ACQUISITION INNOVATION RESEARCH CENTER

Development of an Artificial Intelligence (AI) Test Harness for the Department of Defense (DoD)

EXECUTIVE SUMMARY

SEPTEMBER 2024

PRINCIPAL INVESTIGATOR
Laura Freeman, *Virginia Tech*

CO-PRINCIPAL INVESTIGATOR
Stephen Adams, *Virginia Tech*
Erin Lanus, *Virginia Tech*
Naren Ramakrishnan, *Virginia Tech*

RESEARCH ASSOCIATE
Brian Mayer, *Virginia Tech*
Patrick Butler, *Virginia Tech*

RESEARCH ASSISTANT PROFESSOR
Jaganmohan Chandrasekaran, *Virginia Tech*

RESEARCH ASSOCIATE PROFESSOR
William Headley, *Virginia Tech*

RESEARCH DATA ANALYST
Brian Lee, *Virginia Tech*

SOFTWARE DEVELOPER
Alex Kyer, *Virginia Tech*

SYSTEMS ADMIN
Jared Gregerson, *Virginia Tech*

VIRGINIA TECH

SPONSOR

Mr. Paul Lowe, Deputy Director, Strategic Initiatives, Policy and Emerging Technologies (Acting), Director, Operational Test and Evaulation (DOT&E)
Mr. Nilo Thomas, DOT&E Software and AI Advisor

EXECUTIVE SUMMARY

The development of Artificial Intelligence (AI) has revolutionized the way organizations operate. The Department of Defense (DoD) is no exception to this transformation. AI has the potential to revolutionize military capabilities and reduce human errors. To keep pace with our adversaries, one of the Office of the Director, Operational Test and Evaluation's (DOT&E) core strategic pillars is to pioneer test and evaluation (T&E) methods of weapon systems designed to change over time. Machine learning (ML) and AI models are notably capable of learning and changing over time. Moreover, the stochastic nature of the models that learn based on past data present new challenges for T&E. It is essential to ensure that these systems operate effectively, safely, and securely. A reliable test harness that provides high-quality data, AI models, and test and evaluation capabilities will accelerate and inform the development of new methods. The DOT&E has the responsibility to develop policies for T&E of AI-enabled systems. However, the current state of the AI capabilities and the corresponding T&E methods for AI/ML are evolving. The development of test harnesses has the potential to not only accelerate method development but also inform DOT&E's policy and guidance. Finally, test harnesses can serve as an educational resource for the T&E community where testers can learn T&E for AI-enabled systems by leveraging tools, processes, and methods in the T&E harness.

In this research, the team designed a framework for an AI Test Harness that could be applied to multiple types of AI models. Along with the framework, the research team developed a set of requirements for an AI Test Harness and produced a simple prototype. The research team then applied the developed framework to two use cases. The first is a Radio Frequency Machine Learning (RFML) use case that uses standard classification models. Under this use case, the research team advanced synthetic data generation capability by publicly releasing Python-based Wideband Aggregate SPectrum GENerator (PY-WASPGEN), a toolset for producing radio frequency (RF) data for training and testing AI/ML models. The research team also demonstrated the Coverage of Data Explorer (CODEX) capability on an RFML example. The project developed education and training material on the application of standard T&E methods to classification problems.

The second use case focused on Large Language Models (LLMs), a form of generative AI. LLMs are considered one of the most advanced forms of AI and recently gained popularity. Due to their recent advances, T&E for these types of models is nascent. The research team conducted a survey of the academic literature and industry best practices to assess the current state of T&E for LLMs. The results of this survey led to a framework for the various tasks a LLM can perform and the characteristics of a LLM that should be evaluated. Education and training material for some of these tasks was developed and publicly released. In this work, the research team did not distinguish between a LLM and a LLM-based system. The current version of the proposed harness framework conflates the two, but future work should investigate T&E for the LLM separate from the LLM-based system.

The contrast of these two separate use cases provides context to the value of test harnesses and the challenges in implementing them. The RF use case demonstrates that developing a test harness for standard machine learning models, such as classifiers, is a decently straightforward software engineering task supported by widely available open-source tooling and new capabilities investments that the DoD is making.

However, a test harness for AI-systems and generative AI requires more research. The former must consider the systems interactions with other systems, including human users and AI-enabled systems that may evolve over time. The latter requires everything involved in developing a test harness for an AI-enabled system plus further study on metrics, data sets, and balancing the results of tests on multiple tasks with possibly competing objectives.

CONCLUSIONS AND RECOMMENDATIONS

The rapid evolution of AI has transformed various industries, including defense, where AI capabilities are being integrated into mission-critical systems. This research has demonstrated the potential of an AI Test Harness framework to accelerate the development of T&E methodologies for AI and ML systems. Through two distinct use cases—Radio Frequency Machine Learning (RFML) and Large Language Models (LLMs)—the research team has provided valuable insights into the challenges and opportunities in testing AI.

The framework and prototype developed through this project serve as an important foundation for future work in AI T&E. It also emphasizes the need for continued investment in research, tool development, and educational resources. By establishing a reliable test harness and a robust set of policies, standards, and metrics, the DOT&E can better equip the T&E community to handle the unique challenges posed by AI and ML systems, ultimately ensuring that these technologies can be deployed safely and effectively in defense applications.

Future work should further distinguish between testing individual AI models and more complex AI-enabled systems, particularly generative AI like LLMs, and refine the tools, datasets, and metrics needed to evaluate these emerging capabilities. Specific recommendations include:

Recommendation 1: The DoD and AIRC should continue the development of Test Harnesses to advance T&E of AI-enabled systems.

a. AIRC should serves as a facilitator of test harness models for academic research in T&E of AI models.

b. The DoD should continue to invest in research on AI-enabled systems test capabilities and how they differ from AI model T&E.

Recommendation 2: The DoD should ensure private data sets for testing of LLMs. Public data sets may quickly become ineffective for testing LLMs as all public data will likely be used during training. Private data sets for each task should be developed that are withheld from the public and only used for testing.

Recommendation 3: The DoD or AIRC should develop dashboards for tracking the performance of LLMs on tasks relevant to the DoD. Dashboards should provide the ability to compare holistic evaluation with task-specific evaluation in DoD contexts.

DISCLAIMER

The Acquisition Innovation Research Center (AIRC) is a multi-university partnership led and managed by the Stevens Institute of Technology and sponsored by the U.S. Department of Defense (DoD) through the Systems Engineering Research Center (SERC)—a DoD University-Affiliated Research Center (UARC).

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Under Secretary of Defense for Acquisition and Sustainment (OUSD(A&S)) and the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under Contract HQ0034-19-D-0003, TO#0092.

The views, findings, conclusions, and recommendations expressed in this material are solely those of the authors and do not necessarily reflect the views or positions of the United States Government (including the Department of Defense (DoD) and any government personnel), the Stevens Institute of Technology, or Virginia Tech.

No Warranty.

This Material is furnished on an "as-is" basis. The Stevens Institute of Technology and Virginia Tech make no warranties of any kind—either expressed or implied—as to any matter, including (but not limited to) warranty of fitness for purpose or merchantability, exclusivity, or results obtained from use of the material.

The Stevens Institute of Technology and Virginia Tech do not make any warranty of any kind with respect to freedom from patent, trademark, or copyright infringement.

SYSTEMS ENGINEERING RESEARCH CENTER

AIRC ACQUISITION INNOVATION RESEARCH CENTER

STEVENS INSTITUTE OF TECHNOLOGY 1870