ORCD Project Olympia

Accelerate Your AI/ML and Beyond Research. Spring 2025 Strategic Seed Fund Solicitation

The MIT Office of Research Computing and Data (ORCD) is launching a new strategic seed fund initiative. The seed fund is one part of a strategic effort supporting infrastructure for applied AI/ML research across all of MIT. Under this first round of the program ORCD is planning to support around 6 project opportunities for 3-6 months. Eligible projects will be activities that can usefully leverage ORCD GPU resources to accelerate their research and produce tangible results. Projects of interest to the program will also collaborate with ORCD team members, providing feedback and contributing toward evaluating and improving the ORCD research computing environment for all of MIT. Projects should be led by someone holding PI status in an MIT DLCI.

In this seed fund round, we are seeking projects that can leverage multiple ORCD H200 GPU nodes to bring research close to publication ready. Projects suitable for this seed fund should have already established a basic level of capability and be able to show a promising likelihood of producing impactful results within 6 months or less. Funding support will be provided for 3 months of graduate student time or equivalent (see later details). Projects will be provided with enhanced access to GPU resources and fast storage to help achieve their research goals. Project participants will be expected to interact regularly with members of the ORCD team. This initiative is part of ORCD efforts to create a powerful AI/ML platform for all of MIT.

Successful projects will work with the ORCD team to provide guidance on how to deliver an effective platform service for boosting AI/ML research for all at MIT. Ideal projects will be interested both in advancing their own domain research and in contributing to growing an ecosystem of openly shared applied AI/ML software and tool knowledge. Innovative closed source projects are of interest, but the ORCD team is particularly interested in projects that might include jointly publishable, reproducible software artefacts that can be published and shared in open forums such as Zenodo (https://zenodo.org/), or in online publications like the Journal of Open Source Software (JOSS - https://joss.theoj.org/), or in online technology meetings such as the High-Performance Extreme Computing conference (HPEC - https://ieee-hpec.org/).

Funding and GPU resources are for close to publication ready applied AI projects in any domain. Areas of potential interest include:

- Dataset and database development. Novel multi-modal data curation and assembly; domain specific embedding development and testing.

- Application and/or development of foundation models. Models in new areas, including models that could impact areas of interest to MIT strategic initiatives such as health, climate.
- Model tuning. Novel approaches to imposing domain constraints (physical laws, known axioms, prior knowledge) to applied scenarios, including ideas on how to extend physical process emulation models robustly into "out-of-distribution" scenarios.
- Contextual training. Leveraging context windows to shape generative AI system responses in areas from science and engineering process emulation to decision making.
- Infrastructure innovation. Research directions, with domain science applications, around more efficient gradient descent optimization strategies, ML network optimization, information compression and training and inference efficiency, novel software tool chains and hardware/software technologies.
- Innovations in reduced precision (e.g. FP8 and FP4) AI/ML for applied research.
- Automated testing, checking, analyzing and generation of physical world discrete PDE and ODE models.
- Novel uses of GPU systems for research beyond AI/ML domains.

We encourage both open-science and limited access research.

Application process

Interested applicants should submit a short technical narrative (one to two pages, not including references and links) and budget using the forms at https://orcd.mit.edu/seed_fund/spring2025. In this round funded projects will be provided for up to three months of financial support for a graduate student or equivalent position and priority access to up to twenty-four H200 GPUs or equivalent. Projects should explain how they will realize a goal of accelerating work for a target publication within 6 months and should provide a technical narrative that supports that goal. They should explain the status of their work and plans (including an anticipated timeline) for achieving publication quality results in the period proposed. We expect most projects to complete in less than six months. We anticipate making the first awards from this program to start in June 2025. Two-page proposals should be submitted by Feb 28, 2025. We expect to select approximately 6 projects in this first round. We anticipate inviting a set of roughly 12 final candidate projects to make brief 15-minute presentations sometime in mid-March. Selection decisions will be shared by April 24, 2025.

The project description should ideally include - evidence of multi-GPU scaling of proposed computational work (including ensemble based scaling); evidence of the team's ability to leverage H200 or equivalent based resources; a summary of a proposed publication ot other outcome and the results needed to achieve that goal; convergence and other plots where relevant from standard systems such as weights and biases; an estimated, up to 6 months, timeline for the project; suggestions for how the proposed activity will work with the ORCD team over the project duration. The ORCD team is available to advise groups on how to determine if their project is suited for a seed fund cycle and to provide guidance on how to bring a project to an appropriate level of readiness.

Awarded projects will be expected to participate in 30 minute bi-weekly meetings to review progress, to provide feedback on how the ORCD environment is meeting their needs and to discuss how to ensure progress is satisfactory.

We expect to solicit projects every six months. Renewal of awards that have made progress will be considered in subsequent solicitations. ORCD is actively raising funds to support this program so that, if the approach proves successful, the scope and breadth can grow over time. We are also actively working on access to other GPU resources, including large collections of light-weight inference oriented L4 GPUs and access to next generation NVidia and AMD GPUs.