Standards Body · Foundational source, public edition · Released July 17, 2026
Canonical record: https://standardsbody.ai/library/foundational-source/evidence-standards/

EVIDENCE_STANDARDS.md

Standards Body Evidence Standards

Project: Standards Body
Primary domain: standardsbody.ai
Core line: Foundations for Frontier AI
Document type: Canonical evidence-quality, claims, confidence, sourcing, and decision standard
Version: 1.0
Status: Approved foundational source
Document owner: Standards Body
Applies to: All Standards Body research, foundation papers, evaluation protocols, review findings, standards proposals, institutional claims, public reports, website copy, case studies, source registries, partnerships, contributor work, and future assurance activities
Related canonical sources: `PROJECT_IDENTITY.md`, `TERMINOLOGY.md`, `FOUNDATIONS_APPENDIX.md`, `RESEARCH_METHODOLOGY.md`, `EVALUATION_PHILOSOPHY.md`, `TRANSPARENCY_FRAMEWORK.md`, `VERSION_HISTORY.md`
Research basis reviewed through: July 16, 2026
Review cycle: Annual review, with event-triggered revision after a material change in evaluation science, measurement standards, conformity assessment, legal requirements, research methods, project authority, or public-claims practice

---

Document Purpose

This document defines the evidence standards of Standards Body.

It governs:

- What counts as evidence
- Which evidence is appropriate for different claims
- How evidence quality should be assessed
- How primary and secondary sources should be used
- How technical, organizational, legal, policy, and institutional claims differ
- How uncertainty and confidence should be expressed
- How contradictory evidence should be handled
- How confidential evidence may support public conclusions
- How model-generated analysis may and may not be used

- How causal, comparative, predictive, negative, and absence claims should be supported
- How evidence should be documented, cited, versioned, corrected, withdrawn, and retired
- What level of evidence is required before Standards Body adopts a position, publishes a claim, proposes a standard, supports a threshold, or participates in an assurance activity
- How evidence packages should be structured for review, standards development, and decision-making

This file is the canonical source for evidentiary sufficiency within Standards Body.

It does not replace domain expertise.

It does not create one universal hierarchy in which every randomized study is automatically stronger than every operational record, every public source is stronger than every confidential source, or every quantitative result is stronger than every expert judgment.

Evidence quality depends on:

- The claim
- the decision
- the evaluated object
- the method
- the context
- the consequence of error
- the source's access
- the source's independence
- the age of the evidence
- the ability to verify or challenge it

The central requirement is not maximum volume of evidence.

It is evidence fit for the claim and decision.

Executive Summary

Frontier AI institutions regularly make claims that exceed their evidence.

Examples include:

- A benchmark result is presented as proof of general capability.
- A failed elicitation attempt is presented as proof that a capability is absent.
- An internal review is presented as an independent audit.
- A management-system certificate is presented as proof that a model is safe.
- A policy proposal is presented as though technical evidence uniquely determines it.
- An expert's prediction is presented as established fact.
- A confidential finding is invoked without explaining who reviewed it or what the public can reasonably infer.
- A model-generated summary is cited as an independent source.

- A legal requirement is described from memory rather than the authoritative legal text.
- A historical result is applied to a model or system that has materially changed.
- A public claim omits contrary evidence, uncertainty, or known limitations.

Standards Body rejects these practices.

The project adopts the following core proposition:

> **A claim should be no broader, more certain, more current, or more authoritative than the evidence supporting it.**

Evidence standards should connect four elements:

1. The claim being made
2. The evidence supporting or challenging it
3. The confidence justified by that evidence
4. The decision or public statement the evidence is being used to support

A technically impressive result can still be weak evidence when:

- The construct is unclear
- the system identity is incomplete
- the tasks are contaminated
- the administration conditions are unrealistic
- the result cannot be reproduced
- the evaluator lacks independence
- the sample is unrepresentative
- the evidence is stale
- the conclusion goes beyond the tested conditions

A qualitative record can be strong evidence when:

- The claim is institutional or procedural
- the source had direct access
- the record has reliable provenance
- multiple independent accounts corroborate it
- the limitations are visible
- the evidence is decision-relevant

Standards Body therefore evaluates evidence across multiple dimensions rather than using a single source hierarchy.

The core dimensions are:

- Relevance
- directness
- validity
- reliability
- provenance
- completeness

- independence
- recency
- representativeness
- reproducibility
- security and integrity
- uncertainty
- decision relevance
- consistency with other evidence
- susceptibility to manipulation

The same evidence may receive different weight for different claims.

A developer's internal system log may be strong direct evidence that an event occurred.

It may be weak evidence for the claim that the organization's controls are effective across all deployments.

An external evaluator's report may be strong evidence for performance under one protocol.

It may be weak evidence for broad legal compliance.

A certification may be strong evidence that specified requirements were assessed under a defined scheme.

It is not automatic evidence that a model is safe, unbiased, secure, or suitable for every use.

The Standards Body evidence model contains five evidence levels.

Level E0: Unsupported

The claim rests primarily on assertion, reputation, intuition, marketing, or unverifiable information.

Level E1: Preliminary

Some relevant evidence exists, but it is limited, indirect, weakly controlled, unreplicated, or materially incomplete.

Level E2: Supported

Multiple relevant sources or one strong direct source support a bounded claim, with important limitations remaining.

Level E3: Substantiated

The claim is supported by strong, traceable, appropriately designed evidence, independent challenge, uncertainty analysis, and material contrary-evidence review.

Level E4: Decision-Grade

The evidence is sufficiently mature, current, secure, independently reviewable, and decision-linked for a specified consequential use.

These levels apply to a defined claim, not to a source or organization in the abstract.

Evidence level is not the same as confidence.

Confidence also depends on:

- Agreement among sources
- unresolved uncertainty
- model or system change
- hidden information
- potential bias
- decision sensitivity

Standards Body uses five confidence labels:

- Very low
- low
- moderate
- high
- very high

Very high confidence should be rare in frontier AI work.

A strong evidence package should include:

- The exact claim
- system identity
- protocol or method
- source register
- direct and supporting evidence
- contrary evidence
- uncertainty
- conflict information
- security status
- reviewer findings
- confidence
- claim limitations
- expiration or review date
- correction history

The project should prefer primary sources for:

- Legal claims
- standards claims

- institutional status
- evaluation results
- company policies
- technical methods
- public commitments
- regulatory decisions

Secondary sources remain valuable for:

- Synthesis
- context
- criticism
- comparison
- discovery
- interpretation

They should not displace authoritative primary material when that material is available.

Confidential evidence may be necessary for:

- Model access
- protected task results
- security incidents
- proprietary system details
- personal data
- national-security information

A public conclusion based partly on confidential evidence must state:

- That nonpublic evidence was used
- who had access
- what type of review occurred
- what limitations the confidentiality creates
- what public claim remains justified
- whether an independent reviewer corroborated the evidence

Model-generated text may assist:

- Search planning
- source discovery
- extraction
- comparison
- drafting
- consistency review

It may not serve as independent evidence.

The human or institution publishing the claim remains responsible for verifying every substantive statement.

The final rule is:

> **Show the claim, show the evidence, show the limitations, show the confidence, and show who is responsible.**

---

1. Foundational Propositions

1.1 Claim-Bounded Evidence

> **Evidence should be assessed against a specific claim, not judged as strong or weak in the abstract.**

1.2 Decision-Bounded Sufficiency

> **The amount and quality of evidence required should rise with the consequence, irreversibility, and uncertainty of the decision.**

1.3 Source-Proximity Principle

> **Where available, the source closest to the relevant event, system, method, law, or decision should normally anchor the evidence record.**

1.4 Multiple-Dimension Principle

> **Evidence quality should be assessed across several dimensions rather than reduced to one hierarchy or score.**

1.5 Contrary-Evidence Principle

> **A credible evidence review actively searches for evidence that could weaken or overturn the preferred conclusion.**

1.6 Traceability Principle

> **Every consequential conclusion should be traceable to identifiable sources, methods, assumptions, reviewers, and versions.**

1.7 Uncertainty Principle

> **Uncertainty should be treated as part of the evidence output rather than removed to create a cleaner conclusion.**

1.8 Recency Principle

Evidence should be re-evaluated when the model, system, protocol, deployment, threat, standard, law, or institution materially changes.

1.9 Independence Principle

Evidence produced by an interested party may be important and direct, but its weight should account for incentives, access, conflicts, and external challenge.

1.10 Reproducibility Principle

Evidence becomes more credible when qualified others can reproduce, replicate, reperform, or otherwise independently test the relevant claim.

1.11 Confidentiality Principle

Evidence does not become invalid merely because it is confidential, but confidentiality increases the burden for governance, independent review, traceability, and claim limitation.

1.12 No Evidence Laundering

A citation, expert title, certification, institutional logo, or model-generated summary should not be used to make weak evidence appear stronger than it is.

1.13 Correction Principle

The credibility of an evidence institution depends partly on how visibly and effectively it corrects error.

1.14 Authority Principle

Technical evidence can inform authority, but evidence does not create legal or regulatory authority by itself.

---

2. Scope

2.1 Included Evidence Domains

This standard applies to:

- AI evaluation results

- benchmark and protocol claims
- capability and safeguard findings
- security evidence
- incident evidence
- organizational and governance evidence
- audit and certification evidence
- accreditation and competence evidence
- legal and regulatory claims
- standards and conformity claims
- policy analysis
- forecasts
- expert judgments
- public statements
- historical records
- case studies
- international comparisons
- research synthesis
- website and media claims

2.2 Excluded Functions

This standard does not:

- Replace a court's evidentiary rules
- replace professional legal standards
- establish medical or clinical evidence rules
- establish national-security classification
- guarantee scientific truth
- eliminate expert disagreement
- require public disclosure of all evidence
- create certification or accreditation authority
- determine public policy automatically

2.3 Higher Domain Standards

Where a specialized field has stronger applicable evidence requirements, the stronger requirements should govern.

Examples may include:

- Clinical research
- nuclear safety
- aviation
- laboratory metrology
- cybersecurity incident response
- classified national-security analysis

- financial auditing
- legal proceedings

2.4 Minimum Standard

This document is a floor for Standards Body work.

A project may adopt stricter evidence requirements.

---

3. Canonical Evidence Definitions

Definitions in `TERMINOLOGY.md` govern.

The following definitions are central to this document.

3.1 Evidence

Information, observation, artifact, record, result, or testimony relevant to supporting or challenging a claim.

3.2 Claim

A proposition asserted to be true, supported, justified, or sufficiently reliable for a specified purpose.

3.3 Evidence Standard

The level, type, quality, and sufficiency of evidence required for a defined claim or decision.

3.4 Evidence Package

A structured set of artifacts supporting or challenging a claim, review, standard, certification, recognition, or decision.

3.5 Direct Evidence

Evidence derived directly from the object, event, system, or process at issue.

3.6 Indirect Evidence

Evidence supporting a claim through inference rather than direct observation.

3.7 Primary Source

An original source of data, law, standards text, evaluation results, institutional policy, or first-hand record.

3.8 Secondary Source

A source interpreting, summarizing, comparing, or analyzing primary material.

3.9 Tertiary Source

A source that aggregates or simplifies primary and secondary material, such as a general encyclopedia, directory, overview, or automatically generated summary.

3.10 Corroboration

Independent or meaningfully distinct evidence supporting the same claim.

3.11 Triangulation

Use of multiple methods, sources, or perspectives to assess a claim.

3.12 Provenance

The origin, history, custody, modification, and ownership of evidence or an artifact.

3.13 Chain of Custody

A documented record of possession, transfer, access, and handling.

3.14 Traceability

The ability to connect a claim, result, decision, or artifact to its sources, methods, versions, and responsible parties.

3.15 Reproducibility

The ability to obtain consistent computational or analytical results using the same data, code, methods, and conditions.

3.16 Replicability

The ability of an independent study or implementation to support a finding using new data, tasks, systems, or materially independent execution.

3.17 Reperformance

Independent execution of a specified procedure to verify reported work.

3.18 Confidence

The degree of justified belief in a conclusion based on available evidence.

3.19 Evidence Gap

Missing evidence necessary to support, challenge, or resolve a claim.

3.20 Material Evidence

Evidence that could reasonably change the conclusion, confidence, scope, or decision.

3.21 Contrary Evidence

Evidence that weakens, contradicts, narrows, or creates material uncertainty about a claim.

3.22 Decision-Grade Evidence

Evidence sufficiently mature, current, relevant, independently challengeable, and well-governed for a specified consequential decision.

3.23 Epistemic Status

A concise statement describing how strongly a claim is supported and what uncertainty remains.

---

4. The Evidence Object

Standards Body treats an evidence object as more than a document or result.

A complete evidence object includes:

4.1 Identifier

A unique record or citation identifier.

4.2 Claim Relationship

Which claim the evidence supports, challenges, or contextualizes.

4.3 Source

Who or what produced it.

4.4 Date

When the evidence was created, observed, published, or last verified.

4.5 Method

How it was produced.

4.6 Evaluated Object

The exact model, system, organization, process, law, standard, event, or decision addressed.

4.7 Scope

The boundaries within which the evidence is relevant.

4.8 Result

The observation, finding, record, or conclusion.

4.9 Uncertainty

Known measurement, sampling, model, interpretive, or access limitations.

4.10 Provenance

Origin, modification history, custody, and authenticity.

4.11 Independence

Relevant relationships between source and claim.

4.12 Security Status

Whether the evidence is public, controlled, confidential, restricted, or highly restricted.

4.13 Version and Status

Whether the evidence is current, corrected, superseded, withdrawn, or expired.

4.14 Review Status

Whether it received:

- Internal review
- peer review
- independent expert review
- audit
- validation
- verification
- no formal review

4.15 Citation or Retrieval Location

Where the evidence can be found or how authorized reviewers may access it.

---

5. Claims Architecture

Evidence standards begin with a correctly formed claim.

5.1 Claim Components

A complete claim should identify:

- Subject
- asserted property or event
- scope
- time
- conditions
- comparison where relevant
- degree of certainty
- intended use

5.2 Weak Claim

> Model X is safe.

Problems:

- Undefined system
- undefined harm
- undefined context
- undefined period
- undefined risk tolerance
- no evidence boundary

5.3 Bounded Claim

> Under Protocol Y version 1.3, System X version 4.2 did not exceed the defined critical cyber capability threshold in the assessed environment, using the stated tool access and elicitation budget, as of July 2026.

This claim remains limited.

It does not establish:

- Absence of all cyber capability
- universal safety
- future performance
- performance under other scaffolds
- legal compliance
- deployment suitability

5.4 Claim Decomposition

Broad claims should be decomposed into testable subclaims.

Example:

"The evaluator is independent" may contain:

- No ownership relationship
- no result-dependent fee
- no publication veto
- controlled client concentration
- method freedom
- conflict disclosure
- sufficient access
- governance independence

5.5 Claim Owner

Every material claim should have an identifiable owner responsible for:

- Evidence selection
- wording
- limitations
- correction
- status

5.6 Claim Register

Major projects should maintain a claim register containing:

- Claim ID
- wording
- claim type
- owner
- evidence level
- confidence
- sources
- contrary evidence
- review date
- publication location
- status

---

6. Evidence Levels

Evidence levels express the maturity of support for one defined claim.

They do not rank people, institutions, or publications.

6.1 Level E0: Unsupported

Characteristics:

- Assertion without traceable evidence
- marketing language
- anonymous or unverifiable report
- reputation alone
- speculative statement presented as fact
- model-generated output without source verification
- citation that does not support the claim

Permitted uses:

- Hypothesis generation
- early signal logging
- question formation

Not permitted for:

- Canonical factual claims
- standards requirements
- high-stakes thresholds
- public assurance claims
- certification decisions

6.2 Level E1: Preliminary

Characteristics:

- One limited source
- indirect evidence
- uncontrolled observation
- unpublished internal result
- small or unrepresentative sample
- weak provenance
- limited access
- high uncertainty
- no independent challenge

Permitted uses:

- Exploratory research
- early warning
- pilot design
- provisional hypothesis
- preliminary internal decision with caution

Required language:

- Preliminary
- suggestive
- early evidence
- not independently verified
- subject to material uncertainty

6.3 Level E2: Supported

Characteristics:

- One strong direct source or several relevant sources
- defined method
- identifiable object and version
- reasonable provenance
- bounded conclusion
- uncertainty described
- material limitations remain
- some corroboration or review

Permitted uses:

- Working paper conclusions
- bounded public claims
- pilot decisions

- proposed framework elements
- risk-management action that remains reversible

6.4 Level E3: Substantiated

Characteristics:

- Strong direct evidence
- appropriate design and controls
- reproducible or independently challenged work
- system and protocol identity
- contrary-evidence review
- uncertainty analysis
- traceable provenance
- material conflicts addressed
- current evidence
- clear limitations

Permitted uses:

- Canonical project positions
- mature technical conclusions
- proposed standards provisions
- consequential internal decisions
- public findings with bounded wording

6.5 Level E4: Decision-Grade

Characteristics:

- E3 qualities
- explicit decision link
- consequence-sensitive evidence design
- strong independent review
- adequate security and chain of custody
- decision-relevant thresholds or criteria
- current evaluator competence
- documented false-positive and false-negative implications
- correction and appeal process
- expiration or monitoring
- implementation evidence where relevant

Permitted uses:

- High-stakes deployment decisions
- certification or conformity decisions under a valid scheme
- procurement requirements

- formal standards progression
- recognition decisions
- regulatory support where the authority accepts the evidence

6.6 No Automatic Promotion

Evidence should not be promoted to a higher level because:

- The source is famous
- the organization is large
- the result is quantitative
- the paper is peer reviewed
- the claim is urgent
- many secondary sources repeat it
- the evidence is confidential
- the evidence is public
- the source is a government
- the source is a frontier laboratory

6.7 Evidence-Level Record

Every E2 through E4 claim should record:

- Assigned level
- rationale
- reviewer
- date
- unresolved gaps
- conditions for promotion
- expiration or reassessment trigger

---

7. Evidence Quality Dimensions

Evidence should be assessed dimension by dimension.

A source may be strong on one dimension and weak on another.

7.1 Relevance

Question:

Does the evidence address the actual claim?

Weak relevance examples:

- Using a broad benchmark to support a narrow real-world capability claim

- using management-system documentation to support a technical safety claim
- using policy popularity to support technical validity
- using one legal jurisdiction to describe another

7.2 Directness

Question:

How many inferential steps separate the evidence from the claim?

Direct evidence generally deserves more weight, but indirect evidence can be essential when direct observation is impossible or unsafe.

7.3 Construct Validity

Question:

Does the evidence measure or represent the intended concept?

Required considerations:

- Construct definition
- task coverage
- proxy quality
- alternative explanations
- expert review
- deployment relevance

7.4 Internal Validity

Question:

Does the design support the claimed explanation, comparison, or causal interpretation?

Consider:

- Controls
- confounding
- selection
- leakage
- administration consistency
- scoring
- missing data
- analyst degrees of freedom

7.5 External Validity

Question:

Can the conclusion reasonably generalize beyond the assessed sample, environment, system, or period?

7.6 Reliability

Question:

Would repeated measurement under equivalent conditions produce materially consistent results?

Consider:

- Run-to-run variation
- task sampling
- judge consistency
- model nondeterminism
- environment stability
- human-rater agreement

7.7 Provenance

Question:

Can the origin and history of the evidence be established?

Strong provenance includes:

- Original source
- creation date
- authorship
- modification history
- version
- custody
- signatures or hashes where relevant

7.8 Authenticity

Question:

Is the evidence genuine and unaltered?

Potential controls:

- Digital signatures
- cryptographic hashes
- trusted repositories

- source confirmation
- preserved originals
- tamper-evident logs

7.9 Completeness

Question:

Does the evidence package include the material information needed for interpretation?

Missing information may include:

- Failed runs
- excluded tasks
- model configuration
- elicitation budget
- scoring changes
- contrary findings
- incidents
- conflicts

7.10 Independence

Question:

What interests or relationships could influence production or interpretation?

Independence affects weight, not automatic admissibility.

Developer evidence may be uniquely informative.

It should be identified as first-party evidence and independently challenged where the claim is consequential.

7.11 Recency

Question:

Is the evidence still applicable?

Recency depends on:

- Model change
- system change
- protocol change
- threat change
- law or standard change
- institutional status

- deployment context

7.12 Representativeness

Question:

Does the sample represent the relevant task universe, users, environments, languages, sectors, or threat actors?

7.13 Reproducibility

Question:

Can qualified reviewers recompute or re-execute the work using the same artifacts?

7.14 Replicability

Question:

Can materially independent evidence support the same finding?

7.15 Security and Integrity

Question:

Was the evidence protected against:

- Leakage
- contamination
- tampering
- unauthorized access
- selective deletion
- score manipulation

7.16 Measurement Uncertainty

Question:

Are uncertainty sources identified and, where feasible, quantified?

7.17 Interpretive Uncertainty

Question:

Could qualified reviewers reasonably interpret the same evidence differently?

7.18 Decision Relevance

Question:

Does the evidence reduce uncertainty for the actual decision?

7.19 Timeliness

Question:

Was evidence produced early enough to affect the decision?

7.20 Susceptibility to Gaming

Question:

Could the producer or evaluated party optimize the evidence without improving the underlying property?

7.21 Conflict Exposure

Question:

What financial, organizational, political, professional, or reputational conflicts affect the evidence?

7.22 Transparency and Reviewability

Question:

Can qualified reviewers understand and challenge the evidence?

Full public disclosure is not always required.

Sufficient reviewability is.

7.23 Legal and Ethical Fitness

Question:

Was the evidence collected, processed, and disclosed lawfully and ethically?

7.24 Burden and Proportionality

Question:

Was the evidence burden appropriate to the claim and consequence?

Excessive evidence requirements can:

- Delay beneficial action
- exclude smaller actors
- centralize authority
- create compliance theater
- consume scarce evaluator capacity

7.25 Overall Assessment

Overall quality should be expressed as a reasoned profile.

Do not average all dimensions into one score when a critical weakness is dispositive.

---

8. Source Hierarchy and Source Use

Standards Body uses a claim-specific source hierarchy.

8.1 Tier S1: Authoritative Primary Sources

Examples:

- Enacted law
- official regulation
- court decision
- official standards text
- regulator decision
- original evaluation report
- original dataset or artifact
- signed institutional policy
- direct system logs
- original public filing
- official registry record

Use:

Anchor claims concerning legal status, institutional status, standards content, evaluation methods, and direct events.

Limitations:

An authoritative source may still be incomplete, self-interested, outdated, or narrow.

8.2 Tier S2: Primary Research and Direct Technical Evidence

Examples:

- Peer-reviewed or openly reviewed research paper
- preprint with complete methods
- evaluation dataset
- experiment record
- independent technical report
- incident investigation
- reproducible code
- expert review with direct access

Use:

Support technical, scientific, and evaluative claims.

Limitations:

Publication venue does not eliminate design, conflict, or reproducibility problems.

8.3 Tier S3: High-Quality Institutional Synthesis

Examples:

- National academy report
- systematic review
- standards-body explanatory guidance
- government technical synthesis
- international-organization analysis
- transparent multi-expert report

Use:

Context, synthesis, comparison, and interpretation.

Limitations:

May lag current systems or reflect institutional consensus rather than direct evidence.

8.4 Tier S4: Credible Secondary Analysis

Examples:

- Academic commentary
- specialist journalism
- expert blog with sources
- think-tank report
- legal analysis

- technical explainer

Use:

Discovery, criticism, context, and alternate interpretations.

Limitations:

Should not replace available primary evidence for consequential claims.

8.5 Tier S5: Preliminary or Informal Sources

Examples:

- Conference talk
- interview
- social-media statement
- forum discussion
- anonymous report
- unverified leak
- preliminary slide
- personal communication

Use:

Early signals, hypothesis generation, source discovery.

Limitations:

Require corroboration before factual adoption.

8.6 Tier S6: Tertiary and Generated Sources

Examples:

- General encyclopedia
- unsourced aggregator
- search snippet
- model-generated summary
- automatically generated profile
- citation index without source inspection

Use:

Navigation and discovery only.

Not acceptable as sole support for a material claim.

8.7 Primary-Source Rule

For legal, standards, product-policy, certification, accreditation, and institutional-status claims, inspect the authoritative primary source where available.

8.8 Secondary-Source Rule

Use secondary analysis to:

- Identify disagreement
- expose limitations
- explain context
- compare institutions
- locate primary evidence

8.9 Search-Result Rule

Search-result snippets are not evidence.

The underlying source should be opened and reviewed.

8.10 Citation-Count Rule

Citation count does not establish validity.

It may indicate influence or attention.

8.11 Peer-Review Rule

Peer review is relevant evidence of review.

It is not a guarantee of:

- Correctness
- replicability
- current applicability
- absence of conflict
- policy relevance

8.12 Government-Source Rule

Government sources may be authoritative for:

- Their own rules
- official decisions
- official data
- institutional position

They are not automatically the strongest source for every technical or normative claim.

8.13 Company-Source Rule

Company sources may be authoritative for:

- Product releases
- public policies
- internal commitments
- declared methods
- first-party findings

They require external challenge for consequential claims about effectiveness, safety, independence, or broad impact.

8.14 Anonymous-Source Rule

Anonymous evidence should be used only when:

- The information is material
- identification would create credible risk
- the recipient can assess access and credibility
- corroboration exists where feasible
- the public claim is limited

8.15 Personal-Communication Rule

Record:

- Speaker
- date
- role
- context
- permission
- whether independently corroborated

Avoid relying on unverifiable personal communication for canonical factual claims.

---

9. Claim Types and Required Evidence

Different claim types require different evidence.

9.1 Descriptive Claims

Example:

> The organization published a new evaluation framework on a specified date.

Minimum evidence:

- Primary publication
- date
- identity
- current status

9.2 Technical Performance Claims

Example:

> System X achieved a specified success rate under Protocol Y.

Minimum evidence:

- System identity
- protocol version
- task population
- administration
- scoring
- uncertainty
- raw or reviewable evidence
- date

High-stakes addition:

- Held-out integrity
- independent review
- replication or reperformance
- contrary evidence
- expiration

9.3 Capability Claims

Example:

> System X demonstrated autonomous performance in a defined capability domain.

Minimum evidence:

- Construct
- task family
- elicitation
- system configuration
- success criteria
- reliability
- autonomy

- resource use
- limitations

Avoid:

- Generalizing from knowledge questions to operational capability
- generalizing from one scaffold to every deployment

9.4 Absence Claims

Example:

> The system does not possess Capability C.

This is a strong claim.

Required evidence may include:

- Broad and valid task coverage
- strong elicitation
- multiple methods
- sufficient power
- repeated negative results
- external challenge
- defined upper bounds
- explicit scope

Preferred wording when evidence is weaker:

> The capability was not demonstrated under the assessed conditions.

9.5 Safety Claims

Example:

> The system is safe.

This claim is normally too broad.

Required replacement:

- Define harm
- context
- risk tolerance
- safeguards
- evidence
- time
- residual risk

Preferred form:

The system met the specified safeguard and risk requirements under the assessed deployment conditions.

9.6 Safeguard-Effectiveness Claims

Minimum evidence:

- Threat model
- baseline
- attack methods
- adaptive testing
- bypasses
- operational conditions
- residual risk
- uncertainty
- independent challenge

9.7 Security Claims

Minimum evidence:

- Asset
- threat
- control
- attack surface
- test scope
- vulnerability status
- incident history
- time

Avoid:

- Secure
- unhackable
- failsafe

without defined bounds.

9.8 Reliability Claims

Minimum evidence:

- Task distribution
- repeated trials
- environment variation
- failure definition
- missing or invalid runs

- uncertainty
- distribution-shift analysis

9.9 Comparative Claims

Example:

> System A is more capable than System B.

Minimum evidence:

- Comparable system definitions
- same or mapped protocol
- same administration
- equivalent resources
- uncertainty
- statistical or qualitative comparison
- material configuration differences

9.10 Superiority Claims

Example:

> Method A is the best available evaluation method.

Required evidence:

- Defined alternatives
- comparison criteria
- multiple settings
- limitations
- independent replication
- recency

Preferred wording when incomplete:

> Method A showed stronger performance on the defined criteria in the assessed settings.

9.11 Causal Claims

Example:

> The safeguard caused the reduction in misuse.

Required evidence:

- Causal design or strong causal reasoning
- counterfactual

- confounder assessment
- timing
- mechanism
- sensitivity analysis
- alternative explanations

Correlation alone is insufficient.

9.12 Predictive Claims

Example:

> This evaluation predicts future real-world incidents.

Required evidence:

- Defined prediction
- prospective or historical validation
- outcome measure
- time horizon
- calibration
- base rates
- uncertainty
- out-of-sample testing

9.13 Forecast Claims

Required evidence:

- Forecaster identity
- information date
- target definition
- time horizon
- probability
- rationale or model
- calibration history where available
- update rule

Forecasts should remain distinct from observed facts.

9.14 Organizational-Practice Claims

Example:

> The organization has an effective incident-response program.

Required evidence:

- Policies
- roles
- training
- exercises
- actual incidents
- response records
- corrective actions
- independent assessment

Documentation alone is weak evidence of effectiveness.

9.15 Independence Claims

Required evidence:

- Governance
- funding
- client relationships
- selection process
- access
- method freedom
- publication rights
- conflicts
- actual examples of unfavorable findings

9.16 Competence Claims

Required evidence:

- Scope
- qualifications
- experience
- work samples
- proficiency
- quality controls
- continuing competence

9.17 Audit Claims

Required evidence:

- Criteria
- scope
- period
- auditor

- independence
- procedures
- findings
- limitations
- status

9.18 Certification Claims

Required evidence:

- Scheme
- requirements
- certified object
- certification body
- scope
- decision
- validity
- surveillance
- registry status

Certification does not establish properties outside scope.

9.19 Accreditation Claims

Required evidence:

- Accreditation body
- standard or criteria
- accredited body
- exact scope
- effective date
- status
- suspension or withdrawal information

9.20 Compliance Claims

Required evidence:

- Applicable law, regulation, contract, or rule
- jurisdiction
- date
- facts
- qualified interpretation
- scope
- unresolved questions

Standards Body should not issue legal opinions without authorized legal expertise.

9.21 Policy Claims

Example:

> A requirement should be mandatory.

Required evidence:

- Technical need
- public-interest rationale
- alternatives
- costs
- benefits
- distributional effects
- capacity
- legal authority
- implementation
- uncertainty
- values and tradeoffs

Technical evidence informs policy.

It does not uniquely determine it.

9.22 Consensus Claims

Required evidence:

- Defined expert or stakeholder population
- process
- participation
- agreement level
- substantial objections
- dissent

Avoid:

- Experts agree
- global consensus
- industry consensus

without documentation.

9.23 Historical Claims

Required evidence:

- Contemporaneous primary record where available
- date

- source context
- later correction
- distinction between fact and interpretation

9.24 Reputation Claims

Example:

> Organization X is the leading evaluator.

Required evidence:

- Defined criterion
- comparison set
- period
- source
- conflicts

Prefer measurable descriptions over vague rankings.

9.25 Public-Impact Claims

Example:

> The standard reduced harm.

Required evidence:

- Baseline
- adoption
- implementation
- outcome
- attribution
- displacement
- unintended effects
- distribution

---

10. Technical Evaluation Evidence

10.1 Minimum Technical Record

A technical evaluation record should include:

- Evaluation purpose
- claim
- model and system identity

- protocol identifier and version
- task source and sampling
- held-out status
- elicitation conditions
- tools and scaffolds
- environment
- retries
- parameters
- scoring
- invalid runs
- uncertainty
- baseline
- date
- evaluator
- conflicts
- limitations
- result status

10.2 Raw Evidence

Where security and privacy permit, preserve:

- Inputs
- outputs
- trajectories
- logs
- timestamps
- environment state
- scoring records
- reviewer annotations
- code version
- configuration

10.3 Failed and Excluded Runs

Document:

- Failure reason
- exclusion rule
- count
- effect on result
- whether exclusion was pre-specified

Selective removal of failures is prohibited.

10.4 Task Sampling

Evidence should state:

- Target task universe
- sampling method
- sample size
- exclusions
- difficulty
- representativeness
- exposure status

10.5 Elicitation

Evidence should state:

- Prompting
- examples
- fine-tuning
- tools
- human assistance
- search
- retries
- compute
- time
- optimization effort

10.6 System Configuration

Record all material components.

A base-model result should not be presented as a full-system result unless the configuration is equivalent.

10.7 Scoring Evidence

Scoring should include:

- Rule
- rubric
- judge
- judge version
- human review
- agreement
- ambiguity
- appeal

- score changes

10.8 Model-Based Judges

Model-based scoring should identify:

- Judge model
- version
- prompt
- calibration
- human-validation sample
- conflict or shared lineage
- bias and failure analysis

10.9 Human Judges

Human scoring should identify:

- Qualifications
- training
- blinding
- conflicts
- agreement
- adjudication
- workload
- compensation

10.10 Statistical Evidence

Where applicable, report:

- Sample size
- estimate
- uncertainty interval
- distribution
- repeated-measures structure
- missing data
- multiple comparisons
- effect size
- sensitivity analysis

10.11 Qualitative Technical Evidence

Qualitative evidence may include:

- Failure narratives

- trajectory analysis
- expert review
- red-team observations
- emergent strategy
- unusual behavior

It should be systematically recorded rather than treated as anecdote alone.

10.12 Result Expiration

Technical evidence should expire or be reviewed after:

- Material model update
- system change
- protocol change
- task compromise
- new elicitation method
- major incident
- scheduled period

---

11. Organizational and Institutional Evidence

AI assurance depends on both technical and organizational evidence.

11.1 Policy Versus Practice

A policy proves that an organization documented an intention or requirement.

It does not prove:

- Implementation
- effectiveness
- consistency
- enforcement
- continuous operation

11.2 Organizational Evidence Types

Relevant evidence may include:

- Policies
- governance charters
- role descriptions
- training records
- access logs

- incident records
- audit trails
- budgets
- staffing
- committee minutes
- decision records
- complaints
- corrective actions
- exercises
- deployment records
- performance metrics

11.3 Practice Evidence

Stronger practice evidence includes:

- Repeated operational records
- sampled case files
- observed exercises
- incident handling
- independent interviews
- system logs
- correction history
- evidence of unfavorable decisions

11.4 Governance Claims

Claims about governance should examine:

- Actual decision rights
- escalation
- conflicts
- recusals
- dissent
- oversight
- accountability
- emergency use
- documented outcomes

11.5 Culture Claims

Claims such as "the organization has a strong safety culture" require caution.

Possible evidence:

- Reporting behavior

- retaliation controls
- near-miss disclosure
- staff surveys
- decision records
- resource allocation
- leadership conduct
- treatment of dissent
- correction history

11.6 Resource Claims

A policy may fail because:

- Staffing is insufficient
- expertise is absent
- compute is unavailable
- security is weak
- evaluation time is compressed
- budgets are unstable

Evidence should examine practical capacity.

11.7 Independence Claims

Independence should be evidenced through structure and behavior.

Strong evidence includes:

- Governance separation
- protected publication rights
- diversified funding
- non-result-dependent fees
- conflict controls
- unfavorable findings
- documented recusals
- external oversight

11.8 Competence Claims

Evidence of competence should match scope.

A general AI credential is not sufficient evidence of:

- Cyber expertise
- biological expertise
- standards competence
- statistical competence

- legal competence
- secure evaluation competence

11.9 Organizational Sampling

When evaluating an organization, define:

- Units
- locations
- teams
- time period
- systems
- records
- selection method
- exclusions

11.10 Management-System Evidence

Management-system evidence can support claims concerning:

- Governance
- process
- documented controls
- continuous improvement
- internal review

It should not be treated as direct proof of every product or system outcome.

---

12. Legal, Regulatory, and Standards Evidence

12.1 Authoritative Legal Sources

Prefer:

- Enacted text
- official regulation
- court decision
- agency order
- official guidance
- legislative or treaty record
- official effective-date and status page

12.2 Legal Status

Record:

- Jurisdiction
- authority
- version
- effective date
- amendments
- transition
- current status

12.3 Proposed Law

Clearly distinguish:

- Proposal
- introduced bill
- adopted text
- effective requirement
- guidance
- enforcement action

12.4 Legal Interpretation

Legal interpretation should identify:

- Qualified author
- jurisdiction
- assumptions
- contested questions
- date
- whether the statement is advice, analysis, or official decision

12.5 Standards Evidence

For a standards claim, record:

- Issuing body
- standard number
- title
- edition
- status
- scope
- access date
- whether the standard is voluntary

- whether it has been incorporated into law or contract

12.6 Standards Summaries

An official summary may support general description.

It should not replace the standard text when exact requirements matter.

12.7 Copyrighted Standards

Where full standards text cannot be reproduced:

- Cite the official record
- quote only within lawful limits
- summarize carefully
- identify interpretation
- avoid implying access not possessed

12.8 Certification and Accreditation Status

Use official registries or issuing-body records where available.

Check:

- Scope
- site
- certificate number
- status
- expiration
- suspension
- withdrawal

12.9 Legal Crosswalks

A standards crosswalk does not establish legal equivalence unless the competent authority recognizes it.

12.10 Temporal Stability

Legal and standards claims should be rechecked before publication because:

- Text changes
- implementation dates change
- standards are revised
- certifications expire
- guidance is replaced

- court decisions alter interpretation

---

13. Policy and Normative Evidence

13.1 Fact Versus Value

Policy analysis should distinguish:

- Descriptive fact
- technical inference
- forecast
- value judgment
- institutional preference
- legal constraint
- policy recommendation

13.2 Evidence Does Not Eliminate Values

Questions such as acceptable risk, rights, distribution, and public authority contain normative judgments.

Standards Body should make those judgments visible.

13.3 Policy Option Analysis

A policy recommendation should examine:

- Problem definition
- evidence
- alternatives
- expected benefits
- expected costs
- uncertainty
- distributional effects
- implementation
- enforcement
- institutional capacity
- reversibility
- international effects
- failure modes

13.4 Stakeholder Evidence

Stakeholder experience can provide direct evidence concerning:

- Deployment impact
- access barriers
- operational failure
- rights
- burden
- local context

It should not be dismissed because it is not quantitative.

13.5 Public Opinion

Public opinion is evidence of:

- Preference
- concern
- legitimacy
- perceived impact

It is not direct evidence of technical validity.

13.6 Economic Evidence

Economic claims should identify:

- Model
- assumptions
- counterfactual
- price or cost basis
- time horizon
- sensitivity
- distribution

13.7 International Comparison

Cross-jurisdiction claims should account for:

- Legal differences
- institutional capacity
- market structure
- language
- culture
- enforcement
- data quality

14. Expert Judgment

Expert judgment is necessary where:

- Evidence is incomplete
- systems are novel
- risks are rare
- direct testing is unsafe
- interpretation requires deep domain knowledge
- decisions cannot wait for complete data

14.1 Expert Qualification

Record:

- Domain
- experience
- methods
- relevant work
- current competence
- conflicts
- institutional role

14.2 Structured Judgment

Prefer structured methods with:

- Defined question
- individual estimates before discussion
- assumptions
- evidence access
- confidence
- rationale
- aggregation rule
- disagreement record
- update process

14.3 Panel Composition

Consider diversity in:

- Technical discipline
- institutional role
- geography
- methods

- worldview
- deployment experience
- public-interest expertise

14.4 Consensus

Consensus should not be manufactured through pressure to agree.

Record:

- Areas of agreement
- substantial objections
- minority views
- confidence
- unresolved questions

14.5 Expert Elicitation

Where experts provide estimates, record:

- Target quantity
- units
- time horizon
- probability scale
- calibration support
- decomposition
- aggregation
- update date

14.6 Expert Judgment Limits

Expert status does not eliminate:

- Bias
- groupthink
- overconfidence
- reputation incentives
- incomplete access
- political conflict

14.7 Expert Judgment as Evidence

Expert judgment should be labeled as such.

It should not be silently converted into an observed fact.

---

15. Confidential and Restricted Evidence

15.1 Legitimate Uses

Confidential evidence may be necessary for:

- Held-out evaluations
- security vulnerabilities
- model weights
- system prompts
- internal incidents
- proprietary methods
- personal data
- national-security information
- contractual records
- unpublished investigations

15.2 Confidentiality Does Not Increase Quality Automatically

A claim is not stronger merely because the evidence is secret.

15.3 Confidential Evidence Requirements

Record:

- Evidence class
- owner
- source
- date
- authorized reviewers
- access method
- provenance
- limitations
- retention
- public claim supported
- review date

15.4 Independent Access

For consequential public claims, confidential evidence should normally be reviewed by a qualified independent person or body.

15.5 Public Disclosure Statement

A public report should state:

- That confidential evidence was used
- the general evidence type
- who reviewed it
- the review scope
- why it is not public
- what conclusion is supported
- what cannot be independently verified by the public

15.6 Confidentiality and Dissent

Reviewers should be able to record dissent without disclosing protected details.

15.7 Selective Disclosure

Do not provide favorable confidential evidence while withholding materially adverse evidence from reviewers.

15.8 Security Classification

Use a defined classification and handling system.

15.9 Sunset and Release

Protected evidence should have:

- Review date
- declassification or release criteria
- destruction or archive plan
- successor evidence plan

15.10 Evidence Access Log

Log:

- User
- purpose
- date
- action
- export
- modification
- revocation

---

16. Model-Generated and Tool-Assisted Evidence Work

16.1 Permitted Assistance

AI tools may assist with:

- Search-query generation
- source discovery
- document extraction
- translation support
- citation formatting
- comparison
- classification
- consistency checks
- draft generation
- data analysis under review
- code generation under testing

16.2 Not Independent Evidence

A model-generated statement is not an independent source.

It may reflect:

- Training data
- tool results
- prompt assumptions
- hallucination
- outdated information
- citation fabrication
- hidden transformation

16.3 Verification Requirement

Every material model-generated claim should be checked against:

- Primary source
- trusted data
- reproducible computation
- qualified human judgment

16.4 Citation Rule

Cite the underlying source, not the model output, unless the model interaction itself is the object of study.

16.5 Analysis Rule

Model-assisted analysis should record:

- Tool or model
- version
- prompt or procedure where material
- input data
- human review
- tests
- limitations

16.6 Sensitive Information

Do not submit protected evidence to an external model or service without authorization and appropriate controls.

16.7 Automated Extraction

Automated extraction should be sampled for:

- Omission
- transcription error
- table error
- citation mismatch
- context loss

16.8 Model-Based Scoring

See Section 10.8.

16.9 Human Accountability

The named author, reviewer, or institution remains responsible for the final work.

16.10 Disclosure

Material use of AI should be disclosed when it affects:

- Method
- reproducibility
- evidence interpretation
- confidentiality
- authorship
- public trust

---

17. Uncertainty

17.1 Uncertainty Types

Evidence reviews should consider:

- Measurement uncertainty
- sampling uncertainty
- model uncertainty
- task uncertainty
- system-identity uncertainty
- contamination uncertainty
- elicitation uncertainty
- interpretive uncertainty
- causal uncertainty
- forecast uncertainty
- legal uncertainty
- institutional uncertainty
- unknown unknowns

17.2 Quantitative Uncertainty

Where valid, report:

- Standard error
- confidence or credible interval
- distribution
- sensitivity
- scenario range
- calibration
- probability

17.3 Qualitative Uncertainty

Use structured explanation when quantification would create false precision.

17.4 Sources of Uncertainty

List the material sources separately.

Avoid one vague limitation statement.

17.5 Uncertainty and Decisions

A decision may still be justified under uncertainty.

The decision record should explain:

- What is uncertain
- consequence of delay
- consequence of error
- reversibility
- monitoring
- update triggers

17.6 Uncertainty Reduction

Identify which additional evidence would most improve the decision.

17.7 Irreducible Uncertainty

Some uncertainty may remain because:

- Systems are adaptive
- deployment is open-ended
- harmful events are rare
- evidence is protected
- future actors are unknown
- models change

17.8 False Precision

Do not assign a precise probability or score when the evidence cannot support it.

17.9 Measurement Guidance

Standards Body draws on established measurement practice that treats uncertainty as a required component of credible measurement rather than an afterthought.[^gum]

---

18. Confidence Ratings

Confidence expresses the strength of justified belief in a bounded conclusion.

18.1 Very Low Confidence

Conditions may include:

- E0 or weak E1 evidence
- major evidence gaps
- indirect or unverifiable sources

- unresolved contradictions
- unstable system identity
- severe conflict
- high uncertainty

Permitted wording:

- Speculative
- unverified
- insufficient evidence
- early signal only

18.2 Low Confidence

Conditions may include:

- Preliminary direct evidence
- limited sample
- weak replication
- important access limitations
- substantial alternative explanations

18.3 Moderate Confidence

Conditions may include:

- E2 support
- relevant direct evidence
- some corroboration
- bounded claim
- known limitations
- no decisive contrary evidence

18.4 High Confidence

Conditions may include:

- E3 evidence
- strong method
- current system identity
- independent challenge
- reproducibility or replication
- uncertainty analysis
- contrary-evidence review

18.5 Very High Confidence

Conditions should normally include:

- Extensive E3 or E4 evidence
- repeated independent support
- stable construct
- strong provenance
- low material uncertainty
- no credible unresolved contradiction
- mature operational experience

Very high confidence should be uncommon for frontier AI capability, safety, and forecasting claims.

18.6 Confidence Is Claim-Specific

A report may support:

- High confidence that a test was conducted correctly
- moderate confidence that the test measures the intended capability
- low confidence that the result predicts real-world harm

18.7 Confidence Changes

Confidence should decrease when:

- System changes
- evidence expires
- contamination is suspected
- a contrary study appears
- evaluator status changes
- an incident reveals a gap

18.8 Confidence Record

Record:

- Label
- rationale
- evidence level
- key supporting evidence
- key contrary evidence
- uncertainty
- reviewer
- date

---

19. Evidence Synthesis

19.1 Synthesis Purpose

Evidence synthesis should answer:

- What is supported?
- What is not supported?
- Where does evidence disagree?
- Why?
- What remains unknown?
- What decision follows?

19.2 Synthesis Methods

Methods may include:

- Narrative synthesis
- systematic review
- meta-analysis
- structured expert judgment
- evidence matrix
- causal analysis
- case comparison
- triangulation
- bridge study

19.3 Selection Criteria

State:

- Search scope
- inclusion
- exclusion
- date range
- source types
- languages
- security limitations

19.4 Evidence Table

A synthesis should include a table or structured register with:

- Source
- claim
- method

- system
- date
- quality
- conflict
- result
- limitation
- weight

19.5 Contrary Evidence Search

Actively search for:

- Failed replications
- negative results
- alternative explanations
- incidents
- critiques
- withdrawn work
- conflicting jurisdictions
- evaluator disagreement

19.6 Weighting

Weight should consider quality dimensions, not source count alone.

19.7 Heterogeneity

Do not combine evidence when:

- Constructs differ
- systems differ materially
- conditions differ
- score meaning differs
- legal contexts differ
- evidence status differs

19.8 No Forced Consensus

Preserve unresolved conflict.

19.9 Living Synthesis

High-change topics should use:

- Review dates

- update triggers
- status records
- source monitoring
- version history

---

20. Reproducibility, Replicability, and Independent Challenge

20.1 Reproducibility

Standards Body uses *reproducibility* for obtaining consistent computational or analytical results using the same data, code, methods, and conditions.

A reproducibility package should include, where lawful and safe:

- Data or data-access instructions
- code
- environment
- dependencies
- configuration
- model version
- protocol
- random seeds
- execution instructions
- scoring
- expected outputs
- known nondeterminism

The National Academies distinguishes computational reproducibility from replicability using new data or independent methods, a distinction adopted here because it clarifies what kind of independent support actually exists.[^nasem-repro]

20.2 Replicability

Replicability asks whether materially independent work supports the finding.

Replication may use:

- New task samples
- new environments
- another evaluator
- another implementation
- another system version
- another jurisdiction
- another language

20.3 Reperformance

Reperformance is especially relevant to:

- Audits
- scoring
- evidence extraction
- control testing
- calculations
- certification decisions

20.4 Conceptual Replication

A different method may test the same underlying construct.

This may provide stronger generalization evidence than exact repetition.

20.5 Protected Reproducibility

Held-out or confidential evidence may be reproduced through:

- Secure enclaves
- independent custodians
- controlled-access repositories
- authorized reviewer sessions
- split custody
- signed result artifacts

20.6 Reproducibility Failure

A failure to reproduce may result from:

- Missing dependencies
- model version drift
- undocumented prompts
- stochastic variation
- inaccessible data
- scoring changes
- task leakage
- analyst error

It should trigger investigation, not immediate accusation.

20.7 Replication Failure

A failed replication may reflect:

- Original error
- changed construct
- different population
- underpowered study
- protocol difference
- system change
- publication bias

20.8 Independent Challenge

When exact reproduction is impossible, qualified reviewers should still be able to challenge:

- Method
- assumptions
- evidence
- scoring
- uncertainty
- interpretation
- conflicts
- claim scope

20.9 Minimum Reproducibility Statement

Every major technical paper should state one of:

- Fully reproducible from public artifacts
- reproducible under controlled access
- partially reproducible
- not currently reproducible

and explain why.

---

21. Incident, Near-Miss, and Failure Evidence

21.1 Incident Evidence

Incident evidence may include:

- Logs
- user reports
- system outputs

- access records
- screenshots
- forensic artifacts
- internal communications
- control failures
- response records
- downstream impact

21.2 Near-Miss Evidence

Near misses can reveal:

- Weak controls
- latent pathways
- detection success
- accidental containment
- untested capability
- governance gaps

21.3 Incident Validation

Confirm:

- Event identity
- system identity
- time
- source
- impact
- causal role
- evidence integrity
- duplicate reports
- confidence

21.4 Causation

Do not attribute an incident to AI merely because AI was present.

Assess whether the system:

- Initiated
- enabled
- accelerated
- amplified
- failed to prevent
- merely coincided with

the event.

21.5 Root-Cause Evidence

Root-cause analysis should distinguish:

- Immediate cause
- contributing factors
- organizational conditions
- safeguard failure
- human action
- model behavior
- environmental condition

21.6 Incident Severity

Severity should be based on:

- Actual harm
- potential harm
- scale
- reversibility
- affected parties
- strategic consequence
- recurrence risk

21.7 Incident Confidentiality

Protect:

- Personal data
- exploitable vulnerabilities
- national-security information
- active investigation
- confidential business information

while preserving public learning where possible.

21.8 Incident Feedback

Validated incident evidence should update:

- Protocols
- task banks
- threat models
- safeguards
- standards
- evaluator scopes
- recognition

- public claims

21.9 Failure Database

`FAILURE_DATABASE.md` should record:

- Event
- evidence
- confidence
- source
- system
- failure mode
- root cause
- lessons
- affected files
- correction

---

22. Temporal Validity and Freshness

22.1 Evidence Age

Age alone does not determine quality.

A stable legal principle may remain current for decades.

A model capability result may become obsolete within weeks.

22.2 Freshness Categories

Stable

Expected to remain applicable absent formal change.

Examples:

- Foundational definitions
- established measurement principles

Slow-Changing

Review annually or after institutional change.

Examples:

- Governance structures
- standards status
- accreditation arrangements

Moderate-Change

Review quarterly or after material development.

Examples:

- evaluator landscape
- policy implementation
- deployment practices

Fast-Changing

Verify immediately before use.

Examples:

- Model availability
- system capabilities
- software versions
- legal implementation dates
- certification status
- current officeholders
- active incidents

22.3 Evidence Date Fields

Record separately:

- Event date
- data-collection date
- publication date
- access date
- verification date
- review date

22.4 Freshness Trigger

Recheck evidence after:

- Model release
- protocol revision
- legal amendment
- standard revision
- certification suspension
- major incident
- institutional reorganization
- public correction
- new replication

22.5 Historical Evidence

Historical evidence remains relevant for:

- Trend
- institutional learning
- prior commitments
- failure analysis
- precedent

It should not be presented as current status.

22.6 Stale Evidence Language

Use:

- Historical
- superseded
- current as of
- not reverified
- may no longer apply

22.7 Automated Monitoring

Automated monitoring may flag changes.

A human or authorized process should verify material status changes.

---

23. Citation and Attribution Standards

23.1 Citation Purpose

Citations should allow a reader to:

- Locate the source
- verify the claim
- understand the evidence type
- assess recency
- distinguish source from interpretation

23.2 Citation Coverage

Cite:

- Noncommon factual claims

- statistics
- legal requirements
- standards claims
- technical findings
- institutional status
- direct quotations
- historical events
- contested claims
- forecasts and expert judgments

23.3 Primary Citation

Where possible, cite the primary source directly.

23.4 Secondary Citation

A secondary source may be cited for:

- Interpretation
- synthesis
- criticism
- context
- historical account

23.5 Citation Accuracy

The cited source must support the exact claim.

Prohibited practices:

- Citation dumping
- citing a source that only mentions the topic
- citing an abstract for a detailed claim not contained there
- citing a secondary summary as though it were the original result
- citing a withdrawn source without status

23.6 Quotation

Direct quotation should:

- Be exact
- preserve context
- identify source
- comply with copyright
- avoid selective distortion

23.7 Paraphrase

Paraphrase should preserve:

- Meaning
- uncertainty
- qualifications
- scope

23.8 Page and Section References

Use page, section, table, figure, paragraph, or clause references where precision matters.

23.9 Web Sources

Record:

- Title
- institution
- publication or update date
- URL
- access date
- status where relevant

23.10 Dynamic Pages

For changing pages, preserve:

- Archived copy where lawful
- screenshot or record
- access date
- change note

23.11 Data and Code

Cite:

- Dataset
- version
- repository
- code commit
- license
- access date
- preprocessing

23.12 Standards

Cite the exact standard identifier, edition, title, and status.

23.13 Legal Sources

Cite the authoritative legal text and jurisdiction.

23.14 Personal Communication

Use only with permission and clear labeling.

23.15 AI Tool Attribution

Do not cite an AI assistant as the source of an underlying factual claim.

Disclose material AI assistance separately.

---

24. Public Claims and Communication

24.1 Public-Claim Minimum

A consequential public claim should communicate:

- Subject
- scope
- evidence
- conditions
- uncertainty
- limitations
- date
- status
- responsible institution

24.2 Headline Discipline

Headlines should not exceed the body evidence.

24.3 Summary Discipline

Executive summaries should preserve:

- Qualifications
- uncertainty

- contrary evidence
- claim boundaries

24.4 Visualizations

Charts should include:

- Units
- source
- date
- scale
- uncertainty
- missing data
- comparable categories

Avoid:

- Truncated axes that distort interpretation
- combining incomparable protocols
- decorative precision
- omitted confidence intervals where material

24.5 Rankings

Do not publish rankings when:

- Construct validity is weak
- results are not comparable
- uncertainty overlaps materially
- tasks are easily gamed
- the ranking would create harmful incentives

24.6 Safety Language

Avoid:

- Safe
- risk-free
- harmless
- failsafe
- secure

as absolute labels.

24.7 Legal Language

Avoid:

- Approved
- compliant
- certified
- accredited
- official

unless the relevant authority, scope, and status are clear.

24.8 Confidential-Evidence Language

State the role and limits of nonpublic evidence.

24.9 Preliminary Evidence

Prominently label preliminary findings.

24.10 Corrections

Public corrections should reach the same channels as the original claim where feasible.

---

25. Evidence Contradiction and Dispute

25.1 Contradiction Types

Evidence may conflict because of:

- Different system versions
- different constructs
- different task samples
- different elicitation
- different environments
- different scoring
- different access
- different time
- error
- fraud
- genuine heterogeneity

25.2 Contradiction Procedure

1. Confirm the claims are actually comparable.
2. Verify source authenticity.
3. inspect system and protocol versions.
4. compare methods.
5. compare samples.
6. compare uncertainty.
7. examine conflicts.
8. seek reperformance or replication.
9. preserve unresolved disagreement.
10. update confidence.

25.3 Disputed Status

A claim may be labeled:

- Under review
- disputed
- partially substantiated
- inconclusive
- superseded

25.4 No Majority-by-Citation

Ten weak derivative sources do not outweigh one strong contradictory primary source automatically.

25.5 Institutional Disagreement

When institutions disagree, document:

- Mandates
- evidence access
- methods
- assumptions
- thresholds
- legal context
- dissent

25.6 Fraud or Misconduct Allegation

Do not make misconduct claims without strong evidence and due process.

Distinguish:

- Error
- negligence
- selective reporting
- conflict
- fabrication
- fraud

---

26. Corrections, Retractions, and Withdrawals

26.1 Correction Trigger

Correct when:

- A factual statement is wrong
- a citation does not support a claim
- material evidence was omitted
- system identity was incorrect
- a calculation is wrong
- status changed
- wording overstates evidence

26.2 Correction Types

Minor Correction

Does not change the central conclusion.

Material Correction

Changes scope, confidence, evidence level, or decision implication.

Retraction or Withdrawal

The work should no longer support the claim.

26.3 Correction Record

Record:

- Original
- corrected text
- reason
- date
- responsible owner

- affected documents
- effect on conclusion

26.4 Preservation

Do not silently delete the prior version where a material public record exists.

26.5 Propagation

Review dependent:

- Website pages
- foundation papers
- standards proposals
- source registries
- public statements
- partnership materials

26.6 Timeliness

Correct promptly after verification.

26.7 Incentives

Correction should be treated as an integrity function, not automatic institutional failure.

Concealment and repeated negligence should receive different treatment.

---

27. Evidence Security and Data Governance

27.1 Evidence Security Objectives

Protect:

- Confidentiality
- integrity
- availability
- authenticity
- provenance
- authorized use

27.2 Access

Use:

- Role-based access
- least privilege
- multifactor authentication
- access review
- revocation
- logging

27.3 Storage

Define:

- Approved environment
- encryption
- backup
- geographic or legal constraints
- retention
- deletion

27.4 Transfer

Use:

- Secure channels
- recipient verification
- transfer logs
- file integrity
- onward-disclosure rules

27.5 Sensitive Personal Data

Minimize and protect personal data.

27.6 Evidence Tampering

Controls may include:

- Hashes
- signatures
- immutable logs
- witnessed transfer
- read-only archives
- independent copies

27.7 Provenance Interchange

Machine-readable provenance can draw on general models such as the W3C PROV data model, which represents entities, activities, agents, and their relationships for provenance interchange. [^w3c-prov]

27.8 Incident Response

Evidence compromise should trigger:

- Containment
- impact assessment
- notification
- result review
- correction
- re-evaluation
- control improvement

---

28. Evidence Review Lifecycle

28.1 Question

Define the decision and claim.

28.2 Plan

Define:

- Evidence standard
- source scope
- methods
- reviewers
- security
- timeline

28.3 Collect

Gather primary, secondary, supporting, and contrary evidence.

28.4 Authenticate

Verify source, version, status, and provenance.

28.5 Assess

Evaluate quality dimensions.

28.6 Synthesize

Integrate evidence and preserve conflict.

28.7 Assign

Assign evidence level and confidence.

28.8 Review

Conduct relevant technical, methodological, legal, security, or independent review.

28.9 Decide

State conclusion, limitations, and decision implication.

28.10 Publish or Restrict

Apply disclosure classification.

28.11 Monitor

Track new evidence, incidents, and status changes.

28.12 Correct

Revise, suspend, withdraw, or retire as needed.

---

29. Evidence Governance

29.1 Governing Functions

Standards Body should maintain governance for:

- Evidence definitions
- source selection
- claim approval
- confidence ratings
- confidential access

- corrections
- disputes
- model-assisted research
- citation practice
- evidence retention
- periodic review

29.2 Roles

Claim Owner

Responsible for wording, evidence selection, limitations, and correction.

Research Lead

Responsible for evidence collection and synthesis.

Domain Reviewer

Reviews subject-matter validity.

Methodological Reviewer

Reviews design, measurement, and inference.

Source Reviewer

Checks source authenticity, citation fit, and status.

Security Reviewer

Reviews protected evidence and handling.

Independent Reviewer

Challenges material claims and contrary-evidence handling.

Publication Owner

Approves public wording and status.

Records Custodian

Maintains evidence packages and version history.

29.3 Separation

For high-stakes claims, the person making the original analysis should not be the only person approving:

- Evidence level
- confidence
- public wording
- correction
- withdrawal

29.4 Conflict Disclosure

Reviewers should disclose material:

- Financial
- organizational
- intellectual
- professional
- political
- personal conflicts

29.5 Evidence Committee

A future Evidence Standards Committee may:

- Maintain this file
- resolve disputed evidence classifications
- review E4 claims
- approve high-stakes public language
- supervise corrections
- publish guidance
- audit evidence practice

29.6 Appeals

A contributor or affected party may appeal:

- Evidence exclusion
- confidence rating
- public wording
- correction decision
- source characterization
- conflict handling

29.7 External Challenge

Canonical claims should support external critique through:

- Public source lists
- methods
- claim registers
- comment channels
- correction processes
- controlled access where necessary

29.8 Evidence Audit

Periodic evidence audits should sample:

- Citation accuracy
- source quality
- freshness
- primary-source use
- contrary evidence
- confidential-evidence governance
- AI-tool disclosure
- correction propagation

---

30. Evidence Maturity Model

Level 0: Assertion-Led

Characteristics:

- Claims precede evidence
- sources are absent or decorative
- authority relies on branding or prestige
- no correction process

Level 1: Source-Documented

Characteristics:

- Material claims have sources
- primary and secondary evidence are distinguished
- dates and versions are recorded
- preliminary status is visible

Level 2: Quality-Assessed

Characteristics:

- Evidence dimensions are reviewed
- claim types have minimum requirements
- contrary evidence is sought
- uncertainty and confidence are stated
- model-generated work is verified

Level 3: Independently Reviewed

Characteristics:

- Consequential claims receive methodological and domain review
- confidential evidence is governed
- reproducibility status is declared
- corrections propagate
- evidence packages are maintained

Level 4: Decision-Grade

Characteristics:

- Evidence standards are tied to consequence
- E4 packages support specified decisions
- false-positive and false-negative costs are assessed
- security and chain of custody are strong
- appeals and expiration exist

Level 5: Adaptive Evidence Institution

Characteristics:

- Evidence is monitored continuously
- stale claims are automatically flagged
- incidents update evidence standards
- public claim registers are current
- evidence audits occur
- standards evolve based on measured outcomes

---

31. Implementation Pathway

Phase 1: Claim Inventory

Identify all major claims in:

- Foundation papers
- website copy
- standards proposals
- public descriptions
- institutional documents

Assign claim IDs and owners.

Phase 2: Source Registry

Create `SOURCES.md` with:

- Source ID
- title
- author or institution
- source tier
- date
- version
- subject
- URL or location
- access
- status
- related claims

Phase 3: Evidence-Level Pilot

Select twenty material claims.

Assign:

- Evidence level
- confidence
- supporting evidence
- contrary evidence
- review date

Compare reviewer consistency.

Phase 4: Citation Audit

Audit existing canonical documents for:

- Unsupported facts
- outdated sources
- secondary-source substitution
- citation mismatch
- missing status
- unsupported authority

Phase 5: Evidence Package Pilot

Build one complete evidence package for a high-stakes capability claim.

Phase 6: Confidential-Evidence Procedure

Establish:

- Classification
- access
- independent review
- public disclosure statement
- retention
- incident response

Phase 7: Correction System

Create:

- Correction request
- triage
- severity
- approval
- publication
- propagation
- archive

Phase 8: Research Method Integration

Apply this standard through `RESEARCH_METHODOLOGY.md`.

Phase 9: Standards Process Integration

Require evidence-readiness review in `STANDARDS_DEVELOPMENT_PROCESS.md`.

Phase 10: Public Evidence Register

Publish bounded information concerning:

- Canonical claims
- evidence level
- confidence
- review date
- correction status

Security-sensitive evidence remains controlled.

---

32. Evidence Standards Scorecard

Dimension	Core question
Claim definition	Is the claim specific, bounded, and testable?
Claim owner	Is someone responsible for evidence and correction?
Decision link	Is the intended use clear?
Relevance	Does the evidence address the actual claim?
Directness	Is the inference distance understood?
Construct validity	Does the evidence measure the intended property?
Internal validity	Does the design support the conclusion?
External validity	Is generalization justified?
Reliability	Is the result consistent under equivalent conditions?
Provenance	Can origin and history be traced?
Authenticity	Is the evidence genuine and unaltered?
Completeness	Are failures, exclusions, and contrary results included?
Independence	Are producer interests and external challenge clear?
Recency	Is the evidence current for the object and decision?
Representativeness	Does the evidence cover the relevant population or task universe?
Reproducibility	Can the work be recomputed or re-executed?
Replicability	Does independent evidence support the finding?
Security	Was the evidence protected from leakage and tampering?
Uncertainty	Are material uncertainties stated or quantified?
Contrary evidence	Was disconfirming evidence actively considered?
Conflict	Are material conflicts disclosed and managed?
Source quality	Are authoritative primary sources used where available?
Citation accuracy	Does each citation support the exact claim?

Dimension	Core question
Confidential evidence	Is nonpublic evidence independently governed?
Model assistance	Was AI-assisted work verified and disclosed where material?
Confidence	Is the confidence rating justified?
Evidence level	Is the evidence level appropriate and documented?
Public wording	Does the public claim remain within evidence?
Status	Is the evidence current, corrected, suspended, or withdrawn?
Expiration	Is re-review or expiration defined?
Correction	Can error be corrected and propagated?
Proportionality	Is the burden appropriate to the consequence?
Decision readiness	Is the package sufficient for the specified use?

32.1 Critical Failures

The following normally prevent an E3 or E4 classification:

- Unidentified evaluated object
- material citation mismatch
- unaddressed task compromise
- hidden exclusion of adverse results
- no uncertainty for a sensitive quantitative claim
- decisive conflict without mitigation
- unsupported legal or certification claim
- material contrary evidence ignored
- unverifiable confidential evidence without qualified review
- model-generated claim treated as independent evidence
- stale evidence applied to a materially changed system

32.2 No Universal Evidence Score

Do not convert the scorecard into one universal numerical rating.

A critical failure should remain visible.

---

33. Claim and Evidence Register Template

Project:
Record owner:
Review date:

Field	Entry
Claim ID	
Exact claim	
Claim type	
Intended use	
Consequence if wrong	
Subject or evaluated object	
Scope	
Time period	
Evidence level	
Confidence	
Primary evidence	
Supporting evidence	
Contrary evidence	
Assumptions	
Uncertainty	
Conflicts	
Security status	
Reviewer	
Public wording	
Expiration	
Status	
Correction history	

---

34. Evidence Source Record Template

Source ID:
Title:
Author or institution:
Source type:
Source tier:
Publication or creation date:
Version:
Status:
URL or controlled location:
Access date:

Relevance

Claims Supported

Claims Challenged

Method

Evaluated Object

Provenance

Independence and Conflicts

Limitations

Security Classification

Review Notes

Superseding Source

---

35. Evidence Quality Assessment Template

Evidence object:
Claim:
Reviewer:
Date:

Rate each dimension as:

- Strong
- adequate
- limited
- weak
- not assessed

Dimensions

- Relevance

- directness
- construct validity
- internal validity
- external validity
- reliability
- provenance
- authenticity
- completeness
- independence
- recency
- representativeness
- reproducibility
- replicability
- security
- uncertainty
- decision relevance
- conflict exposure
- gaming susceptibility
- legal and ethical fitness

Critical Limitations

Contrary Evidence

Overall Weight

Evidence Level Recommendation

Additional Evidence Needed

---

36. Confidential Evidence Review Template

Evidence package ID:
Classification:
Owner:
Claim supported:
Review purpose:

Authorized Reviewers

Reviewer Qualifications

Access Method

Evidence Types

Provenance and Custody

Completeness Statement

Material Contrary Evidence

Security Limitations

Independent Findings

Publicly Supportable Conclusion

Publicly Unverifiable Components

Retention and Review Date

Release, Destruction, or Archive Plan

---

37. Confidence Assessment Template

Claim:
Evidence level:
Reviewer:
Date:

Supporting Evidence

Contrary Evidence

Material Uncertainty

System or Context Stability

Reproducibility and Replication

Conflict Assessment

Confidence Label

- Very low
- low
- moderate
- high
- very high

Rationale

Conditions That Would Increase Confidence

Conditions That Would Decrease Confidence

Review Trigger

---

38. Correction and Withdrawal Template

Record or publication:
Version:
Claim ID:
Date identified:
Reporter:

Issue

Evidence

Severity

- Minor
- material
- invalidating

Original Wording or Result

Corrected Wording or Result

Effect on Evidence Level

Effect on Confidence

Effect on Decisions

Affected Documents

Public Notice

Approval

Completion Date

---

39. Decision-Grade Evidence Package Template

A. Decision

- Decision owner
- authority
- question
- alternatives
- consequence of error
- timing

B. Claim

- Exact wording
- scope
- system
- period
- evidence standard

C. Technical Evidence

- Protocol
- result
- raw records
- uncertainty
- reproducibility
- integrity

D. Organizational Evidence

- Governance
- controls
- competence
- operations
- incidents

E. Independent Review

- Mandate
- reviewers
- access
- findings
- dissent
- conflicts

F. Contrary Evidence

- Sources
- findings
- resolution
- remaining disagreement

G. Legal and Policy Context

- Jurisdiction
- applicable requirements

- authority
- normative considerations

H. Evidence Rating

- Evidence level
- confidence
- rationale
- expiration

I. Decision Implications

- Supported decision
- conditions
- monitoring
- appeal
- reversal

J. Public Record

- Public claim
- confidential-evidence statement
- limitations
- status
- correction channel

---

40. Canonical Standards Body Positions

Standards Body adopts the following working positions.

1. Evidence quality is claim-specific.
2. Evidence sufficiency is decision-specific.
3. A claim should not exceed the scope, certainty, recency, or authority of its evidence.
4. Primary sources should anchor legal, standards, institutional-status, and direct technical claims when available.
5. Secondary sources are valuable for synthesis and criticism but should not replace authoritative primary material without reason.
6. Search snippets, unsourced aggregators, and model-generated summaries are not sufficient evidence for material claims.

7. Peer review increases confidence in process but does not guarantee correctness or replicability.
8. Quantitative evidence is not automatically stronger than qualitative evidence.
9. Qualitative evidence should be collected and analyzed systematically.
10. Developer evidence may be direct and important but requires conflict-aware interpretation and external challenge for consequential claims.
11. External evidence is not automatically independent evidence.
12. Confidential evidence can support a public claim only under stronger governance and review.
13. Confidentiality should not be used to launder unsupported conclusions.
14. Model-generated content may assist analysis but is not an independent factual source.
15. Human or institutional authors remain accountable for model-assisted work.
16. Capability claims should identify the system, protocol, conditions, resources, and limitations.
17. A failed capability evaluation normally supports "not demonstrated," not "absent."
18. Safety claims should be replaced by bounded risk and safeguard claims.
19. Security claims should state the threat model.
20. Comparative claims require comparable objects, protocols, resources, and conditions.
21. Causal claims require causal evidence or explicit causal reasoning.
22. Forecasts should be labeled as forecasts and include time horizon, probability, and information date.
23. Organizational policies are evidence of documented intent, not automatic evidence of effective practice.
24. Certification is evidence only within the certification scheme and scope.
25. Accreditation is evidence of conformity-assessment competence within scope, not universal institutional quality.
26. Legal and regulatory claims should be checked against current authoritative sources.
27. Policy recommendations should distinguish technical evidence from value and authority judgments.
28. Contrary evidence should be actively sought and preserved.
29. Source count should not substitute for source quality.

30. Evidence contradiction should be investigated before results are averaged or one source is dismissed.
31. High-stakes claims should use evidence portfolios rather than one metric.
32. Material failed and excluded evaluation runs should be documented.
33. Elicitation conditions are part of capability evidence.
34. System configuration is part of result identity.
35. Evidence should carry uncertainty.
36. Very high confidence should be rare in frontier AI claims.
37. Evidence should expire or be reviewed after material change.
38. Reproducibility status should be declared for major technical work.
39. Replication and independent challenge should be proportionate to consequence.
40. Provenance should be preserved in human-readable and, where useful, machine-readable form.
41. Incident and near-miss evidence should update protocols, standards, and safeguards.
42. Public summaries should preserve material limitations.
43. Headlines and visualizations should not overstate the underlying evidence.
44. Rankings should not be published when comparability and construct validity are weak.
45. Corrections should be visible, timely, and propagated.
46. Withdrawal is appropriate when evidence no longer supports the claim.
47. Evidence governance should include conflicts, appeals, security, and records.
48. Decision-grade evidence should include the consequence of false positives and false negatives.
49. Evidence burden should remain proportionate and should not unnecessarily exclude smaller actors.
50. Standards Body should evaluate its own evidence practices through recurring audits and public correction.

---

41. Relationship to the Eight Foundations

Foundation 1: Dynamic Evaluation Protocols

This file defines the evidence required to justify protocol design, change, comparison, and retirement.

Foundation 2: Held-Out Evaluations

This file defines evidentiary treatment of protected tasks, confidential findings, provenance, chain of custody, and compromise.

Foundation 3: High-Stakes Capability Evaluation

This file defines stronger evidence levels, confidence, decision-grade packages, and absence-claim limits.

Foundation 4: Independent Expert Review

This file defines what reviewers should inspect and how independent findings affect evidence weight.

Foundation 5: Third-Party Auditor Ecosystem

This file distinguishes testing, evaluation, audit, certification, accreditation, and competence evidence.

Foundation 6: Progressive Standards and Requirements

This file supplies evidence-readiness requirements for moving from research to guidance, standards, procurement, or binding rules.

Foundation 7: Incentives and Prestige

This file prevents evidence metrics, recognition, and correction systems from rewarding superficial claims.

Foundation 8: Global Interoperability

This file defines the metadata, provenance, confidence, status, and recognition information needed for evidence portability.

---

42. Final Evidence Position

Standards Body is an evidence institution before it is a standards institution.

A standards project that cannot distinguish:

- Direct evidence from interpretation
- evaluation from assurance
- preliminary findings from decision-grade findings
- absent from not demonstrated
- confidential evidence from unverifiable assertion
- legal authority from technical recommendation
- reproducibility from repetition
- confidence from certainty
- correction from concealment

cannot credibly develop standards for frontier AI.

The purpose of evidence standards is not to create a ritual of citations or an appearance of scientific caution.

It is to make claims reviewable.

Every consequential claim should allow a qualified reader to determine:

- What exactly is being claimed?
- Which object and version does it concern?
- What evidence supports it?
- What evidence challenges it?
- How was the evidence produced?
- Who had access?
- Who reviewed it?
- What conflicts exist?
- What uncertainty remains?
- How current is it?
- What decision can it support?
- When should it be revisited?
- Who will correct it if it is wrong?

The defining evidence rule of Standards Body is:

> **Make the claim bounded, make the evidence traceable, make the uncertainty visible, and make correction possible.**

References and Research Basis

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

[^nist-genai]: National Institute of Standards and Technology, **Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile**, NIST AI 600-1, 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

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

[^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

[^nist-airc]: National Institute of Standards and Technology, **AI RMF Core**, NIST AI Resource Center. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

[^iso-17025]: International Organization for Standardization, **ISO/IEC 17025:2017, General Requirements for the Competence of Testing and Calibration Laboratories**. https://www.iso.org/ISO-IEC-17025-testing-and-calibration-laboratories.html

[^iso-17029]: International Organization for Standardization, **ISO/IEC 17029:2019, General Principles and Requirements for Validation and Verification Bodies**. https://casco.iso.org/bodies.html

[^iso-17020]: International Organization for Standardization, **ISO/IEC 17020:2026, Requirements for Bodies Performing Inspection**. https://www.iso.org/standard/17020

[^iso-17065]: International Organization for Standardization, **ISO/IEC 17065:2012, Requirements for Bodies Certifying Products, Processes and Services**. https://www.iso.org/standard/46568.html

[^iso-17011]: International Organization for Standardization, **ISO/IEC 17011:2017, Requirements for Accreditation Bodies Accrediting Conformity Assessment Bodies**. https://www.iso.org/standard/67198.html

[^gum]: Joint Committee for Guides in Metrology, **JCGM 100:2008, Evaluation of Measurement Data, Guide to the Expression of Uncertainty in Measurement**. https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf

[^nasem-repro]: National Academies of Sciences, Engineering, and Medicine, **Reproducibility and Replicability in Science**, 2019. https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science

[^w3c-prov]: World Wide Web Consortium, **PROV-DM: The PROV Data Model** and **PROV-O: The PROV Ontology**, 2013. https://www.w3.org/TR/prov-dm/ and https://www.w3.org/TR/prov-o/

[^oecd-haip]: OECD, **Hiroshima AI Process Reporting Framework**. https://oecd.ai/en/hiroshima

[^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 canonical Standards Body evidence standard. Defines foundational propositions, evidence objects, claim architecture, evidence levels, quality dimensions, source hierarchy, claim-specific requirements, technical and organizational evidence, legal and policy evidence, expert judgment, confidential evidence, model-assisted work, uncertainty, confidence, synthesis, reproducibility, incident evidence, freshness, citations, public claims, disputes, corrections, security, lifecycle, governance, maturity, implementation, scorecards, operational templates, canonical positions, foundation interfaces, and primary research basis.

Status: Approved foundational source.