

NIST AI Technology Evaluation (AITE) Overview & Road Map (Version 1.0)

July 24, 2026

1 Introduction

The Technology Test and Evaluation Division at the National Institute of Standards and Technology (NIST) has launched a new program to provide researchers with a sequestered testbed environment for the evaluation of AI model performance in a variety of meaningful tasks across diverse datasets, modalities, and domains. The Artificial Intelligence Technology Evaluation (AITE), currently in its initial phase, provides volunteer testing of AI models on blind data and its sequestered environment mitigates the risk of train/test data contamination to ensure rigorous, objective assessment. The infrastructure provided by NIST will provide common data, metrics and scoring to help developers understand the performance of their models.

The purposes of the effort include to:

- serve as a neutral 3rd party for hosting such tests,
- be able to utilize data that is not publicly released in order to remove the potential for train/test data contamination and enable the use of data where public distribution is undesired, and
- measure the state-of-the-art in myriad tasks and domains.

AITE will rely on engagement from participants in two different tracks, each offering distinct advantages:

- *Data providers* submit an original dataset in their domain that is inaccessible to others and a meaningful task to be performed on that dataset. Data providers will receive careful measurements of top models on their data conducting their task.
- *Model providers* submit AI models to be tested on the datasets and tasks. Model providers will learn how their models perform on an increasing number of datasets and tasks, and how their models perform relative to others on the same data using the same metrics, improving comparability while ensuring the evaluation data is not used for training any model.

Participation is open to all who wish to engage in one of the ways described above, and who can abide by the AITE Participation Agreement and rules. To request participation or ask questions, contact us at aite-poc@list.nist.gov.

The AITE is being developed in phases, and additional tasks/datasets/modalities/domains will be added over time. In its full realization, the intention for AITE is to: evaluate leading academic- and industry-provided, as well as open-source, models; include an ever growing set of AI technology tests on the order of hundreds of tasks/datasets/domains; incorporate sufficient data security for controlled (operational) but unclassified information (CUI) data; be fully automated; be an exemplary evaluation that makes use of novel, insightful, and/or interesting metrology.

This document introduces the NIST AITE program and provides a road map for its development. Information regarding the current phase of AITE's development, which is the initial phase, and an overview of the initial AI Technology Tests themselves, can be found here:
<https://ai-challenges.nist.gov/aite>

2 Road Map

The NIST AITE program has reached its launch point, consisting of three tests and results from a single model, and is being developed in four phases:

- Phase 1: (*Initial phase*) AITE will accept a very limited number of additional AI application tests and models for testing. The focus of this phase will be to exercise and hone the processes for working with external collaborators to develop tests and test external models.
- Phase 2: (*Early growth and innovation phase*) AITE will accept a small number of additional AI application tests and models for testing. The focus of this phase will be to increase the types and varieties of tasks/modalities and domains and corresponding evaluation techniques.
- Phase 3: (*Accelerated growth and innovation phase*) AITE will accept a large number of additional AI application tests and models for testing. The focus of this phase will be scaling the number of tests and models, as well as the evaluation techniques.
- Phase 4: (*Continuous growth and innovation phase*) AITE will continue to accept a large and growing number of additional AI application tests and models for testing. The focus of this phase will be on long term growth and alignment with emerging needs.

The precise scope and timing of each of the above phases will be contingent on capacity and levels of engagement. In the early phases, data providers will be selected based on a combination of the anticipated levels of ease and impact of integrating the AI application tests into AITE. Throughout all phases, all models submitted to AITE will be queued and tested on a first come, first served basis, as capacity allows.

3 Public Display of Results

NIST will publish and periodically update a leaderboard and other reports summarizing results, including performance charts. Results displayed will specify, for each eligible model, the dataset, task, model (identified by name), metric, scores, and measures of uncertainty.

4 Disclaimer

Certain commercial equipment, instruments, software, or materials are identified in this document to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor necessarily the best available for the purpose. The descriptions and views contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST or the U.S. Government.

Change Logs

- Version 1; July 24, 2026; Initial Release