

Public Comment on NIST AI 200-2 ipd

The TEVV-Athlon Framework for Evaluating AI Systems

To TEVV-Athlon@nist.gov

Subject NIST AI 200-2 ipd - Comments on Evidence Traceability, System-State Binding, and Bounded Interpretation

Dear NIST AI 200-2 authors and reviewers,

ODA3 Institute appreciates the opportunity to comment on the Initial Public Draft of NIST AI 200-2, "The TEVV-Athlon Framework for Evaluating AI Systems." The draft provides a useful and flexible structure for translating evaluation objectives into Metrology Blocks, Events, Tools, evidence, and synthesized organizational insight. In particular, the separation of Articulate & Organize, Define & Construct, Apply & Measure, and Synthesize & Interrogate (§2.1-§2.4, pp. 3-7), together with the framework's emphasis on measurement validity, uncertainty, realistic testing, and risk-proportionate rigor, provides a strong foundation for practical AI evaluation.

Our comments focus on a narrower question: how evidence generated through a TEVV-Athlon remains attributable to the evaluated system and how subsequent conclusions remain bounded to what that evidence actually establishes.

1. Make evidence provenance and traceability explicit

Relevant provisions: §2.3, p. 6; §2.4, p. 7.

The draft defines Events as activities that produce evidence for Metrology Blocks, and the Toolbox as the methods and instruments used to elicit, collect, and analyze information. Synthesize & Interrogate subsequently works back through the evaluation structure to interpret the collected information.

We recommend making the traceability of that evidence chain more explicit. For evidence produced during an Event, an evaluator should be able to determine the relevant Metrology Block, Event, Tool or instrument, evaluated system state, execution context, and any material transformation or aggregation through which the evidence passed before being relied upon in synthesis. This need not prescribe a specific database, schema, storage technology, or evidence format.

2. Bind evaluation evidence to the material system state actually tested

Relevant provisions: §2.1, p. 3; §4.5, p. 15.

The draft recognizes that lifecycle stage and operational context affect evaluation design, and cautions that results may not generalize beyond the conditions under which they were obtained. It also notes that changing deployment settings, populations, or system capabilities may require periodic review and revision of measurement approaches.

We recommend that TEVV execution records identify the material system and configuration state under evaluation - for example, model or system version, relevant configuration, enabled

capabilities or tools, operating environment, material dependencies, and execution period. The level of detail should remain risk-based; the objective is not exhaustive configuration capture, but enough state information for a later reviewer to determine what was actually tested and whether a subsequent system state remains within the evaluated boundary. Stage 4 conclusions should remain bounded to that recorded state unless additional evidence supports broader generalization.

3. Distinguish valid measurement production from evidentiary reliance

Relevant provisions: §4.4-§4.5, p. 15.

The draft contains strong guidance on measurement validity, assumptions, limitations, experimental design, uncertainty, and calibration. We recommend making one additional distinction explicit: producing a valid measurement is not necessarily the same as establishing the broader proposition or conclusion for which that measurement is later relied upon.

Apply & Measure addresses how evidence is generated; Synthesize & Interrogate addresses how that evidence is interpreted. Between those activities, it may be useful to establish an explicit analytical checkpoint requiring the evaluator to identify the evidence actually relied upon, the proposition or conclusion to which it bears, material assumptions, relevant alternative explanations, and limitations constraining the inference.

4. Preserve materially different negative and indeterminate outcomes

Relevant provisions: §2.4, p. 7; §4.4, p. 15.

TEVV can produce outcomes not faithfully represented by a simple positive/negative result - for example: an expected behavior tested and not observed; a procedure that failed to complete; unavailable evidence; a system's explicit refusal of an operation; contradictory observations; or a measurement that cannot resolve the question examined. These carry different evidentiary meanings and should not collapse into one generic failure or limitation category.

We do not propose a universal taxonomy - only that a later reviewer be able to distinguish what was observed, what was not observed under a defined test, what could not be evaluated, and what remained unresolved.

5. Make bounded interpretation an explicit Stage 4 reporting expectation

Relevant provisions: §4.5, p. 15; §4.6, p. 16.

The draft already states that measurement results can be misleading when methodological limitations are ignored, and warns that strong benchmark performance does not always indicate broader system quality or deployment suitability. We recommend operationalizing this directly within Synthesize & Interrogate: in addition to describing what the evidence supports, a report should identify material boundaries on interpretation and, where relevant, state what the evaluation does not establish - untested environments, configurations, populations, capabilities, operational paths, system states, or time periods.

6. Clarify the current scope for physical or embodied AI evaluation

Relevant provisions: Appendices A-C, pp. 21-26.

The Toolbox appendices provide substantial coverage for software- and model-mediated evaluation. We did not find corresponding evaluation categories addressing systems that can

initiate or influence physical action - for example, actuator behavior, physical state transitions, degraded or safe-state operation, or intervention/override mechanisms - anywhere in the draft. We do not suggest the generic architecture is inapplicable to such systems; NIST may wish to identify this as an explicit scope boundary, or consider additional illustrative guidance in a future revision.

7. Expand the existing treatment of agentic-system evaluation

Relevant provisions: Appendix B, Table 4, p. 24.

The draft already recognizes agent and tool abuse within its red-teaming methods - unsafe tool selection, excessive agency, unauthorized action, harmful task execution. NIST may wish to expand this existing treatment to address evaluation patterns specific to longer-duration or stateful agentic execution: sequences of tool use, delegated or changing authority, cross-step dependencies, persistent state, intervention or escalation, and multi-agent interaction where relevant.

Cross-cutting observation

These recommendations do not call for a more prescriptive TEVV-Athlon; the draft's flexibility is a strength. They concern preserving the relationship between what was intended to be measured, how it was measured, which system state was actually evaluated, what evidence was produced, what that evidence supports, and what remains outside the resulting conclusion. Making that chain more explicit would strengthen the framework for technical evaluation, organizational decision-making, procurement, and independent review, without requiring a uniform methodology across diverse AI systems.

Thank you for the opportunity to comment on the Initial Public Draft.

Sincerely,

ODA3 Institute
ODA3 Pvt Ltd
India
contact@oda3.org