Subject: Hugging Face Response to NIST AI Standards "Zero Drafts" TEVV RFI

Dear NIST AI Standards Team,

Hugging Face appreciates the opportunity to provide input on NIST's proposed Zero Draft for a Standard on AI Testing, Evaluation, Verification, and Validation (TEVV). As a community-driven platform democratizing responsible AI through open-source and open-science initiatives, serving over 7 million users and hosting over 2 million public models, we bring extensive experience with large-scale AI evaluation, community-driven benchmarking, and collaborative model development.

Executive Summary

- **Evaluation paradigms:** AI evaluation is not limited to TEVV. Benchmarking, safety-focused evaluations, construct-oriented measurement, exploratory testing, real-world impact studies, and TEVV all play vital roles. The framework should explicitly situate TEVV within this broader landscape.
- **Measurement science:** Clause 4 should embed measurement science foundations, drawing from ongoing psychometrics and social science to ensure validity, reliability, and construct rigor.
- **Evaluation integrity:** Benchmark contamination and saturation threaten validity. The standard should recommend detection, monitoring, reporting, and prevention strategies.
- **Concept map refinements:** Clarify AI-specific evaluation challenges, expand to include construct-oriented paradigms, and set explicit standards for reliability and validity.
- **Governance:** Clause 5 should emphasize diverse evaluation teams, multi-stakeholder participation, community reporting, and risk-based prioritization.
- **Documentation:** TEVV results should integrate with widely used documentation formats such as Model and Dataset Cards, or move to a standard, with requirements for machine-readable and human-readable outputs.
- **Appendices:** Include case studies and methodological examples covering collaborative efforts, contamination detection, dynamic evaluation, and behavioral assessment.
- **Implementation priorities:** Immediate enhancements, pilot testing approaches, and long-term resource development should be specified to ensure effective adoption.

General Comments

We commend NIST's framework-based approach, which avoids prescriptive methods in a rapidly evolving field. However, the Zero Draft currently frames evaluation exclusively through the TEVV perspective. In practice, **AI evaluation is pluralistic, with multiple paradigms**. Benchmarking provides comparative metrics across models; safety evaluations stress-test systems under adversarial or high-risk conditions; construct-oriented approaches measure latent abilities that transfer across contexts; exploratory evaluations probe for unexpected or emergent behaviors; and real-world impact studies assess systems in deployment. TEVV is an important paradigm, but it is only one among many. A robust standard should situate TEVV within this broader landscape, offering guidance on how organizations can integrate multiple paradigms to meet their goals.

Clause-Specific Comments

Clause 4 – Measurement Science Foundations

We recommend that Clause 4 include a dedicated subsection on **measurement science foundations**. Established principles from social science measurement offer rigorous methods for defining constructs, measuring latent capabilities, and ensuring validity and reliability. Incorporating concepts such as **construct validity**, **internal consistency**, and **test–retest reliability** would bring much-needed rigor to AI evaluation. By grounding TEVV in these principles, NIST can ensure that AI evaluation results are not only robust within a single context but also generalizable across domains and tasks. The framework should clearly distinguish between continuous evaluation and point-in-time assessment, particularly given the dynamic nature of AI systems in deployment. Organizations must establish protocols for when and how frequently to conduct re-evaluations.

We further recommend integrating considerations around **benchmark contamination and saturation**. Contamination occurs when test data leak into training sets, inflating model performance without reflecting genuine generalization. Recent analyses show that roughly 3% of benchmark questions can be retrieved directly from public sources, with performance dropping by 15% when this leakage is removed. Saturation arises when benchmarks approach ceiling performance, obscuring meaningful differences between systems. To address these challenges, the framework should define contamination and saturation as threats to validity, recommend systematic audits for leakage, encourage dynamic or temporally gated benchmarks, provide venues for reporting contamination, and provide criteria for retiring or rotating benchmarks that have saturated.

Clause 4 – Concept Map and Definitions

The proposed concept map provides a useful foundation and we recommend further refinements. First, the framework could better clarify the **distinction between software testing and AI evaluation**. While traditional software testing verifies deterministic outputs against specifications, AI evaluation must grapple with probabilistic, context-dependent, and emergent behaviors. Second, the concept map should incorporate **capability-oriented and construct-oriented paradigms**, which focus on measuring underlying cognitive abilities rather than just surface-level task performance. This means testing whether models truly understand concepts or are simply memorizing patterns - for instance, evaluating whether a model maintains performance when question formats change (from multiple choice to fill-in-the-blank) or when answer options are reordered. Third, the framework should provide more detailed guidance on **construct/context validity and reliability specifications**, including accepted thresholds, reporting practices, and statistical measures such as confidence intervals and interannotator agreements. These refinements will ensure the concept map reflects both the realities of AI evaluation and the scientific standards necessary for robust practice and implementation.

Clause 5 – Governance and Organizational Requirements

Clause 5 should be expanded to emphasize the importance of **governance structures and multi-stakeholder engagement**. Evaluation is strongest when it incorporates interdisciplinary perspectives and expertise. As referenced in Clause 4, multidisciplinary teams including technical experts, domain specialists, and impacted communities are essential to identifying risks and blind spots. Moreover, organizations should be encouraged to adopt mechanisms for **community input and feedback**, such as bug bounties, user reporting, and open discussion forums, which have proven effective in surfacing evaluation issues in practice.

Resource allocation should also be guided by a **risk-based prioritization framework**. Borrowing from fields such as transportation, aerospace, and pharmaceuticals, evaluation resources should be proportionally directed toward the highest-stakes applications where risks to safety and society are greatest. Finally, given that modern AI development often involves distributed supply chains and third-party dependencies, the framework should provide explicit guidance on **evaluating components across opaque or fragmented supply chains**, ensuring that risks introduced by external dependencies are properly managed. Specific appendices could illustrate **supply chain TEVV management** through real-world case studies.

Clause 6 – Documentation

Documentation is critical for transparency, accountability, and interoperability. We recommend that TEVV documentation explicitly build on established standards such as Model Cards and Dataset Cards rather than creating parallel systems. TEVV results should be presented in both **human-readable and machine-readable formats**, enabling stakeholders to understand evaluation outcomes while also supporting integration into automated monitoring pipelines. The framework should also provide specific guidance on documenting **probabilistic findings and uncertainty estimates**, including statistical reporting standards, confidence intervals, and uncertainty communication protocols ensuring that decision-makers interpret results responsibly rather than as deterministic facts.

Appendices – Recommended Additions

To ground the framework in practice, we encourage NIST to include appendices with **detailed case studies and technical methods**. Examples might include collaborative evaluation efforts such as **BigScience**, which successfully coordinated over 1,000 researchers across 60 countries. Additional case studies should illustrate **contamination detection** in benchmarks such as MMLU and GSM8K, where models were able to reproduce test questions, accompanied by **step-by-step procedures** such as perplexity analysis and n-gram overlap detection. It is also important to include **guidance for resource-constrained organizations**, showing how community platforms and shared infrastructure can support robust TEVV even under limited budgets, with specific cost-benefit analyses and resource allocation strategies.

The appendices should highlight **dynamic evaluation approaches** like LiveCodeBench, which introduce temporally gated tasks to prevent contamination, and **behavioral and real-world assessment methods**, which evaluate deployed systems in dynamic contexts. rather than relying solely on static test sets. Approaches to **multi-modal and interactive evaluation** are especially important as AI systems increasingly integrate text, image, and audio. The framework should also include examples of **evaluating evaluation methods themselves**, such as validating LLM-as-judge systems and synthetic benchmarks.

Implementation Recommendations

In addition to structural improvements, the framework should provide practical implementation guidance, such as:

- **Pilot Testing Approach:** Test framework elements with existing evaluation platforms such as Hugging Face community infrastructure, openly available evaluation tools, and academic initiatives before finalization. Engage with evaluation practitioners in research groups and in other safety-critical domains to validate applicability and identify gaps. Coordinate with international standards bodies to ensure compatibility and mutual recognition.
- **Resource Development Support:** Invest in shared, open infrastructure that democratizes access to TEVV tools, develop training and education resources to assist organizations in implementation, and establish continuous improvement mechanisms so the framework evolves alongside evaluation science.

Hugging Face stands ready to contribute our expertise in **community-driven evaluation, open infrastructure, and collaborative science** to support NIST in developing TEVV standards that are both practical and scientifically grounded.

Respectfully,

Avijit Ghosh, Applied Policy Researcher
Irene Solaiman, Chief Policy Officer
Hugging Face