PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-x968-ukea
Comments Due: May 28, 2025 **Submission Type:** Web

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0274
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Open Forum for AI

General Comment

See attached file(s)

Attachments

AI-RD-Strategy-OFAI

May 29, 2025

Faisal D'Souza
Networking and Information Technology
Research and Development (NITRD)
National Coordination Office (NCO)
National Science Foundation

Re: Docket ID No. NSF-2025-OGC-0001

Dear Mr. D'Souza:

The Open Forum for Artificial Intelligence (OFAI) is a university-led, collaborative initiative that aims to bend the arc toward human-centered, responsible, transparent, and ethical AI. The OFAI seeks to bring an academic and nonprofit perspective to critical conversations, working alongside industry and government to foster innovation such that everyone benefits from the use of AI. OFAI is led by Carnegie Mellon University (CMU) and includes the open source program offices from George Washington University, Georgia Institute of Technology, University of Texas at Austin, and North Carolina A&T State University. OFAI also has voices from the nonprofit sector including the Open Source Initiative (OSI), Creative Commons, Conscience, and the Atlantic Council as well as individual fellows from industry and government. The OFAI is pleased to provide recommendations to the Office of Science & Technology Policy (OSTP) and the National Science Foundation that highlight how openness can be leveraged in the country's AI R&D strategy to accelerate discovery and advance AI.

Openness Unlocks AI Innovation

Openness is a bedrock principle in research and development that has made the U.S. science and technology apparatus the envy of the world. It is a key characteristic with application to many parts of the innovation ecosystem—from open science and open education to open data and open source software. Openness is the mechanism by which researchers, developers, entrepreneurs, and investors can share and collaborate on emerging technology and new knowledge, just as they did with the advent of the Internet and the Web. The complex nature of AI systems and the speed at which they are being applied in our everyday lives makes openness even more important today.

Openness in AI enhances both non-commercial and commercial applications by allowing the brightest minds to contribute to and evaluate systems so that the best ones can be developed and deployed. It breaks down silos and accelerates discovery in ways closed approaches cannot. For these reasons, we believe openness should be

foundational in the administration's AI R&D strategy. We offer recommendations in key areas at the intersection of AI R&D and openness.

Expand Access to Open Datasets for AI Training and Evaluation

Open Corpus of Knowledge for AI Training
Openly sharing data is a means by which researchers can show their work and build trust in their conclusions. The benefits of open data in the research context can also be applied to AI systems and the data they are trained on. While the legal aspects of using copyrighted content to train AI systems is playing out in U.S. courts, there remains a need to expand the corpus of knowledge available for AI training and evaluation to increase discoverability of such knowledge. The benefits of expanding access to datasets are multifold. Creating more "open" datasets ensures innovators of all stripes—from startups and university research teams to large tech firms—can evaluate and fine-tune new models and systems. It also builds trust and allows for better evaluations of the systems themselves, enabling researchers to root out low-quality training data. And finally, it provides clarity to developers who wish to use the data but are unsure if they are legally allowed to do so.

For example, rich open datasets in pharmaceutical R&D have given way to groundbreaking tools like AlphaFold where experts can predict protein structures and the Human Genome Project which generated the first sequence of the human genome. These accomplishments would not have been possible without access to large, high-quality, open data. Yet, researchers predict that current efforts by private companies to create closed data for use in AlphaFold2 will not generate the breadth of data needed to truly power AI.[1] Tools like AlphaFold2 will be less successful and feats like the Human Genome Project out of reach without more open and standardized datasets.

There are many efforts underway to develop open datasets and connect existing ones so that information for AI training and evaluation is easier to find and use. However, these efforts may not be of immediate interest to commercial entities because the content by its nature will be open to everyone. That is precisely why government investment in such initiatives is so important—it provides the entire AI community with openly available resources to accelerate innovation. This will ensure big and small players have access to high-quality data, making the systems they create more accurate and useful.

[1] Richard Gold and Robert Cook-Deegan, *AI drug development's data problem* (April 2025)

Federal investment in and coordination of these efforts would be beneficial to extend their impact and address the growing need for AI training datasets. We highlight a few initiatives below and welcome the opportunity to work with the administration on strategic investments in this area.

- The Public Interest Corpus seeks to create high-quality AI training data from memory organizations (e.g., libraries, archives, museums) and their partners (e.g., publishers).
- The Institutional Data Initiative is a team of data scientists and community builders working to make knowledge collections at universities, libraries, and government agencies available as open datasets that can be used to train AI models.
- The U.S. Repository Network aims to create a more interoperable network of open repositories (government and non-government) in the U.S. so that the information in such repositories can be reused by others.

Along with the creation of more open data, governance and community standards are critical to facilitate access to complex data sources. Civil society groups have suggested that such a governance structure should be viewed as the "Data Commons."[2] That is, a governance structure that is flexible enough for varying use cases. It is crucial for the U.S. to be engaged in such global and domestic governance discussions.

Expand Access to AI R&D Resources and Develop AI Literacy

Universities across the U.S. are rapidly adopting AI tools and platforms for faculty, students, and researchers to work with models and learn valuable AI literacy skills. There are also a growing number of schools developing their own AI platforms. The Dietrich Analysis & Research Education (DARE) platform, built as an Open Source project at CMU, promotes human-centered AI by allowing students, faculty, and staff to leverage multiple large language models, transform data and experiment through a locally controlled LLM gateway. This enables faculty to augment research capabilities and develop curriculum in a platform that adapts to their pedagogical needs rather than adapting their pedagogy to fit available tools. DARE puts AI in the loop with humans in control, promoting human agency and interaction transparency while empowering students to use AI responsibly. Sage, a tool at the University of Texas at Austin, is an AI teaching and learning guide that draws on the LLM Claude and established principles of learning experience design and responsible AI adoption. With Sage, faculty at UT Austin can design tutoring sessions for students on any topic.

[2] Alek Tarkowski, *Data Governance in Open Source AI: Enabling Responsible and Systematic Access* (February 2025)

While initiatives like those at CMU and UT Austin are growing across the country, the experts and technologists that build them need resources beyond just open data. They need access to computing power and Open Source software as well as training and openly licensed educational materials. Computing power in particular is often out of reach for tool builders at universities or nonprofit research centers due to prohibitive costs. Further, they need a sustainable funding environment for these resources such that they can take advantage of the latest model developments and immediately deploy them.

The government's AI R&D strategy should include investment in these resources to ultimately build a *public* infrastructure for AI as well as sustained funding for a coordinated network of universities building cutting-edge AI tools. The federal program supporting the network of universities should be flexible enough to enable experts at those universities to adapt to and leverage new information about AI instead of waiting months or years for another funding cycle.

Invest in the Development of Open Source AI

Open Source software is another crucial component of any AI system. And while an AI system is much broader than the software code, the ubiquitous nature of Open Source software provides a clear example of what can happen when technology is shared without restriction and creators are given the freedom to innovate. The Open Source Definition, maintained by the Open Source Initiative (OSI), removes barriers to learning, using, sharing, and improving software systems. Today, Open Source software accounts for more than 97 percent of applications we use[3] and 90 percent of companies report using open source software in some way.[4] Recognizing the need to apply Open Source principles to AI, OSI co-developed the Open Source AI Definition (OSAID), releasing version 1.0 in October 2024. In 2025, OSI is leading a community effort to evaluate the definition and identify models that meet it. This work will inform future iterations of the definition and best practices for developing truly Open Source AI.

The Open Source AI Definition

An *Open Source AI* is an AI system made available under terms and in a way that grants the freedoms to:

- **Use** the system for any purpose and without having to ask for permission.
- **Study** how the system works and inspect its components.

[3] Black Duck, Open Source Security and Risk Analysis (2025)
[4] Github, Octoverse Report (2022)

- **Modify** the system for any purpose, including to change its output.
- **Share** the system for others to use with or without modifications, for any purpose.

The preferred form of making modifications to a machine-learning system must include all the elements below:

- **Data Information**: Preferably the original data, or if it is not legally possible, sufficiently detailed information about the data used to train the system so that a skilled person can build a substantially equivalent system. Data Information shall be made available under OSI-approved terms.
- **Code**: The complete source code used to train and run the system. The code shall represent the full specification of how the data was processed and filtered, and how the training was done. Code shall be made available under OSI-approved licenses.
- **Parameters**: The model parameters, such as weights or other configuration settings. Parameters shall be made available under OSI-approved terms.

The administration's AI R&D strategy should include an investment in the development and deployment of Open Source AI in alignment with OSI's evolving definition. Investments in Open Source AI should focus on systems that can provide real-world solutions to public challenges such as those in healthcare, education, agriculture, and research itself. For example, truly Open Source AI models like the Allen Institute's (Ai2) OLMo provide researchers with a large language model that can be just as powerful as many of the proprietary ones but with access to the components and without the price tag. Ai2's most recent release, OLMo 2 32B, outperforms GPT3.5-Turbo and GPT-4o mini on a variety of benchmarks.[5] Especially at a time when new knowledge is being learned everyday about how to improve AI systems, researchers and developers need access that grants them the freedom to use, study, modify, and share the system and its components.

Further, countries around the world are trying desperately to catch up to America's lead in AI by developing innovative models that share some of their components openly. China's DeepSeek is one such example that has demonstrated just how powerful an open model can be. The U.S. needs more Open Source AI systems to provide researchers, developers, entrepreneurs, and investors options for using and building on models that have greater transparency and lower barriers to adoption.

[5] OLMo 2 32B Release Notes

Conduct Research into Mechanisms for Openness

The benefits of openness are vast but the research community still lacks consensus on *how* to evaluate AI systems. A recent survey of academics and corporate researchers conducted by the Association for the Advancement of Artificial Intelligence (AAAI) found that a lack of suitable evaluation methodologies and the black-box nature of AI systems were the biggest challenges to evaluating them.[6] The administration should conduct research, through intramural and extramural grants, into methodologies for evaluating AI systems and the level of openness needed to do so. It should also consult AI researchers and drive consensus around such methods through the National Institute of Standards and Technology (NIST) and Federal science agencies.

Research and consultations in this area should address the following:

- How should AI systems be evaluated for risk—from personal safety to national security in presence of access barriers?
- What level of openness is needed to evaluate systems for such risks?
- What methodologies exist for providing that transparency?
    - For example, OSI's Open Source AI Definition (OSAID) is an existing framework that ensures key information about AI systems are shared openly.

Track Research into Downstream Impacts of Openness Policies

Regulatory frameworks impacting Open Source AI are being implemented or are under consideration in many jurisdictions around the world. Understanding the implications of such proposals is critical. Members of the OFAI are conducting research into the economic impacts of various openness regulations. Their research aims to address the following:

- When do openness regulations enhance model access, transparency, and innovation, and when might they hinder these goals?
- How should openness regulation be designed to encourage greater competition and investment in AI development?
- How should the concept of "openness" in AI be defined for regulatory purposes?

We urge the administration to review this research and engage with members of the OFAI to understand its implications. We would welcome the opportunity to meet with you and discuss this research further.

[6] AAAI's Presidential Panel on the Future of AI Research (March 2025)

We thank the Office of Science & Technology Policy and the National Science Foundation for the opportunity to contribute ideas to the administration's 2025 National AI R&D Strategic Plan. We look forward to working together to accelerate AI-driven innovation.

Sincerely,

Sayeed Choudhury
Executive Director
Open Forum for AI (OFAI)

Contact
Katie Steen-James
Policy Working Group Lead for OFAI
Senior U.S. Policy Manager for the Open Source Initiative

Note: This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the 2025 National AI R&D Strategic Plan and associated documents without attribution.