Computing Community Consortium
1828 L St., NW, Ste. 800
Washington DC, 20036
ccc@cra.org

CCC's response to the Notice of Request for Information (RFI) on Frontiers in AI for Science, Security, and Technology (FASST) Initiative

This response is prepared by the Computing Research Association (CRA)'s Computing Community Consortium (CCC). CRA is an association of over 270 North American computing research organizations, both academic and industrial, and partners from six professional computing societies.

The CCC's mission, a CRA subcommittee, is to enable the pursuit of innovative, high-impact computing research that aligns with pressing national and global challenges. Please note any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the authors' affiliations, or of the National Science Foundation, which funds the CCC.

November 11th, 2024

Written by: *David Danks (University of California, San Diego), Catherine Gill (Computing Community Consortium), Ian Foster (Argonne National Laboratory), William Gropp (University of Illinois, Urbana Champaign), Daniel Lopresti (Lehigh University), Manish Parashar (University of Utah), Matthew Turk (Toyota Technological Institute at Chicago), and Katherine Yelick (University of California, Berkeley).*

III. Questions

1. Data

(a) What kinds of data governance practices, risks, and opportunities should DOE take into consideration, particularly for open sourcing scientific corpuses to the community or interested parties?

Complementing open data, there is significant support within the computing community for open-source tools in AI development. Open-source models act as powerful catalysts for research collaboration and foster inclusivity by lowering the barriers to entry, enabling a wider range of researchers and institutions to engage with cutting-edge AI systems. DOE can strengthen the research ecosystem and enable impactful contributions that are both accessible and widely beneficial by emphasizing open-source development. It would be advantageous for the DOE to recognize the value of open-source frameworks and explore partnerships that facilitate open collaboration within the community.

(b) What types of scientific and energy data should DOE prioritize for large-scale tokenization?

A priority for the FASST program should be the availability of diverse data. Ensuring that a variety of datasets—representative of different domains, scales, and complexities—are made accessible will support robust model training and testing. Additionally, these datasets should be AI-ready, meaning they should be preprocessed, annotated, and structured in ways that facilitate immediate use in machine learning applications. Metadata about each dataset should also be included to aid AI systems in training on these datasets. To address instances where data is classified or sensitive, the program should develop proxy datasets that approximate the structure or characteristics of restricted data, allowing researchers to experiment and refine models without compromising national security.

DOE has significant datasets from simulations (materials, chemistry, climate, cosmological, etc.), non-human genomics, environmental sensors, physics experiments, and its user facilities. These are complementary to the large datasets of interest in the commercial world. Scientific literature should also be made AI-ready to allow models to read and train on it effectively. This would involve reformatting papers and articles into machine-readable formats, enabling natural language processing models to extract key insights, identify trends, and generate new hypotheses. Such a resource could become invaluable for researchers across disciplines, providing a vast knowledge base that AI models can use to facilitate discoveries, optimize processes, and support evidence-based policymaking, serving both scientific and societal needs.

(c) Are there partner organizations with relevant scientific or energy-related data that DOE should work with?

The FASST program would further benefit from a strong, cooperative relationship with the NAIRR infrastructure and its data resources/providers. As both agencies fund and support AI research, it is important to minimize overlap and focus on complementary strengths. Through collaboration, a fully-developed NAIRR could offer general data preparation, storage, and computational resources, allowing FASST to focus more on domain-specific data curation, including proxy data for classified information. This would enable DOE researchers to access an array of prepared data, leveraging NAIRR's resources while maintaining security protocols, and allow NAIRR to draw from FASST's more specialized data assets. NAIRR can provide training programs in AI basics and technical skills, helping to build a larger, more diverse AI workforce. Meanwhile, FASST can focus on advanced, domain-specific training programs in areas like AI-driven energy modeling or security analysis. Collaboration between the two programs could facilitate cross-training, offering researchers and students exposure to both general AI techniques and specialized applications. NAIRR could prioritize foundational AI research and broader public applications, while FASST focuses on highly specialized projects aimed at national security and scientific breakthroughs. The agencies could jointly support interdisciplinary research initiatives, particularly where foundational advances in AI (from NAIRR) intersect with applied research in DOE's specialized fields (from FASST). This collaboration would build a more integrated AI ecosystem, maximizing the effectiveness of federal resources while fostering innovation that addresses both scientific and societal needs.

2. Compute

(b) How can DOE improve awareness of existing allocation processes for DOE's AI-capable supercomputers and AI testbeds for smaller companies and newer research teams? [3] How should DOE evaluate compute resource allocation strategies for large-scale foundation-model training and/or other AI use cases?

DOE should prioritize transparency and collaboration with the broader AI research community. Because the FASST program will likely have the capacity to support only a limited number of projects, it is essential that DOE provides clear criteria for how projects will be selected and why. By establishing and communicating a transparent selection framework, the DOE can ensure that researchers and institutional leaders understand the priorities driving project funding, fostering trust and alignment between the DOE and the research community. Other federal agencies may provide valuable insight in determining which projects are funded by FASST, and we encourage the DOE to include other agencies in this decision making process. We particularly suggest coordinating with those leading the NAIRR initiative. Doing so will ensure that chosen projects serve a broad range of national interests and reduce potential duplication of efforts across agencies.

5. Workforce

(a) DOE has an inventory of AI workforce training programs underway through our national labs.[4] What other partnerships or convenings could DOE host or develop to support an AI ready scientific workforce in the United States?

The program should also adopt a broadened approach to workforce development, recognizing that AI expertise extends beyond developers and researchers. DOE must consider everyone involved in the AI lifecycle, from scientists who interpret AI-driven outcomes to policymakers who need to understand AI's implications for public welfare. Building a workforce that is well-equipped to engage with and benefit from AI requires cross-training in both technical and interpretive skills. In this respect, initiatives like Schmidt Sciences' AI cross-training for post-doctoral researchers can serve as models, helping DOE prepare individuals to work confidently across disciplines and to apply AI insights effectively in varied scientific and policy contexts.

6. Governance

(a) How can DOE effectively engage and partner with industry and civil society? What are convenings, organizational structures, and engagement mechanisms that DOE should consider for FASST?

To support long-term program success, DOE should encourage interdisciplinary collaboration among institutions and universities across scientific areas. Current programs likely emphasize direct funding to universities or student placements within DOE labs, but incentivizing cooperation between departments in DOE labs and institutions would build bridges across scientific fields and foster innovation. Additionally, commercial-scale AI datacenters are extremely expensive to build from the ground up. It may be more efficient for the DOE to pursue partnerships with industry rather than replicate or replace such infrastructure internally. This approach would allow the DOE to access advanced capabilities in a cost-effective manner, supporting sustainable growth in AI research and development.

Continuous collaboration with experts in adversarial testing is also crucial for AI safety and red-teaming. With advancements in AI, the range of models and frameworks available for red-teaming has expanded, offering robust tools that can be directly leveraged for DOE's purposes. Leveraging "off-the-shelf" models for safety and adversarial testing can significantly reduce costs and improve DOE's ability to identify and address security vulnerabilities early in development. Maintaining proactive partnerships with research institutions and industry experts focused on safety will be essential for DOE to meet security and ethical standards, especially as AI models become more complex.