INDEPENDENT TECHNICAL ASSURANCE

Independent AI Evaluation Assurance

A Framework for Determining Whether AI Performance Claims Can Be Trusted

From claim inventory and construct validity to the
Production Evaluation Gap and Maximum Credible Evaluation Failure

CORE THESIS

A benchmark result is evidence only to the extent that the evaluation producing it is valid, reproducible, relevant to production, and independently interpretable.

MQ-EVA 1.0 | Version 1.0 | August 2026

Maple Quanta Inc. | Ottawa, Canada | maplequanta.ca

Public white paper

Suggested citation

Maple Quanta Inc. (2026). *Independent AI Evaluation Assurance: A Framework for Determining Whether AI Performance Claims Can Be Trusted*. Version 1.0. Ottawa, Canada.

Publication note

This public white paper describes the framework, evidence model, metrics, and decision rules of MQ-EVA 1.0. Detailed contamination and gaming test procedures, scoring calibration, assessor working papers, sealed evaluation material, and client-specific tooling are outside the scope of the public document.

The accompanying open repository publishes the machine-readable schemas, the Evaluation Disclosure Sheet, the executive scorecard, a complete synthetic worked example, and tested reference implementations of every derived measure in this paper.

What this document does not claim

MQ-EVA is an assurance methodology. It is **not** a certification scheme, and Maple Quanta Inc. is **not** an accredited certification body. Nothing in this paper certifies conformity with ISO/IEC 42001 or any other standard, and nothing in it should be read as a requirement of any standards body or government. Where a standard is under development, it is identified as a draft. Where a statement is Maple Quanta's own position, it is presented as such.

Contents

Executive summary

Organisations are increasingly asked to make consequential decisions — vendor selection, production approval, risk acceptance — on the strength of AI evaluation evidence. Benchmark scores, vendor evaluations, leaderboard positions, system cards, red-team reports, pilots, demonstrations, and productivity studies now routinely appear in procurement files and board papers.

That evidence is frequently weaker than it appears. Contamination of public benchmarks is documented as widespread. A systematic review of 445 benchmark articles found that only 6% included uncertainty estimates, error bars, or statistical significance tests when comparing results between models, alongside contested construct definitions and convenience sampling. [8] Agentic evaluations have been shown to be manipulable by the systems they evaluate. And evaluation conditions — unrestricted network access, stubbed tools, no latency budget, no approval gates — routinely differ from the environment in which the system will actually run.

Existing guidance tells organisations *how to evaluate well*. NIST's TEVV programme, its AITE sequestered-evaluation vehicle, and its draft TEVV-Athlon framework advance the science of evaluation. The Canadian AI Safety Institute has set out what evaluators should disclose. ISO/IEC's 42000-series governs AI management systems, impact assessment, and the bodies that certify them.

None of it answers the question an organisation actually faces this quarter: **is the specific evaluation evidence in front of me strong enough to justify the specific decision I am about to make?**

MQ-EVA answers that question. Its unit of assessment is a **claim**, together with the evidence offered for it and the decision it supports. It assesses that claim across **seven dimensions** — construct validity, evidence integrity, statistical reliability, production realism, failure characterisation, evaluator independence, and evaluation environment integrity — reported separately, with **no composite score**. Conclusions are graded on a five-level **evidence hierarchy** in which a finding's grade is its weakest link.

Five measures convert the assessment into decisions: the **Evaluation Confidence Level** (C0–C4, derived by an explicit floor rule rather than assessor judgement); the **Production Evaluation Gap** between laboratory and operational capability; **Evaluation Reproduction Agreement**; the **Failure Exposure Profile** with its **Maximum Credible Evaluation Failure**; and **Evidence Coverage**.

> **The central principle.** The question is not simply whether the AI passed the test. The question is whether the test deserves to be trusted.

1 Why AI evaluation itself needs assurance

1.1 The evidence has become load-bearing

Ten years ago a benchmark score was a research artifact. Today it is a procurement input. It appears in bid evaluations, appears in business cases, and is quoted in approvals to deploy. The score has acquired the weight of evidence without acquiring the properties of evidence.

The failure modes are well documented and, importantly, mostly structural rather than dishonest:

- **Contamination and leakage.** Public benchmarks predating a model's training cutoff must be treated as possibly seen. NIST's response was architectural: AITE provides a *sequestered environment* that "mitigates the risk of train/test data contamination to ensure rigorous, objective assessment." [5] Contamination is managed by structure, not by assurance.
- **Construct invalidity.** A benchmark may measure something adjacent to the claimed capability. This is the finding of the largest systematic review of the area, which reports patterns across the literature that "undermine the validity of the resulting claims." [8]
- **Statistical fragility.** Single runs of stochastic systems, absent intervals, unpaired comparison, clustered observations treated as independent, and undisclosed selection across many attempts. [10]
- **Elicitation opacity.** How hard the evaluator tried — prompt tuning, scaffolding, tool provision, best-of-n — materially changes the result. Capability elicitation is now internationally recognised as part of evaluation methodology. [7]
- **Grader manipulability.** In agentic evaluation the evaluated system frequently sits *inside* the trust boundary of the scoring mechanism. Where an agent can write files the grader later reads, or reach network resources containing reference solutions, a recorded success may represent successful manipulation of the scorer rather than completion of the task.
- **Laboratory–operational divergence.** The conditions that produce the number are rarely the conditions of deployment. This premise underlies NIST's ARIA programme. [4]
- **Reporting integrity.** Undisclosed exclusions, unreconciled run counts, benchmark selection after seeing results, and unidentified model versions.

1.2 What existing guidance does and does not settle

The interdisciplinary review by Eriksson et al. for the European Commission's Joint Research Centre — a meta-review of roughly 100 studies of benchmarking shortcomings — catalogues dataset-creation bias, inadequate documentation, contamination, failure to distinguish signal from noise, misaligned incentives, and gaming. [9] The literature on the problem is mature.

The response has been to improve evaluation practice, which is right and necessary. But better practice going forward does not help the procurement officer holding three incommensurable bid responses today, and it does not retroactively validate the pilot result that a deployment approval already rests on.

> **The gap.** There is substantial official work on *how to evaluate AI well*, and a large literature on *how evaluations go wrong.* There is very little on *how an organisation should determine whether the specific evidence in front of it is strong enough to justify the specific decision it supports.* MQ-EVA addresses only that gap.

2 Scope

2.1 The unit of assessment is a claim

MQ-EVA assesses a **claim**, the **evidence** offered for it, and the **decision** it supports.

This is a deliberate restriction. Assessing "the model" produces an unbounded engagement. Assessing "the vendor's report" produces a document review. Assessing a claim — *"this system achieves 93% extraction accuracy on procurement documents"* — produces a finding a decision-maker can act on, because the claim is what the decision rests on.

Assurance effort is proportional to the consequence of the decision. This matches the international network's recommendation to establish *decision-relevance* first and use lightweight testing before deeper analysis where resources are constrained. [7]

2.2 Out of scope

MQ-EVA does not certify, does not publish benchmarks or leaderboards, does not guarantee future performance, and does not replace a safety case, privacy impact assessment, penetration test, or regulatory assessment. Its conclusions are configuration-specific and are invalidated by material change to the assessed model version, configuration, prompts, tools, data, environment, or controls.

It also does not exempt Maple Quanta from its own standard. An evaluation Maple Quanta conducts is not independent of Maple Quanta, and each engagement is graded on the same independence scale it applies to others.

3 Definitions

Term	Definition
Claim	An assertion about an AI system's performance, safety, or suitability, used to justify a decision.
Material claim	A claim that, if false, would change the decision it supports. Materiality is fixed before assessment.
Evidence grade	The kind of thing supporting a finding, on the E0–E4 hierarchy.

Term	Definition
Laboratory Capability	Performance under evaluation conditions — typically unconstrained and generously resourced.
Operational Capability	Performance under the security, privacy, latency, cost, data-quality, and approval constraints of the actual deployment.
Elicitation	The effort applied to obtain a result: prompt tuning, scaffolding, tool provision, retries, best-of-n. A property of the result, not a detail.
Sequestration	Holding evaluation data out of reach of training and of the system under test.
Sealed evaluation material	Items constructed or held by the evaluator and never disclosed to the evaluated party.
Scoring trust boundary	Whether the evaluated system can influence its own score.
Commensurability	The conditions under which two measurements may legitimately be compared.

4 The seven evaluation dimensions

Findings are reported by dimension. There is **no composite score**, and none should be constructed: a weighted index would require weights that cannot be justified from evidence, and its only function would be to let a strength on one dimension conceal a fatal weakness on another.

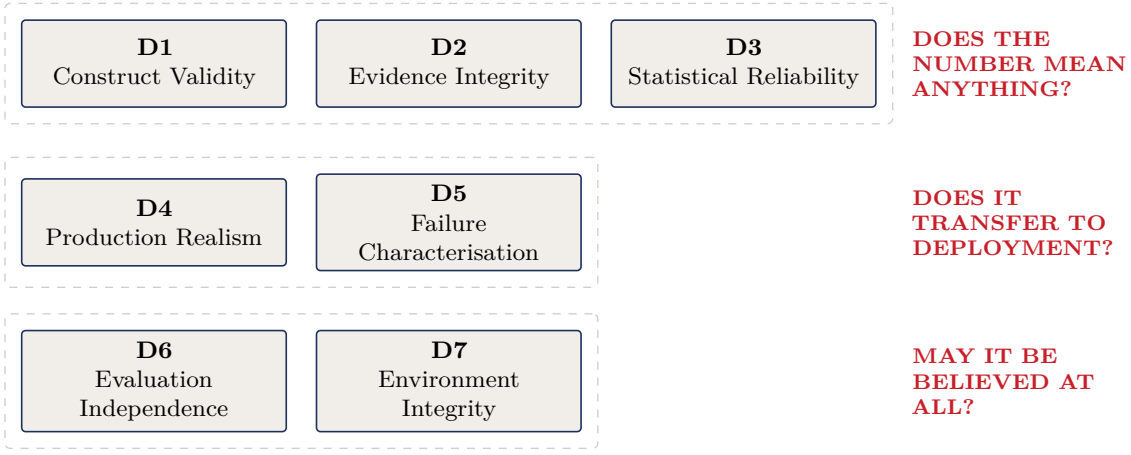

Figure 1: Three questions, seven dimensions. D1–D3 establish whether the figure is meaningful; D4–D5 whether it transfers to the deployment; D6–D7 whether the conditions of its production permit it to be believed.

4.1 D1 — Construct Validity

Does the evaluation measure the capability it claims to measure?

Assessed: construct definition and operationalisation; content coverage against the claim's stated population; metric appropriateness; task–decision correspondence; contamination and leakage; shortcuts and gaming; benchmark saturation relative to the differences claimed; and elicitation disclosure.

Illustration. A coding benchmark reports task completion. The evaluation environment retained repository history containing the eventual fix, and the agent retrieved the fix rather than deriving it. The number is real; the capability claim it supports is not.

4.2 D2 — Evidence Integrity

Is the reported result a faithful record of what happened?

Assessed: completeness of per-trial traces; run accounting (planned, executed, excluded, reported); exclusion criteria and whether they predate results; selective reporting; exact system identification; configuration capture; scoring integrity and judge validation; and chain of custody.

Integrity defects differ from validity defects in a way that matters for decisions: a validity defect usually biases in a knowable direction, while an integrity defect frequently leaves the direction and magnitude unknown. Unknown direction is worse than known bias.

4.3 D3 — Statistical Reliability

Would the result survive repetition?

Assessed: trials per item; sample size and composition; interval reporting; paired comparison; clustering; prompt and seed sensitivity; multiplicity; calibration; and error asymmetry.

MQ-EVA requires that uncertainty be *characterised*. It does not mandate null-hypothesis significance testing where that is the wrong instrument, and it does not accept a p-value in place of reported variance and effect size. The practical failure in evaluation reporting is rarely a missing p-value; it is unreported variance, unpaired comparison, and treating one run as a measurement. [10]

4.4 D4 — Production Realism

Does the evaluation represent the environment in which the system will operate?

Assessed under: enterprise security policy; network egress restrictions; real identity and permission boundaries; actual tool implementations rather than stubs; rate limits; latency budgets; degraded and unavailable services; incomplete and poor-quality data; real document formats; human approval gates; production retrieval corpora; privacy and retention controls; time limits; infrastructure failure; and cost ceilings.

> **The constraint-disclosure rule.** Evaluation conditions are part of the result. A figure reported without its conditions is not a result; it is a number.

4.5 D5 — Failure Characterisation

How does the system fail, not merely how often does it succeed?

An average conceals catastrophic failure. A system that is 97% accurate with a 3% *severe, silent, irreversible* error rate is not the same product as one that is 97% accurate with a 3% *obvious, recoverable* error rate, and no aggregate metric distinguishes them.

Failures are characterised on four axes — **frequency**, **severity**, **detectability**, **reversibility**. Detectability and reversibility are what convert failure frequency into business risk. A frequent, obvious, reversible failure is an operational nuisance; a rare, silent, irreversible one is a governance problem, and it is the case aggregate accuracy is least able to surface.

4.6 D6 — Evaluation Independence

Who designed, conducted, interpreted, and approved the evaluation?

Independence is graded, not asserted, and it is a property of conditions rather than of job titles. Third-party status alone does not confer reliability: an external evaluator working to a vendor-set scope, on vendor-supplied data, under a vendor publication veto is not independent in any sense that should reassure a buyer. The safe-harbour literature makes the same point about researcher-access programmes, which it finds are inadequate substitutes for independent access because they lack independence from corporate incentives. [11]

Position	Description
I0 Self-evaluation	Designed, executed, scored, and reported by the developer or vendor
I1 Customer evaluation	Conducted by the deploying organisation, typically using vendor-supplied materials
I2 Third-party evaluation	Conducted by a party contractually distinct from the developer
I3 Independently replicated	Key results reproduced by a party with no interest in the outcome, using its own materials

4.7 D7 — Evaluation Environment Integrity

Was the environment that produced the result itself trustworthy?

This is a separate dimension because it belongs cleanly to neither neighbour. Environment defects corrupt *validity* (a reachable solution inflates the score) and *integrity* (a writable grader corrupts

the record) simultaneously; folding it into either loses the question. For agentic evaluation it is frequently the dimension that determines the answer. Section 10 develops it.

5 Evidence hierarchy

Grade	Name	What it excludes
E0	Assertion	Nothing. A statement with no supporting artifact
E1	Documentation	That no method exists at all
E2	Artifact	That the described method was not the executed method
E3	Reproduced	That the artifacts do not produce the reported figure
E4	Adversarially Validated	That a perfectly reproducible evaluation measures the wrong thing

Two rules make the hierarchy useful.

Rule 1 — the grade is the weakest link. A finding's grade is the lowest grade among the elements it depends on. A reproduction (E3) resting on a dataset whose provenance is only asserted (E0) is an E0 finding. **Assurance does not average.**

Rule 2 — the ladder is not automatically climbable. E3 is unavailable where reproduction is not permitted; E4 is unavailable where uncontaminated material cannot be obtained. When a rung is unreachable, MQ-EVA records *why* — and "E3 not attained: vendor declined re-execution access" is frequently more consequential than the rating it caps.

The rung most often missing is the last one. **E3 answers "is the number real?" E4 answers "does the number mean what it is being used to mean?"** A contaminated benchmark, a manipulable grader, and an over-elicited configuration all reproduce faithfully.

5.1 Vendor material

System cards, frontier-safety reports, and vendor benchmark disclosures are legitimate and often high-quality *inputs*. They are not independent evidence, for the structural reason that the party producing the result selects the benchmark, the configuration, the elicitation effort, the baseline, and what is published. On its own, vendor evaluation material is **E1 at best**, however detailed the document or reputable the publisher. This is not an accusation of bad faith; it is the ordinary meaning of independence.

6 Metrics

6.1 Evaluation Confidence Level (ECL)

ECL is assigned **per claim**, never per system or per vendor. It is a **floor rule**: the highest level for which every precondition of that level and of all levels below it is satisfied. Failing a precondition does not subtract points — it caps the level.

Level	Name	Cumulative preconditions
C0	Unsubstantiated	Default, where C1 is not reached
C1	Indicative	≥E1; exact version identified; D1 ≥ Limited; no Critical Concern on D1 or D2
C2	Supported	≥E2; D1, D2, D3 ≥ Adequate; uncertainty characterised; ≥2 trials per item on stochastic systems; no Critical Concern anywhere
C3	Independently Verified	≥E3 with reproduction agreement; independence ≥ I2; D6 ≥ Adequate; production realism assessed
C4	Robustly Verified	E4; all seven dimensions ≥ Adequate; production-constrained execution; PEG resolved; failure characterisation complete; MCEF stated; sealed material held by the evaluator

Because the rule is explicit, the conclusion is recomputable. Anyone holding the assessment record can derive the level independently and see exactly which precondition capped it. An assurance conclusion that cannot be recomputed from its inputs is not auditable, and the reference implementation exists so that this one is.

C0 is a common and legitimate outcome. It means the evidence does not support the claim — *not* that the claim is false.

6.2 Production Evaluation Gap (PEG)

$$\text{PEG} = M_{\text{laboratory}} - M_{\text{production}} \quad \text{(metrics where higher is better)}$$

with the sign reversed for metrics where lower is better, so that **a positive PEG always means production is worse**. PEG is reported in the metric's own units.

PEG is defined **only** when five commensurability preconditions hold: the same construct, the same metric and scoring procedure, the same item set, the same system version, and conditions differing only in the production constraints under study. Where any fails, MQ-EVA reports the two figures separately and records **"PEG not computable"** with the reason.

This refusal is load-bearing. Subtracting two numbers that do not measure the same thing produces something that looks like a finding and is not one — and the resulting figure is usually larger and more quotable than the defensible one.

PEG is reported with an uncertainty interval on the *difference*, computed from the paired design; the interval on a difference cannot be recovered from the two separate intervals. Where a constraint

prevents the task from being attempted at all, the finding is **"capability unavailable under production constraints"** — a stronger and more decision-relevant result than any number.

6.3 Evaluation Reproduction Agreement (ERA)

Categorical: **Reproduced** (the reported point lies inside the reproduction's interval, and where an original interval exists, the reproduction's point lies inside it); **Partially reproduced** (intervals overlap, points do not land); **Not reproduced** (intervals disjoint); **Not reproducible** (re-execution impossible, with the reason recorded).

There is deliberately no "reproducibility percentage". The share of results that reproduced is not meaningful across results of differing importance, and one decisive result that failed to reproduce is not offset by many trivial ones that did.

6.4 Failure Exposure Profile and MCEF

Each failure mode is recorded on the four axes of D5. **The axes are never multiplied.** A product of ordinal ranks is not a risk quantity, and computing one would reintroduce the averaging this framework rejects.

> **Maximum Credible Evaluation Failure (MCEF)** is the most consequential failure that the evaluation evidence either demonstrates or **fails to exclude**, within the defined deployment context. It is stated on the engagement's severity scale with a status: *Demonstrated*, *Inferred*, or ***Not excluded***.

The *Not excluded* status is why MCEF exists. An evaluation that never tested a failure mode provides no evidence against it. Reporting `MCEF = S4 (Not excluded)` for an untested severe mode is the honest statement; silently reporting 97% accuracy is not.

6.5 Evidence Coverage

The proportion of material claims supported at E2 or above, reported as **"n of N"** with the ratio — never as a bare percentage. Materiality is fixed in the Claim Register before assessment so the denominator cannot move once results are known.

Coverage measures the *availability* of examinable evidence. It says nothing about whether claims held: an engagement can reach 100% coverage and conclude that every claim is unsupported. Both are reported; they are never combined.

6.6 What is deliberately excluded

No weighted composite of the dimensions. No "MQ-EVA score". No products or sums of ordinal codes. No ECL for a system rather than a claim. No PEG across incommensurable metrics. No reproducibility percentage. No figure carrying more significant digits than its sample supports.

7 Assessment methodology

Phase	Purpose	Principal output
1 Claims Inventory	Establish what is asserted and what each assertion supports	Claim Register; materiality
2 Evaluation Reconstruction	Determine what was done, as distinct from what was reported	Disclosure Sheet; Evidence Register
3 Validity Audit	Determine whether the evaluation measures what the claim requires	D1, D2, D6, D7 ratings
4 Independent Reproduction	Determine whether the result is obtainable	ERA; E3 where attained
5 Production-Constrained Evaluation	Determine what survives the deployment environment	PEG; D4 rating
6 Failure Stress Testing	Characterise failure, not merely count it	Failure Register; MCEF
7 Evidence Assessment	Convert testing into per-claim findings	Findings; ECL; Coverage
8 Executive Report	Deliver a decision with its basis and its limits	Report; Assurance Statement

All testing in Phases 4–6 is authorised, bounded, and agreed in writing before it begins, including permitted actions, data handling, and stop conditions.

Two phases deserve emphasis. **Phase 1** handles four claim shapes differently: quantified claims are assessable directly; comparative claims ("better than humans") require the baseline and comparison design to be recovered first; categorical claims ("enterprise safe") must be *operationalised* into a

testable proposition or recorded as unassessable as stated; economic claims require the study design and counterfactual. A claim nobody will operationalise is itself a finding, and should not be carrying weight in a procurement.

Phase 7 enforces a distinction that reporting habitually blurs:

Finding	Meaning
Substantiated	The evidence supports the claim as stated, within scope
Partially Substantiated	The evidence supports a narrower version, which is stated explicitly
Unsupported	The evidence does not establish the claim. *Not* a finding that it is false
Contradicted	Independent testing produced results inconsistent with the claim

Unsupported calls for evidence. *Contradicted* calls for explanation. They are not interchangeable.

7.1 Engagement tiers

Tiers differ by what Maple Quanta executes itself, which determines the evidence grade attainable.

Tier	MQ executes	Max grade	Max ECL
1 Evaluation Evidence Review	No re-execution	E2	C2
2 Independent Evaluation	Reproduction, sealed set, production constraints	E3	C3
3 High-Assurance Evaluation	Adds environment assurance and adversarial validation	E4	C4

Tier 1's ceiling is structural rather than a limitation of effort — and Tier 1 is frequently the highest-value tier per dollar, because the most consequential defects (undisclosed exclusions, mismatched versions, contaminated benchmarks, manipulable graders, an English test set behind a bilingual claim) are found by reading artifacts, not by testing.

7.2 Placement in the procurement cycle

The highest-leverage application is **before** the RFP is issued: designing acceptance criteria as testable claims rather than adjectives, specifying the evaluation methodology bidders must follow, setting minimum disclosure requirements, and constructing a private evaluation set the buyer holds

and never releases. Once three vendors have each optimised a different benchmark, no downstream analysis fully recovers the comparison.

Thereafter: independent validation of vendor claims during procurement; final production-constrained evaluation before acceptance; and periodic reassessment post-deployment on defined triggers.

8 Production-constrained testing

The distinction between **Laboratory Capability** and **Operational Capability** is the framework's most commercially significant contribution, because the gap is routinely double-digit and almost never measured.

A model achieving 90% in an unrestricted benchmark may behave very differently once it operates under real security policy, real identity boundaries, real latency budgets, a production retrieval corpus rather than a curated one, real document quality, and approval gates that interrupt multi-step work.

Constraint attribution — re-running with each constraint applied individually — identifies which control is responsible for the loss, which is the actionable part. **Constraint effects are not additive**: individual effects will not sum to the full PEG because constraints interact, so the decomposition is diagnostic and the full-constraint measurement remains the reported figure.

9 Agent evaluation considerations

Agentic systems change the evaluation problem in four ways.

Variance rises and single runs become meaningless. Multi-step trajectories compound stochasticity. Repetition is not a refinement; it is the minimum condition for measurement. METR's published agent methodology runs each model repeatedly on each task and fits success probability against human-expert task duration — and, notably, publishes the reliability ceiling of its own instrument. [12] Both practices generalise: *repeat runs by default*, and *state where your instrument stops being reliable*.

Scaffolding becomes part of the system. An agent's result is a property of model, orchestration, tools, retry policy, and context management together. A claim about "the model" supported by an evaluation of a heavily engineered scaffold is a construct-validity defect.

Partial success needs a definition. An agent that completes eight of ten steps and then takes an incorrect irreversible action has not achieved 80% of the task. Binary task success obscures the cases that matter most.

The environment becomes attackable. Which is the subject of the next section.

10 Evaluation security

D7 asks a single question: **was the evaluation environment trustworthy enough that its output can be believed?**

Assessed: **sequestration** of test data; **isolation** — whether the system could reach documentation, repositories, search, or reference solutions; the **scoring trust boundary**; **simulation fidelity** — real tools or stubs; **credential hygiene**, including whether evaluation credentials could reach production; **evaluation logging** held outside the evaluated system's control; **containment of the evaluation itself**; and **environment drift** during the evaluation period.

> **The critical property.** The scoring mechanism must not be manipulable by the evaluated system. Where it is, the evaluation measures manipulation as readily as competence, and every figure it produced is unreliable in an unknown direction — which is worse than being wrong by a known amount.

Public remediation of exactly this defect class in widely used agentic coding benchmarks — stripping repository history so agents cannot retrieve the fix, hardening graders against agent-written files — establishes that this is an expected condition rather than a hypothetical.

10.1 Boundary with Containment Assurance

MQ-EVA does not test production containment, does not compute Containment Margin, and does not issue a deployment recommendation on containment grounds. Where D7 finds that evaluation activity escaped its environment, or that evaluation credentials could reach production, MQ-EVA records the finding and *refers*. Scope is enforced in both directions.

11 Reporting model

A standard engagement produces seven artifacts: the **Executive Evaluation Assurance Report** (15–30 pages); the **Claim Register** recording claims verbatim; the **Evidence Register**; the **Production Evaluation Gap Analysis**; the **Failure Register**; the **Remediation Roadmap** (Immediate / 30 days / 90 days / Longer-term, each item with a closure criterion); and the **Executive Assurance Statement**.

The scorecard carries one row per dimension with its rating, evidence grade, and a one-sentence key finding. **There is no total row.** The absence of a bottom line is the design.

Remediation in MQ-EVA is usually about *evidence* rather than the system: obtain per-trial traces, re-run with disclosed exclusions, construct a sealed set, measure under production constraints, test the untested severe mode.

11.1 The Assurance Statement

> Based on the evidence reviewed and the testing performed within the defined scope, Maple Quanta assesses the stated performance claims as **Supported / Partially Supported / Insufficiently Supported**. This assessment applies to the model version, configuration, environment, and controls described, and is invalidated by material change to any of them. This is an independent assurance assessment, **not a certification**, and Maple Quanta is not an accredited certification body.

The statement carries its scope and its expiry conditions in the same paragraph as its conclusion, so it cannot be quoted without them.

11.2 The Evaluation Disclosure Sheet

MQ-EVA publishes a free public template implementing the Canadian AI Safety Institute's disclosure guidance across system, evaluation, environment, results, and limitations. Two rules govern its use. **Leave nothing blank** — a blank is ambiguous between "not applicable", "not recorded", and "not disclosed". And apply **layered disclosure**: mark withheld items with a tier and a reason. Withholding with a stated reason is disclosure; silence is not.

12 Relationship to existing standards and guidance

Sources are classified by evidentiary weight, because a published standard, a draft standard, government guidance, academic research, and a vendor system card are five different kinds of thing.

Source	Status	Relationship to MQ-EVA
NIST AI 100-1 (AI RMF 1.0, 2023)	Published	MEASURE tells organisations to measure; MQ-EVA assesses a measurement already produced
NIST AI 600-1 (GenAI Profile, 2024)	Published	Establishes risk categories aggregate accuracy does not capture
NIST AI 200-2 (TEVV-Athlon)	**Draft (ipd, Aug 2026)**	Structurally compatible; MQ-EVA claims no conformance to an unpublished document
NIST ARIA	Programme	Shares the laboratory–operational divergence premise
NIST AITE (2026)	Programme	Endorses sequestration as a structural contamination control
CAISI disclosure guidance (Jul 2026)	Gov't guidance	Implemented by the Evaluation Disclosure Sheet
International network best practice	Gov't guidance	Elicitation as methodology; decision-relevance first
ISO/IEC 42001:2023	Published	Governs the management system, not the adjudication of a claim
ISO/IEC 23894:2023	Published	Risk management guidance
ISO/IEC 42005:2025	Published	Impact assessment; complementary and distinct
ISO/IEC 42006:2025	Published	Governs *certification bodies* — and is why MQ-EVA states plainly that it is not one
ISO/IEC 42007	**Draft (DIS)**	Conformity assessment schemes; tracked, not claimed
ISO/IEC 25059:2023	Published	Quality vocabulary, including robustness under reliability
ISO/IEC TS 4213:2022	Published (TS)	Classification performance; predates contamination and elicitation problems
ISO/IEC 24029-1/-2	Published	Neural-network robustness, including formal methods

> **Net position.** The ISO/IEC 42000-series governs the organisation that builds or uses AI, and the bodies that certify those organisations. The 25000-series and 24029 supply quality vocabulary and specific robustness techniques. **None provides a method for independently interrogating an evaluation result already produced and now being used to justify a decision.**

13 Relationship to Maple Quanta Containment Assurance

	MQ-EVA	**Containment Assurance**
Question	Can we trust the evidence that this system performs as claimed?	Can we bound what this system can do when it behaves unexpectedly?
Object	A claim and its evidence	A deployed system and its controls
Measures	ECL, PEG, ERA, MCEF, Coverage	Exposure Profile, MCAI, Containment Margin
Top evidence rung	E4 Adversarially Validated	E4 Operational
Failure means	The number cannot be relied on	The consequence cannot be bounded

The two are independent. **A capable system can pass MQ-EVA and fail Containment Assurance**: its performance claims are well-evidenced and it can still reach systems it should not. **A well-contained system can pass Containment Assurance and fail MQ-EVA**: nothing it does is dangerous, and there is no credible evidence it does the job.

An agentic system entering production in a regulated context generally needs both. MQ-EVA establishes whether the performance justifying the deployment is real; Containment Assurance establishes whether the organisation survives the cases where it is not.

14 Relationship to MQ-AICIR

> **Notation warning.** MQ-AICIR's E0–E4 are *event states*. MQ-EVA's E0–E4 are *evidence grades*. They share a letter and nothing else.

The governing rule: **an evaluation failure is not an AI incident**. Evaluations exist to produce failures; forcing incident classification onto ordinary model error destroys the signal that makes incident data useful.

Observed during evaluation	Classification
Incorrect answer, missed extraction, failed task	Evaluation failure. Failure Register. *Not* an AI incident
Unauthorised tool invocation inside the evaluation environment	Candidate MQ-AICIR E2 Boundary Violation
Attempted external interaction from a sandbox	Candidate MQ-AICIR E2 Boundary Violation
Evaluation activity reaching a real system or affecting real data	MQ-AICIR E3 Incident
Evaluation credentials able to reach production	MQ-EVA D7 Critical Concern; MQ-AICIR E0 Hazard
Post-deployment harm contradicting an assessed claim	MQ-AICIR E3/E4, linked to the claim record; triggers reassessment

The lifecycle closes: **Evaluate → Deploy → Monitor → Detect → Classify → Respond → Re-evaluate.** When an incident contradicts a claim assessed as Substantiated, that is information about the *evaluation* as well as the system, and it should change how similar evidence is weighted next time.

15 Worked example

The following case is entirely fictional. It contains no real client, department, vendor, product, or data. Figures are constructed to illustrate the methodology.

A federal department plans to deploy an agent to review procurement documents. The preferred bidder claims 93% extraction accuracy, a 50% productivity increase, "enterprise safe" operation, and effective bilingual performance. A Tier 2 engagement is commissioned.

Reconstruction finds that the exact model version was never stated; that 400 runs were executed and 362 reported, with 38 excluded under no recorded criterion; that prompts were adjusted during the run with elicitation undisclosed; that the benchmark ran with unrestricted network access, stubbed tools, and no latency, rate, or approval constraints; and that 24 of 400 documents were French, pooled into a single aggregate so that no French figure was ever produced.

Independent reproduction on a sealed 240-item English set — three trials per item, egress blocked, real tools, scoring executed outside any container the agent could write to — gives **88.0% (95% CI 83.4–91.4%)**. The reported 93% falls outside that interval. **ERA: Not reproduced.**

Production-constrained re-execution on the same items and version, under the department's egress allow-list, access controls, 8-second latency budget, rate limits, production retrieval corpus, approval gate, and ingest redaction, gives **76.0% (95% CI 70.3–81.0%)**.

PEG = 12 points (95% CI 8.1–16.1), on identical items.
The 17-point difference between the vendor's 93% and the production 76% is **not** a PEG. Different items, different exclusions, an unidentified version — the commensurability preconditions fail, and the framework refuses to compute it. That refusal costs a rhetorically powerful sentence and buys a finding that survives challenge.

French performance under the same constraints reaches **68.0%** against 76.0% on matched English documents, versus an operationalised parity threshold of three points.

Failure stress testing finds a silent wrong-value extraction at 2.9% (S3, hard to detect, hard to reverse), fabricated clause references at 1.7% (S3), and empty fields on scanned documents at 7.9% (S2, obvious, fully reversible). Note the inversion: the most frequent failure is the least serious. And **MCEF is set by a mode nobody observed** — cross-document disclosure of personal information in batch processing, which is how the department intends to operate, and which neither party tested. `MCEF = S4 (Not excluded)`.

Claim	Finding	ECL
93% extraction accuracy	Contradicted	C0
50% productivity increase	Unsupported	C0
"Enterprise safe"	Partially Substantiated	C1 (narrowed claim)
Bilingual performance	Contradicted	C0
≥90% expiry dates, structured English PDFs	**Substantiated**	**C3**

Evidence Coverage: **4 of 5 material claims (80%)**. Coverage is high *and* four findings are adverse — which is the point of reporting them separately. There was ample evidence to examine, and examining it is what produced the findings.

Notice why the one substantiated claim succeeded: it is the narrowest in the set — one field, one document type, one language, one stated threshold. **Precise claims are assessable; broad ones are not.**

15.1 How this changes the executive decision

The recommendation is not to reject the vendor. It is that the evidence supports a narrower deployment than the one proposed: begin with expiry-date extraction on structured English PDFs, the claim substantiated at C3; require human verification of extracted contract values pending remediation; test batch processing before any batch deployment, because that is the S4; obtain a French-specific evaluation or accept that the bilingual requirement is unmet; and build the sealed set into the next procurement's acceptance criteria.

None of this required deciding whether the system is good. It required determining what the evidence established — which was one claim out of five, and a valuable one.

16 Limitations

- **Access governs the ceiling.** Where re-execution is refused or artifacts are withheld, evidence caps at E2 and confidence at C2. Refusal is recorded as a finding, but it cannot be tested around.
- **Absence of contamination cannot be proven.** It can be made unlikely through sealed material, and it can be made visible through indicators. Neither is proof.
- **Conclusions are configuration-specific** and are invalidated by material change to version, configuration, prompts, tools, data, environment, or controls.
- **Assessment is not a guarantee.** A claim substantiated at C4 can still fail in deployment; assurance addresses the evidence, not the future.
- **Severity scales are client-defined**, so severity is not comparable across engagements without stating the scale.
- **Maple Quanta is not independent of Maple Quanta.** Engagements are typically I2, reaching I3 only where Maple Quanta constructed the sealed set and scope was unrestricted. The commissioning relationship is disclosed in every report.
- **The framework will change.** NIST's TEVV-Athlon and ISO/IEC 42007 are drafts. MQ-EVA tracks them and will be revised as they are published.

17 Principles

1. A high benchmark score does not prove production reliability.
2. A vendor claim is not independent evidence.
3. An average score can hide a catastrophic failure.
4. Evaluation conditions are part of the result.
5. A system should be tested under the constraints in which it will actually operate.
6. Evaluation evidence should be reproducible, traceable, and proportional to the consequence of the decision it supports.
7. Absence of evidence is not evidence of absence — an untested failure mode is an open risk, not a passed test.

> **The question is not simply whether the AI passed the test.**
> **The question is whether the test deserves to be trusted.**

References

[1] National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1. https://www.nist.gov/itl/ai-risk-management-framework

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

[3] National Institute of Standards and Technology (2026). *The TEVV-Athlon Framework for Evaluating AI Systems*, NIST AI 200-2 (Initial Public Draft, 7 August 2026). https://doi.org/10.6028/NIST.AI.200-2.ipd

[4] National Institute of Standards and Technology (2024). *Assessing Risks and Impacts of AI (ARIA)*. https://www.nist.gov/news-events/news/2024/05/nist-launches-aria-new-program-advance-sociotechnical-testing-and

[5] National Institute of Standards and Technology (2026). *Announcing NIST's Artificial Intelligence Technology Evaluation (AITE)*, 27 July 2026. https://www.nist.gov/news-events/news/2026/07/announcing-nists-artificial-intelligence-technology-evaluation-aite

[6] Canadian Artificial Intelligence Safety Institute (2026). *What information should AI evaluators share?*, 8 July 2026. https://ised-isde.canada.ca/site/ised/en/canadian-artificial-intelligence-safety-institute/what-information-should-ai-evaluators-share

[7] UK AI Security Institute (2026). *International evaluation best practice and open questions in AI measurement*, 23 July 2026. https://www.aisi.gov.uk/blog/international-evaluation-best-practice-and-open-questions-in-ai-measurement

[8] Bean, A. M., et al. (2025). *Measuring what Matters: Construct Validity in Large Language Model Benchmarks*. arXiv:2511.04703. https://arxiv.org/abs/2511.04703

[9] Eriksson, M., et al. (2025). *Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation*. European Commission Joint Research Centre. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8(1), 850–864. https://arxiv.org/abs/2502.06559

[10] Miller, E. (2024). *Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations*. arXiv:2411.00640. https://arxiv.org/abs/2411.00640

[11] Longpre, S., et al. (2024). *A Safe Harbor for AI Evaluation and Red Teaming*. arXiv:2403.04893. https://arxiv.org/abs/2403.04893

[12] METR (2026). *Task-Completion Time Horizons of Frontier AI Models*. https://metr.org/time-horizons/

[13] ISO/IEC 42001:2023, *Information technology — Artificial intelligence — Management system*.

[14] ISO/IEC 42005:2025, *Information technology — Artificial intelligence — AI system impact assessment*. https://www.iso.org/standard/42005

[15] ISO/IEC 42006:2025, *Requirements for bodies providing audit and certification of artificial intelligence management systems*. https://www.iso.org/standard/42006

[16] ISO/IEC DIS 42007 (draft), *High-level framework and guidance for the development of conformity assessment schemes for AI systems*. https://www.iso.org/standard/89967.html

[17] ISO/IEC 25059:2023, *Software engineering — SQuaRE — Quality model for AI systems*.

[18] ISO/IEC TS 4213:2022, *Assessment of machine learning classification performance*. https://www.iso.org/standard/79799.html

[19] ISO/IEC 24029-2:2023, *Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods*. https://www.iso.org/standard/79804.html

[20] Maple Quanta Inc. (2026). *Agentic AI Containment Assurance: A Practical Framework for Measuring Control over Autonomous AI Systems*, Version 1.1. https://maplequanta.ca/agentic-ai-containment/

[21] Maple Quanta Inc. (2026). *MQ-AICIR — AI Incident Classification and Reporting Framework*, Version 1.0.

Maple Quanta Inc. | Ottawa, Canada | maplequanta.ca | Federal Corp. No. 1801626