Received: May 29, 2025
Tracking No. mba-72xy-1acq
Comments Due: May 28, 2025 **Submission Type:** Web

PUBLIC SUBMISSION

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-0333
Comment on FR Doc # 2025-07332

Submitter Information

Organization: University of Missouri-Columbia

General Comment

Regional AI Hubs for Developing Foundational AI and Region-Specific AI Applications in Domain Sciences

Attachments

UMissouri-Comments-AI-RandD-RFI-2025

<u>Title</u>: Regional AI Hubs for Developing Foundational AI and Region-Specific AI Applications in Domain Sciences

<u>Comment Submission by</u>: University of Missouri-Columbia (MU)

<u>Contributors</u>: *Prasad Calyam, Tanu Malik, Clintin Davis-Stober, Filiz Bunyak, Yi Shang, Dong Xu, Derek Anderson, Rohit Chadha, Kannappan Palaniappan, Praveen Rao, Sachin Handa, Chris Wikle, Mihail Popescu, Jianfeng Zhou, Taesic Kim, Xiaoqin Zou, Shi-Jie Chen*

Main-Theme: Regional AI Hubs

As a main theme, we request prioritizing funding for "Regional AI Hubs" that can support and sustain AI-based research, education, and economic development at flagship universities such as the University of Missouri-Columbia (MU). As artificial intelligence (AI) research and development accelerate at an unprecedented rate, meaningful participation in this AI revolution demands significant investment in both cyber-physical and human infrastructure. However, these resources such as GPU and high-performance computing, skilled workforce, or advanced training are currently not readily available, which limits innovation in regional industry, municipal and state governments, and educational institutions of all levels.

While educational institutions, regional industries, or local governments may independently choose to deploy AI resources, we think that consolidation of resources in the form of regional hubs will lead to better allocation of the cyber-physical and human infrastructure needed for supporting and sustaining long-term AI-based research, education, and economic development. Such consolidation has been shown through numerous examples in computing such as the invention of the Internet which started with local funding at educational institutions, or the sustaining of HPC facilities which have been consolidated into a few national supercomputing centers and regional network providers across the nation.

We advocate that large, public land-grant research universities, such as the University of Missouri-Columbia, are well positioned to serve as anchor institutions for the development of <u>Regional AI Hubs</u> - organizations that harness AI expertise and technology to support regional industry, enterprise, education, and local/state governments. Regional AI Hubs can then be directed toward providing infrastructure, expertise, tools/technologies for domain-specific application purposes at a regional level.

The development of Regional AI Hubs will enable efficient usage of public and private monies by pooling and organizing AI technologies and expertise. The return on investment for supporting such infrastructure could be considerable. The funding and development of Regional AI Hubs would be a major financial investment in the future of AI technology and productivity. For the US to maintain its leadership in this area, major, coordinated investments, such as the development of Regional AI Hubs are needed, as other global organizations are making investments at considerable scale, e.g., in February of this year, the EU announced a 200 billion euro investment in AI research.

Regional AI hubs represent a strategic investment in national competitiveness by democratizing access to AI capabilities and by ensuring that the transformative benefits of AI reach all geographic regions of the country, not just a few establishments with existing infrastructure and capital.

Technical and Economic Impact through AI-adoption in Regions within the US:

- AI-ready infrastructure investments for research and education, regional economic development
- AI literacy investments to grow workforce, with more instructors, skilled students
- Usable AI in Specific Domains to advance e.g., food, healthcare, smartgrid, drug discovery, materials, public safety
- Foundational AI R&D to accelerate scientific discovery and to maintain technological and economic leadership in AI, while guiding long-term research directions and ensuring safety of AI progress
- Enabling responsible, trustworthy, regulations and energy-efficiency in AI advances

Mission of Regional AI Hubs: The core mission of these regional AI hubs would be to: 1) provide educators, researchers, developers, and local enterprises with access to state-of-the-art AI infrastructure; 2) facilitate education, training, and workforce development programs, and sharing of technical resources; 3) promote interdisciplinary collaboration, connecting academia, government, and industry; and 4) serve as centers of education, innovation, and economic development tailored to regional needs and strengths while contributing to global AI ecosystems.

Justification and Benefits: 1) Different geographical regions offer different strengths in talent pools, academic institutions, and industry concentrations. These regional hubs will enable specialization in AI applications most relevant to their local economies, such as agriculture, manufacturing, healthcare, drug discovery, etc. 2) These regional hubs will enable decentralization of power & innovation, and will lead to a more balanced AI ecosystem that supports AI advances beyond a few large coastal centers. 3) These regional hubs will promote economic growth and resilience by enabling workforce development and retention, by reducing talent migration, and creating stable career pathways in AI-related fields; and by supporting academic-industry collaborations to accelerate research translation and commercialization.

Rationale for a Regional AI Hub lead by University of Missouri: A Regional AI Hub led by the University of Missouri would be a strategically compelling initiative with strong institutional, geographic, and socio-economic justifications.

The University of Missouri (MU) is a public land-grant research university in Columbia, Missouri and is also one of only six public universities nationwide that can claim schools or colleges of medicine, veterinary medicine, agriculture, engineering and law all on the same campus. An MU-led Regional AI Hub can play a critical role in democratization of AI in other smaller higher education institutions in the region that lack expertise and research infrastructure.

- MU's geographic and economic position provides natural access to a multi-state region that hosts key industries including agriculture, healthcare, manufacturing, finance, geospatial technologies and biotechnology that offer significant opportunities for AI-enabled innovation and transformation. This region also connects urban and rural communities, an MU-based AI hub can expand access to AI technologies and resources for rural populations and smaller higher education institutions in the regions.
- Faculty across the MU System are actively engaged in foundational and applied AI research. Units such as College of Engineering, institutes such as NextGen Precision Health, MU Center for Cyber Education, Research and Infrastructure (CERI), MU Center for Geospatial Intelligence (CGI), MU Materials Science & Engineering Institute (MUMSEI), MU Institute for Data Science & Informatics (MUIDSI) provide robust infrastructure for advancing AI innovation. MU's strategic partnerships with organizations

such as the Donald Danforth Plant Science Center and the Taylor Geospatial Institute (TGI) further strengthen its capacity to lead in AI across key sectors including healthcare, biomedical research, material science, agriculture, plant science, and geospatial analytics.

- Because of its land-grant mission, MU has been serving a longstanding role in training, education, and technical support across Missouri and the broader Midwest. MU's existing programs and connections can be scaled to include AI technical training, and workforce development, especially for rural and underserved communities. MU's strong extension network that supports statewide outreach is ideal for distributing AI tools and knowledge across regions.

Next, we outline specific funding priorities that would be supported by a Regional AI Hub model. These funding priorities are organized along two major sub-themes: *Usable AI for Domain Science Applications* and *Advances in Responsible and Trustworthy AI Foundations.*

Sub-Theme1: Usable AI for Domain Science Applications

One of our compelling recommendations is to prioritize Usable AI to enable science-based decision making as a foundational pillar of the national AI strategy. While significant advancements have been made in AI theory, their real-world impact is often limited by challenges in human-AI interaction and relevant cloud architecture designs involving hardware and software integration. Usable AI focuses on designing systems that are not only technically proficient but also intuitive, transparent, and aligned with human cognitive, economic and social contexts. This approach ensures that AI technologies are accessible and practically beneficial across diverse application sectors and populations.

In the following, we provide several domain science research areas at MU that can be supported by federal AI funding in order to expand the advances in Usable AI for improving science decision making.

Usable AI for Precision Agriculture

MU, research in Usable AI for agriculture is involving advancing explainable and human-centered systems that bridge the gap between AI innovation and community adoption. One core initiative focuses on developing explainable AI models that enable domain experts—such as agronomists, animal scientists, farmers, and extension specialists—to understand, validate, and trust AI-generated decisions. Given the complexity of agricultural systems and the limited availability of labeled training data, direct application of generic pre-trained models often results in poor performance or misleading outcomes. To address this, MU researchers are embedding transparent reasoning mechanisms within AI systems, allowing users to inspect the basis of AI outputs and understand the rationale behind recommendations. This fosters trust and encourages adoption, especially among users who may have limited technical familiarity with AI.

In parallel, MU experts are working to lower barriers to data acquisition—a persistent challenge in agricultural AI. The lack of standardized instruments, inconsistent data protocols, and heterogeneous field conditions hinder the creation of robust datasets. To address this, researchers are exploring interactive AI tools, including large language models (LLMs) and chatbot interfaces, that can guide users—especially farmers and technicians—through structured data collection processes in real time. These tools offer conversational support and real-time validation, helping users collect high-quality, usable data for next-generation AI training.

Usable AI for Geospatial Analytics: Disaster Risk Assessment and Community Resilience
At MU, researchers are exploring Usable AI frameworks to transform how communities anticipate and respond to extreme events such as mega-wildfires, hurricanes, earthquakes, and large-scale flooding. While traditional statistical models for these spatio-temporal phenomena are often complex, brittle, and limited in realism and scalability, early applications of AI-based estimators have shown strong potential to support uncertainty-aware modeling at operational scales. However, unlocking the full value of AI in this context requires a human-centered approach that emphasizes transparency, interactivity, and domain-aligned interpretation.

A central challenge lies in modeling community-level responses to disasters—particularly at the urban-wildland interface, where real-time questions arise around optimal fire suppression strategies and dynamic evacuation routes. Existing agent-based models rely on fixed behavioral rules that lack nuance, scale poorly, and often fail to reflect the socio-demographic complexity of human populations. Usable AI can address these limitations by generating synthetic populations with realistic demographic attributes, learning adaptive behaviors from prior events, and simulating decision-making at both individual and collective levels under various disaster scenarios.

To ensure trust, interpretability, and actionable insight, MU researchers are developing hybrid AI-statistical-demographic models with user-facing interfaces that allow emergency managers, local leaders, and community stakeholders to explore, query, and visualize alternative outcomes. These tools prioritize explainability and accessibility, enabling stakeholders with limited technical expertise to engage meaningfully in risk planning, scenario evaluation, and resilience-building. This work exemplifies Usable AI as a critical enabler of equitable, informed, and adaptive disaster preparedness at scale.

Usable AI for Precision Medicine, Healthcare, and Life Sciences
We highlight two use cases of usable AI in precision medicine, healthcare, and life sciences.

<u>Protein Interaction Modeling and Structure-Based Drug Design:</u> The accurate prediction of protein-protein interactions is foundational for understanding biological mechanisms and advancing rational drug design. While recent breakthroughs—such as AlphaFold2 and AlphaFold3—have dramatically improved protein structure prediction, key challenges remain in usability, scalability, and interpretability. For instance, the prediction of disordered or transient protein-protein interactions remains out of reach, and the high computational cost of deep learning models prevents their application in large-scale peptide screening or mutation scanning workflows.

To address these gaps, emerging research at MU is integrating deep learning with physicochemical and bioinformatics models to predict whether two proteins will bind, validate interaction interfaces, and assess how mutations affect binding affinity. However, to be broadly usable, these AI systems must evolve beyond black-box outputs: users need intuitive interfaces, human-interpretable predictions, and workflows that guide decision-making in drug discovery. This calls for the development of interactive, explainable AI platforms that enable biomedical researchers—who may not be AI experts—to query, visualize, and validate predictions with confidence.

<u>RNA Structural Biology and Therapeutic Innovation:</u> RNA-based therapeutics represent a rapidly growing sector, yet key bottlenecks remain in predicting RNA 3D structures, designing RNA-targeted small molecules, and optimizing genome-editing tools like CRISPR-Cas. Traditional AI models often underperform in RNA applications due to sparse training data and the intrinsic flexibility of RNA structures.

To overcome these challenges, research efforts at MU are focusing on hybrid AI-physics approaches that combine deep learning with molecular dynamics and de novo computation. These models not only improve prediction accuracy but also enhance interpretability by embedding physical principles that domain experts can understand and trust. In practice, such tools are helping researchers reconstruct RNA structures from cryo-EM data, simulate RNA-ligand interactions, and fine-tune guide RNA designs for CRISPR through kinetic modeling.

For these systems to adopt Usable AI, they must support interactive exploration, uncertainty quantification, and visual feedback—capabilities that make AI tools transparent and actionable for molecular biologists, chemists, and drug developers. These efforts exemplify how Usable AI can bridge the gap between complex machine learning models and real-world applications in life sciences, supporting the next wave of precision medicine and bioengineering.

Usable AI for Sustainable Energy Futures

The futures of AI and energy are deeply intertwined, presenting both a transformative opportunity and a pressing responsibility. AI has the potential to revolutionize energy systems by enabling smarter grid management, optimizing energy use in buildings and transportation, accelerating the discovery of next-generation energy materials, enhancing energy market operations, and safeguarding critical infrastructure through intelligent cybersecurity. At the same time, the rapid growth of AI—particularly in large-scale deep learning and foundation models—poses significant energy demands, with data centers emerging as major power consumers. To ensure AI remains a net positive force, it is essential to develop energy-efficient and interpretable AI systems that can both advance sustainable energy solutions and reduce their own environmental footprint. This calls for a co-evolutionary approach: AI must help optimize energy systems, while energy-conscious design must shape the future of AI itself. Achieving this vision requires collaborative investment in research, policy, and open infrastructure—prioritizing not just performance, but also usability, transparency, and ecological responsibility. Usable AI, when designed with energy in mind, becomes a critical enabler of equitable, efficient, and resilient energy futures.

Usable AI for Material Sciences

In materials science scientific discovery remains slow—often requiring decades to bring new materials to market—due to reliance on expert intuition and incremental experimentation. National initiatives like the Materials Genome Initiative, the $100 million AI-Semiconductor Materials Initiative, and the $45 million HITS program, to name a few, reflect increasing federal recognition of AI's potential to dramatically advance this domain. At MU, we are addressing this challenge in part through an NSF Research Traineeship (NRT) program (started in 2024) that is preparing interdisciplinary graduate researchers at the intersection of AI, data science (math, statistics, computer science), and materials science (physics, chemistry). A primary emphasis of this program is to prepare undergraduate and graduate students at the convergence of data science, AI and material science. Students learn how to develop usable interfaces and usable and interpretable analysis. Learning how to communicate uncertainties and provide rationale behind the decisions can go a long way to promote Usable AI for science decision making.

Key Recommendations:

Our key recommendations for promoting Usable AI to enable science-based decision making are as follows:

1. **Promote Transparent and Explainable AI**: Encourage the development of AI models that can provide understandable explanations for their decisions, fostering trust and facilitating better human-AI collaboration.

2. **Enhance Workforce Training and Education**: Implement programs that equip the current and future workforce with the skills necessary to effectively interact with AI systems, emphasizing usability and ethical considerations.
3. **Support Open-Source Usable AI Tools**: Fund the creation and dissemination of open-source platforms and tools that prioritize usability, enabling broader adoption and adaptation across various industries and communities.

Sub-Theme2: Advances in Responsible and Trustworthy AI Foundations

The promise of AI to accelerate discovery, broaden the scale of adoption and drive economic growth will not be fully realized unless society can trust the systems built using AI. As we deploy, scale, and sustain such systems, it will be important to support efforts for Responsible AI that is focused on building technical and social foundations that make AI safe, reliable, secure, and energy-aware from first principles, while also creating the standards, evaluation frameworks, and cyberinfrastructure needed for wide adoption of AI innovations. By embedding these principles early alongside excellence in software and hardware, we can ensure that breakthroughs in foundational and applied AI translate into benefits for every community and every scientific field.

In the following, we highlight foundational research thrusts that exemplify the funding priority needs to broaden the scale of AI adoption:

Human-AI Teaming and AI Agents for Scientific Discovery

Federal AI R&D investments could be prioritized to support Human-AI teaming for scientific discovery tasks. AI agents offer transformative potential to address complex scientific challenges across domains. These interactive systems can perceive multimodal inputs, comprehend natural language, and perform context-aware actions to support users in reasoning, decision-making, and communication. In industry, AI agents have become essential tools, streamlining operations in sectors such as customer service, manufacturing, and education. However, their adoption in academic science, particularly in data-intensive fields such as bioinformatics and drug development, remains limited. Scientific discovery tasks in such fields e.g., material science, chemistry and biomedical science will involve AI assistants that have to aid scientists/researchers in experiments that are time-consuming, siloed across various scientific instruments, and data scarce.

This gap presents a critical opportunity: by developing intelligent, adaptive agents capable of integrating heterogeneous data sources, synthesizing scientific literature and databases, and optimizing analytic workflows, we can significantly accelerate the pace of discovery. The “AI-Labmates” can work similar to “Co-Pilot” AI technologies used in industry for coding productivity but will need to be customizable to serve various scientific domain needs. They can function as domain-customized assistants, analogous to coding co-pilots, but tailored to support the long-horizon, iterative nature of scientific inquiry. In bioinformatics, for instance, AI agents could autonomously annotate datasets, recommend pipelines for multi-omics integration, simulate cellular systems, and continuously learn from expert feedback. More broadly, AI agents can fuse experimental data with literature-derived knowledge to generate novel hypotheses, refine predictive models, and guide decision-making with greater speed and precision. We recommend sustained federal investment in the foundational research, infrastructure, and open-source platforms required to develop these next-generation agents that aid as “AI Labmates”. Doing so will empower researchers across disciplines, democratize access to advanced scientific tools, and transform the future of discovery through effective human-AI collaboration.

To accelerate foundational scientific discovery and reduce the time to bring new materials to market, a team of interdisciplinary researchers at the MU spanning computer science, electrical, chemical, and mechanical engineering, alongside our cross-campus materials science institute are advancing new theories of human–AI teaming that tightly couple physical experimentation with simulation and human knowledge. Our research spans three integrated thrusts: (1) agent-directed oxidative molecular layer deposition (oMLD) to understand how monomer sequencing affects optical and electrical properties; (2) AI-guided 2-photon polymerization and electromagnetic simulation to design, fabricate, and understand the effects of complex 3D infrared metasurfaces; and (3) multi-scale integration of these physical systems for enhanced benefits like conformal oMLD coatings to tune spectral responses, enabling applications such as thermal imaging and radiative cooling. Together, these foundational investigations form an end-to-end, multi-scale (atomic to macro level), AI-driven materials discovery process. At the core is a flexible, hierarchical agentic AI framework that supports multi-objective constrained optimization, active learning, inverse design, and natural language summarization to interface with large language models for scientific reasoning. The broader opportunity and challenge lies in interfacing these multi-disciplinary AI concepts with high-cost, experimental and automated simulation tools, requiring novel, explainable optimization strategies to enhance both machine performance and human insight. However, realizing this vision demands increased and sustained national investment in multidisciplinary human-centric AI that is purpose-built for advancing discovery across heterogeneous scientific environments.

Trustworthy and Standards-based AI

In a world increasingly shaped by algorithms, the future of trustworthy AI hinges not just on innovation, but on integrity. Reproducible and accountable systems are the scaffolding on which responsible AI must be built; systems where every decision, prediction, and output can be traced, understood, and validated. Imagine an AI ecosystem where models are not black boxes but living documents annotated, transparent, and open to scrutiny. Where provenance is as fundamental as performance, and where every dataset, hyperparameter, and preprocessing step is logged with the rigor of a scientific experiment. In this vision, accountability is not an afterthought or compliance checkbox, but an embedded principle: AI systems that can explain themselves, audits that can replay decisions, and infrastructures that support dynamic trust not just static validation. These systems will empower not only regulators and researchers, but citizens themselves to ask: “Why did the algorithm decide this?” and receive a meaningful, verifiable answer. They will allow researchers to build on each other's work, fostering a cumulative AI science that accelerates innovation without compromising integrity. The road to such systems requires sustained federal investment in open standards, robust tooling, and interdisciplinary thinking.

As AI profoundly reshapes all aspects of human endeavors, including education, the economy, and society, it presents both exciting new opportunities and potential risks. To derive the maximum benefits of AI, the associated risks must be mitigated. AI is susceptible to a variety of safety and security risks, including hallucinations (where AI produces incorrect answers), prompt injection attacks (where malicious attackers attempt to manipulate generative AI systems, revealing private information and/or producing potentially harmful output), and data poisoning attacks (where malicious attackers attempt to manipulate training data). It is critical to have federal funding support for the development of trustworthy AI frameworks that address these security and privacy issues. Such frameworks can provide the highest assurance to AI data providers that individual privacy is maintained and to AI consumers that the AI-generated models and code will not cause them inadvertent harm. We call for sustained federal funding to develop frameworks for trustworthy AI systems that uphold the highest standards for safety, reliability, and security, not only in their design but also in the mechanisms that ensure these standards are met. The mechanisms may take several forms (or a combination of them), including formal methods, explainable AI, establishing legal standards for accountability of decision-making, and human oversight.

We, at MU, call for several directions in which sustained federal funding has the potential to advance the trustworthy and standards-based AI efforts: (1) Invest in the development and maintenance of open-source platforms, audit kernels, and reproducibility toolchains that enable full lifecycle tracking of AI experiments from data collection to model deployment. (2) Develop and enforce national standards for documenting datasets, model architectures, training procedures, and evaluation metrics. These standards should support interoperability and enable independent verification of results. (3) Tie funding, publication, and evaluation criteria for federally funded AI research to clear reproducibility benchmarks e.g., availability of code, data, and execution environments with version control and audit trails. (4) Establish a government-backed digital archive where AI models, datasets, notebooks, and their provenance metadata can be deposited, accessed, and cited similar to clinical trial registries in medicine. (5) Require reproducibility and accountability checks for AI systems used in critical domains (e.g., healthcare, criminal justice, finance) to ensure decisions can be independently verified. (6) Invest in curricula and training programs that teach best practices in reproducible and accountable AI, targeting both the research community and government agencies deploying AI systems. (7) Encourage participatory design and public oversight mechanisms e.g., citizen panels and red teaming that enhance trust and hold systems accountable in democratic, socially responsive ways.

Frameworks for Evaluating AI Decision Making

AI systems are increasingly being used for decision making, both as recommendation systems for enhancing human decision makers and for replacing human decision makers. While the speed and scale of AI decision making systems is attractive for many application areas, many questions remain regarding the suitability and quality of AI decisions. Given that AI systems tend to be “black box” processes and are fundamentally probabilistic, it is not always clear how to evaluate their decisions and, subsequently, how to improve them. It is critical to prioritize federal funding support for the development of theoretic frameworks for evaluating the quality and consistency of AI decision making systems. Such frameworks would necessarily bring together computer scientists, engineers, decision theorists, philosophers, and behavioral scientists. It is crucial to invest in development of frameworks that can be applied to complex decision environments where there is not necessarily a ground truth (e.g., subjective preferences) or a ground truth is not knowable (e.g., prediction, decision making under uncertainty/risk).

Energy-efficient AI

Energy-efficient AI innovations are necessary to reduce the computational resources and energy consumption of AI models and systems. The current generation of AI hyperscalar data centers for generative AI and LLM require a gigawatt of power and cost about $60 billion to build and support several hundred thousand water-cooled Nvidia Blackwell Graphics Processing Units (GPUs). Scaling up to support next-generation Agentic-AI and Physical-AI for scientific discovery, business workflows, advanced manufacturing, improved healthcare delivery, etc. which will be even more compute intensive and necessitate real time interactions with people and the environment. Consequently, it will require breakthroughs in energy efficient AI for broad scale adoption and economic feasibility.

Previous algorithmic approaches for energy efficient AI have explored model pruning, neural network architecture compression or simplification using knowledge distillation, deep network weight quantization, linear or reduced accuracy multiplication, dynamic computation for resource allocation and related approaches. For next-generation AI Engines with a greater compute capability per watt of energy, we call for funding that will invest in exploring new approaches to realize more efficient AI including hardware and system level approaches. These include jointly optimizing specialized hardware and software for AI models like GPUs, Tensor Processing Units (TPUs), field-programmable gate arrays (FPGAs), System-on-a-Chip (SoC) AI building blocks, integrating AI into the edge sensor at the device level for intelligence at the edge, neuromorphic, analog and biologically inspired architectures. In terms of system level optimization for energy efficient AI approaches to build upon, research can involve federated learning to avoid large/centralized data repositories, automated optimization compiler technologies, distributed computing and dynamic workload management. General approaches can include sparsity exploitation in model weights and in activations during inference, continuous learning with efficiency to avoid retraining from scratch with massive datasets that grow incrementally, selectively update parts of the model using meta-learning approaches, shift more intelligence to the edge, build physics-aware and model-based AI that incorporate science and engineering knowledge to enable more data-efficient learning.

Advancing Cyberinfrastructure for Scalable AI Training and Inference

AI has become ubiquitous and drives a plethora of real-world applications in information retrieval (IR), natural language processing (NLP), computer vision, speech recognition, e-commerce, healthcare, defense, and so on. Generative AI services (e.g., ChatGPT, DALL.E), which can generate new content of different modalities such as text and images, have had an explosive growth in recent years. It is predicted that generative AI will become a $1.3 trillion market by 2032. Large language models (LLMs) (a.k.a. foundation models), which are trained on a large corpus of unlabeled data via self-supervised learning, have become the bedrock of generative AI. Several LLMs has been proposed in recent years, namely, GPT-3, Jurassic, Gopher, Megatron-Turing NLG 530B, OPT, and LaMDA. Google's PaLM, Meta's Llama, NVIDIA's NVLM, and Anthropic's Claude family of models are other notable contributions. Indeed, LLMs have taken over the world of AI like a storm.

LLMs are pretrained on massive datasets using 100's of graphics processing units (GPUs) costing millions of dollars. It is claimed that OpenAI's GPT-3 (with 175 billion parameters) cost more than $4.6 million to train. Databricks spent $10 million to build their LLM called DBRX. Meta used clusters with 24K GPUs to train Llama3. While the race for building larger LLMs continues to drive technology companies, there is a growing need in academia to empower users to pretrain LLMs to foster new scientific discoveries as well as prepare the next generation of scientists, engineers, and educators. Furthermore, specialized LLMs trained on domain-specific datasets are becoming more attractive as they tend to be more accurate than generalized LLMs.

Although it is impossible to level the playing field for academic users in terms of access to massive compute resources and datasets for training LLMs, we posit that new advances and federal investments in cost-effective cyberinfrastructure can narrow the gap for academic users. As a result, academic users are empowered with generative AI capabilities at no charge to foster innovations in computing, informatics, sciences, and beyond. New research directions include development of novel algorithms, techniques, and systems for scalable AI training and inference on heterogeneous computing frameworks. Access to massive data repositories is also essential for pretraining LLMs. Enabling high utilization of computing resources during training and achieving low response time during inference would be key factors for success.

Key Recommendations:

Our key recommendations for promoting Responsible and Trustworthy AI foundations for broader scale of adoption are as follows:

1. **Operationalize Human-AI Teaming Paradigms**: Fund cross-disciplinary collaborations that pair cognitive scientists, domain experts, and AI engineers to create interactive agents whose reasoning processes are visible, contestable, and aligned with human goals especially for scientific discoveries.
2. **Promote Open Evaluation Frameworks & Public Benchmark Repositories**: Ensure that fairness, accountability, transparency, and safety metrics are maintained as community resources, lowering barriers for small labs, startups, and public agencies.
3. **Prioritize funding of AI Risk Management Frameworks and Standards:** Support efforts on auditable lifecycle documentation, red-team testing, and third-party certification for high-impact models before deployment.
4. **Invest in Energy-Aware AI Hardware, Software, and Metrics**: Provide matching grants for academia–industry consortia that advance low-power architectures, carbon-aware scheduling, and public dashboards reporting the environmental/energy footprint of large-scale training.
5. **Expand Secure, Scalable Cyberinfrastructure for AI R&D**: Establish regional AI testbeds with high-speed networks, confidential-computing enclaves, and federated-data services to democratize access to state-of-the-art training and inference capabilities.