Academic Grant Program

Call for Proposals: AI Inference, Agents, and Systems Software

About This Call for Proposal (CFP)

NVIDIA has been transforming accelerated computing for more than 25 years. We're defining the next era of scientific computing and supporting top researchers doing compelling, computationally intense work to solve some of the world's most challenging problems. Developer programs support researchers in using our SDKs, frameworks, and web services by providing training and creating online communities to help researchers do their life's work using NVIDIA technologies.

NVIDIA solicits proposals for innovative projects related to generative AI, particularly agentic systems, distributed inference, language models, diffusion models, and multimodal models. Projects leveraging AI models must incorporate NVIDIA models, such as Nemotron™, Cosmos™, or Omniverse™ wherever possible. All others must make extensive use of NVIDIA software and AI distributions, such as NVIDIA NeMo™ microservices, NVIDIA Dynamo, NVIDIA FLARE™, or CUDA-X™ GPU-accelerated libraries. Proposals should be attached to one or more of the following themes:

Techniques for Customizing, Enhancing, and Operating Gen AI

- **Targeted Improvement:** Propose and develop novel methods for targeted model quality improvement, including those in the direction of automated synthetic data generation, iterative self-distillation, or self-play.
- **Explainability:** Investigate new approaches to improving model explainability, be it through architectural modifications, model tool use, or automated extrinsic verification.
- **Reliability:** Propose innovations targeting the inherent unreliability of generative AI systems, be it the occurrences of factual hallucinations, failures to follow complex instructions, lack of robustness with respect to input parameter perturbation, or general output non-determinism.
- **Safety, Security, Privacy:** Investigate model alignment techniques and architectural innovations designed to enhance the explainability, robustness, and overall trustworthiness of models and agents. Research areas might include jailbreaking, prompt injection, data exfiltration, memory poisoning, reconstruction attacks, and model inversion.
- **Inference Efficiency:** Propose innovations in inference implementations to ensure efficient generation of high-quality outputs while maintaining model interpretability and safety. This includes model compression methods, such as quantization and pruning, efficient decoding systems, efficient serving and scheduling approaches, novel disaggregation approaches, and quality-enhancing output post-processing techniques. This also includes improvements to fault-tolerance techniques to reduce downtime when a software or hardware error is encountered.

Exploring Compound AI Systems

- **Reasoning:** Introduce new methodologies for enabling generative AI systems to solve problems on their own accord, either with the help of external tools (e.g., SAT solvers, logic/arithmetic verifiers), aided by reward models, or through self-reflection/self-consistency). The focus is on the research of new techniques, not application of existing techniques to new modalities/fields/problems.
- **AI Agent Systems and Architectures:** Advance the operation of agentic model systems through improvements to efficiency, reliability, or safety. Propose targeted adjustments to existing AI systems to improve efficiency, reliability, or problem-solving performance.

Exploring Systems Software for AI

- **Scalable AI Infrastructure:** Propose scalable inference systems, including distributed runtimes, scheduling techniques for performance or energy efficiency, resource management frameworks, and fault tolerance techniques. Develop systems for edge AI and/or systems support for AI safety, security, and privacy.
- **Compilation and Optimization:** Introduce optimization techniques or agentic AI flows for parallelization, tiling, fusion, etc.
- **Data Management:** Develop optimizations for KVcache management, RAGs, long-term memory, and other states in distributed inference systems.
- **AI for Systems:** Introduce agentic AI techniques to advance beyond the limitations of manual design.

Award Details

Selected principal investigators (PIs) may receive:

- Up to two NVIDIA DGX™ Spark
- Up to 30,000 H100 80 GB hours or equivalent (max. of eight concurrent GPUs)

Not all projects that meet eligibility requirements will be selected for an award. The final award amount will be determined by the NVIDIA awards panel. GPU hours provided to the PI will expire six months after the award; unused GPU hours will be forfeited. Physical hardware will be shipped to the PI.

Applicant Eligibility

Full-time faculty at accredited academic institutions that award research degrees to Ph.D. students are eligible. Postdocs and graduate students must work with a full-time faculty member to submit on their behalf.

Each person can submit one proposal per quarter, a maximum of four proposals annually. For example, one submission is allowed in the first quarter, January–March. Each individual applicant is eligible to receive one award per calendar year.

Proposal Requirements

Proposals **must** follow the proposal template and should not exceed **four pages**, not including appendices.

Expectations of Recipients

Award recipients should make reasonable efforts to acknowledge the support of NVIDIA Corporation and reference how specific hardware and software contributed to project results. Recipients will inform NVIDIA of publications, presentations, open-source code and data releases, and speaking engagements that reference the supported project via the NVIDIA academic grant portal. Failure to report in the portal will influence future award selection.

Please review NVIDIA Academic Grant Program terms and conditions.

NVIDIA