119TH CONGRESS
1ST SESSION

# S. 3336

To require the Secretary of Homeland Security to carry out prize competitions to advance the science of interpretability and to develop adversarial robustness with respect to artificial intelligence products, and for other purposes.

---

IN THE SENATE OF THE UNITED STATES

DECEMBER 3, 2025

Ms. HASSAN (for herself and Mr. BANKS) introduced the following bill; which was read twice and referred to the Committee on Homeland Security and Governmental Affairs

---

# A BILL

To require the Secretary of Homeland Security to carry out prize competitions to advance the science of interpretability and to develop adversarial robustness with respect to artificial intelligence products, and for other purposes.

*Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,*

**SECTION 1. SHORT TITLE.**

This Act may be cited as the ''Reliable Artificial Intelligence Research Act of 2025''.

**SEC. 2. DEFINITIONS.**

In this Act:

(1) ADVERSARIAL ROBUSTNESS.—The term ''adversarial robustness'' means the degree to which an artificial intelligence model is able to resist attacks that would induce it to produce incorrect, restricted, or harmful outputs, while maintaining integrity, reliability, and privacy.

(2) ARTIFICIAL INTELLIGENCE.—The term ''artificial intelligence'' has the meaning given the term in section 5002 of the National Artificial Intelligence Initiative Act of 2020 (15 U.S.C. 9401).

(3) INTERPRETABILITY.—The term ''interpretability'' means the degree to which humans are able to accurately understand how an artificial intelligence model makes decisions and considers inputs and how the outputs or behaviors of the model respond to a change in the inputs.

(4) RED-TEAMING.—The term ''red-teaming'' means a structured, interactive, and adversarial process to test an artificial intelligence system by simulating real-world actions to find vulnerabilities or flaws in the system.

(5) SECRETARY.—The term ''Secretary'' means the Secretary of Homeland Security.

**SEC. 3. PRIZE COMPETITION FOR ARTIFICIAL INTELLIGENCE INTERPRETABILITY RESEARCH.**

(a) PRIZE COMPETITION REQUIRED.—Not later than 270 days after the date of enactment of this Act, the Secretary of Homeland Security shall commence carrying out at least 1 prize competition under section 24 of the Stevenson-Wydler Technology Innovation Act of 1980 (15 U.S.C. 3719) to advance the science of interpretability in a manner relevant to commercially available or widely used artificial intelligence products.

(b) CONSULTATION.—In carrying out the prize competition required by subsection (a), the Secretary shall consult with—

(1) the Secretary of Commerce;

(2) the Director of the National Institute of Standards and Technology;

(3) the National Cyber Director;

(4) the Director of the National Science Foundation; and

(5) any industry expert from the artificial intelligence sector in the United States that the Secretary considers relevant.

(c) STRUCTURE AND EVALUATION CRITERIA.—

(1) IN GENERAL.—The Secretary shall develop the structure and evaluation criteria for a prize competition carried out under subsection (a) in accord-

ance with the primary purpose described in that subsection.

(2) COMPETITION STRUCTURE.—The Secretary may—

(A) structure a competition under subsection (a) into 1 or more phases, including submission of interpretability frameworks, submission of interpretable artificial intelligence models, and unique basic research; and

(B) open these phases to the same, or to distinct, contestant pools.

(3) EVALUATION CONSIDERATIONS.—In developing the evaluation criteria for the frameworks, models, or methods submitted for a prize competition under subsection (a), the Secretary shall consider—

(A) the degree to which a submission advances broadly applicable principles of artificial intelligence interpretability;

(B) the practical value of a submission in making artificial intelligence more understandable and reliable in high-risk, high-value use cases; and

(C) the likelihood that the unique research submitted will create standards for artificial in-

telligence interpretability in the government or industry.

(d) PROGRAM ADMINISTRATION.—The Secretary may enter into contracts, cooperative agreements, or other agreements with for-profit or nonprofit entities or State, territorial, local, or Tribal agencies to design and administer any prize competition carried out under subsection (a).

**SEC. 4. PRIZE COMPETITION FOR ARTIFICIAL INTELLIGENCE ADVERSARIAL ROBUSTNESS RESEARCH.**

(a) PRIZE COMPETITION REQUIRED.—Not later than 270 days after the date of enactment of this Act, the Secretary shall commence carrying out at least 1 prize competition under section 24 of the Stevenson-Wydler Technology Innovation Act of 1980 (15 U.S.C. 3719) to develop capable artificial intelligence models that are designed to exhibit adversarial robustness in circumstances necessary for at least 1 high-impact, high-risk application in government or industry.

(b) CONSULTATION.—In carrying out a prize competition required by subsection (a), the Secretary shall consult with—

(1) the Secretary of Commerce;

(2) the Director of the Institute of Standards and Technology;

(3) the National Cyber Director;

(4) the Director of the National Science Foundation;

(5) any industry expert from the artificial intelligence sector in the United States that the Secretary considers relevant; and

(6) the head of any Federal agency who has authority or expertise in a high-impact, high-risk application of artificial intelligence that could be an appropriate subject for a prize competition under subsection (a).

(c) STRUCTURE AND EVALUATION CRITERIA.—

(1) IN GENERAL.—The Secretary shall develop the structure and evaluation criteria for a prize competition carried out under subsection (a) in accordance with the primary purpose described in that subsection.

(2) COMPETITION STRUCTURE.—The Secretary may—

(A) structure a competition under subsection (a) into 1 or more phases, including submission of adversarial robustness frame-

works, submission of artificial intelligence models, and red-teaming; and

(B) open these phases to the same, or to distinct, contestant pools.

(3) EVALUATION CONSIDERATIONS.—In developing the evaluation criteria for the frameworks, models, or methods submitted for a prize competition under subsection (a), the Secretary shall consider—

(A) the degree to which a submission advances broadly applicable principles of artificial intelligence robustness; and

(B) the practical value of the submission in reducing the risk of adversarial attacks in high-risk, high-value use cases of artificial intelligence.

(d) PROGRAM ADMINISTRATION.—The Secretary may enter into contracts, cooperative agreements, or other agreements with for-profit or nonprofit entities or State, territorial, local, or Tribal agencies to design and administer any prize competition carried out under subsection (a).

**SEC. 5. TRACKING AND REPORTING.**

(a) IN GENERAL.—Not later than 180 days after the date on which the first prize competition concludes, the

Secretary shall submit to the appropriate congressional committees a report that includes—

(1) an evaluation of how the results of the competitions inform the fields of interpretability and adversarial robustness;

(2) an assessment of any gaps in these fields identified by the Secretary over the course of the competitions; and

(3) any suggested action that Congress should take to advance the fields of interpretability, adversarial robustness, and any related research.

(b) APPROPRIATE CONGRESSIONAL COMMITTEES DEFINED.—In this section, the term ''appropriate congressional committees'' means—

(1) the Committee on Homeland Security and Governmental Affairs of the Senate; and

(2) the Committee on Homeland Security of the House of Representatives.

**SEC. 6. APPROPRIATIONS.**

There is authorized to be appropriated to the Secretary to carry out this section $10,000,000 for the period of fiscal years 2026 through 2030.

○