STRIDE Ventures
AI Efficiency Challenge Solicitation

1. THE AI EFFICIENCY CHALLENGE

Construction of large, inefficient AI data centers is an increasing problem for U.S. capacity and competitiveness. Although significant AI/ML efficiency gains are possible with existing technologies, a gap exists between these innovations and their large-scale deployment.

STRIDE aims to dramatically improve the efficiency of at-scale AI/ML systems and data centers in order to improve the effective capacity and competitiveness of the U.S.-based companies. The primary mechanism for attaining this goal through this Challenge will be the accelerated commercial adoption of translation-ready solutions. It is anticipated that the program as a whole will yield substantial efficiency improvements, although specific quantitative goals will be determined by each individual project.

The premise of this Challenge is that

- Researchers have already created technologies having the potential to collectively deliver multiple orders of magnitude in improvements to the efficiency with which AI/ML workloads are fielded, and
- Rapid adoption of these technologies can accelerate U.S. competitiveness in the AI space by reducing the cost of training and/or inference, addressing near term limitations on the buildout of data center capacity and/or infrastructure to power those data centers, and accelerating time to market of new models.

Given the above premise, a goal of this Challenge is to have awarded teams achieve preliminary deployments at scale within one (1) year after receiving funding, with additional, more impactful deployments occurring in the second year of the program.

This Challenge intends to support technologies that are translation-ready and primarily software-based, in order to achieve more rapid deployment. In this context, translation-ready technologies are those that have demonstrated promising research results in a realistic setting, and that the work remaining to deploy them at scale is engineering, integration, and validation. That is, translation-ready

technologies should not rely on or need new research or long lead-time hardware/infrastructure to be deployed at scale.

2. TEAM STRUCTURE

Two types of proposals are targeted for this Challenge: Solution Teams, and Benchmarking Teams.

The first type of proposals are from Solution Teams. Solution Teams are teams developing and deploying technologies to improve efficiency at various layers of the AI software stack. Solution Teams can be multidisciplinary teams that combine expertise across research, development, and the deployment/operation of at-scale AI/ML systems.

Importantly, Solution Teams **must** include two types of participants:

- Technology developers – Researchers, innovators, and/or others with or developing translation-ready technologies,

- Problem owners – People who come from industry and/or organizations that deploy AI/ML systems at scale and seek to improve the efficiency and competitiveness of their offerings.

In baseball jargon, technology developers are “pitchers”, and problem owners are “catchers”. Pitchers are the entrepreneurial researchers in academia, government, ventures, corporate research and/or non-profits with translation-ready technologies seeking to deploy and implement their solutions at-scale in real-world deployments. Catchers are people at organizations that (i) operate AI/ML workloads at a meaningful scale, (ii) have the technical capability and willingness to integrate technologies proposed by solution developers, and (iii) commit to measuring and reporting the resulting efficiency gains.

Solutions Teams **must** include both solution developers and one or more problem owners in order to be eligible for this Challenge.

The second type of proposals are from Benchmarking Teams. In similar baseball jargon, a Benchmarking Team is an “umpire” with expertise in benchmarking AI/ML software-based systems. This team will be expected to develop one or more

industry benchmarks to assess the efficiency of the AI/ML software stack, and to potentially assess the efficiency gains of one or more of the Solution Teams.

3. SCOPE

This Challenge is a translational activity targeting dramatic near-term improvements in AI/ML efficiency at scale.

For Solution teams
In-scope applications are only those that are translation-ready and primarily software-based. That is, those software-based technologies that have demonstrated promising research results in a realistic setting, and where the remaining work to deploy them at scale is engineering, integration, and validation. In-scope technologies should not depend on long lead-time or slow-moving infrastructure or hardware – such as energy generation, batteries, power grid management, materials research, etc. – to be deployed at scale.

Some examples of topics that are in scope are:

- Software implementation *of AI/ML systems:* Software and algorithms that lead to efficiency improvements at one or more points across the AI/ML pipeline from data preparation to training to inference-serving to application/agentic architecture. These techniques could focus on the efficient utilization of computational resources and/or of memory/storage resources. Importantly, the goal is not to change the design of the models (or sub-models) themselves but rather to improve the overall efficiency of their implementation and deployment. Automated techniques that involve mixtures of experts, the automated distillation/compression of models, cloud-edge partitioning, etc., are within scope.

- *Tools to guide the creation of efficient AI-related code:* Tools that dramatically improve the ability to create code that executes highly efficiently are in scope, as are tools, which may themselves be AI-based that generate highly efficient code and/or transform existing code. These tools may be informed through either offline or online measurement of program performance, e.g., using instrumentation that is built into GPUs/accelerators. They should ensure the correctness and/or explainability of the resultant code (or guidance) and attain

dramatic net improvements in end-to-end computational efficiency over the entire life cycle, including software development, of at-scale AI systems.

- *MLOps and distributed system software:* Also in scope are improvements and new paradigms for system software (scheduling, placement, runtime systems, orchestration, etc.).

- *Efficient agentic orchestration:* Techniques that substantially improve the end-to-end efficiency of systems in which AI-based systems invoke agents, coordinate their actions, etc. are in scope.

- *Edge computing:* Approaches that off-load computation from the cloud to edge devices are in-scope. These can include hybrid approaches, such as those in which parts of the inference are performed in the cloud while other parts are performed at (or near) the edge. Edge approaches are particularly relevant where power/energy, runtime/throughput, connectivity or privacy considerations make fully cloud-side inference impractical.

- *Energy and/or Thermal Management:* Software-enabled demand response flexibility and/or efficiency gains in energy distribution and/or thermal management are also in scope, provided the overall efficiency gains are meaningful relative to other software-enabled opportunities and there is a path through which the innovations can be deployed within the timeframe of this Challenge.

- *AI/ML algorithms:* Although the primary focus of this Challenge is on the efficient training and operation of AI Models (vs the design of the models themselves), new or improved types of models and/or training algorithms that are translation-ready and demonstrably more efficient will be considered, provided there is a path to their at-scale deployment within the timeframe of this Challenge.

- *Hardware supporting AI/ML:* Although the primary scope of this Challenge is software or algorithms, translation-ready improvements in the efficiency of the compute triad (logic, memory/storage, and communications) at any level (device, processor, rack, datacenter, cloud/edge) will be considered, provided there is a path through which the innovations can be deployed within the timeframe of this Challenge.

For Benchmarking Teams
Although the bulk of the funded teams will be Solution Teams, this Challenge may fund up to two teams whose focus is on benchmarking the efficiency of AI/ML systems. A Benchmarking Team will be expected to develop one or more industry benchmarks for use by commercial entities to assess the efficiency of the AI/ML software stack, and to drive adoption of that benchmark in the market. Additionally, the Benchmarking Team(s) will be expected to assist one or more of the Solution teams with the quantification and validation of their efficiency gains.

Up to eight (8) Solution Teams and up to two (2) Benchmarking Teams will be selected to participate in this Challenge.

4. PROGRAM DESIGN

Application Type, Funding Level, and Schedule Track
Teams will submit a written application outlining: the sources of the inefficiency they have identified; their translation-ready technology; how the technology can be used to address sources of inefficiency; and their plan for at-scale deployment. See section 5 and the application form for specific details and questions.

Multiple Principal Investigators (PIs) from the same university can apply to this Challenge, and a given PI can be a team member or the lead on multiple applications. However, a given PI will be awarded as the lead for only one (1) funded project.

Proposals will be evaluated based on their potential for disruptive efficiency gains, feasibility, and alignment with the program's goals.

On the application form, applicants will choose their desired application type: Solution Team or Benchmarking Team.

Additionally, applicants can apply for different funding amounts. On the application form, applicants will choose their desired funding level: Large funding level of $3.5M, or Medium funding level of $1.75M in total to achieve the project goals. For both funding levels, award recipients are expected to apply a meaningful amount of the budget directly toward deployment-related activities such as integration, instrumentation, and at-scale operation. In addition to the funding requested through

this Challenge, teams may also benefit from in-kind resources (such as staff and/or cloud services, etc.) provided by their commercial team members or partners.

Lastly, since the goal of this Challenge is fast execution to accelerate solution deployment at scale, applicants will select their desired schedule track: Fast Track or Regular Track. Execution timelines for the Fast Track are compressed relative to the Regular Track, as described below. The Fast Track exists to provide an alternative for faster deployment of funds to support teams that can operate at an advanced pace. Note: Selecting Fast Track is **not** considered to be a positive or negative during the review of applications, and it does **not** impact the budget level selection. Teams selecting either schedule track (Fast Track or Regular Track) may opt for either funding level (Large or Medium level).

Applicants will be expected to provide project plans demonstrating that they can quickly and cost-effectively ramp-up resources dedicated to the proposed activities in order to achieve the proposed results on a timely basis.

Activity Stages, Schedules, and Payments
This Challenge is structured as a two-year milestone-driven program. The process is designed to be fast, focused, and flexible, supporting progress of translation-ready technologies towards deployment at scale, while ensuring accountability.

The Challenge is structured in three (3) Stages:

- Stage 1 (2 months for Regular Track, 1 month for Fast Track): Each team will conduct a detailed analysis of the efficiency issues they will address, refine the specification of the metrics they will use for measuring progress, and elaborate on their plans for a series of at-scale deployments. For Solution Teams, this includes identifying / measuring inefficiencies and validating translation-ready technologies through prototype deployments. For Benchmarking Teams, this includes defining the benchmark scope, design principles, target adopters, and path to adoption.

    - Deliverables for Solution Teams:
        - A baseline quantification of the specific sources of inefficiency they are addressing,
        - The metrics through which efficiency gains will be measured,

    - The current baselines for both research/benchtop tests and the at-scale deployment(s) that will be used to measure progress,
    - The team's goals (both benchtop and at-scale) for efficiency gains (relative to the baselines) at the end of each of Stage 2 and Stage 3,
    - The plan for a series of at-scale deployments that will demonstrate progress towards and eventual attainment of the team's defined goals.

  - Deliverables for Benchmarking Teams:
    - A proposed set of deliverables that demonstrate the viability of the proposed benchmark development activity and its path to industry adoption.

- Stage 2 (10 months for Regular Track, 5 months for Fast Track): Each Solution Team will engage in spiral development-deployment cycles and will demonstrate progress through micro-benchmark achievements and at-scale deployments towards their stated Stage 2 goals. It is expected that some initial level of deployment will be attained within the first 5 months of this stage, for Regular Track projects and earlier for Fast Track projects. Specific development and deployment-related activities may include: necessary enhancements to the translation-ready technology to prepare it for at-scale deployment; prototypes to demonstrate the benefits of the technology in realistic settings; at-scale-deployment and operation of the technology; and measurement of the resultant efficiency gains. A Benchmarking Team will develop and prototype the benchmark framework, conduct validation studies, demonstrate tangible progress towards stakeholder adoption, and incorporate stakeholder feedback.

  - Specific deliverables for Solution Teams and Benchmarking Teams will be defined by each team at the end of Stage 1.

- Stage 3 (12 months for Regular Track, 6 months for Fast Track): Each Solution Team that meets Stage 2 goals and is selected to enter Stage 3, will continue spiral development-deployment cycles towards their stated Stage 3 goals. Teams will demonstrate significant further improvement in efficiency and engage in more ambitious at-scale deployments. Each Benchmarking Team will finalize benchmark specifications and promote industry adoption.

- Specific deliverables for Solution Teams and Benchmarking Teams will be defined by each team at the end of Stage 1 and updated at the end of Stage 2.

Payments are made retrospectively, and are tranched based upon milestone achievement/accomplishment.

Support and Evaluation

During each stage, selected teams will be supported by the U.S. National Science Foundation (NSF), STRIDE, and industry experts across two primary engagement types: weekly coaching and quarterly mentoring sessions. Weekly coaching will take place virtually on a one-on-one basis, focusing on each team's milestones and progress towards achieving those within the stages. Mentoring will take place on a quarterly basis, with a total of three mentoring sessions per Challenge year, where selected teams will gather in-person for up to three days to attend both individual and group sessions focused on supporting the team's stage goals.

At the completion of Stage 1, milestone reports will be submitted to and certified by STRIDE with the assistance of expert coaches. Each team's report will include a rigorous description of the team's Stage 2 and 3 goals in addition to the milestones by which interim progress will be evaluated and funds will be disbursed.

The completion of the Stage 2 and Stage 3 milestones and attainment of the agreed stage goals will be assessed by a jury of experts, drawing on submitted reports, demonstrations and in-person presentations (described in section 7: Selection and Contracting Process, below). Failure to achieve milestones could trigger a no-cost funding extension, and/or could result a team being deemed ineligible to progress to the next Stage.

Application Timeline

Date	Event
May 18, 2026	Call for submissions launched
May 28, 2026	Information Webinar #1
June 11, 2026	Information Webinar #2
June 25, 2026	Information Webinar #3
July 7, 2026	Information Webinar #4
July 13, 2026	Application deadline (11:59PM PT)

July 27, 2026	Notification of invitation to Pitch Day
August 13-14, 2026	Pitch Day for invited teams
September 8, 2026	Communication of awarding decision to teams
September 14, 2026	Start of Stage 1

Regular Track Timeline

Date	Event
September 2026	Start of Stage 1 (2 months)
November 2026	Start of Stage 2 (10 months)
December 2026	Quarterly Meeting #1
March 2027	Quarterly Meeting #2
July 2027	Quarterly Meeting #3
August 2027	End of Stage 2 Jury Meeting
September 2027	Start of Stage 3 (12 months)
September 2028	Challenge End, Jury Meeting, Winner Announcement

Fast Track Timeline

Date	Event
September, 2026	Start of Stage 1 (1 month)
October 2026	Start of Stage 2 (5 months)
December 2026	Quarterly Meeting #1
February 2027	End of Stage 2 Jury Meeting
March 2027	Quarterly Meeting #2
July 2027	Quarterly Meeting #3
March 2027	Start of Stage 3 (6 months)
September 2027	Challenge End, Jury Meeting, Winner Announcement

5. ELEGIBILITY

The Challenge is open to a range of U.S.-based applicants. The following are eligible to participate as applicants and team members:

- Academic institutions and research organizations,
- For-profit companies of any size, including startups and established enterprises,
- Nonprofit organizations,
- Consortia.

The lead applicant must be a U.S.-based entity and will serve as the primary point of contact and contractual partner that receives funds from STRIDE.

Development work funded through this Challenge must be done either in the U.S. or by non-U.S.-based employees of the lead U.S.-based entity. International collaborators can be team members or sub-contractors, but funding will be directly awarded only to U.S.-based organizations. Deployments funded through this Challenge can be done either inside or outside the U.S.

Federally Funded Research and Development Centers (FFRDCs) are eligible to participate as lead applicants, team members, or subcontractors, but are subject to applicable direct competition limitations and cannot respond to this solicitation unless they submit a letter drafted and signed by their sponsoring organization that (a) cites the specific authority establishing their eligibility to propose to Government solicitations and compete with industry, and (b) certifies the FFRDC's compliance with the associated FFRDC sponsor agreement's terms and conditions.

The following are not eligible to participate as applicants or team members:

- Organizations on the U.S. Department of Commerce Bureau of Industry and Security (BIS) Entity List
- Entities identified under Section 1260H of the William M. (Mac) Thornberry National Defense Authorization Act ("NDAA") for Fiscal Year 2021 (P.L. 116-283)
- Foreign entities of concern as defined in the CHIPS Act, P.L. 117-167 § 10638(3) (42 U.S.C. § 19237(3)); and
- Individuals participating in a Malign Foreign Talent Recruitment Program, as defined in the CHIPS Act, P.L. 117-167 § 10638(4) (42 U.S.C. § 19237(4)).

Award recipients will be required to provide initial and annual certifications on these above points.

Team Composition Requirements

There is no required/recommended team size. However, teams must demonstrate that all team members are essential to the success of the project and that they have or can assemble the full range of expertise needed to execute the proposed work. Specifically, each Solution Team must include at least one problem owner (catcher) organization that will engage in the at-scale deployment of the technology and the measurement of the efficiency gains. In the special case of a benchmarking

proposal, the Benchmarking Team should include an organization with a successful record of leading industry-facing benchmarking activities.

6. APPLICATION AND APPLICATION PROCESS

Applicants must submit a complete application package through the online portal available on the STRIDE website, by the deadline listed on the STRIDE website.

Applicants will select their team type (Solution Team or Benchmarking Team), which will lead to separate application questions, summarized below. Additionally, applicants will select their schedule track (Fast Track or Regular Track) and their funding level (Large or Medium).

Solution Teams' application will include:

- A brief description of the source and scale of the inefficiency to be addressed,
- A summary of (or links to descriptions of) the translation-ready technologies that will be deployed and how they are differentiated from existing solutions,
- A description of how the technology will be applied to address the inefficiency and an analysis of the potential end-to-end efficiency gains that can be achieved,
- A description of the at-scale deployment opportunity, an explanation as to why the approach has not been deployed in this environment to date and why it is feasible to deploy it now, and a description of the top 1-3 remaining obstacles to deployment,
- A work plan covering all three stages of the Challenge, with emphasis on intermediate steps that will build confidence in the technology, inform further refinement of the plan, etc.,
- Team bios and organizational information, and
- Milestone plan and Budget estimate, including proposed milestones in the work plan mapped to resources to be funded by STRIDE and a description of any resources that will be contributed by the participants.
- A Letter of Intent from the catcher organization describing their commitment to deploy the Solution team's technology at scale.

Benchmarking Teams' application will include:

- A brief description of the benchmark to be created and how efficiency measurement will figure prominently in it,

- A description of how industry participation in the development and adoption of the benchmark will be encouraged,

- An explanation as to why the benchmark doesn't exist today and how past barriers to its development and adoption by industry will be overcome,

- A work plan covering all three stages of the Challenge, with emphasis on intermediate steps that will build confidence in the benchmark, inform further refinement of the plan, and drive industry adoption,

- Team bios and organizational information, with specific emphasis on the team's qualifications to lead the creation of a benchmark that will garner industry acceptance, and

- Milestone plan and budget estimate, including proposed milestones and resources, and a description of any resources that will be contributed by the participants. The plan and budget should include a task related to assisting the other teams with the quantification and validations of their efficiency gains.

Applicants are also encouraged to leverage existing federally supported computing and data resources where applicable. Researchers and educators across all disciplines may apply for allocations of computing or data resources through the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program; additional information, including allocation request procedures, is available at https://access-ci.org/. Those engaged in AI research or research that employs AI methods may also access a broad suite of federally funded and private sector-contributed resources through the National Artificial Intelligence Research Resource (NAIRR) program; information on the NAIRR and currently available opportunities can be found at https://nairrpilot.org.

Each application will require two self-certifications before submission: (a) that the lead applicant has read and understood the STRIDE Ventures AI Efficiency Challenge Participant Agreement, and (b) that the applicant is eligible to receive federal funding under this Challenge, including compliance with the Foreign Entities of Concern restrictions described above. Applications without these certifications will not be considered.

Application materials will be treated as confidential. They will be reviewed by Start2 Group, domain experts engaged under NDA to support the down-select shortlisting, and by the Jury for shortlisted applicants. Summary information about the applicant pool, the applications, and the basis for selection will be provided to NSF as part of the selection process of pitching companies. Teams selected to pitch will be required to submit additional information to NSF.

7. SELECTION AND CONTRACTING PROCESS

Applications that are responsive to the Challenge goals will be reviewed by experts with experience across research, development, AI/ML, cloud operations, commercialization, etc. The first-round review will assess feasibility, potential for large efficiency gains, and alignment with the program's goals.

From this pool, a subset of applicants will be invited to pitch to a jury of interdisciplinary experts. This will be the team's opportunity to make the case for its vision: to show how its approach challenges the status quo, redefines what's possible, and will result in at-scale deployments that yield significant improvements in the efficiency of at-scale AI/ML systems.

The Pitch Days event will be a two-day in-person event where selected applicants will bring two representatives (minimum) to present for 45 minutes.

Final selection of applicants invited to Pitch Days, and the final award decisions following Pitch Days are subject to approval by NSF.

A Principal Investigator (PI) can be the lead on only one (1) funded award, but multiple PIs from the same university or academic medical center can apply to this Challenge.

NSF Eligibility Review
Once applicants are notified of their selection to pitch, they will be required to submit evaluation materials that will be submitted to the NSF for review and approval. These include the following:

- Current & Pending (Other) Support (CPoS) Forms
- Biographical Sketches
- Collaborators and Other Affiliations (COA)

These documents are expected to be submitted for each key senior personnel on the applicant's team. Please note that AI Efficiency Challenge funds should be listed as pending.

Templates, instructions for using SciENcv, and definitions of key senior personnel are available at the following links:

- SciENcv Overview: https://www.ncbi.nlm.nih.gov/sciencv/
- SciENcv FAQs: https://nsf-gov-resources.nsf.gov/files/SciENcvFAQs.pdf
- Biosketch: https://www.nsf.gov/funding/senior-personnel-documents#biographical-sketch-0bd
- COA FAQ: https://www.nsf.gov/funding/senior-personnel-documents/faq/coa
- COA template: https://nsf-gov-resources.nsf.gov/files/coa_template.xlsx

NSF retains authority to determine the eligibility of applicants and selected Performers, including review of key personnel, organizational affiliations, and other relevant information. NSF may, at its discretion, remove proposals from consideration or decline to fund selected Performers based on research security concerns.

Participant Agreement
The Challenge is administered by Start2 Group as the OT Contractor under an Other Transaction (OT) Agreement with NSF. Selected teams ('Performers') will sign a Performer Agreement directly with Start2 Group, which incorporates required terms from the NSF agreement.

The AI Efficiency Challenge is built to move quickly and keep administration light. To support this, the Participant Agreement will be shared shortly after the application

opens, allowing applicants time to review and signal their agreement to the contract terms prior to Jury decisions. This will help ensure that funds can be released quickly once Stage 1 companies are selected.

Applicants are expected to familiarize themselves with the agreement which, once available, will be available on the STRIDE website and at the following link: STRIDE Ventures AI Efficiency Challenge Participant Agreement Link.

For questions regarding the solicitation, please:

- Reach out to info@stride-ventures.com
- Review FAQs and register to the Information Webinar available at https://stride-ventures.com/ai-efficiency-challenge/