purchasers, governments, insurers, and developers need to know what evaluator claims mean.

4.7 Innovation

Different evaluators can develop:

- New methods
- domain protocols
- secure infrastructure
- testing tools
- reporting models

4.8 Resilience

A plural ecosystem can continue operating if one evaluator:

- Fails
- loses access
- becomes captured
- suffers a security incident
- leaves the market

4.9 Trust

Trust becomes stronger when competence and impartiality are verified through shared institutions.

4.10 Limits

An ecosystem can also create:

- Fragmentation
- inconsistent quality
- duplicated audits
- market confusion
- client capture
- high compliance cost
- barriers to entry

The ecosystem must be designed rather than assumed.

---

5. Functional Architecture

A mature ecosystem should distinguish its major functions.

5.1 Standard Setter

Defines requirements or methods.

Risks:

- Incumbent capture
- slow revision
- self-serving standards
- insufficient technical evidence

5.2 Scheme Owner

Defines how conformity assessment is performed against requirements.

Responsibilities may include:

- Scope
- evidence
- decision rules
- evaluator requirements
- certification rules
- marks
- surveillance
- complaints

5.3 Test or Evaluation Laboratory

Produces technical evidence.

Relevant analogies come from ISO/IEC 17025, which sets requirements for competence, impartiality, and consistent operation of testing and calibration laboratories.[^iso-17025]

5.4 Inspection Body

Examines systems, processes, or deployments against specified or professional criteria.

5.5 Audit Body

Assesses records, processes, controls, and compliance.

5.6 Certification Body

Issues formal attestations under a certification scheme.

ISO/IEC 17065 covers bodies certifying products, processes, and services.[^iso-17065]

ISO/IEC 17021-1 covers bodies auditing and certifying management systems.[^iso-17021]

5.7 Validation and Verification Body

Evaluates claims or statements according to defined criteria.

5.8 Accreditation Body

Assesses evaluator competence and impartiality.

5.9 Proficiency-Testing Provider

Runs exercises that compare evaluator performance.

5.10 Reference-Asset Custodian

Maintains:

- Task banks
- reference systems
- challenge environments
- calibration artifacts
- known incident cases

5.11 Government Institute

May:

- Conduct evaluations
- recognize evaluators
- fund research
- support secure access
- coordinate international work
- enforce requirements

5.12 Developer

Provides system access, technical information, internal evidence, and remediation.

5.13 Purchaser or User

Defines assurance needs and relies on results.

5.14 Insurer

May use assurance evidence for underwriting, controls, or exclusions.

5.15 Regulator

May define legal requirements, recognize schemes, inspect bodies, or use results.

5.16 Public-Interest Organization

Can contribute:

- Independent oversight
- affected-party knowledge
- public accountability
- neglected-risk research

5.17 Registry Operator

Maintains verified information about:

- Evaluators
- scopes
- accreditation
- sanctions
- certificates
- current status

5.18 Appeals Body

Reviews disputed decisions independently.

---

6. Separation of Roles

6.1 Why Separation Matters

An organization that:

- Writes the standard
- owns the scheme
- evaluates the system
- certifies the system
- accredits itself
- sells consulting

- controls the registry

has concentrated incompatible powers.

6.2 Standard Setting and Certification

A certification body should not unilaterally write requirements to fit its clients.

6.3 Consulting and Certification

Consulting can create self-review.

Potential controls:

- Structural separation
- personnel separation
- time separation
- disclosure
- prohibition for defined engagements

6.4 Evaluation and Accreditation

An organization should not accredit itself.

6.5 Testing and Final Certification

A certification body may use test evidence from laboratories.

It should preserve independent certification decision authority.

6.6 Scheme Ownership and Evaluation

A scheme owner may operate evaluation activities, but risks should be disclosed and governed.

6.7 Government Roles

Government may act as:

- Regulator
- funder
- evaluator
- accreditation body
- purchaser

Role conflicts should be explicit.

6.8 Small Ecosystem Reality

Early ecosystems may require combined roles.

Combined roles should be treated as temporary or controlled rather than invisible.

6.9 Incompatible Functions Register

Every recognized body should disclose which functions it performs and how conflicts are controlled.

---

7. Types of Third-Party AI Assurance Providers

7.1 Technical Evaluation Laboratory

Focus:

- Benchmarks
- held-out tests
- capability measurement
- safeguard tests
- agent environments

7.2 Domain-Specialist Laboratory

Focus:

- Cyber
- biology
- critical infrastructure
- finance
- persuasion
- scientific systems

7.3 Security Assessment Firm

Focus:

- Weight security
- access controls
- infrastructure
- red teaming
- threat modeling
- incident response

7.4 Governance and Compliance Auditor

Focus:

- Frontier safety frameworks
- policies
- approvals
- records
- oversight
- implementation

Research on third-party compliance review has proposed models ranging from minimalist to comprehensive assessment of company adherence to frontier safety frameworks.[^compliance-reviews]

7.5 System Inspection Body

Focus:

- Deployment configuration
- operating environment
- monitoring
- human oversight
- change control

7.6 Certification Body

Focus:

- Formal determination against a defined scheme

7.7 Government Evaluator

Focus:

- Public-interest risk
- national security
- policy support
- international cooperation

7.8 Academic Evaluation Consortium

Focus:

- Research methods
- peer review
- open publication

- novel evaluation

7.9 Nonprofit Public-Interest Evaluator

Focus:

- Independent challenge
- neglected risks
- public accountability
- open infrastructure

7.10 Commercial Assurance Provider

Focus:

- Professionalized repeated assessment
- scalable operations
- client services

7.11 Open Evaluation Community

Focus:

- Reproducibility
- public benchmarks
- open-weight systems
- distributed expertise

7.12 Hybrid Body

Combines several roles under controlled governance.

---

8. Competence Framework

Evaluator recognition should be based on demonstrated competence.

8.1 Organizational Competence

Includes:

- Governance
- management
- quality system
- personnel
- infrastructure

- security
- method control
- records
- complaints
- corrective action

8.2 Personnel Competence

Includes:

- Technical knowledge
- domain expertise
- evaluation design
- statistics
- security
- review judgment
- professional conduct
- communication

8.3 Method Competence

The evaluator should demonstrate competence in the exact methods used.

Examples:

- Agent evaluation
- biological uplift study
- cyber range
- safeguard evaluation
- compliance review
- safety-case assessment

8.4 Scope-Specific Competence

An evaluator competent in language-model benchmarking is not automatically competent in:

- CBRN
- cyber operations
- model-weight security
- financial auditing
- certification

8.5 System Competence

Competence may differ across:

- Closed API models
- open-weight models
- multimodal systems
- autonomous agents
- deployed products
- fine-tuned systems

8.6 Security Competence

Sensitive work may require:

- Threat modeling
- secure access
- dual-use handling
- protected disclosure
- incident response
- sandboxing

8.7 Decision Competence

Evaluators should understand the decision their evidence supports.

8.8 Competence Evidence

Can include:

- Education
- experience
- validated work
- proficiency testing
- witnessed assessments
- peer review
- publications
- training
- references
- security record

8.9 Continuing Competence

Require:

- Ongoing work

- training
- updated methods
- performance monitoring
- periodic reassessment

8.10 Loss of Competence

Scope should narrow when:

- Key personnel leave
- methods become obsolete
- infrastructure changes
- performance declines
- security fails

---

9. Scope of Recognition

9.1 Why Scope Matters

"Accredited AI evaluator" is too broad to be meaningful.

9.2 Scope Dimensions

A scope may specify:

- Activity
- domain
- model type
- system type
- method
- risk level
- access mode
- deployment context
- assurance level
- jurisdiction
- security classification

9.3 Example Scope

> Evaluation of autonomous cyber capabilities in tool-using language-model systems using controlled challenge environments under Protocol DEP-CYBER versions 1.x.

This does not authorize:

- Biological evaluation

- governance certification
- open-ended product safety claims
- later major protocol versions without review

9.4 Flexible Scope

A flexible scope can allow validated method updates within controlled boundaries.

Risk:

- Hidden scope expansion

Control:

- Method-validation procedure
- change notification
- surveillance
- public scope detail

9.5 Scope Extension

Require:

- Application
- competence evidence
- method review
- witnessed activity
- decision
- registry update

9.6 Scope Suspension

May apply to:

- One domain
- one method
- one site
- one assurance level
- one key person

9.7 Public Scope Statement

Registries should display current and historical scopes clearly.

---

10. Quality Management System

A third-party evaluator should operate a documented quality system.

10.1 Governance

Define:

- Legal responsibility
- leadership
- impartiality
- decision authority
- escalation
- oversight

10.2 Document Control

Control:

- Protocols
- methods
- software
- task banks
- forms
- reports
- policies
- versions

10.3 Record Control

Retain:

- Contracts
- access
- runs
- logs
- findings
- reviews
- complaints
- personnel competence
- incidents
- corrective actions

10.4 Personnel Control

Maintain:

- Roles
- qualifications
- authorization
- supervision
- training
- performance
- conflicts

10.5 Method Control

Require:

- Selection
- validation
- verification
- versioning
- uncertainty
- change management

10.6 Equipment and Infrastructure

Control:

- Compute
- software
- sandboxes
- networks
- storage
- clocks
- logs
- environments

10.7 External Providers

Manage:

- Cloud services
- model APIs
- contractors
- annotation platforms
- security vendors
- domain laboratories

10.8 Nonconforming Work

Define response when:

- Wrong model is tested
- protocol deviates
- scoring fails
- task leaks
- infrastructure breaks
- conflict emerges

10.9 Corrective Action

Address root cause rather than only the observed error.

10.10 Internal Audit

Evaluator should audit its own quality system.

Internal audit is not a substitute for accreditation surveillance.

10.11 Management Review

Leadership should review:

- Performance
- complaints
- impartiality
- security
- resources
- improvement
- scope

10.12 Improvement

Use evidence to revise methods and systems.

---

11. Method Validation

11.1 Standard Methods

Use published methods where fit for purpose.

11.2 Modified Methods

Document modifications and their effects.

11.3 Laboratory-Developed Methods

Require full validation.

11.4 Validation Questions

- Does the method measure the intended construct?
- Is it reliable?
- Is it sensitive to meaningful differences?
- Is scoring correct?
- Are tasks representative?
- Does configuration matter?
- Can another evaluator run it?
- Are uncertainty and limitations known?

11.5 Software Validation

Validate:

- Harness
- scorers
- judges
- task loaders
- agent environments
- logging
- report generation

11.6 Model-Judge Validation

Assess:

- Bias
- stability
- manipulation
- model-family effects
- domain competence
- disagreement with humans

11.7 Change Control

Revalidate after material changes.

11.8 Method Registry

Maintain public or controlled information about:

- Method
- owner
- version
- validation
- approved scope
- known limitations
- incidents

11.9 Method Portability

Test whether methods work across evaluators and infrastructure.

---

12. Proficiency Testing and Inter-Evaluator Comparison

12.1 Purpose

Proficiency testing asks whether different evaluators can produce credible results on common or comparable cases.

12.2 Why It Matters

Without comparison:

- Evaluator bias remains hidden
- scoring differences grow
- infrastructure errors persist
- market claims are difficult to compare

12.3 AI Proficiency-Testing Models

Common Reference Model

All evaluators assess the same frozen or reproducible system.

Common Task Package

All evaluators use a protected task package.

Common Incident Case

Evaluators review the same evidence and produce findings.

Blind Synthetic Organization

Evaluators audit a controlled fictional company.

Distributed Challenge

Different evaluators receive equivalent forms.

Cross-Evaluator Replication

One evaluator repeats another's work.

12.4 Scoring Proficiency

Compare:

- Task outcomes
- confidence
- thresholds
- findings
- severity
- uncertainty
- report claims

12.5 Security

Proficiency exercises must not leak active tasks or sensitive methods.

12.6 Poor Performance

Possible response:

- Investigation
- training
- corrective action
- repeated exercise
- scope limitation
- suspension

12.7 Publication

Aggregate proficiency results can improve market trust.

Individual results may need controlled disclosure.

12.8 Proficiency Provider Independence

The provider should be competent and impartial.

12.9 Reference Stability

Reference models and tasks can become obsolete.

Use dynamic and versioned exercises.

---

13. Impartiality and Independence

13.1 Impartiality Risk Analysis

Every evaluator should identify risks arising from:

- Ownership
- governance
- funding
- clients
- consulting
- personnel
- partnerships
- intellectual property
- repeat engagements
- political interests

13.2 Client Concentration

A body dependent on one developer may be formally separate but practically constrained.

13.3 Result-Dependent Compensation

Should be prohibited.

13.4 Consulting

An evaluator should not certify a system it designed or remediated without robust separation and scheme permission.

13.5 Staff Rotation

Can reduce familiarity and capture.

13.6 Independent Decision Function

The person authorizing a certification or final attestation should not simply be the person who performed the evaluation.

13.7 Impartiality Committee

A body may use an independent committee to review:

- Conflicts
- market pressure
- complaints
- funding
- scope
- controversial decisions

13.8 Transparency

Disclose material:

- Ownership
- funders
- client concentration
- related services
- governance
- conflicts

13.9 Structural Limits

Disclosure does not resolve every conflict.

13.10 Foundation 4 Integration

Use the multidimensional independence profile established in Foundation 4.

---

14. Business Models and Market Incentives

14.1 Fee-for-Service

Developer or purchaser pays per engagement.

Strengths:

- Direct demand
- scalability

Risks:

- Client capture
- evaluator shopping
- pressure for speed

14.2 Subscription or Retainer

Provides ongoing assurance.

Strengths:

- Continuous access
- institutional knowledge

Risks:

- Dependency
- reduced challenge
- high switching cost

14.3 Pooled Industry Funding

Strengths:

- Shared infrastructure
- reduced direct leverage

Risks:

- Collective capture
- incumbent dominance

14.4 Public Funding

Strengths:

- Public-interest orientation
- support for neglected work

Risks:

- Political change
- procurement delay
- national priorities

14.5 Philanthropic Funding

Strengths:

- Early ecosystem support
- public goods

Risks:

- Funder agenda
- instability
- concentration

14.6 Accreditation and Registry Fees

Can support infrastructure.

Risk:

- Incentive to approve more bodies
- barriers to small entrants

14.7 Insurance-Funded Assessment

Insurer pays or requires evaluation.

Risk:

- Narrow underwriting focus
- confidentiality
- insurer influence

14.8 Mixed Model

Diversification can reduce dependence.

14.9 Nonfinancial Incentives

Evaluators may seek:

- Prestige
- early access
- publication
- influence
- future employment
- government relationships

14.10 Market Design Principle

Revenue should support rigorous work without making favorable outcomes commercially necessary.

---

15. Evaluator Selection and Procurement

15.1 Selection Criteria

Purchasers should consider:

- Scope
- accreditation
- competence
- independence
- method
- security
- access
- track record
- liability
- international recognition

15.2 Lowest Price Risk

Evaluation quality can be undermined by procurement focused only on price.

15.3 Framework Agreements

Standing contracts can speed access.

Risk:

- Entrenchment
- reduced competition

15.4 Rotating Panels

Purchasers may rotate qualified evaluators.

15.5 Multi-Evaluator Engagement

High-consequence work can use:

- Lead evaluator
- independent replicator

- domain specialist
- security reviewer

15.6 Developer Choice

Developers should not have unlimited ability to choose the most favorable evaluator for mandatory or public-interest claims.

15.7 Scheme Allocation

A scheme can assign evaluators based on:

- Rotation
- randomization
- scope
- capacity
- conflicts
- jurisdiction

15.8 Procurement Transparency

Disclose selection logic for material public claims.

---

16. Assurance Engagement Types

16.1 Agreed-Upon Procedures

Evaluator performs specified procedures and reports factual findings.

No broad assurance conclusion.

16.2 Limited Assurance

Provides moderate confidence based on narrower evidence.

16.3 Reasonable Assurance

Provides higher confidence based on deeper evidence.

It is not absolute assurance.

16.4 Capability Evaluation

Measures defined capability under specified conditions.

16.5 Safeguard Evaluation

Tests controls against a threat model.

16.6 Compliance Review

Assesses adherence to a framework or policy.

16.7 System Audit

Examines technical and organizational controls.

16.8 Certification Assessment

Determines conformity with a defined certification scheme.

16.9 Continuous Assurance

Uses ongoing access, monitoring, repeated tests, and incident evidence.

16.10 Frontier AI Assurance Levels

Recent frontier AI auditing research proposes staged assurance levels ranging from time-bounded audits to continuous, deception-resilient verification.[^frontier-auditing]

Standards Body should study such frameworks without adopting any one level system automatically.

---

17. Access to Frontier Systems

17.1 Access as a Market Constraint

Evaluator quality depends on:

- Model access
- information access
- time
- compute
- personnel support
- tools
- logs

17.2 Developer Gatekeeping

Developers may control which evaluators can participate.

This can affect market independence.

17.3 Standardized Access Agreements

A shared access agreement can define:

- Scope
- security
- model identity
- testing
- logs
- publication
- incidents
- termination

17.4 Access Tiers

Possible tiers:

- Public
- API
- controlled black box
- grey box
- evaluator-operated
- weight access
- continuous monitoring

17.5 Access Portability

Qualified evaluators should not need to rebuild every integration from zero.

17.6 Secure Evaluation Facilities

Shared facilities can support smaller bodies.

17.7 Access Denial

If a developer denies adequate access, reports should state the resulting limitation.

17.8 Access Retaliation

An evaluator should not lose future access solely because of an evidence-based unfavorable result.

17.9 Shared Playbook

OpenAI's 2026 playbook emphasizes that trustworthy third-party evaluations should report the system tested, tool access, harness, elicitation methods, available resources, and validity checks. [^openai-playbook]

---

18. Security

18.1 Assets

Evaluators may handle:

- Model weights
- system prompts
- vulnerabilities
- held-out tasks
- incidents
- user data
- developer controls
- national-security information
- harmful outputs

18.2 Security Management

Require:

- Threat model
- access control
- secure storage
- secure compute
- logging
- personnel security
- vendor management
- incident response
- retention
- secure deletion

18.3 Evaluation-Induced Risk

An evaluator may create:

- Exploit code
- biological guidance
- safeguard bypasses
- adversarial fine-tunes
- sensitive datasets

18.4 Security Scope

Evaluator recognition should specify the sensitivity level it can handle.

18.5 Security Incident

A serious breach may trigger:

- Suspension
- notification
- investigation
- scope withdrawal
- re-accreditation

18.6 Evidence Preservation

Security controls should preserve auditability.

18.7 Developer Security Requirements

Developers should also protect evaluator tasks and methods.

18.8 Shared Infrastructure Risk

Centralized secure facilities create efficiency and concentration risk.

---

19. Accreditation Architecture

19.1 Accreditation Purpose

Accreditation provides independent recognition of evaluator competence and impartiality within scope.

It does not endorse every result.

19.2 Accreditation Body Requirements

A credible accreditation body should demonstrate:

- Competence
- impartiality
- consistent operation
- transparent process
- complaints
- appeals
- peer evaluation
- international cooperation

ISO/IEC 17011 provides a general model for accreditation-body competence and impartiality. [^iso-17011]

19.3 Initial Assessment

May include:

- Document review
- personnel review
- method review
- security assessment
- witnessed evaluation
- proficiency evidence
- office assessment
- conflict review
- report review

19.4 Accreditation Decision

Decision should be independent from the assessment team.

19.5 Scope Publication

Registry should show:

- Body
- scope
- locations
- methods
- status
- date
- conditions
- suspensions

19.6 Surveillance

Use:

- Periodic assessment
- witnessed work
- proficiency testing
- complaint review
- incident review
- remote monitoring
- report sampling

19.7 Reassessment

Comprehensive reassessment after a defined period.

19.8 Event-Triggered Assessment

Trigger after:

- Key personnel change
- method change
- security incident
- major complaint
- ownership change
- scope expansion
- repeated error

19.9 Suspension

Temporary restriction pending correction.

19.10 Withdrawal

Remove accreditation after serious or unresolved failure.

19.11 Appeal

Evaluator can challenge accreditation decisions.

19.12 Accreditation Competition

Multiple accreditation bodies may exist.

They themselves require peer recognition and oversight.

20. Surveillance and Continuous Competence

20.1 Initial Approval Is Insufficient

AI methods and systems change quickly.

20.2 Surveillance Inputs

- Evaluation reports
- raw evidence samples
- proficiency results
- complaints
- security incidents
- staff changes
- client concentration
- corrective action
- method updates
- public claims

20.3 Witnessed Assessment

Accreditation assessors observe an evaluator performing work.

20.4 Remote Surveillance

Can review:

- Logs
- method versions
- access records
- report metadata
- competence records

20.5 Unannounced Review

May be justified for high-risk scopes.

20.6 Performance Indicators

Track:

- Error rate
- correction rate

- reproducibility
- complaint rate
- security
- timeliness
- scope compliance
- independence

20.7 Dynamic Scope

Surveillance frequency should increase when:

- Methods change quickly
- risks are high
- performance is uncertain
- access is sensitive
- evaluator is new

20.8 Continuous Accreditation

Continuous data can support more responsive oversight.

It should not become opaque algorithmic governance.

---

21. Complaints, Appeals, and Whistleblowing

21.1 Complaint Sources

- Developers
- purchasers
- employees
- reviewers
- affected parties
- regulators
- other evaluators
- public researchers

21.2 Complaint Topics

- Incorrect result
- conflict
- security
- scope overreach
- retaliation
- misleading claim

- method failure
- discrimination
- bribery
- data misuse

21.3 Complaint Process

- Intake
- triage
- conflict check
- evidence preservation
- investigation
- decision
- correction
- appeal
- public status where appropriate

21.4 Whistleblower Protection

Personnel should be able to report serious issues safely.

21.5 Frivolous Complaints

Processes should resist harassment and strategic delay.

21.6 Public Notice

Material sanctions should be reflected in registries.

21.7 Cross-Border Complaints

Define jurisdiction and cooperation.

---

22. Liability and Accountability

22.1 Why Liability Matters

Evaluator error can contribute to:

- Unsafe deployment
- unnecessary restriction
- financial loss
- reputational harm

- regulatory action
- security exposure

22.2 No Absolute Guarantee

Assurance reports should not imply certainty.

22.3 Professional Liability

Evaluator organizations may need:

- Professional indemnity
- cyber insurance
- contractual liability
- defined caps
- exclusions

22.4 Liability Balance

Excessive liability can eliminate small evaluators.

Insufficient liability can weaken accountability.

22.5 Standard of Care

A future ecosystem should define reasonable professional practice.

22.6 Developer Responsibility

Third-party review should not transfer all responsibility away from the developer.

22.7 Accreditation-Body Responsibility

Accreditation confirms competence within scope, not correctness of every engagement.

22.8 Public Claims

Misleading use of evaluator reports should create consequences.

---

23. Public Registries and Marks

23.1 Registry Purpose

Enable users to verify:

- Evaluator identity
- accreditation
- scope
- status
- sanctions
- locations
- methods
- validity

23.2 Certificate Registry

Can reduce fraudulent certification claims.

23.3 Required Metadata

- Body
- accreditation body
- scope
- scheme
- version
- issue date
- expiry
- suspension
- withdrawal
- report identifier

23.4 Public Report Link

Where appropriate, link to:

- Summary
- limitations
- assurance level
- system version
- decision scope

23.5 Marks

Marks can communicate assurance.

They can also oversimplify.

23.6 Mark Rules

Specify:

- Meaning
- scope
- permitted use
- prohibited claims
- expiry
- withdrawal
- enforcement

23.7 No "Safe AI" Mark

A broad safety mark is likely to overclaim unless tightly scoped.

Preferred claims should identify:

- System
- version
- scheme
- requirements
- deployment
- date

23.8 Registry Governance

Registry operator should be neutral and secure.

---

24. Market Concentration and Access

24.1 Concentration Risk

A small number of evaluators may control:

- Model access
- certification
- government contracts
- task banks
- standards
- accreditation

24.2 Causes

- High security cost
- scarce expertise
- developer preference
- accreditation burden
- procurement barriers
- data access
- brand trust

24.3 Consequences

- High prices
- weak challenge
- capture
- slow innovation
- exclusion
- single points of failure

24.4 Small-Evaluator Participation

Support through:

- Shared secure facilities
- grants
- subsidized accreditation
- open methods
- modular scopes
- supervised access
- proficiency pathways
- consortium models

24.5 Open-Source Infrastructure

Open evaluation tools can reduce entry costs.

Open tooling does not eliminate the need for quality or security.

24.6 Antitrust and Competition

Market structure should be monitored where assurance becomes required.

24.7 Evaluator Portability

Clients should be able to change evaluators without losing all institutional evidence.

24.8 Multi-Sourcing

Critical decisions may benefit from more than one evaluator.

---

25. Evaluator Shopping and Inconsistent Results

25.1 Evaluator Shopping

Occurs when organizations seek favorable conclusions.

25.2 Controls

- Assignment
- rotation
- disclosure of prior engagements
- central registry
- standardized scope
- independent certification decision
- result portability

25.3 Opinion Shopping

A client may terminate after adverse preliminary findings and seek another evaluator.

Require disclosure of:

- Prior evaluator
- termination reason
- unresolved findings

25.4 Inconsistent Results

Differences can arise from:

- Method
- access
- configuration
- expertise
- thresholds
- interpretation
- conflicts

25.5 Reconciliation

Use:

- Meta-review
- joint technical session
- additional evidence
- reference protocol
- published difference analysis

25.6 Pluralism Versus Confusion

Different conclusions can be legitimate.

The ecosystem should explain why they differ.

---

26. International Recognition

26.1 Why Recognition Matters

Without recognition, organizations may face repeated assessment in every market.

26.2 IAF MLA

The IAF Multilateral Recognition Arrangement aims to support mutual recognition and acceptance of accredited certification and validation or verification statements among signatories.[^iaf-mla]

26.3 ILAC MRA

The ILAC Mutual Recognition Arrangement supports acceptance of accredited testing, calibration, medical testing, inspection, proficiency-testing, and reference-material activities.[^ilac-mra]

26.4 Peer Evaluation

Accreditation bodies participating in recognition arrangements are evaluated by peers.

26.5 AI Application

A future AI recognition system may cover:

- Technical evaluation laboratories
- inspection bodies
- safety-framework auditors
- certification bodies

- validation and verification bodies

26.6 Recognition Conditions

Require:

- Shared base standard
- comparable accreditation
- scope clarity
- peer evaluation
- complaint cooperation
- registry transparency
- security compatibility

26.7 National Differences

Jurisdictions may impose additional requirements.

26.8 No Automatic Policy Equivalence

Recognition of technical competence does not force identical legal decisions.

26.9 Global Accreditation Cooperation

IAF and ILAC have been moving toward a unified global accreditation cooperation structure, reflecting the long-term value of coordinated recognition infrastructure.[^ilac-home]

26.10 Early AI Path

Begin with:

- Bilateral recognition
- pilot peer reviews
- shared proficiency testing
- common metadata
- recognition of limited scopes

---

27. Government and Regulatory Relationships

27.1 Voluntary Ecosystem

Can develop before legal mandates.

27.2 Government Recognition

Government may recognize:

- Accreditation bodies
- evaluator scopes
- schemes
- reports
- certificates

27.3 Procurement

Government can require qualified evaluation for purchases.

27.4 Regulatory Reliance

Regulators may use third-party evidence while retaining authority.

27.5 Regulatory Capture

Risks:

- Incumbent-dominated requirements
- excessive burden
- closed evaluator market
- weak public oversight

27.6 Government Evaluator Conflict

A government institute may evaluate and regulate.

Separation of functions may be needed.

27.7 Innovation Support

Public funding can build:

- Testbeds
- reference models
- secure facilities
- proficiency programs
- evaluator training

27.8 NIST TEVV

NIST's TEVV work supports evaluation tasks, testbeds, tools, datasets, guidelines, and standards development as part of a broader AI measurement ecosystem.[^nist-tevv]

---

28. Insurance, Procurement, and Private Ordering

28.1 Insurance

Insurers may require:

- Security audits
- capability assessments
- incident controls
- monitoring

28.2 Procurement

Purchasers may require evaluation before deployment.

28.3 Contracts

Contracts can define:

- Access
- evidence
- re-evaluation
- incidents
- liability
- audit rights

28.4 Investor and Lender Expectations

Financial actors may rely on assurance reports.

28.5 Private Standardization

Private requirements can spread before formal regulation.

28.6 Risks

- Uncoordinated requirements
- opaque methods
- market exclusion

- assurance theater
- confidentiality

28.7 Interoperability

Shared evaluator standards can reduce duplicated burden.

---

29. Accreditation Lifecycle

Stage 1: Ecosystem Need

Identify a defined evaluation or audit scope requiring recognition.

Stage 2: Base Requirements

Establish:

- Evaluator standard
- scheme
- methods
- competence
- security
- reporting

Stage 3: Accreditation-Body Readiness

Confirm that the accreditation body understands the scope.

Stage 4: Application

Evaluator submits:

- Organization
- scope
- personnel
- methods
- quality system
- security
- conflicts
- performance

Stage 5: Document Review

Assess system readiness.

Stage 6: Assessment Planning

Select qualified assessors and witness activities.

Stage 7: Office and System Assessment

Review governance, records, quality, security, and impartiality.

Stage 8: Witnessed Evaluation

Observe real or controlled evaluator work.

Stage 9: Proficiency Evidence

Review comparison performance.

Stage 10: Nonconformities

Evaluator corrects identified issues.

Stage 11: Independent Decision

Separate decision authority determines scope and status.

Stage 12: Registry Publication

Publish recognition.

Stage 13: Surveillance

Monitor continued competence.

Stage 14: Scope Change

Extend, reduce, suspend, or withdraw.

Stage 15: Reassessment

Conduct periodic comprehensive review.

Stage 16: International Recognition

Accreditation body participates in peer evaluation and mutual recognition.

---

30. Maturity Model

Level 0: Unstructured Evaluator Market

Characteristics:

- Self-described auditors
- no shared definitions
- no scope
- no registry
- no surveillance
- one-off reports

Level 1: Professional Evaluator Practice

Characteristics:

- Documented methods
- competence records
- conflict disclosure
- security
- quality control

Level 2: Scheme-Recognized Evaluators

Characteristics:

- Defined schemes
- reviewed methods
- approved scopes
- complaints
- public registry

Level 3: Accredited Evaluator Ecosystem

Characteristics:

- Independent accreditation
- surveillance
- proficiency testing
- scope control

- sanctions
- multiple qualified bodies

Level 4: Internationally Interoperable Assurance Regime

Characteristics:

- Peer-evaluated accreditation bodies
- mutual recognition
- shared metadata
- cross-border acceptance
- continuous assurance
- incident coordination
- mature market oversight

---

31. Implementation Pathway

Phase 1: Role Clarification

Publish common definitions for:

- Evaluation
- audit
- inspection
- certification
- accreditation

Phase 2: Evaluator Registry

Create a voluntary registry with:

- Competence
- methods
- scope
- independence
- security
- references

Registry entry is not accreditation.

Phase 3: Common Evaluator Requirements

Develop baseline requirements for:

- Quality

- competence
- impartiality
- records
- security
- reporting
- complaints

Phase 4: Pilot Scopes

Begin with bounded scopes such as:

- Agentic cyber evaluation
- held-out protocol administration
- frontier safety-framework compliance review

Phase 5: Proficiency Pilots

Run inter-evaluator comparisons.

Phase 6: Scheme Development

Create transparent rules for selected assurance activities.

Phase 7: Accreditation Pilot

Partner with experienced accreditation institutions or build a peer-reviewed pilot process.

Phase 8: Surveillance

Monitor recognized evaluators.

Phase 9: Public Recognition

Publish verified scopes and status.

Phase 10: International Pilots

Run bilateral recognition and shared proficiency exercises.

Phase 11: Procurement and Insurance Integration

Test reliance on evaluator evidence.

Phase 12: Continuous Improvement

Use complaints, incidents, and performance data to revise the system.

---

32. Proposed Standards Body Pilot

32.1 Pilot Name

Frontier AI Evaluator Qualification and Proficiency Pilot

32.2 Purpose

Test whether multiple evaluator organizations can competently and consistently administer a defined Standards Body protocol.

32.3 Initial Scope

> Administration and reporting of the Standards Body held-out autonomous cyber capability protocol under controlled conditions.

32.4 Participants

Invite:

- Academic laboratory
- nonprofit evaluator
- commercial evaluator
- government institute
- open evaluation consortium

32.5 Baseline Requirements

Participants submit:

- Competence
- independence profile
- quality procedures
- security
- method
- personnel
- conflicts

32.6 Common Exercise

Each evaluates:

- The same reference model
- equivalent held-out tasks
- standardized environment
- common reporting minimum

32.7 Comparison

Assess:

- Results
- uncertainty
- configuration
- scoring
- findings
- security
- time
- cost
- reporting

32.8 Witnessed Assessment

Independent assessors observe each evaluator.

32.9 Remediation

Participants correct issues and repeat selected tasks.

32.10 Outputs

- Pilot evaluator requirements
- scope template
- proficiency report
- quality checklist
- registry prototype
- accreditation gap analysis
- recommendations for Foundation 5 revision

32.11 Success Criteria

The pilot succeeds if it:

- Distinguishes competent from weak practice

- improves evaluator consistency
- preserves methodological diversity
- supports smaller participants
- identifies security needs
- produces a usable public scope
- demonstrates a path toward independent accreditation
- avoids implying premature certification authority

---

33. Metrics for Evaluating the Ecosystem

33.1 Capacity

- Qualified evaluators
- domain coverage
- geographic coverage
- evaluation throughput
- wait times

33.2 Competence

- Proficiency performance
- witnessed findings
- report quality
- method validation
- staff continuity

33.3 Independence

- Client concentration
- funding diversity
- consulting conflicts
- selection control
- publication rights

33.4 Consistency

- Inter-evaluator agreement
- threshold consistency
- repeatability
- comparable reporting

33.5 Security

- Incidents
- task leakage
- model exposure
- corrective action
- security scope

33.6 Market Health

- New entrants
- concentration
- pricing
- small-body participation
- evaluator switching
- innovation

33.7 Accountability

- Complaints
- appeals
- suspensions
- withdrawals
- corrections
- registry accuracy

33.8 International Recognition

- Peer-evaluated accreditation bodies
- mutual-recognition agreements
- cross-border accepted reports
- duplicated-assessment reduction

33.9 Decision Utility

- Decisions informed
- safeguards improved
- incidents anticipated
- false claims corrected
- procurement use
- insurer use

33.10 Burden

- Cost

- time
- duplicated work
- access delay
- developer burden
- evaluator burden

---

34. Failure Modes and Safeguards

34.1 Auditor Branding Without Competence

Failure: Organizations call themselves AI auditors without demonstrated scope-specific competence.

Safeguard: Public scope, competence evidence, accreditation, proficiency testing.

34.2 Accreditation Theater

Failure: Recognition is treated as a marketing badge rather than continuous oversight.

Safeguard: Surveillance, witnessed assessments, sanctions, public status.

34.3 One Scope Becomes Universal

Failure: Cyber competence is used to claim biology or governance competence.

Safeguard: Granular scopes and claim enforcement.

34.4 Client Capture

Failure: Evaluators depend on favorable results for revenue.

Safeguard: Payment separation, disclosure, pooled funding, rotation.

34.5 Evaluator Shopping

Failure: Clients seek favorable assessors.

Safeguard: Assignment, registry, prior-engagement disclosure, scheme rules.

34.6 Consulting Conflict

Failure: Evaluator certifies its own design or remediation.

Safeguard: Separation, prohibition, independent decision.

34.7 Scheme Capture

Failure: Incumbents write requirements that entrench their practices.

Safeguard: balanced governance, public consultation, small-actor participation.

34.8 Accreditation Monopoly

Failure: One accreditation body controls entry and methods.

Safeguard: peer recognition, multiple bodies, appeals, public oversight.

34.9 Fragmented Methods

Failure: Results cannot be compared.

Safeguard: core metadata, proficiency testing, method registries.

34.10 False Certification Claims

Failure: Advisory review is marketed as proof of safety.

Safeguard: terminology rules, registry, mark enforcement.

34.11 Stale Accreditation

Failure: Recognition continues after methods and personnel change.

Safeguard: event-triggered surveillance, expiration, scope reduction.

34.12 Security Failure

Failure: Evaluator leaks models, tasks, incidents, or harmful methods.

Safeguard: security-scoped competence and sanctions.

34.13 Excessive Barrier to Entry

Failure: Accreditation excludes smaller innovative evaluators.

Safeguard: modular scopes, grants, shared infrastructure, supervised pathways.

34.14 Race to the Bottom

Failure: Price competition reduces depth.

Safeguard: minimum scheme requirements and procurement quality criteria.

34.15 Race to Complexity

Failure: Evaluators create unnecessarily elaborate processes to justify fees.

Safeguard: proportionality and decision-utility metrics.

34.16 Public Confusion

Failure: Users cannot distinguish testing, audit, certification, and accreditation.

Safeguard: canonical terminology and precise claims.

34.17 International Duplication

Failure: Same system is repeatedly evaluated under incompatible regimes.

Safeguard: mutual recognition and crosswalks.

34.18 Regulatory Outsourcing

Failure: Government delegates policy responsibility to private evaluators.

Safeguard: retain accountable public decision authority.

34.19 Developer Responsibility Erosion

Failure: Certification is used to disclaim developer responsibility.

Safeguard: explicit responsibility allocation.

34.20 Evaluator Retaliation

Failure: Unfavorable evaluators lose access.

Safeguard: access agreements, public reporting, multiple access channels.

---

35. Serious Objections

Objection 1: Frontier AI Evaluation Science Is Too Immature for Accreditation

This objection is partly correct.

Response:

- Begin with narrow scopes

- accredit competence and process, not scientific certainty
- use dynamic scope
- conduct frequent surveillance
- avoid broad safety certification

Residual concern:

Premature formalization can freeze weak methods.

Objection 2: Accreditation Creates Bureaucracy

It can.

Response:

- Proportional scopes
- modular requirements
- shared infrastructure
- risk-based surveillance
- digital records

Residual concern:

Some burden is unavoidable.

Objection 3: Market Competition Already Disciplines Evaluators

Competition can also produce:

- Client capture
- evaluator shopping
- opaque methods
- weak claims

Independent oversight remains necessary.

Objection 4: Government Should Perform All High-Stakes Evaluations

Government offers authority and resources.

It may lack:

- Capacity
- speed
- specialized depth
- international reach
- market responsiveness

A mixed ecosystem is stronger.

Objection 5: Commercial Auditors Cannot Be Independent

Commercial incentives create risk but do not make impartial work impossible.

Existing assurance fields use structural controls, accreditation, surveillance, liability, and recognition.

Residual concern:

Client dependence remains material.

Objection 6: Certification Will Create False Assurance

Yes, if scope and claim are broad.

Response:

- Narrow certification claims
- explicit limitations
- system version
- expiry
- continuous surveillance

Residual concern:

Public interpretation may still oversimplify.

Objection 7: Small Evaluators Cannot Meet Security Requirements

Some cannot.

Response:

- Shared facilities
- grants
- supervised scopes
- consortium models
- open tools

Objection 8: International Recognition Is Politically Unrealistic

Full global recognition may be slow.

Bilateral and domain-specific recognition can begin earlier.

Objection 9: Evaluator Ecosystems Will Be Captured by Large Laboratories

This is a central risk.

Response:

- Diversified funding
- independent selection
- public-interest governance
- open-source participation
- concentration monitoring

Objection 10: A Model Changes Too Quickly for Certification

Response:

- Version-specific certification
- configuration identity
- change triggers
- continuous assurance
- narrow claims

Residual concern:

Some systems may change too frequently for meaningful static certification.

Objection 11: Mutual Recognition Can Spread Weak Assurance

Correct.

Recognition requires strong peer evaluation and scope control.

Objection 12: Existing ISO Structures Are Too Slow

They may be too slow for some frontier changes.

The ecosystem should retain:

- Dynamic protocols
- emergency updates
- modular schemes
- rapid technical guidance

without abandoning rigor.

---

36. Evidence Gaps

36.1 Evaluator Competence

Which competence indicators predict high-quality frontier evaluations?

36.2 Proficiency Testing

How should inter-evaluator comparison work for stochastic, dynamic systems?

36.3 Accreditation Scope

How granular should scopes be?

36.4 Surveillance

Which ongoing data best predicts evaluator failure?

36.5 Market Structure

What degree of concentration is efficient versus dangerous?

36.6 Funding

Which business models preserve independence and capacity?

36.7 Liability

What standard of care is appropriate?

36.8 Certification

Which AI claims are mature enough for certification?

36.9 Continuous Assurance

How should ongoing assessment work for frequently updated systems?

36.10 International Recognition

Which AI assurance activities can be mutually recognized first?

36.11 Security

How can small evaluators securely access frontier systems?

36.12 Evaluator Shopping

Which controls work without eliminating legitimate choice?

36.13 Accreditation-Body Competence

Who can assess frontier AI evaluators competently?

36.14 Decision Impact

Does accreditation improve outcomes enough to justify cost?

---

37. Research Agenda

Priority 1: Evaluator Taxonomy

Create a clear map of testing, inspection, audit, certification, validation, verification, and accreditation.

Priority 2: Scope Design

Develop granular scope templates for frontier AI activities.

Priority 3: Competence Standards

Define role, team, and organizational competence.

Priority 4: Proficiency Testing

Pilot common model, task, incident, and audit exercises.

Priority 5: Method Validation

Create requirements for dynamic and held-out methods.

Priority 6: Quality Management

Adapt laboratory and conformity-assessment quality systems to AI.

Priority 7: Impartiality

Measure client dependence, consulting conflicts, and capture.

Priority 8: Market Design

Study competition, concentration, entry, and pricing.

Priority 9: Secure Access

Build shared facilities and portable access agreements.

Priority 10: Surveillance

Develop event-triggered and continuous accreditation.

Priority 11: Public Registries

Design verified, machine-readable scope and status records.

Priority 12: Liability

Study professional responsibility and insurance.

Priority 13: International Recognition

Pilot peer evaluation and mutual recognition.

Priority 14: Certification Claims

Identify claims mature enough for formal attestation.

Priority 15: Ecosystem Effectiveness

Measure whether third-party assurance improves deployment, safeguards, and public trust.

---

38. Near-Term Experiments

Experiment 1: Scope Classification

Have multiple institutions classify evaluator scopes and compare consistency.

Experiment 2: Common Model Exercise

Run a shared proficiency test across evaluator types.

Experiment 3: Common Audit Case

Provide a synthetic frontier safety framework and evidence package for review.

Experiment 4: Witnessed Assessment

Observe evaluators performing the same protocol.

Experiment 5: Client-Concentration Disclosure

Test a standardized financial independence metric.

Experiment 6: Evaluator Assignment

Compare client choice, rotation, and random assignment.

Experiment 7: Shared Secure Facility

Pilot access for small evaluators.

Experiment 8: Registry Prototype

Create machine-readable evaluator and scope records.

Experiment 9: Surveillance Dashboard

Track method, personnel, complaints, security, and performance changes.

Experiment 10: Cross-Border Replication

Have evaluators in two jurisdictions repeat an assessment.

Experiment 11: Mark Comprehension

Test whether users understand narrow certification claims.

Experiment 12: Suspension Drill

Simulate a major evaluator security incident and registry response.

39. Implications for Future Standards

A future evaluator accreditation standard could require:

39.1 Legal and Governance Identity

Defined responsibility and authority.

39.2 Impartiality

Risk analysis, conflict controls, and independence profile.

39.3 Competence

Personnel, team, organizational, domain, method, and security competence.

39.4 Scope

Precise activities, domains, methods, systems, and limits.

39.5 Quality Management

Documents, records, internal audit, management review, corrective action.

39.6 Method Validation

Selection, validation, versioning, and uncertainty.

39.7 Infrastructure

Secure compute, software, environments, logging, and external providers.

39.8 Evaluation Process

Contract review, planning, administration, scoring, review, reporting.

39.9 Decision Independence

Separation of evaluation and final attestation where applicable.

39.10 Reporting

Required metadata, scope, confidence, limitation, and expiry.

39.11 Complaints and Appeals

Independent and documented process.

39.12 Security

Sensitivity-scoped controls and incident response.

39.13 Proficiency Testing

Participation and response to poor performance.

39.14 Surveillance

Periodic and event-triggered monitoring.

39.15 Suspension and Withdrawal

Clear rules and public status.

39.16 International Recognition

Peer evaluation and compatible scopes.

These requirements should be developed in `EVALUATOR_ACCREDITATION_FRAMEWORK.md` through the future `STANDARDS_DEVELOPMENT_PROCESS.md`.

---

40. Relationship to the Other Foundations

Foundation 1: Dynamic Evaluation Protocols

Evaluators must maintain competence as protocols change.

Foundation 2: Held-Out Evaluations

Third-party bodies need secure access, custody, administration, and compromise response.

Foundation 3: High-Stakes Capability Evaluation

Risk domain and consequence determine evaluator scope, rigor, and assurance level.

Foundation 4: Independent Expert Review

Foundation 4 defines meaningful review. Foundation 5 scales and institutionalizes it.

Foundation 6: Progressive Standards and Requirements

Third-party evaluation can move from voluntary review to procurement, certification, insurance, and formal requirements.

Foundation 7: Incentives and Prestige

Market recognition should reward competence and integrity rather than marketing.

Foundation 8: Global Interoperability

Accreditation, proficiency, registries, and mutual recognition support cross-border acceptance.

---

41. Canonical Standards Body Positions

Standards Body adopts the following working positions.

1. Third-party AI assurance should be treated as an ecosystem, not a single service category.
2. Testing, evaluation, inspection, audit, certification, and accreditation should not be used interchangeably.
3. Accreditation recognizes competence within scope. It does not guarantee every result.
4. No evaluator should claim competence across all AI domains by default.
5. Evaluator scopes should identify activity, domain, method, system type, assurance level, and limits.
6. Recognition should depend on demonstrated competence, impartiality, consistent operation, security, and performance.
7. Initial approval should be followed by surveillance, proficiency testing, and reassessment.
8. The bodies accrediting evaluators should themselves be competent, impartial, and peer reviewed.
9. Scheme owners, evaluators, certification bodies, and accreditation bodies should be separated where incompatible conflicts exist.

10. Developer-funded evaluation can be legitimate only with strong impartiality controls and disclosure.
11. Result-dependent compensation should be prohibited.
12. Evaluator client concentration should be disclosed and monitored.
13. Evaluator shopping and opinion shopping should be controlled.
14. A third-party review should not be marketed as certification unless a valid certification scheme exists.
15. Broad "safe AI" certification claims should be avoided.
16. Marks and certificates should identify system, version, scope, scheme, and expiry.
17. High-consequence evaluation requires secure access and evidence preservation.
18. Smaller evaluators and open communities should have credible entry pathways.
19. Open tools and shared facilities can reduce barriers but do not replace quality assurance.
20. Proficiency testing and cross-evaluator replication are core infrastructure.
21. Evaluator security incidents should affect status and scope.
22. Public registries should make legitimate scope and current status verifiable.
23. International mutual recognition should begin with narrow, comparable scopes.
24. Technical recognition should not force identical policy decisions across jurisdictions.
25. Government, private, academic, nonprofit, and open evaluators can coexist.
26. No evaluator or accreditation body should become an unaccountable permanent monopoly.
27. Certification should not transfer responsibility away from developers or deployers.
28. Evaluator evidence should expire after material system, method, or personnel changes.
29. The assurance ecosystem should be monitored for capture, concentration, burden, and real-world effectiveness.
30. Formalization should proceed carefully enough to avoid freezing immature evaluation science.

---

42. Decision Rules

An evaluator should be recognized only when:

- Its legal and governance responsibilities are clear
- Its competence matches the requested scope

- Its methods are validated
- Its personnel are authorized and current
- Its independence risks are controlled
- Its security matches the sensitivity of work
- Its quality system supports consistent operation
- It can handle complaints and corrective action
- It participates in proficiency activities
- Its claims remain within scope

Recognition should be limited or suspended when:

- Key personnel leave
- methods change materially without validation
- security controls fail
- conflicts become unmanageable
- proficiency performance is poor
- reports repeatedly overclaim
- records are incomplete
- complaints reveal systemic failure
- scope is exceeded
- ownership or funding creates new risk

A certification claim should not be permitted when:

- Requirements are undefined
- the scheme is not governed
- evaluator competence is unverified
- the assessed system version is unclear
- surveillance is absent
- the claim implies broad safety beyond evidence

Mutual recognition should not be granted when:

- Accreditation systems are not peer evaluated
- scopes are incompatible
- security is inadequate
- complaints cannot be shared
- registry status is unreliable
- political pressure overrides technical requirements

---

43. Evaluator Organizational Profile Template

A. Identity

- Legal name
- ownership

- locations
- leadership
- jurisdiction
- contact

B. Activities

- Testing
- evaluation
- inspection
- audit
- certification
- validation
- verification
- consulting
- training

C. Scope

- Domains
- systems
- methods
- assurance levels
- security levels
- limitations

D. Governance

- Decision authority
- oversight
- impartiality
- complaints
- appeals

E. Competence

- Personnel
- qualifications
- experience
- proficiency
- training

F. Quality System

- Method control
- records
- internal audit
- corrective action
- management review

G. Security

- Access
- infrastructure
- incident response
- sensitivity scope

H. Independence

- Ownership
- funding
- client concentration
- consulting
- related bodies
- publication rights

I. Performance

- Evaluations
- corrections
- complaints
- incidents
- proficiency results

J. Recognition

- Accreditation body
- scope
- status
- date
- expiry
- conditions

---

44. Accreditation Scope Template

Evaluator:
Accreditation body:
Scope identifier:
Status:
Effective date:
Expiry or reassessment date:

Activity

Domain

System Type

Method or Protocol

Approved Versions

Access Mode

Assurance Level

Security Classification

Locations

Key Personnel or Competence Requirements

Limitations

Conditions

Suspensions or Exclusions

Recognition Arrangements

---

45. Evaluator Engagement Template

A. Client and Evaluated Party

B. Evaluator

C. Activity Type

- Test
- evaluation
- inspection
- audit
- validation
- verification
- certification assessment

D. Scope

E. Criteria or Protocol

F. System and Version

G. Access

H. Methods

I. Security

J. Independence and Conflicts

K. Personnel

L. Outputs

M. Developer Review

N. Complaints and Appeal

O. Publication

P. Liability

Q. Expiration and Re-Evaluation

---

46. Surveillance Report Template

Evaluator:
Scope:
Assessment date:
Assessors:

Organizational Changes

Personnel Changes

Method Changes

Security Changes

Client Concentration

Proficiency Performance

Sampled Reports

Complaints and Appeals

Incidents

Corrective Actions

Scope Compliance

Nonconformities

Decision

- Maintain
- maintain with conditions
- reduce
- suspend
- withdraw
- reassess

Next Review

---

47. Evaluator Complaint Template

Complaint identifier:
Complainant:
Evaluator:
Engagement or certificate:
Date:

Allegation

Evidence

Conflict Check

Immediate Risk

Investigation

Evaluator Response

Finding

Corrective Action

Status Change

Public Notice

Appeal

Closure

---

48. Ecosystem Scorecard

Dimension	Core Question
Role clarity	Are testing, audit, certification, and accreditation distinguished?
Scheme quality	Are requirements and decision rules defined?

Dimension	Core Question
Competence	Does the evaluator demonstrate scope-specific ability?
Scope	Are claims limited to approved activities?
Impartiality	Are ownership, funding, consulting, and client risks controlled?
Quality system	Can the body operate consistently and correct errors?
Method validity	Are evaluation methods fit for purpose and versioned?
Security	Can sensitive models, tasks, and evidence be protected?
Access	Does the evaluator receive enough access for its claims?
Proficiency	Has performance been compared with peers?
Surveillance	Is competence monitored after initial recognition?
Decision independence	Is final attestation separated from incompatible roles?
Complaints	Can stakeholders raise and resolve material concerns?
Appeals	Can recognition and certification decisions be challenged?
Sanctions	Can scope be reduced, suspended, or withdrawn?
Registry	Can users verify status and scope?
Mark control	Are certification claims precise and enforceable?
Market health	Is there competition without a race to the bottom?
Small-actor access	Can qualified smaller evaluators participate?
International recognition	Can credible results cross borders?
Liability	Are responsibility and professional consequences appropriate?
Dynamic adaptation	Can scopes and methods evolve with AI systems?
Public interest	Does the ecosystem serve more than client marketing?
Decision utility	Does assurance improve real-world decisions?

---

49. Final Perspective

Independent evaluation cannot become durable public infrastructure if every engagement begins from zero.

A one-off review may produce insight.

An ecosystem produces continuity.

It can preserve methods.

It can train people.

It can compare performance.

It can discipline claims.

It can investigate failure.

It can recognize competence.

It can withdraw recognition.

It can allow one jurisdiction to trust evidence produced in another.

But an ecosystem can also become a market for reassurance.

Commercial pressure can reward favorable reports.

Accreditation can become a badge.

Standards can entrench incumbents.

Security requirements can exclude smaller experts.

Certification can be mistaken for proof of safety.

A few organizations can accumulate excessive power over access, evaluation, and public legitimacy.

The design must therefore begin with clear distinctions.

Testing is not certification.

Review is not accreditation.

Accreditation is not proof that every result is correct.

A certificate is not a universal statement about a system.

Third party does not automatically mean independent.

Prestige does not automatically mean competent.

Formal process does not automatically mean useful.

The purpose of the ecosystem is not to produce more labels.

It is to produce more trustworthy evidence at scale.

That requires institutions capable of evaluating the evaluators.

It requires narrow scopes.

It requires quality systems.

It requires proficiency.

It requires surveillance.

It requires complaints and appeals.

It requires consequences for failure.

It requires pathways for new entrants.

It requires international recognition without global monopoly.

It requires the willingness to say that some forms of AI assurance are not yet mature enough for certification.

The fifth foundation of Standards Body is therefore accountable evaluation capacity.

The future of frontier AI assurance will depend not only on which systems are tested, but on whether the institutions conducting the tests deserve to be trusted.

References and Research Basis

[^iso-17011]: International Organization for Standardization, **ISO/IEC 17011:2017, Conformity assessment, Requirements for accreditation bodies accrediting conformity assessment bodies**. https://www.iso.org/standard/67198.html

[^iso-17025]: International Organization for Standardization, **ISO/IEC 17025, Testing and calibration laboratories**. https://www.iso.org/ISO-IEC-17025-testing-and-calibration-laboratories.html

[^iso-17020]: International Organization for Standardization, **ISO/IEC 17020:2026, Conformity assessment, Requirements for the operation of various types of bodies performing inspection**. https://www.iso.org/standard/17020

[^iso-17065]: International Organization for Standardization, **ISO/IEC 17065:2012, Conformity assessment, Requirements for bodies certifying products, processes and services**. https://www.iso.org/standard/46568.html

[^iso-17021]: International Organization for Standardization, **ISO/IEC 17021-1:2015, Conformity assessment, Requirements for bodies providing audit and certification of management systems**. https://www.iso.org/standard/61651.html

[^iso-casco-bodies]: International Organization for Standardization, **CASCO Toolbox: Bodies**. https://casco.iso.org/bodies.html

[^iso-casco-examples]: International Organization for Standardization, **CASCO Toolbox: Examples**. https://casco.iso.org/examples.html

[^iso-recognition]: International Organization for Standardization, **CASCO Toolbox: Recognition of conformity-assessment bodies**. https://casco.iso.org/recognition-of-cabs.html

[^iso-building-trust]: International Organization for Standardization, **Building Trust: The Conformity Assessment Toolbox**. https://www.iso.org/iso/casco_building-trust.pdf

[^iso-certification]: International Organization for Standardization, **Certification**. https://www.iso.org/certification.html

[^iaf-mla]: International Accreditation Forum, **Purpose of the IAF Multilateral Recognition Arrangement**. https://iaf.nu/en/mla-purpose/

[^iaf-about-mla]: International Accreditation Forum, **About the IAF MLA**. https://iaf.nu/en/about-iaf-mla/

[^iaf-peer]: International Accreditation Forum, **IAF ML 4:2025, Policies and Procedures for the MLA**. https://iaf.nu/iaf_system/uploads/documents/IAF_ML_4_Issue_9_04072025.pdf

[^ilac-mra]: International Laboratory Accreditation Cooperation, **ILAC MRA and Signatories**. https://ilac.org/ilac-mra-and-signatories/

[^ilac-home]: International Laboratory Accreditation Cooperation, **Global Accreditation Cooperation**. https://ilac.org/

[^ilac-facts]: International Laboratory Accreditation Cooperation, **Facts and Figures**. https://ilac.org/about-ilac/facts-and-figures/

[^nist-tevv]: National Institute of Standards and Technology, **AI Test, Evaluation, Validation and Verification**. https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv

[^nist-rmf]: National Institute of Standards and Technology, **Artificial Intelligence Risk Management Framework 1.0**, 2023. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[^nist-zero-draft]: National Institute of Standards and Technology, **Outline: Proposed Zero Draft for a Standard on AI Testing, Evaluation, Verification, and Validation**, 2025. https://www.nist.gov/document/outline-proposed-zero-draft-standard-ai-testing-evaluation-verification-and-validation

[^openai-playbook]: OpenAI, **A Shared Playbook for Trustworthy Third-Party Evaluations**, May 29, 2026. https://openai.com/index/trustworthy-third-party-evaluations-foundations/

[^openai-external]: OpenAI, **Strengthening Our Safety Ecosystem with External Testing**, November 19, 2025. https://openai.com/index/strengthening-safety-with-external-testing/

[^frontier-auditing]: Miles Brundage et al., **Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies**, 2026. https://arxiv.org/abs/2601.11699

[^compliance-reviews]: Aidan Homewood et al., **Third-Party Compliance Reviews for Frontier AI Safety Frameworks**, 2025. https://arxiv.org/abs/2505.01643

[^framework-evaluation]: Lily Stelling et al., **Evaluating AI Companies' Frontier Safety Frameworks: Methodology and Results**, 2025. https://arxiv.org/abs/2512.01166

[^external-access]: Jacob Charnock et al., **Expanding External Access to Frontier AI Models for Dangerous Capability Evaluations**, 2026. https://arxiv.org/abs/2601.11916

[^aisi-lessons]: UK AI Security Institute, **Early Lessons from Evaluating Frontier AI Systems**, 2024. https://www.aisi.gov.uk/blog/early-lessons-from-evaluating-frontier-ai-systems

[^aisi-research]: UK AI Security Institute, **Research Agenda**. https://www.aisi.gov.uk/research-agenda

[^inspect]: UK AI Security Institute, **Inspect AI**. https://inspect.aisi.org.uk/

Revision Record

Version 1.0

Date: July 16, 2026

Change type: Complete foundational edition

Summary: Establishes the fully developed canonical working white paper for Foundation 5. Defines the third-party auditor ecosystem, core assurance roles, testing, inspection, audit, certification, accreditation, evaluator types, competence, scope, quality systems, method validation, proficiency testing, impartiality, business models, selection, access, security, accreditation architecture, surveillance, complaints, liability, registries, market concentration, evaluator shopping, international recognition, government and insurance relationships, maturity model, implementation pathway, Standards Body evaluator pilot, metrics, failure analysis, objections, evidence gaps, research agenda, standards implications, operational templates, scorecard, and primary-source research basis.

Status: Ready for internal review and future expert critique.