MLTE: Machine Learning Test and Evaluation

ARE YOU DEVELOPING A MACHINE LEARNING (ML) MODEL and unsure about what constitutes proper testing? Are you acquiring an ML model and want to know how it was tested? Machine Learning and Test Evaluation (MLTE), pronounced "melt," can help you develop or acquire ML models that are production ready by providing a comprehensive process and tool for testing ML models and reporting test results.

Current practice for ML model testing during ML model development prioritizes testing model properties, such as model performance (e.g., accuracy). Without adequate consideration of system requirements—such as throughput, resource consumption, or robustness—this practice can lead to failures in model integration, deployment, and operations. In many cases, model developers lack the skills or background to test beyond model performance. In other cases, model developers receive very little system context that informs design decisions.

The Software Engineering Institute (SEI) and the Army Artificial Intelligence Integration Center (AI2C) co-developed MLTE, a system-centric, semi-automated process and infrastructure that enables negotiation, specification, and testing of ML model and system qualities. MLTE produces evidence of testing activities—in the form of test report, test code, and test data—that can be shared with model acquirers and integrators to inform integration and further test and evaluation activities.
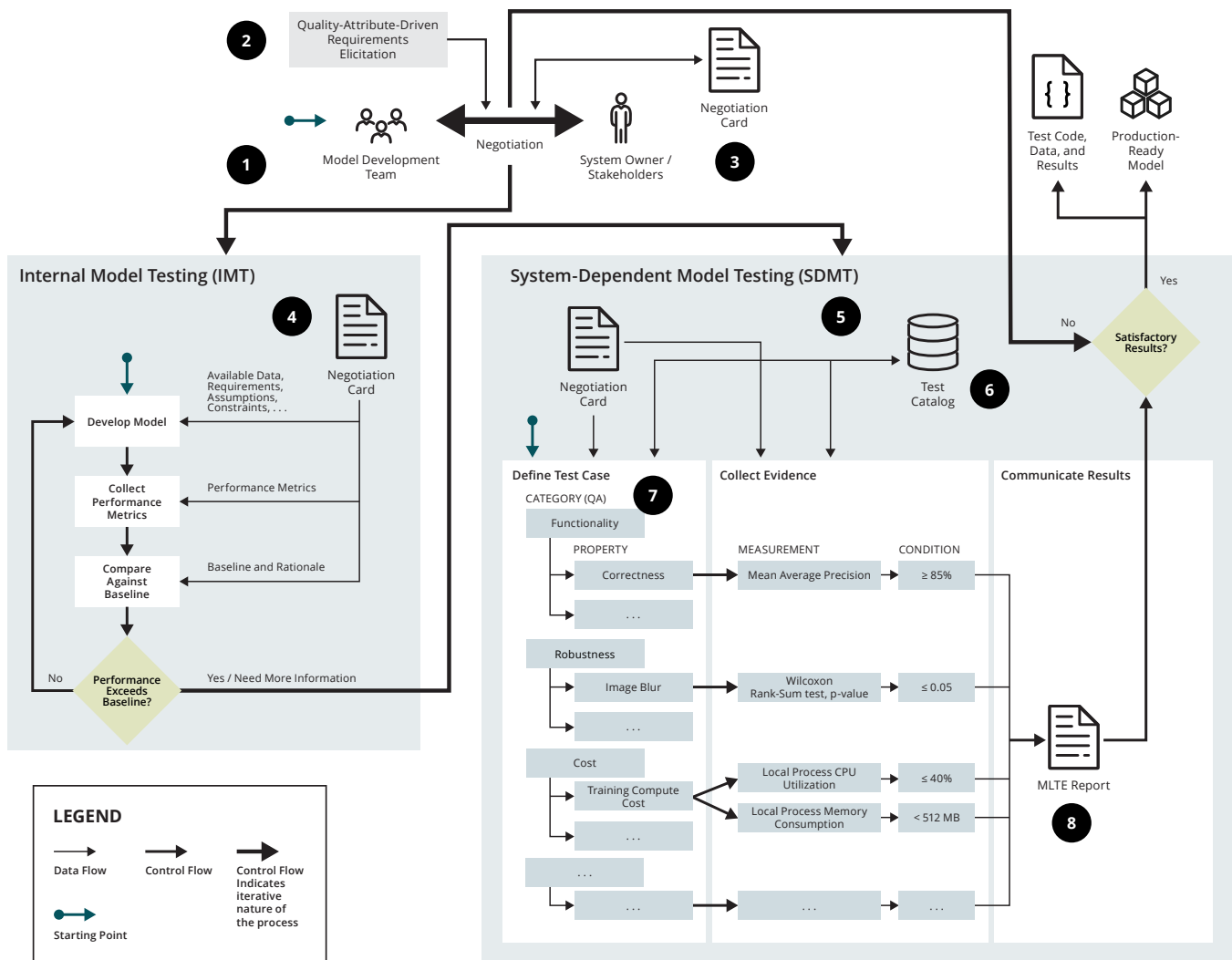
How It Works

Key elements of MLTE promote production readiness of ML models, labeled in the process diagram:

1. Stakeholders negotiate continuously throughout model development to share system context.

2. Model requirements elicitation is driven by quality attribute (i.e., non-functional requirement) scenarios that represent concrete system needs.

3. Negotiation Card records agreements and drives model development and testing.

4. Internal Model Testing recognizes the experimental nature of model development.

5. System-Dependent Model Testing focuses on obtaining evidence of satisfaction of model and system requirements.

6. Test Catalog contains reusable (organizational) examples of test cases organized by quality attribute.

7. Test Cases are first-class artifacts that define metrics, measurement methods, and passing conditions.

8. MLTE Report records and communicates test results.

Reported Results

Evaluations of MLTE with AI2C show that

- Negotiation Card identifies a larger number of model and systems requirements early in the development process.

- Test Catalog supports the development and reuse of test code for validating these requirements.

- Test Code and MLTE Report provide evidence of testing that increases trustworthiness of ML models.

2 Quality-Attribute-Driven Requirements Elicitation

1 Model Development Team

Negotiation

System Owner / Stakeholders

Negotiation Card

3

Test Code, Data, and Results

Production-Ready Model

Internal Model Testing (IMT)

4 Negotiation Card

Available Data, Requirements, Assumptions, Constraints, . . .

Develop Model

Performance Metrics

Collect Performance Metrics

Baseline and Rationale

Compare Against Baseline

No

Performance Exceeds Baseline?

Yes / Need More Information

System-Dependent Model Testing (SDMT)

5 Negotiation Card

Test Catalog

6

No

Yes

Satisfactory Results?

Define Test Case **7**

Collect Evidence

Communicate Results

CATEGORY (QA)

	MEASUREMENT	CONDITION
Functionality		
PROPERTY		
Correctness	Mean Average Precision	≥ 85%
. . .		
Robustness		
Image Blur	Wilcoxon Rank-Sum test, p-value	≤ 0.05
. . .		
Cost		
Training Compute Cost	Local Process CPU Utilization	≤ 40%
	Local Process Memory Consumption	< 512 MB
. . .		
. . .		
. . .	. . .	. . .

MLTE Report **8**

LEGEND

→ Data Flow

➤ Control Flow

➤ Control Flow Indicates iterative nature of the process

●→ Starting Point

The SEI can help you instantiate MLTE in your organizations and integrate it into your model development and model T&E processes.

Contact Us

Download MLTE and get how-to guides and tutorials at **github.com/mlte-team/mlte**

Learn more about the MLTE framework in our papers:

• MLTEing Models: Negotiating, Evaluating, and Documenting Model and System Qualities **insights.sei.cmu.edu/library/mlteing-models-negotiating-evaluating-and-documenting-model-and-system-qualities**

• Using Quality Attribute Scenarios for ML Model Test Case Generation **arxiv.org/abs/2406.08575**

About the SEI

Always focused on the future, the Software Engineering Institute (SEI) advances software as a strategic advantage for national security. We lead research and direct transition of software engineering, cybersecurity, and artificial intelligence technologies at the intersection of academia, industry, and government. We serve the nation as a federally funded research and development center (FFRDC) sponsored by the U.S. Department of Defense (DoD) and are based at Carnegie Mellon University, a global research university annually rated among the best for its programs in computer science and engineering.

Contact Us