

October 6, 2026

P. Jonathon Phillips
Technology Test and Evaluation Division
Information Technology Laboratory
National Institute of Standards and Technology

Submitted by email to TEVV-Athlon@nist.gov

**RE: Comments of the Association for Competitive Technology on NIST AI 200-2 ipd,
The TEVV-Athlon Framework for Evaluating AI Systems**

The Association for Competitive Technology (ACT) appreciates the opportunity to submit comments to the National Institute of Standards and Technology (NIST) on the initial public draft of NIST AI 200-2, The TEVV-Athlon Framework for Evaluating AI Systems. ACT supports NIST's effort to provide a structured but flexible approach for assessing the real-world performance and impacts of artificial intelligence (AI) systems, an effort that advances the call in America's AI Action Plan for building an AI evaluations ecosystem and supporting the science of measuring and evaluating AI systems.¹ The framework's emphasis on tailoring test, evaluation, verification, and validation (TEVV) activities to organizational goals, system characteristics, lifecycle stage, risk, and operational context provides a useful foundation for practical AI evaluation.²

ACT is a not-for-profit global policy trade association for the small business technology developer community. Our members are entrepreneurs, innovators, and independent developers who build products and services across every industry and who create and deploy AI technologies in both consumer and enterprise settings. We work with our members and policymakers to promote a policy environment that rewards and inspires innovation while providing resources that help developers raise capital, create jobs, and continue to build technology.

The small businesses and startups we represent take part in international standards development (and many more would like to) because they leverage the efficiency of technical standards to build their products, including in the context of AI. Smaller firms nonetheless remain underrepresented in AI policy, standards, and evaluation processes and face practical barriers to meaningful participation, among them cost and the staff time

¹ The White House, America's AI Action Plan 10 (July 2025), <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.

² P. Jonathon Phillips et al., NIST, The TEVV-Athlon Framework for Evaluating AI Systems, NIST AI 200-2 ipd (Aug. 2026), <https://doi.org/10.6028/NIST.AI.200-2.ipd>.

these processes demand. Standards bodies and governments should work to lower those barriers so that the resulting approaches reflect the full range of firms that implement and are impacted by them.

The TEVV-Athlon Framework's four stages, Articulate & Organize, Define & Construct, Apply & Measure, and Synthesize & Interrogate, provide a clear structure for translating high-level goals into evidence that can inform real-world decisions. ACT recommends that NIST preserve the framework's risk-based and customizable character while strengthening guidance for small developers, shared responsibility across the AI value chain, agentic and composite systems, technology-neutral evaluation, and interoperability with existing standards and references.

I. NIST Should Preserve the Framework's Flexible, Risk-Based, and Context-Specific Approach

ACT strongly supports NIST's decision not to prescribe a rigid, one-size-fits-all evaluation regime. The U.S. posture regarding AI generally should be to rely on risk-based methods to ensure that the use of AI aligns with any relevant recognized standards of safety and efficacy. Small software and device companies benefit from understanding the distribution of risk and liability in building, testing, and using AI tools. ACT's global policy principles for AI similarly call for leveraging risk-based quality assurance and oversight with responsibility allocated to those best placed to mitigate risk; promoting human-centered and thoughtful design; prioritizing access and affordability; maintaining an ethical and inclusive approach to research and development; and respecting intellectual property, among other recommendations.³

NIST should therefore make explicit in the final framework that a TEVV-Athlon is not a universal checklist and that not every benchmark, red-teaming method, or user-testing technique will be necessary or appropriate for every AI system. Evaluation should be calibrated to intended and reasonably expected uses, lifecycle stage, the nature and magnitude of potential harms, and the evaluator's role and capabilities.

The framework already recognizes in Section 4.2 that systems with higher potential risk may require more rigorous measurement. NIST should elevate that principle throughout the final document and make equally clear that lower-risk systems and uses may appropriately rely on narrower, less resource-intensive TEVV approaches. This would reinforce the framework's core strength: flexibility that supports meaningful evidence rather than process for its own sake.

II. NIST Should Make TEVV Scalable and Usable for Small Developers

³ Association for Competitive Technology, Global Policy Principles for AI, <https://actonline.org/post/ai-policy-principles/>.

NIST should prioritize scalable, risk-based guidance and adoption-ready outputs for smaller developers. For emerging technologies, the policy design choices that agencies make can determine whether best practices become widely adopted or remain aspirational artifacts for only the largest firms. We encourage NIST to ensure its outputs are adoptable by smaller teams that do not have dedicated evaluation staff or large testing budgets.

Small businesses often devote most of their limited resources to product development and day-to-day operations. While larger companies may be more involved and educated about all stages of evaluation and standards processes, many small and medium-sized U.S. innovators face resource asymmetries that prevent them from tracking, designing, and conducting every potentially relevant TEVV activity. These limitations do not mean small developers take evaluation any less seriously; rather, the framework should help them direct scarce resources to the measurements that are most probative for the actual system and risk.

To make the framework more practically useful to smaller firms, NIST should expand the discussion in Section 4.2 of the minimum resources needed for a TEVV-Athlon into guidance on a scaled-down approach for small teams, supported by reusable templates, and should add an example TEVV-Athlon spotlighting a small developer with limited personnel and compute resources. The example in Section 3 of the draft concerns a chatbot, so the most useful addition could be a small developer building an agentic application on a third-party foundation model, which would also show how the framework handles the value-chain and agentic-system issues discussed in Sections III and IV of these comments.

III. NIST Should Reflect Shared Responsibility Across the AI Value Chain

ACT urges NIST to align the TEVV-Athlon Framework with the concept of shared responsibility across the AI value chain. The AI ecosystem includes infrastructure providers, model developers, application developers, deployers, and end users. Not all actors have visibility into or control over how systems are used in practice. Obligations and evaluation expectations should be allocated across the AI value chain based on who can mitigate risk.⁴ Specifically, those in the value chain who are best placed to minimize risks, given what they know and control, should have appropriate incentives to do so. The current framework appropriately begins with entities, stakeholders, organizational goals, and the operational environment. NIST should build on that structure by asking evaluators during the Articulate & Organize stage to identify their role in the AI value chain, what components they control, what information they can reasonably access, what risks they can meaningfully mitigate, and what evidence may need to come from other actors.

⁴ For a fuller treatment of how roles and responsibilities can be allocated among actors in the AI value chain, see the AI Roles and Interdependency Framework attached to the Association for Competitive Technology's comments on the Development of an Artificial Intelligence (AI) Action Plan (Mar. 2025).

This is especially important for small application developers that build on top of third-party foundation models, cloud services, APIs, datasets, or tools.

A small downstream developer should not be expected to recreate tests that require access to model internals it does not possess, just as an upstream model provider may not be positioned to evaluate every downstream deployment context. The framework should expressly support appropriate reliance on upstream documentation, third-party evaluations, contractual information, deployment-specific testing, and post-deployment evidence, while encouraging each actor to conduct the TEVV activities that correspond to its knowledge and control. The framework should also make clear that evidence shared for this purpose need not include trade secrets, proprietary model details, or other confidential business information. This approach would better reflect how modern AI systems are actually developed and deployed.

IV. NIST Should Expand Practical Guidance for Agentic and Composite AI Systems

NIST states that the TEVV-Athlon Framework is intended to accommodate agentic systems, and Appendix B appropriately includes “Agent / tool abuse testing,” including unsafe tool selection, excessive agency, unauthorized action attempts, and harmful task execution, which is an important starting point. The final framework would nonetheless benefit from integrating agentic AI more directly into the core four-stage methodology rather than addressing it primarily as one red-teaming category.

Agentic systems can differ from conventional model evaluation because the object under evaluation may be a dynamic system composed of a model, orchestration logic, memory or state, connected tools and APIs, permissions, external data sources, and human approval points. Harm or failure may emerge from a sequence of individually reasonable actions, from changes in an external tool, or from the interaction among components rather than from a single model output.

NIST should therefore show how each stage of the framework applies to agentic systems: At Articulate & Organize, evaluators should define the system boundary, including the tools, permissions, data sources, and human approval points within scope. At Define & Construct, adherence to those permissions and approval points can be defined as a Block, so that measurement covers whether the system stays within its authorized scope. At Apply & Measure, Events should test multi-step sequences of actions rather than only single outputs, including how the system responds when a connected tool changes or fails. The example TEVV-Athlon suggested in Section II would give NIST a natural place to illustrate these steps.

V. NIST Should Maintain Technology Neutrality Across Open and Closed AI Models

Dynamic competition across the AI ecosystem, including among foundation models, cloud infrastructure providers, specialized applications, edge computing solutions, domain-specific AI systems, and other market segments, is a powerful driver of innovation, affordability, and rapid capability gains. The AI ecosystem benefits from competition among large firms, small businesses and startups, open-source communities, and others. To preserve this dynamic, the TEVV-Athlon Framework should remain neutral among open-source, open-weight, proprietary, and closed AI models and systems.

The availability of source code, model weights, training information, or internal telemetry may enable particular forms of evaluation, but those forms of access should not become proxies for trustworthiness or prerequisites for meaningful TEVV. White-box techniques may be useful where internal access is available; black-box testing, interface-level testing, user testing, field testing, deployment-specific evaluation, and post-deployment monitoring may be more appropriate in other circumstances. The framework should recognize that meaningful evidence can come from different methods depending on the evaluator's goals, role, access, and context.

NIST should also clarify that the relevant evaluation target may be the deployed AI system rather than a foundation model in isolation. Many products combine third-party models with retrieval systems, application logic, safety controls, external tools, and user-interface choices. Evaluation should therefore focus on the actual system and intended use, while recognizing that responsibilities can shift when an actor modifies, fine-tunes, redistributes, or independently deploys a model.

VI. NIST Should Promote Interoperability with Voluntary Consensus Standards and Avoid Duplicative Processes

The TEVV-Athlon Framework should align with approaches reflected in widely used international standards and references. Key resources include the standards work of the International Organization for Standardization and the International Electrotechnical Commission, including ISO/IEC 42001, ISO/IEC 23894, the ISO/IEC 29119 software testing series, and ISO/IEC TS 42119-2 on the testing of AI systems, as well as the NIST AI Risk Management Framework (AI RMF). Unlike larger firms, small developers cannot easily navigate separate evaluation and assurance programs for every market they seek to enter or operate in. Where guidance maps cleanly to recognized international references, small businesses can build once and serve customers worldwide.

NIST should therefore continue mapping the TEVV-Athlon Framework to the AI RMF and, where useful, provide crosswalks or explanatory material showing how TEVV-Athlon evidence can complement existing voluntary consensus standards and assurance activities. The framework should help organizations reuse relevant evidence rather than

create duplicative processes, including by allowing the same evaluation results to support the public-facing AI documentation addressed in NIST AI 300-1.⁵

Continued support for private-sector and multistakeholder standards development serves the same end, as does resisting the temptation to convert voluntary standards wholesale into rigid regulatory checklists that may become outdated as the technology evolves. NIST should make clear that the TEVV-Athlon Framework is a flexible methodology for structuring evaluation evidence, not a substitute for sector-specific standards and requirements.

VII. NIST Should Continue to Emphasize Human-Centered and Real-World Evaluation

ACT supports design and evaluation of AI systems that are informed by real-world workflows, human-centered design and usability principles, and end-user needs. AI systems should facilitate changes in the delivery of goods and services that benefit consumers and businesses. The design, development, and success of AI should leverage collaboration and dialogue among users, AI technology developers, and other stakeholders to have all perspectives reflected in AI solutions.

NIST’s emphasis on user testing, field testing, measurement validity, uncertainty, and the limitations of benchmarks is therefore well placed. The final framework should reinforce that benchmark scores and individual metrics are context-bound pieces of evidence, not universal grades of safety or suitability. Evaluation reports should clearly communicate the objective, system configuration, intended use, relevant baseline, assumptions, uncertainty, known limitations, and the extent to which results are expected to generalize beyond the conditions tested.

This is especially important because TEVV results may be consumed by business decision-makers, procurement specialists, deployers, and other stakeholders who did not design the evaluation. Clear reporting of what an assessment does—and does not—establish will help prevent strong conclusions from being drawn from weak or incomplete evidence, while preserving the practical value of the framework for real-world decision-making.

VIII. Conclusion

The TEVV-Athlon Framework can be a useful bridge between high-level AI risk-management principles and practical evaluation. By preserving its flexible and risk-based approach, ensuring that it is workable for small developers, reflecting shared responsibility across the AI value chain, expanding practical treatment of agentic systems, maintaining technology neutrality, and promoting interoperability with existing standards and

⁵ NIST, Guidance and Templates for Public-Facing AI Documentation: An AI Standards “Zero Draft,” NIST AI 300-1 ipd (2026).

references, NIST can strengthen AI evaluation while preserving space for small business innovation and diverse technical approaches.

This document is approved for public dissemination. The document contains no business-proprietary or confidential information.

Sincerely,

Brian Scarpelli
Senior Global Policy Counsel

Kedharnath Sankararaman
Policy Counsel

Association for Competitive Technology
1401 K St NW, Suite 501
Washington, DC 20005