



October 6, 2026

National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Re: NIST AI 200-2 Initial Public Draft, The TEVV-Athlon Framework for Evaluating AI Systems

To Whom It May Concern:

TechNet is the national, bipartisan network of technology CEOs and senior executives that promotes the growth of the innovation economy. Our diverse membership includes American businesses ranging from startups to the world's leading companies and represents five million employees across information technology, artificial intelligence, e-commerce, advanced energy, transportation, cybersecurity, venture capital, finance, and other sectors. Given the central role our members play in developing, deploying, and using AI systems, TechNet appreciates the opportunity to comment on NIST AI 200-2, *The TEVV-Athlon Framework for Evaluating AI Systems*. We support NIST's effort to provide organizations with a flexible method for translating high-level objectives into evaluations tailored to an AI system, its intended use, and its operating environment.

The draft makes important progress toward a practical TEVV framework. It recognizes that measurement should reflect organizational objectives and real-world use, integrates the NIST AI Risk Management Framework, and acknowledges the limits of benchmarks and individual metrics. It also includes red teaming, user testing, field testing, measurement validation, and scientific measurement practices as elements that organizations may draw upon. These additions address several issues TechNet raised in response to NIST's earlier proposed standard on AI TEVV.

The next draft should retain the framework's flexible, risk-based foundation while providing clearer and more practical guidance on how organizations should design evaluations, interpret and report results, allocate responsibilities, test safeguards, and determine when reevaluation is necessary. Specifically, TechNet recommends that NIST:

- Preserve a risk-based and scalable framework;
- Clarify core TEVV terminology;
- Connect evaluation results to decisions and acceptance criteria;
- Make results reproducible and fit for comparison;
- Protect the integrity of the evaluation process;
- Clarify shared responsibility across the AI value chain;

- Expand agentic and system-level evaluation;
- Evaluate the effectiveness of safeguards;
- Establish change triggers for lifecycle reevaluation;
- Promote interoperability and evidence reuse; and
- Provide additional templates and worked examples.

These recommendations would improve the framework's practical utility without turning it into a certification regime.

Preserve a Risk-Based and Scalable Framework

TechNet supports the draft's central premise that an effective TEVV process must be adaptable to the system, use case, lifecycle stage, and potential impact at issue. NIST should preserve this risk-based structure and state more directly that the framework is not intended to require every organization to use every event or tool. The rigor, scope, frequency, and formality of an evaluation should be proportionate to the reasonably foreseeable risks and the significance of the decision the evaluation will inform.

This clarification is particularly important for startups, small businesses, and organizations deploying lower-risk systems. A TEVV process that assumes extensive compute, specialized personnel, or access to model internals could be impractical for many organizations and may discourage adoption without producing commensurate risk reduction. The framework should recognize several valid levels of evaluation, ranging from routine product testing and monitoring to more intensive assessment for high-risk uses. It should also distinguish between recommended practices and minimum elements needed to describe an assessment credibly.

NIST should further clarify that a TEVV-Athlon produces evidence for a defined purpose. It does not establish that a system is universally safe, trustworthy, or suitable for every context. Evaluation findings are bounded by the system version, tested conditions, measurement methods, assumptions, and intended use. An explicit statement to this effect would reduce the risk that a result is treated as a general approval or certification beyond the scope of the assessment.

Clarify the Core TEVV Terms

The draft uses TEVV as an umbrella term for methods that collect information about system performance, but the practical distinctions among testing, evaluation, verification, and validation remain important. NIST should define each activity and explain its purpose, typical evidence, expected output, and relevant lifecycle stages. Testing generally examines system behavior under specified conditions; evaluation interprets evidence against defined objectives or criteria; verification determines whether the system meets established requirements or design specifications; and validation determines whether it is suitable for its intended use and operating environment. NIST should also distinguish TEVV from related activities such as measurement, assessment, monitoring, and auditing. Although these activities may

overlap, they answer different questions and should not be treated as interchangeable.

The glossary should separately clarify the relationships among system attributes, trustworthiness characteristics, metrology blocks, events, tools, metrics, measures, benchmarks, and assessment results. The “Athlon” metaphor is accessible, but the number of defined components may create uncertainty for practitioners mapping existing evaluation programs to the framework. NIST should provide a single worked example that follows an evaluation from the selected system attribute through the final assessment result and identifies the role of each component along the way. This would help organizations apply the framework without renaming or recreating mature internal processes.

Connect Evaluation Results to Decisions and Acceptance Criteria

The “Articulate and Organize” stage asks evaluators to state the goals of an assessment and identify who will care about the results. NIST should go one step further by asking what decision the evaluation is intended to inform. Examples could include whether to deploy a system, expand it to a new use, modify safeguards, select among vendors, continue monitoring, or pause operation. Identifying the decision at the outset would help evaluators select evidence that is relevant to an actual organizational need.

Where appropriate, organizations should also identify acceptance criteria, baselines, or escalation thresholds before testing begins. These criteria need not take the form of universal numerical thresholds. They may be quantitative, qualitative, comparative, or process-based depending on the system and context. Defining them in advance can reduce post hoc interpretation and make clear how results will affect deployment or risk management decisions.

The “Synthesize and Interrogate” stage should then direct evaluators to document the resulting decision, material limitations, residual risks, and any follow-up action. The framework currently explains that TEVV findings inform the AI RMF Manage function, but practitioners would benefit from a more explicit bridge. NIST could provide a simple decision record that links each material finding to an owner, mitigation, monitoring plan, or reason for accepting the remaining risk. This would keep the framework focused on evidence and accountability without prescribing the organization’s substantive risk tolerance.

Make Results Reproducible and Fit for Comparison

The draft appropriately emphasizes scientific measurement, uncertainty, and the limits of benchmarks. NIST should translate those principles into a minimum set of information that accompanies a TEVV-Athlon report. At a minimum, the report should identify the system and version evaluated, relevant configuration and sampling settings, the tested use and environment, evaluation data or prompts, selection and exclusion criteria, tools and scoring methods, evaluator roles, dates, known

limitations, uncertainty, and material deviations from the planned procedure. Where disclosure of sensitive or proprietary details would create security or intellectual property risks, organizations should be able to provide protected summaries or other evidence appropriate to the intended audience.

This documentation is necessary to interpret results over time and determine whether comparisons are valid. Two systems tested with different prompts, user populations, safeguards, or operating conditions may produce scores that look comparable but answer different questions. NIST should caution against ranking systems or combining results unless the evaluation objectives, conditions, measurement methods, and uncertainty are sufficiently aligned. The framework should similarly encourage reporting distributions, confidence intervals, and failure patterns when a single aggregate score would conceal meaningful variation.

NIST should also encourage versioning for evaluation artifacts, including datasets, prompts, rubrics, tools, and reports. Versioning supports repeatability, helps organizations trace why results changed, and allows an assessment to remain useful as the system evolves. A standardized report template and machine-readable fields for core metadata could improve evidence reuse and enable automation without imposing a single metric or assessment method.

Clarify Shared Responsibility Across the AI Value Chain

Many AI systems are assembled from models, data, cloud services, retrieval components, third-party tools, and application-specific controls supplied by different organizations. No single participant will always possess the access or information needed to evaluate the entire system. The framework should therefore direct organizers to identify the relevant roles at the beginning of a TEVV-Athlon and determine which participant is best positioned to test each component, provide evidence, or monitor performance.

This shared responsibility model should distinguish among information available to model developers, system integrators, deployers, operators, and downstream users. A developer may be best positioned to evaluate general model capabilities or model-level security, while a deployer may be best positioned to evaluate performance in a specific workflow, user population, or operational environment. The final system assessment may need evidence from both. NIST should avoid implying that downstream organizations require unrestricted access to model weights, training data, or other sensitive materials in order to conduct meaningful TEVV.

The framework should instead support proportionate evidence-sharing mechanisms, such as evaluation summaries, documentation, technical interfaces, contractual assurances, or independent assessments where appropriate. It should also encourage reuse of credible evidence when the system, use, and evaluation context are sufficiently aligned. Requiring each participant to recreate the same test can increase cost, delay deployment, and expose sensitive information without improving the quality of the final decision.

Expand Agentic and System Level Evaluation

The draft states that the framework can apply to agentic systems, and Appendix B includes agent or tool abuse within its red teaming examples. NIST should expand this treatment in the main body and add a worked example focused on an agent operating across tools or services. Model-level output testing alone may not capture risks that arise when a system plans across multiple steps, retains memory, invokes external tools, changes data, communicates with other agents, or takes actions with limited human involvement.

For agentic systems, relevant blocks and events may include adherence to authorization boundaries, resistance to direct and indirect prompt injection, safe tool selection, identity and permission handling, maintenance of the intended objective over long task sequences, detection of conflicting instructions, recovery from tool or network failures, and the ability to stop, defer, or request human review. Evaluations should also consider cumulative effects that may not appear in a single interaction, including repeated actions, memory contamination, cascading errors, and interactions among multiple components or agents.

A system-level example should show how to test both successful task completion and compliance with constraints. It should also address realistic sandboxing, logging, rollback, rate limits, access controls, and human override mechanisms. These measures are often implemented outside the underlying model, but they materially affect the safety and reliability of the deployed system and therefore belong within a complete evaluation plan.

Use Change Triggers for Lifecycle Reevaluation

The draft recognizes that measurement validation is ongoing and that changing capabilities or environments may require review. NIST should make the framework explicitly iterative by connecting the “Synthesize and Interrogate” stage back to “Articulate and Organize”. The result of one TEVV-Athlon may identify new risks, change organizational objectives, or reveal that the original blocks and tools no longer capture the relevant outcomes.

Rather than suggesting continuous retesting in every case, NIST should recommend risk-based triggers for reevaluation. Relevant triggers may include a material model or system update, a new use or user population, a change in tools or data sources, evidence of performance drift, a significant incident or near miss, a newly identified attack technique, or a material change in the operating environment. Organizations should determine the scope of reevaluation based on the nature of the change. Some changes may require only targeted regression testing, while others may warrant a broader reassessment.

The framework should also distinguish ongoing monitoring from periodic or event-driven evaluation. Monitoring can identify trends, anomalies, or threshold breaches in operation, while a structured TEVV-Athlon can investigate those signals and support a

decision. Clarifying this relationship would help organizations integrate the framework into existing development, deployment, incident response, and change-management processes.

Promote Interoperability and Evidence Reuse

The framework's close connection to the NIST AI RMF is a strength. NIST should also map the framework's concepts and reporting fields to widely used standards and sector-specific risk management regimes, including ISO/IEC 42001 on AI management systems, ISO/IEC 23894 on AI risk management, ISO/IEC 25059 on the quality model for AI systems, and applicable cybersecurity, privacy, safety, and model risk management frameworks. Crosswalks would allow organizations to use existing governance records and evaluation evidence rather than maintain parallel processes that request substantially similar information in different formats.

NIST should treat these mappings as aids to interoperability rather than statements of equivalence. Different frameworks have different purposes and scopes. A practical crosswalk should identify where existing documentation can serve as an input to a TEVV-Athlon, where additional evidence may be needed, and how the final report can support other assurance or governance activities. NIST should develop and update these resources through an open process with industry, civil society, academia, and sector-specific experts.

Provide Additional Templates and Worked Examples

The query-violation example explains the framework's basic mechanics, but it focuses on a relatively narrow chatbot scenario. NIST should add examples that demonstrate how the same four stages scale across different systems and levels of risk. Useful examples could include a lower-risk enterprise productivity tool, a consequential decision-support application, a computer vision system, and an agent that can act through external tools. At least one example should involve a system assembled from multiple providers so that NIST can illustrate shared responsibility and evidence reuse.

Practical templates would also improve adoption. NIST could provide an optional scoping worksheet, evaluation plan, report structure, change-trigger log, and decision record. These materials should be modular and editable so organizations can incorporate them into existing processes. NIST should pilot the templates with organizations of different sizes and sectors before finalizing them and should maintain the toolbox as a living resource as evaluation methods and AI capabilities change.

Conclusion

The TEVV-Athlon Framework offers a useful structure for turning broad organizational goals into context-specific AI evaluations. Its value will depend on preserving flexibility while giving practitioners enough guidance to produce evidence that is interpretable, repeatable, and relevant to a decision. Clearer terminology, decision criteria, reporting expectations, shared responsibility, agentic system examples,

change triggers, and cross-framework mappings would strengthen the draft without converting it into a rigid compliance checklist.

TechNet appreciates NIST's continued leadership in advancing practical, consensus-based approaches to AI measurement and risk management. We look forward to working with NIST and other stakeholders to refine the TEVV-Athlon Framework and support flexible, risk-based TEVV practices that strengthen trust in AI systems while enabling continued innovation and adoption.

Sincerely,

A handwritten signature in blue ink that reads "Linda Moore". The signature is fluid and cursive, with the first name "Linda" and last name "Moore" clearly distinguishable.

Linda Moore
President and CEO