FRONTIER
MODEL FORUM

September 12 2025

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

To Whom It May Concern:

The Frontier Model Forum (FMF) is grateful to the NIST Zero Draft project for its effort to develop a standard on AI testing, evaluation, verification, and validation (TEVV). We share NIST's view that rigorous TEVV practices are essential for building trust in AI systems and ensuring their safe and responsible deployment across sectors.

As an industry-supported non-profit dedicated to the safe development of advanced general-purpose AI, we welcome the opportunity to provide input on the draft outline of a TEVV standard. Drawing on the technical expertise of our member firms and the consensus we have developed through our technical report series on frontier AI frameworks,[1] our comment focuses on the unique challenges of TEVV for advanced general-purpose AI systems, the importance of offering separate detailed guidance for those systems, and the critical role the U.S. can play in establishing and harmonizing global standards.

Main Comments on the TEVV Outline

Unique Challenges of TEVV for Advanced General-Purpose AI

The draft TEVV outline rightly recognizes many of the unique challenges posed by advanced general-purpose AI. For instance, the draft notes that formal verification and other common methods for TEVV are often not feasible for what it terms 'large AI models,' and observes that evaluations of large models typically yield probability estimates rather than discrete outcomes or definite measures. Likewise, as the draft outline also underscores, the complexity of large model architectures and supply chains makes it difficult to decompose AI systems into components that are more easily evaluated.

Yet advanced general-purpose AI poses other challenges too. Most notably, the capabilities and risks for which advanced general-purpose AI should be tested and evaluated are not as well specified as they are for narrow models. For comparison, consider a machine learning classifier used to diagnose pneumonia from an X-ray. Even though such a classifier is probabilistic and cannot be formally verified, the input and output space of such a model is tightly constrained and its capabilities and associated risks are well understood. Specifying TEVV requirements for such a narrow AI model is thus relatively straightforward. By contrast, the inputs and outputs of an advanced general-purpose AI model or system are much less tightly bounded and their capabilities and risks are far more challenging to specify ex ante.

Establishing suitable TEVV requirements and processes for such a model is far less straightforward.

The challenge with specifying the potential capabilities and risks of general-purpose AI has significant implications for testing and evaluation in high-risk domains. As we have noted elsewhere, advanced AI models and systems may exhibit capabilities in biology and cybersecurity that have the potential to lead to harmful outcomes, including outcomes that are irreversible and that materialize rapidly and at large-scale.[2] The unique nature of those risks has led to the development of a novel risk management instrument – frontier AI frameworks – whose TEVV processes differ in significant ways from those of either narrow AI or less advanced general-purpose AI.[3] For example, frontier AI frameworks often include far more sophisticated threat modeling in the design and development of safety tests and evaluations than risk management processes for other forms of AI.[4]

We recommend that the draft TEVV standard highlight more clearly the challenge associated with specifying capabilities and risks associated with advanced general-purpose AI, as well as its implications for TEVV processes and risk management.

Importance of Distinct Guidance for Advanced General-Purpose AI

Given the unique challenges listed above, we recommend including separate guidance for the testing and evaluation of advanced general-purpose AI in an appendix of the draft TEVV standard. The guidance should outline TEVV practices that have emerged to address the potential capabilities and risks of advanced general-purpose models and systems, and that are less relevant for narrow AI or other forms of general-purpose AI.

The FMF recently documented many of those practices in its recent technical report series on the implementation of frontier AI frameworks.[5] Each report focuses on a key element of frontier AI frameworks:

- Risk Taxonomy and Thresholds for Frontier AI Frameworks
- Frontier Capability Assessments
- Frontier Mitigations
- Third-Party Assessments for Frontier AI Frameworks

While every report refers to TEVV practices, we would highlight the *Frontier Capability Assessments* and *Third-Party Assessments* reports. The former details testing and evaluation processes tailored to the unique capabilities and risks from advanced general-purpose AI, including relative capability, bottleneck, and threat-simulation assessments. The latter documents key functions performed by third-party assessments, such as confirming internal safety testing, leveraging independent testing and evaluation methodologies, and supplementing internal capacity for testing and evaluation.

We recommend that the draft TEVV standard draw on these and similar emerging practices to develop distinct guidance for advanced general-purpose AI.

Establishing and Harmonizing International Standards

We strongly support incorporating TEVV for advanced general-purpose AI models and systems into international standards processes. Given the unique capabilities and global distribution of those models and systems, global standards are needed that are both tailored to advanced general-purpose AI and consistent with existing ISO/IEC standards.

We recommend that the draft TEVV standard include distinct guidance for general-purpose AI that extends, rather than displaces, established standards for testing and evaluation. Feeding such a draft into ISO/IEC JTC 1/SC 42 would help ensure that TEVV practices for advanced general-purpose AI are globally interoperable and harmonized with existing international standards.

Additional Comments

In addition to the suggestions above, we also recommend that the draft TEVV standard underscore the importance of the following:

- **Distinguishing between models and systems**. Although the draft TEVV outline references both AI models and systems, it does not explicitly define these terms or directly address how TEVV for systems may differ from TEVV for models, including for downstream developers across the AI value chain. Since deployed systems typically add mitigations, safeguards, and other safety interventions to an underlying model or set of models, the behaviors of such systems are often different and require different forms of TEVV as a result. Additionally, some risks are context-dependent and can only be meaningfully evaluated at the application layer. A TEVV standard should provide clear, tailored evaluation approaches for both models and systems that reflect their different characteristics and deployment contexts.

- **Distinguishing between TEVV for the development and deployment of AI systems vs. TEVV for the conformity assessment of such systems**. Although the draft outline references conformity assessment, the primary focus seems to be on TEVV approaches in the development and deployment process. While these have some overlap, particularly in methodology, the desired outcome and goals are different. The outline should make clear whether TEVV as part of conformity assessment will be addressed. If so, clear guidelines should be established, aligning to existing guidance from CASCO.

- **Marginal risk**. Evaluating advanced AI models and systems in terms of marginal risk is often crucial for understanding their overall impact on public safety and security. When evaluations are intended to directly evaluate the risk posed by a system, they should in many cases consider evaluating the marginal risk relative to other available applications and resources, including non-AI tools. By comparing the capabilities and behaviors of an AI model or system to existing alternatives, marginal risk approaches can provide understanding of how much additional risk the AI model or system introduces to the broader threat landscape. This approach is particularly valuable when evaluating potential large-scale risks to public safety and critical infrastructure, such as chemical, biological, radiological, and nuclear (CBRN) threats. A TEVV standard should elaborate on marginal risk approaches to TEVV for advanced general-purpose AI systems.

- **Domain expertise**. For some domain areas, effective testing, evaluation, verification, and validation of advanced AI models and systems necessitates collaboration with subject matter experts across relevant technical domains. When evaluating AI models and systems for potential threats in areas such as synthetic biology, cybersecurity, or chemical engineering, it is essential to engage domain experts with formal academic backgrounds and practical experience in the relevant disciplines. For example, assessments of AI capabilities related to biological risks should involve consultation with microbiologists, virologists, biochemists, and other life sciences professionals with advanced degrees and hands-on laboratory experience. These experts possess the specialized knowledge necessary to evaluate the technical feasibility of potential misuse scenarios and to identify risk vectors that may not be apparent to AI researchers alone. A TEVV standard should establish clear protocols for identifying, engaging, and integrating domain expertise throughout the evaluation process.

Conclusion

We welcome the efforts of the NIST Zero Project to develop a draft TEVV standard. The draft outline is a constructive start. As the NIST moves forward, we recommend that the draft TEVV standard explicitly addresses the unique challenges of evaluating large general-purpose AI systems, provides separate guidance tailored to those systems, and is developed in a way that is consistent with existing ISO/IEC processes and standards. These steps would strengthen the credibility and effectiveness of TEVV practices for advanced AI while helping to harmonize global standards.

We appreciate the opportunity to comment and look forward to engaging further with NIST and the Zero Draft effort as the project advances.

ENDNOTES

[1] Frontier Model Forum, “Introducing the FMF’s Technical Report Series on Frontier AI Frameworks.” April 22, 2025.
[2] For more on risk identification for frontier AI frameworks, see our issue brief on Components of Frontier AI Safety Frameworks or section 1.3 of our technical report on Risk Taxonomy and Thresholds for Frontier AI Frameworks.
[3] Frontier AI frameworks are also referred to as responsible scaling policies or frontier safety policies. For more on their origin and development, see METR, "Common Elements of Frontier AI Safety Policies." March 26, 2025.
[4] For more on this point, see both the “Threat Modeling” section of our report on *Risk Taxonomy and Thresholds*, and the section on “Common Assessment Approaches” in our report on *Frontier Capability Assessments*.
[5] Frontier Model Forum, “Introducing the FMF’s Technical Report Series on Frontier AI Frameworks.” April 22, 2025.