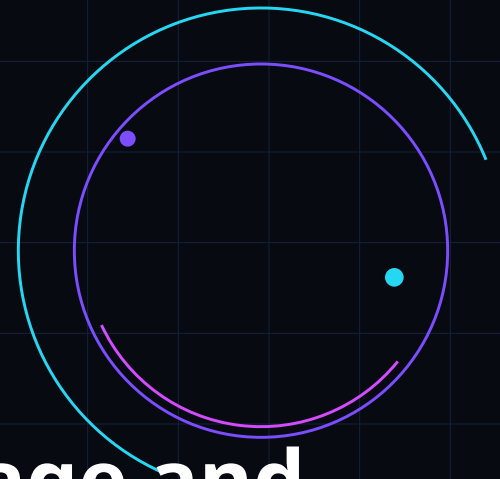




Applied Natural Language and LLM Evaluation Standard

Defines grounded language-task evaluation, retrieval, citation, correctness, uncertainty, robustness, latency, and operational use.

MYTHOS-AI-0003

Candidate Mythos Standard / 2.0.0-candidate

Published 2026-07-26 | Review 2026-10-26 | Public

Mythos Systems

Permanent Identity and Edition Record

Permanent document ID	MYTHOS-AI-0003
Canonical title	Applied Natural Language and LLM Evaluation Standard
Publication class	Standards & Benchmarks
Primary domain	MYTHOS-AI - Artificial Intelligence & Agentic Systems
Version	2.0.0-candidate
Status	Candidate Mythos Standard
Publication date	2026-07-26
Scheduled review	2026-10-26
Publisher	Mythos Systems
Owner	Aaron Stovall
Classification	Public
Prior edition	1.0.0-candidate / 2026-07-24
Supersession	This edition supersedes the prior candidate while preserving the permanent identity and historical source.

Identity doctrine

The permanent document ID is immutable. Production sequence labels are not part of the public identity. Atomic standards are authoritative; portfolios, workbooks, and machine-readable exports are generated views.

Normative Language and Applicability

The key words MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, MAY, and OPTIONAL indicate the strength of provisions in this Mythos standard. The publication is a Mythos-authored candidate standard and does not claim approval by the cited external authorities.

Applicability rule

Every control is applicable unless the organization documents a specific architectural or lifecycle reason that the subject is absent. N/A decisions MUST name the decision owner, evidence, affected scope, review date, and triggers that invalidate the exclusion.

Conformance evidence hierarchy

Direct deterministic proof: schemas, tests, signatures, state reconciliation, access decisions, build or runtime evidence.

Reproducible technical evidence: benchmark fixtures, traces, artifacts, profiles, logs, metrics, and environment manifests.

Independent assessment: qualified human or separate evaluator using an explicit rubric and disclosed uncertainty.

Attestation or supplier claim: informative unless independently verified and current for the deployed configuration.

Demonstration or narrative: insufficient by itself for a conforming score above 2.

Critical-control doctrine

Critical controls cannot be averaged away. A failed or stale critical control constrains the maximum conformance level regardless of the aggregate score.

Scope, Audience, and Engineering Principles

Purpose. Defines grounded language-task evaluation, retrieval, citation, correctness, uncertainty, robustness, latency, and operational use.

In-scope systems. language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output.

Primary audience. AI evaluators, language-system engineers, RAG teams, product teams, domain experts, safety teams, and quality engineers.

Required engineering principles

Model capability and complete system behavior **MUST** be measured separately.

Human, legal, production, financial, identity, and governance authority **MUST** remain external to model confidence or conversational fluency.

Evaluation **MUST** include expected, prohibited, adversarial, degraded, recovery, and lifecycle-change conditions.

Source authority, data rights, permissions, versions, and freshness **MUST** be visible in evidence.

Critical outcomes **MUST** be supported by deterministic validators or independent review wherever feasible.

The organization **MUST** preserve rollback, revocation, correction, deletion, and retirement paths.

Optimization for score, latency, throughput, tokens, or cost **MUST NOT** hide safety, quality, privacy, accessibility, or reliability regressions.

Exclusions

This standard does not certify a model, provider, framework, benchmark, dataset, or product. Conformance is limited to the declared system, use case, version, environment, evidence period, and assessment scope.

Conformance and Scoring Model

Level	Weighted threshold	Critical-control condition	Meaning
No conformance	< 50	Any state	Materially incomplete, unverified, or unsafe.
Foundation	>= 50	No critical control scored 0	Defined baseline with partial implementation and evidence.
Operational	>= 65	All applicable critical controls >= 3	Implemented and operationally tested.
Assured	>= 80	All applicable critical controls >= 4	Independently verified with adversarial and recovery evidence.
Continuously Assured	>= 90	All applicable critical controls = 5	Continuous regression, monitoring, freshness, and incident learning.

Weighted score

For applicable controls: weighted contribution = (control score / 5) x control weight. N/A controls are removed from the denominator only after an approved applicability decision. The final score is normalized to 100.

A conformance claim MUST include scope, system and model versions, assessment date, evidence window, evaluator identity, scores by family, critical-control status, unresolved findings, exceptions, limitations, and next review.

Control Score Interpretation

Score	Maturity	Required evidence state
0	Not implemented	Not implemented: the assessed control is absent, materially unknown, or contradicted by evidence.
1	Initial	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Nonconformity classes

Critical: credible risk of serious harm, unauthorized high-impact action, material security or privacy failure, false assurance, or inability to recover.

Major: requirement absent or ineffective across a meaningful part of scope, with no sufficient compensating control.

Minor: localized or low-impact deficiency that does not invalidate the overall control objective.

Observation: improvement opportunity or emerging risk without current nonconformity.

Score validity

A score expires when evidence exceeds its approved freshness period or a material change invalidates the assessment. Current operation is not inferred from a historic assessment.

Standard Architecture and Control Model

01	Govern	Scope, ownership, authority, policy, impacts, risk tolerance, exceptions, and lifecycle decisions.
02	Map	System boundaries, actors, models, data, prompts, tools, interfaces, populations, dependencies, and failure domains.
03	Measure	Representative and hidden evaluation, deterministic validation, human or model judging, uncertainty, performance, and cost.
04	Manage	Release gates, least privilege, monitoring, incident response, correction, rollback, revocation, deletion, and retirement.
05	Prove	Traceable evidence bundles connecting claims to configurations, tests, artifacts, effects, decisions, and user outcomes.

Assessment unit

The assessment unit is the declared AI system or workflow, not the model name alone. It includes prompts, retrieval, memory, tools, policies, interface, evaluators, infrastructure, human oversight, and operating environment.

Benchmark Dimensions and Weights

Dimension	Weight	Assessment emphasis
Task and dataset integrity	18%	Control effectiveness, evidence quality, critical failures, operational limits, and lifecycle behavior for task and dataset integrity.
Correctness and grounding	25%	Control effectiveness, evidence quality, critical failures, operational limits, and lifecycle behavior for correctness and grounding.
Uncertainty and structure	14%	Control effectiveness, evidence quality, critical failures, operational limits, and lifecycle behavior for uncertainty and structure.
Population and language coverage	10%	Control effectiveness, evidence quality, critical failures, operational limits, and lifecycle behavior for population and language coverage.
Adversarial robustness	12%	Control effectiveness, evidence quality, critical failures, operational limits, and lifecycle behavior for adversarial robustness.
Human and automated grading	10%	Control effectiveness, evidence quality, critical failures, operational limits, and lifecycle behavior for human and automated grading.
Operational performance	6%	Control effectiveness, evidence quality, critical failures, operational limits, and lifecycle behavior for operational performance.
Lifecycle	5%	Control effectiveness, evidence quality, critical failures, operational limits, and lifecycle behavior for lifecycle.

Benchmark scores **MUST** be reported by dimension and by critical-control status. A single aggregate ranking **MUST NOT** be used to conceal weak safety, authority, privacy, recovery, or evidence performance.

Contents and Control Index

[Permanent identity and edition record](#)

[Normative language and applicability](#)

[Scope, audience, and engineering principles](#)

[Conformance and scoring model](#)

[Standard architecture and control model](#)

[Benchmark dimensions and weights](#)

[Control summary matrix](#)

MYTHOS-AI-0003-C01 - Language Task and Intended-Use Profile

MYTHOS-AI-0003-C02 - System Boundary and Configuration Identity

MYTHOS-AI-0003-C03 - Corpus, Source Authority, and Rights

MYTHOS-AI-0003-C04 - Evaluation Set Design and Holdout Protection

MYTHOS-AI-0003-C05 - Prompt, Template, and Context Reproducibility

MYTHOS-AI-0003-C06 - Correctness and Task-Specific Scoring

MYTHOS-AI-0003-C07 - Factuality, Grounding, and Claim Support

MYTHOS-AI-0003-C08 - Retrieval and Reranking Evaluation

MYTHOS-AI-0003-C09 - Citation Accuracy and Source Presentation

MYTHOS-AI-0003-C10 - Uncertainty, Calibration, and Abstention

MYTHOS-AI-0003-C11 - Structured Output and Schema Conformance

MYTHOS-AI-0003-C12 - Multilingual and Cross-Cultural Evaluation

MYTHOS-AI-0003-C13 - Bias, Accessibility, and Affected-Population Slices

MYTHOS-AI-0003-C14 - Prompt Injection and Adversarial Robustness

MYTHOS-AI-0003-C15 - Long-Context and Position-Sensitivity Evaluation

MYTHOS-AI-0003-C16 - Human Evaluation and Judge Validation

MYTHOS-AI-0003-C17 - Latency, Throughput, Cost, and Resource Evaluation

MYTHOS-AI-0003-C18 - Production Monitoring, Drift, and Regression

MYTHOS-AI-0003-C19 - Records, Reassessment, and Retirement

Control Summary Matrix

Control	Requirement	Family	Wt.	Crit.
C01	Language Task and Intended-Use Profile	Task and dataset integrity	6	YES
C02	System Boundary and Configuration Identity	Task and dataset integrity	6	-
C03	Corpus, Source Authority, and Rights	Task and dataset integrity	6	-
C04	Evaluation Set Design and Holdout Protection	Correctness and grounding	5	YES
C05	Prompt, Template, and Context Reproducibility	Correctness and grounding	5	-
C06	Correctness and Task-Specific Scoring	Correctness and grounding	5	YES
C07	Factuality, Grounding, and Claim Support	Correctness and grounding	5	YES
C08	Retrieval and Reranking Evaluation	Correctness and grounding	5	YES
C09	Citation Accuracy and Source Presentation	Uncertainty and structure	5	YES
C10	Uncertainty, Calibration, and Abstention	Uncertainty and structure	5	YES

Control Summary Matrix - Continued

Control	Requirement	Family	Wt.	Crit.
C11	Structured Output and Schema Conformance	Uncertainty and structure	4	YES
C12	Multilingual and Cross-Cultural Evaluation	Population and language coverage	5	-
C13	Bias, Accessibility, and Affected-Population Slices	Population and language coverage	5	-
C14	Prompt Injection and Adversarial Robustness	Adversarial robustness	6	YES
C15	Long-Context and Position-Sensitivity Evaluation	Adversarial robustness	6	-
C16	Human Evaluation and Judge Validation	Human and automated grading	5	YES
C17	Latency, Throughput, Cost, and Resource Evaluation	Human and automated grading	5	-
C18	Production Monitoring, Drift, and Regression	Operational performance	6	-
C19	Records, Reassessment, and Retirement	Lifecycle	5	-

Weight integrity

The 19 applicable control weights sum to 100. Family weights match the benchmark dimension model. Critical controls are highlighted in the atomic control pages and assessment workbook.

MYTHOS-AI-0003-C01

Language Task and Intended-Use Profile

Family: Task and dataset integrity | Weight: 6 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain language task and intended-use profile for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. The control must identify accountable ownership, authoritative inputs, expected outputs, operating limits, dependencies, failure behavior, evidence, and reassessment triggers. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that language task and intended-use profile is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for language task and intended-use profile covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

The control must identify accountable ownership, authoritative inputs, expected outputs, operating limits, dependencies, failure behavior, evidence, and reassessment triggers.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C01

Language Task and Intended-Use Profile

Family: Task and dataset integrity | Weight: 6 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for language task and intended-use profile.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

- approved control design or operating contract with accountable owner
- versioned configuration, schema, policy, model, prompt, dataset, or architecture record
- reproducible test plan with expected and observed results
- traceable decision, exception, approval, and remediation records
- negative, boundary, degraded, and recovery test results
- operational telemetry and current-state verification

Required metrics

- applicable control coverage
- critical-failure count
- evidence freshness
- open nonconformities

Score anchors

0	Not implemented: Language Task and Intended-Use Profile is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for language task and intended-use profile.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C02

System Boundary and Configuration Identity

Family: Task and dataset integrity | Weight: 6 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain system boundary and configuration identity for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. The inventory must identify versions, owners, suppliers, deployment locations, interfaces, dependencies, trust boundaries, and supported lifecycle states. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that system boundary and configuration identity is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for system boundary and configuration identity covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

The inventory must identify versions, owners, suppliers, deployment locations, interfaces, dependencies, trust boundaries, and supported lifecycle states.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C02

System Boundary and Configuration Identity

Family: Task and dataset integrity | Weight: 6 | Critical: NO

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for system boundary and configuration identity.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner
versioned configuration, schema, policy, model, prompt, dataset, or architecture record
reproducible test plan with expected and observed results
traceable decision, exception, approval, and remediation records
negative, boundary, degraded, and recovery test results
operational telemetry and current-state verification

Required metrics

applicable control coverage
critical-failure count
evidence freshness
open nonconformities

Score anchors

0	Not implemented: System Boundary and Configuration Identity is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for system boundary and configuration identity.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C03

Corpus, Source Authority, and Rights

Family: Task and dataset integrity | Weight: 6 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain corpus, source authority, and rights for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Authority must be enforceable outside the model through stable identities, least privilege, scoped credentials, separation of duties, revocation, and auditable approval. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that corpus, source authority, and rights is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for corpus, source authority, and rights covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Authority must be enforceable outside the model through stable identities, least privilege, scoped credentials, separation of duties, revocation, and auditable approval.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C03

Corpus, Source Authority, and Rights

Family: Task and dataset integrity | Weight: 6 | Critical: NO

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for corpus, source authority, and rights.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner

versioned configuration, schema, policy, model, prompt, dataset, or architecture record

reproducible test plan with expected and observed results

traceable decision, exception, approval, and remediation records

dataset or source manifest with rights, lineage, quality, and split records

privacy, retention, deletion, and contamination review

Required metrics

lineage completeness

rights-review coverage

duplicate or contamination rate

quality-slice pass rate

deletion-verification rate

Score anchors

0	Not implemented: Corpus, Source Authority, and Rights is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for corpus, source authority, and rights.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C04

Evaluation Set Design and Holdout Protection

Family: Correctness and grounding | Weight: 5 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain evaluation set design and holdout protection for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that evaluation set design and holdout protection is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for evaluation set design and holdout protection covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C04

Evaluation Set Design and Holdout Protection

Family: Correctness and grounding | Weight: 5 | Critical: YES

Assessment procedure

- Examine the approved design, applicability decision, responsible roles, and current implementation for evaluation set design and holdout protection.
- Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.
- Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.
- Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.
- Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.
- Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

- approved control design or operating contract with accountable owner
- versioned configuration, schema, policy, model, prompt, dataset, or architecture record
- reproducible test plan with expected and observed results
- traceable decision, exception, approval, and remediation records
- negative, boundary, degraded, and recovery test results
- operational telemetry and current-state verification

Required metrics

- task success with confidence interval
- critical-failure rate
- slice variance
- reproducibility delta
- human-review burden

Score anchors

0	Not implemented: Evaluation Set Design and Holdout Protection is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

- The organization cannot identify an authoritative owner, current scope, or complete implementation for evaluation set design and holdout protection.
- Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.
- Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.
- Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C05

Prompt, Template, and Context Reproducibility

Family: Correctness and grounding | Weight: 5 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain prompt, template, and context reproducibility for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Context state must preserve authority, chronology, source identity, permissions, unresolved work, decisions, evidence, cache identity, and restoration fidelity under compression or handoff. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that prompt, template, and context reproducibility is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for prompt, template, and context reproducibility covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Context state must preserve authority, chronology, source identity, permissions, unresolved work, decisions, evidence, cache identity, and restoration fidelity under compression or handoff.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C05

Prompt, Template, and Context Reproducibility

Family: Correctness and grounding | Weight: 5 | Critical: NO

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for prompt, template, and context reproducibility.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner

versioned configuration, schema, policy, model, prompt, dataset, or architecture record

reproducible test plan with expected and observed results

traceable decision, exception, approval, and remediation records

context assembly and token accounting trace

checkpoint, summary-fidelity, stale-state, and restoration tests

Required metrics

decision and evidence retention

stale-context rate

tokens per verified outcome

restore fidelity

cross-tenant leakage rate

Score anchors

0	Not implemented: Prompt, Template, and Context Reproducibility is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for prompt, template, and context reproducibility.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C06

Correctness and Task-Specific Scoring

Family: Correctness and grounding | Weight: 5 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain correctness and task-specific scoring for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that correctness and task-specific scoring is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for correctness and task-specific scoring covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C06

Correctness and Task-Specific Scoring

Family: Correctness and grounding | Weight: 5 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for correctness and task-specific scoring.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

- approved control design or operating contract with accountable owner
- versioned configuration, schema, policy, model, prompt, dataset, or architecture record
- reproducible test plan with expected and observed results
- traceable decision, exception, approval, and remediation records
- negative, boundary, degraded, and recovery test results
- operational telemetry and current-state verification

Required metrics

- task success with confidence interval
- critical-failure rate
- slice variance
- reproducibility delta
- human-review burden

Score anchors

0	Not implemented: Correctness and Task-Specific Scoring is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for correctness and task-specific scoring.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C07

Factuality, Grounding, and Claim Support

Family: Correctness and grounding | Weight: 5 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain factuality, grounding, and claim support for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Claims must be distinguishable as sourced fact, calculation, inference, uncertainty, or recommendation, with source-support validation and abstention when evidence is insufficient. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that factuality, grounding, and claim support is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for factuality, grounding, and claim support covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Claims must be distinguishable as sourced fact, calculation, inference, uncertainty, or recommendation, with source-support validation and abstention when evidence is insufficient.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C07

Factuality, Grounding, and Claim Support

Family: Correctness and grounding | Weight: 5 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for factuality, grounding, and claim support.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner

versioned configuration, schema, policy, model, prompt, dataset, or architecture record

reproducible test plan with expected and observed results

traceable decision, exception, approval, and remediation records

negative, boundary, degraded, and recovery test results

operational telemetry and current-state verification

Required metrics

applicable control coverage

critical-failure count

evidence freshness

open nonconformities

Score anchors

0	Not implemented: Factuality, Grounding, and Claim Support is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for factuality, grounding, and claim support.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C08

Retrieval and Reranking Evaluation

Family: Correctness and grounding | Weight: 5 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain retrieval and reranking evaluation for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that retrieval and reranking evaluation is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for retrieval and reranking evaluation covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C08

Retrieval and Reranking Evaluation

Family: Correctness and grounding | Weight: 5 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for retrieval and reranking evaluation.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner

versioned configuration, schema, policy, model, prompt, dataset, or architecture record

reproducible test plan with expected and observed results

traceable decision, exception, approval, and remediation records

negative, boundary, degraded, and recovery test results

operational telemetry and current-state verification

Required metrics

task success with confidence interval

critical-failure rate

slice variance

reproducibility delta

human-review burden

Score anchors

0	Not implemented: Retrieval and Reranking Evaluation is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for retrieval and reranking evaluation.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C09

Citation Accuracy and Source Presentation

Family: Uncertainty and structure | Weight: 5 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain citation accuracy and source presentation for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Data must have documented origin, rights, purpose, classification, quality, lineage, transformation history, access restrictions, retention, and deletion obligations. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that citation accuracy and source presentation is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for citation accuracy and source presentation covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Data must have documented origin, rights, purpose, classification, quality, lineage, transformation history, access restrictions, retention, and deletion obligations.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C09

Citation Accuracy and Source Presentation

Family: Uncertainty and structure | Weight: 5 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for citation accuracy and source presentation.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner

versioned configuration, schema, policy, model, prompt, dataset, or architecture record

reproducible test plan with expected and observed results

traceable decision, exception, approval, and remediation records

dataset or source manifest with rights, lineage, quality, and split records

privacy, retention, deletion, and contamination review

Required metrics

lineage completeness

rights-review coverage

duplicate or contamination rate

quality-slice pass rate

deletion-verification rate

Score anchors

0	Not implemented: Citation Accuracy and Source Presentation is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for citation accuracy and source presentation.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C10

Uncertainty, Calibration, and Abstention

Family: Uncertainty and structure | Weight: 5 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain uncertainty, calibration, and abstention for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. The control must identify accountable ownership, authoritative inputs, expected outputs, operating limits, dependencies, failure behavior, evidence, and reassessment triggers. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that uncertainty, calibration, and abstention is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for uncertainty, calibration, and abstention covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

The control must identify accountable ownership, authoritative inputs, expected outputs, operating limits, dependencies, failure behavior, evidence, and reassessment triggers.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C10

Uncertainty, Calibration, and Abstention

Family: Uncertainty and structure | Weight: 5 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for uncertainty, calibration, and abstention.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

- approved control design or operating contract with accountable owner
- versioned configuration, schema, policy, model, prompt, dataset, or architecture record
- reproducible test plan with expected and observed results
- traceable decision, exception, approval, and remediation records
- negative, boundary, degraded, and recovery test results
- operational telemetry and current-state verification

Required metrics

- applicable control coverage
- critical-failure count
- evidence freshness
- open nonconformities

Score anchors

0	Not implemented: Uncertainty, Calibration, and Abstention is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for uncertainty, calibration, and abstention.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C11

Structured Output and Schema Conformance

Family: Uncertainty and structure | Weight: 4 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain structured output and schema conformance for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. The control must identify accountable ownership, authoritative inputs, expected outputs, operating limits, dependencies, failure behavior, evidence, and reassessment triggers. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that structured output and schema conformance is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for structured output and schema conformance covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

The control must identify accountable ownership, authoritative inputs, expected outputs, operating limits, dependencies, failure behavior, evidence, and reassessment triggers.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C11

Structured Output and Schema Conformance

Family: Uncertainty and structure | Weight: 4 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for structured output and schema conformance.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner
versioned configuration, schema, policy, model, prompt, dataset, or architecture record
reproducible test plan with expected and observed results
traceable decision, exception, approval, and remediation records
negative, boundary, degraded, and recovery test results
operational telemetry and current-state verification

Required metrics

applicable control coverage
critical-failure count
evidence freshness
open nonconformities

Score anchors

0	Not implemented: Structured Output and Schema Conformance is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for structured output and schema conformance.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C12

Multilingual and Cross-Cultural Evaluation

Family: Population and language coverage | Weight: 5 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain multilingual and cross-cultural evaluation for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that multilingual and cross-cultural evaluation is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for multilingual and cross-cultural evaluation covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C12

Multilingual and Cross-Cultural Evaluation

Family: Population and language coverage | Weight: 5 | Critical: NO

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for multilingual and cross-cultural evaluation.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner

versioned configuration, schema, policy, model, prompt, dataset, or architecture record

reproducible test plan with expected and observed results

traceable decision, exception, approval, and remediation records

negative, boundary, degraded, and recovery test results

operational telemetry and current-state verification

Required metrics

task success with confidence interval

critical-failure rate

slice variance

reproducibility delta

human-review burden

Score anchors

0	Not implemented: Multilingual and Cross-Cultural Evaluation is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for multilingual and cross-cultural evaluation.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C13

Bias, Accessibility, and Affected-Population Slices

Family: Population and language coverage | Weight: 5 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain bias, accessibility, and affected-population slices for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Security must isolate tenants, files, processes, networks, secrets, models, caches, tools, and external effects according to an explicit threat model and negative tests. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that bias, accessibility, and affected-population slices is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for bias, accessibility, and affected-population slices covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Security must isolate tenants, files, processes, networks, secrets, models, caches, tools, and external effects according to an explicit threat model and negative tests.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C13

Bias, Accessibility, and Affected-Population Slices

Family: Population and language coverage | Weight: 5 | Critical: NO

Assessment procedure

- Examine the approved design, applicability decision, responsible roles, and current implementation for bias, accessibility, and affected-population slices.
- Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.
- Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.
- Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.
- Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.
- Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

- approved control design or operating contract with accountable owner
- versioned configuration, schema, policy, model, prompt, dataset, or architecture record
- reproducible test plan with expected and observed results
- traceable decision, exception, approval, and remediation records
- negative, boundary, degraded, and recovery test results
- operational telemetry and current-state verification

Required metrics

- applicable control coverage
- critical-failure count
- evidence freshness
- open nonconformities

Score anchors

0	Not implemented: Bias, Accessibility, and Affected-Population Slices is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

- The organization cannot identify an authoritative owner, current scope, or complete implementation for bias, accessibility, and affected-population slices.
- Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.
- Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.
- Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C14

Prompt Injection and Adversarial Robustness

Family: Adversarial robustness | Weight: 6 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain prompt injection and adversarial robustness for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. The control must test realistic adversaries, direct and indirect manipulation, multi-turn escalation, cross-component attacks, prohibited outcomes, control bypass, and safe containment. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that prompt injection and adversarial robustness is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for prompt injection and adversarial robustness covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

The control must test realistic adversaries, direct and indirect manipulation, multi-turn escalation, cross-component attacks, prohibited outcomes, control bypass, and safe containment.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C14

Prompt Injection and Adversarial Robustness

Family: Adversarial robustness | Weight: 6 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for prompt injection and adversarial robustness.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner
versioned configuration, schema, policy, model, prompt, dataset, or architecture record
reproducible test plan with expected and observed results
traceable decision, exception, approval, and remediation records
adversarial case corpus with campaign provenance
critical-failure findings, control response, and retest evidence

Required metrics

attack success rate
critical control bypass rate
safe abstention or containment rate
time to detection
retest closure rate

Score anchors

0	Not implemented: Prompt Injection and Adversarial Robustness is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for prompt injection and adversarial robustness.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C15

Long-Context and Position-Sensitivity Evaluation

Family: Adversarial robustness | Weight: 6 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain long-context and position-sensitivity evaluation for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that long-context and position-sensitivity evaluation is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for long-context and position-sensitivity evaluation covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C15

Long-Context and Position-Sensitivity Evaluation

Family: Adversarial robustness | Weight: 6 | Critical: NO

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for long-context and position-sensitivity evaluation.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner

versioned configuration, schema, policy, model, prompt, dataset, or architecture record

reproducible test plan with expected and observed results

traceable decision, exception, approval, and remediation records

context assembly and token accounting trace

checkpoint, summary-fidelity, stale-state, and restoration tests

Required metrics

task success with confidence interval

critical-failure rate

slice variance

reproducibility delta

human-review burden

Score anchors

0	Not implemented: Long-Context and Position-Sensitivity Evaluation is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for long-context and position-sensitivity evaluation.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C16

Human Evaluation and Judge Validation

Family: Human and automated grading | Weight: 5 | Critical: YES

Normative requirement

The organization **MUST** establish, implement, verify, and maintain human evaluation and judge validation for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that human evaluation and judge validation is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for human evaluation and judge validation covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Evaluation must use representative and hidden cases, reproducible configuration, deterministic validators where possible, statistical uncertainty, critical-failure gates, and slice-level reporting.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C16

Human Evaluation and Judge Validation

Family: Human and automated grading | Weight: 5 | Critical: YES

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for human evaluation and judge validation.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

- approved control design or operating contract with accountable owner
- versioned configuration, schema, policy, model, prompt, dataset, or architecture record
- reproducible test plan with expected and observed results
- traceable decision, exception, approval, and remediation records
- negative, boundary, degraded, and recovery test results
- operational telemetry and current-state verification

Required metrics

- task success with confidence interval
- critical-failure rate
- slice variance
- reproducibility delta
- human-review burden

Score anchors

0	Not implemented: Human Evaluation and Judge Validation is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for human evaluation and judge validation.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C17

Latency, Throughput, Cost, and Resource Evaluation

Family: Human and automated grading | Weight: 5 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain latency, throughput, cost, and resource evaluation for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Data must have documented origin, rights, purpose, classification, quality, lineage, transformation history, access restrictions, retention, and deletion obligations. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that latency, throughput, cost, and resource evaluation is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for latency, throughput, cost, and resource evaluation covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Data must have documented origin, rights, purpose, classification, quality, lineage, transformation history, access restrictions, retention, and deletion obligations.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C17

Latency, Throughput, Cost, and Resource Evaluation

Family: Human and automated grading | Weight: 5 | Critical: NO

Assessment procedure

Examine the approved design, applicability decision, responsible roles, and current implementation for latency, throughput, cost, and resource evaluation.

Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.

Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.

Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.

Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.

Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

approved control design or operating contract with accountable owner
versioned configuration, schema, policy, model, prompt, dataset, or architecture record
reproducible test plan with expected and observed results
traceable decision, exception, approval, and remediation records
dataset or source manifest with rights, lineage, quality, and split records
privacy, retention, deletion, and contamination review

Required metrics

p50/p95/p99 end-to-end latency
successful units per second
saturation and rejection rate
cost per verified outcome
resource efficiency

Score anchors

0	Not implemented: Latency, Throughput, Cost, and Resource Evaluation is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

The organization cannot identify an authoritative owner, current scope, or complete implementation for latency, throughput, cost, and resource evaluation.

Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.

Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.

Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C18

Production Monitoring, Drift, and Regression

Family: Operational performance | Weight: 6 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain production monitoring, drift, and regression for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. The control must define production indicators, drift thresholds, incident triggers, review frequency, change detection, owner notification, corrective action, and retirement criteria. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that production monitoring, drift, and regression is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for production monitoring, drift, and regression covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

The control must define production indicators, drift thresholds, incident triggers, review frequency, change detection, owner notification, corrective action, and retirement criteria.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C18

Production Monitoring, Drift, and Regression

Family: Operational performance | Weight: 6 | Critical: NO

Assessment procedure

- Examine the approved design, applicability decision, responsible roles, and current implementation for production monitoring, drift, and regression.
- Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.
- Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.
- Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.
- Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.
- Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

- approved control design or operating contract with accountable owner
- versioned configuration, schema, policy, model, prompt, dataset, or architecture record
- reproducible test plan with expected and observed results
- traceable decision, exception, approval, and remediation records
- negative, boundary, degraded, and recovery test results
- operational telemetry and current-state verification

Required metrics

- applicable control coverage
- critical-failure count
- evidence freshness
- open nonconformities

Score anchors

0	Not implemented: Production Monitoring, Drift, and Regression is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

- The organization cannot identify an authoritative owner, current scope, or complete implementation for production monitoring, drift, and regression.
- Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.
- Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.
- Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

MYTHOS-AI-0003-C19

Records, Reassessment, and Retirement

Family: Lifecycle | Weight: 5 | Critical: NO

Normative requirement

The organization **MUST** establish, implement, verify, and maintain records, reassessment, and retirement for all in-scope language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output. Recovery must cover safe stop, checkpoint integrity, rollback or compensation, credential revocation, artifact restoration, evidence preservation, requalification, and controlled retirement. The control **MUST** define acceptance criteria, prohibited outcomes, evidence, exception authority, and reassessment triggers. Where automated enforcement is feasible, the organization **SHOULD** enforce the requirement outside the model or agent. Any residual manual dependency **MUST** be documented, assigned, time-bounded, and tested.

Control objective

Objective

Ensure that records, reassessment, and retirement is explicit, bounded, testable, observable, recoverable, and governed throughout the lifecycle.

Implementation requirements

Establish a versioned control contract for records, reassessment, and retirement covering language models and systems used for generation, transformation, extraction, classification, retrieval, grounded question answering, and structured output and name the accountable owner, approver, assessor, and affected stakeholders.

Recovery must cover safe stop, checkpoint integrity, rollback or compensation, credential revocation, artifact restoration, evidence preservation, requalification, and controlled retirement.

Implement the control through machine-enforceable policy, typed schemas, isolated runtime boundaries, validated procedures, or an approved combination appropriate to the risk.

Define explicit nominal, negative, adversarial, malformed, degraded, overloaded, recovery, and rollback tests; critical cases must be hidden from the system under evaluation where feasible.

Correlate evidence to stable system, model, data, prompt, repository, tool, environment, user or tenant, and release identities; protect records against unauthorized modification.

Set review cadence and event-driven reassessment triggers for material changes, incidents, supplier updates, drift, failed controls, new affected populations, and retirement.

Applicability

Applicable unless a documented, approved, evidence-supported exclusion demonstrates that the capability, data, decision, interface, population, or lifecycle stage is absent.

MYTHOS-AI-0003-C19

Records, Reassessment, and Retirement

Family: Lifecycle | Weight: 5 | Critical: NO

Assessment procedure

- Examine the approved design, applicability decision, responsible roles, and current implementation for records, reassessment, and retirement.
- Select representative normal, boundary, high-impact, and adversarial cases, including at least one prohibited outcome and one dependency or recovery failure.
- Verify the exact model, data, prompt, tool, policy, runtime, repository, environment, and evaluator versions used during assessment.
- Execute deterministic validators first, then calibrated model or human judgments only where deterministic proof is not available.
- Reconcile expected behavior with raw outputs, trajectories, tool effects, telemetry, artifacts, user-visible outcomes, and unresolved contradictions.
- Assign a 0-5 score, record evidence links and confidence, open nonconformities, define remediation ownership, and set the next verification date.

Minimum evidence

- approved control design or operating contract with accountable owner
- versioned configuration, schema, policy, model, prompt, dataset, or architecture record
- reproducible test plan with expected and observed results
- traceable decision, exception, approval, and remediation records
- negative, boundary, degraded, and recovery test results
- operational telemetry and current-state verification

Required metrics

- recovery success rate
- duplicate-effect rate
- time to safe state
- residual-effect count
- evidence preservation rate

Score anchors

0	Not implemented: Records, Reassessment, and Retirement is absent, materially unknown, or contradicted by evidence.
1	Initial: intent is documented, but ownership, repeatability, implementation, or evidence is incomplete.
2	Defined: approved procedure and design exist, but implementation or representative testing is partial.
3	Implemented: control operates for the declared scope and passes normal and basic negative tests.
4	Verified: independent assessment, hidden or adversarial cases, current evidence, and recovery validation demonstrate effectiveness.
5	Continuously assured: automated monitoring and regression, current evidence, incident learning, and timely reassessment sustain verified effectiveness.

Failure indicators

- The organization cannot identify an authoritative owner, current scope, or complete implementation for records, reassessment, and retirement.
- Only a successful demonstration, vendor claim, aggregate score, or model-generated judgment is available; raw evidence and negative tests are absent.
- Critical cases can be hidden by averages, unsupported N/A designations, stale evidence, or changes in model, data, prompt, tool, policy, or environment.
- Failure, rollback, revocation, deletion, containment, or retirement behavior has not been exercised under realistic conditions.

Required Benchmark Scenarios

MYTHOS-AI-0003-B01 - Representative production task

Demonstrate the declared use case using current configuration and representative inputs. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

MYTHOS-AI-0003-B02 - Hidden critical holdout

Execute high-impact cases withheld from development, tuning, and prompt iteration. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

MYTHOS-AI-0003-B03 - Malformed and adversarial inputs

Test schema violations, ambiguity, direct or indirect manipulation, and unsafe instructions. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

MYTHOS-AI-0003-B04 - Long-horizon or multi-step execution

Measure compounding error, context loss, looping, premature completion, and recovery. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

MYTHOS-AI-0003-B05 - Dependency and tool failure

Inject model, network, data, tool, credential, evaluator, or service failure. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

Required Benchmark Scenarios - Continued

MYTHOS-AI-0003-B06 - Resource pressure and overload

Exercise concurrency, long inputs, queueing, rate limits, memory, tokens, and cost limits. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

MYTHOS-AI-0003-B07 - Privacy and cross-boundary test

Attempt cross-tenant, secret, personal-data, restricted-source, cache, and egress violations. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

MYTHOS-AI-0003-B08 - Rollback, revocation, or restore

Prove safe stop, compensation, revocation, prior-state restoration, and residual-effect reconciliation. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

MYTHOS-AI-0003-B09 - Human intervention and disagreement

Test approval, interruption, correction, appeal, assessor disagreement, and minority findings. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

MYTHOS-AI-0003-B10 - Material change and regression

Change model, prompt, dataset, tool, provider, runtime, or policy and run requalification gates. Apply the scenario specifically to language and llm evaluation and preserve raw evidence.

Conformance Report and Evidence Bundle

Permanent document ID, standard version, assessment version, system and use-case scope.

System, model, provider, prompt, data, retrieval, memory, tool, evaluator, runtime, repository, and infrastructure identities.

Applicability decisions and normalized denominator.

Control-level scores, weights, critical status, evidence links, assessor, date, confidence, and finding disposition.

Benchmark scenario results with raw artifacts, deterministic validators, judge or expert records, uncertainty, failures, and limitations.

Family and total score, achieved conformance level, unresolved critical or major findings, approved exceptions, expiry, and next review.

Release, rollback, revocation, deletion, and retirement evidence where applicable.

Evidence bundle integrity

The evidence bundle SHOULD use content hashes, signatures or protected repository history, stable artifact IDs, access controls, retention policy, and independent verification. Evidence links alone are insufficient when the referenced object can change without detection.

Exceptions, Waivers, and Compensating Controls

An exception **MUST NOT** be used to relabel a failed requirement as conforming. Exceptions document accepted residual risk for a bounded scope and time; the underlying control score remains evidence-based.

Identify the exact control, system scope, reason, affected users or assets, and current score.

Describe the missing requirement, residual risk, foreseeable failure, and affected decisions.

Define compensating controls, validation evidence, monitoring, owner, approver, start date, expiry, and remediation plan.

Obtain approval from an authority independent of the implementation owner for critical or high-impact exceptions.

Reassess on incident, material change, evidence expiry, control failure, or exception expiration.

Publish exceptions in the conformance report; hidden exceptions invalidate the claim.

Critical exception cap

An unresolved exception against a critical control prevents Assured or Continuously Assured conformance and may prevent Operational conformance when the control score is below 3.

Source Authority and Freshness Register

NIST - Artificial Intelligence Risk Management Framework 1.0

Cross-sector AI risk governance, mapping, measurement, and management.

<https://doi.org/10.6028/NIST.AI.100-1>

NIST - NIST AI 600-1 - Generative Artificial Intelligence Profile

Generative-AI risks, controls, evaluation, and lifecycle actions.

<https://doi.org/10.6028/NIST.AI.600-1>

NIST - AI Resource Center and AI RMF Playbook

Current implementation resources and revision notices for the AI RMF.

<https://airc.nist.gov/>

ISO/IEC - ISO/IEC 42001:2023 - AI management systems

AI management-system requirements and continual improvement.

<https://www.iso.org/standard/42001>

ISO/IEC - ISO/IEC 23894:2023 - Guidance on AI risk management

AI risk-management guidance across lifecycle and stakeholders.

<https://www.iso.org/standard/77304.html>

ISO/IEC - ISO/IEC 42005:2025 - AI system impact assessment

Impact-assessment guidance for individuals and societies.

<https://www.iso.org/standard/42005>

MLCommons - MLCommons Benchmarks

Open performance, quality, and safety benchmark programs.

<https://mlcommons.org/benchmarks/>

Source Authority and Freshness Register - Continued

MLCommons - AI Risk and Reliability

Safety-test methodology and risk-oriented evaluation work.

<https://mlcommons.org/working-groups/ai-risk-reliability/ai-risk-reliability/>

Stanford CRFM - HELM - Holistic Evaluation of Language Models

Scenario- and metric-based model evaluation framework.

<https://arxiv.org/abs/2211.09110>

EleutherAI - Language Model Evaluation Harness

Open evaluation harness and reproducible task definitions.

<https://github.com/EleutherAI/lm-evaluation-harness>

OWASP - Top 10 for LLM and GenAI Applications 2025

LLM application security and prompt-injection risk baseline.

<https://genai.owasp.org/lm-top-10/>

Research - Lost in the Middle

Long-context position-sensitivity evidence.

<https://arxiv.org/abs/2307.03172>

Freshness rule

Sources were verified for this candidate edition on 2026-07-26. The standard **MUST** be revalidated when a cited authority changes, becomes superseded, is withdrawn, or materially alters the control interpretation.

Limitations, Revision History, and Adoption Position

This candidate Mythos standard is designed for engineering governance and assessment. It is not a legal opinion, regulatory certification, safety guarantee, or substitute for domain-specific professional judgment. Model behavior remains probabilistic and system risk depends on data, users, tools, deployment, incentives, and operating context.

Revision history

Version	Date	Change
2.0.0-candidate	2026-07-26	Expanded normative edition with 19 weighted controls, benchmark scenarios, conformance levels, machine-readable catalogs, assessment workbook integration, source freshness, and explicit lineage.
1.0.0-candidate	2026-07-24	Earlier permanent-identity candidate retained in the release lineage archive.

Adoption position

Use this edition as a candidate control and benchmark baseline. Organizations SHOULD pilot it on bounded systems, retain assessment evidence, submit corrections, and avoid representing candidate conformance as external certification.