Creating a Resilient AI Ecosystem with NDIF and NNsight

David Bau

July 4, 2024

The current AI landscape is shaped by two dominant approaches: closed models with opaque APIs, and open models with published parameters. Both approaches have drawbacks that hinder the development of a resilient[1] AI ecosystem. This brief memo introduces the National Deep Inference Fabric (NDIF) and NNsight as technologies designed to address these long-term challenges and create a more resilient, adaptable AI research and development environment.

Current Ecosystem Vulnerabilities

The closed model approach, while protecting proprietary information, severely limits scientific progress. Independent researchers cannot scrutinize model internals, leading to a lack of diverse perspectives in understanding the largest AI systems[2]. This opacity raises the risk that society will be ill-equipped to anticipate, explain, and mitigate new types of problems when they arise.

The open model approach, while promoting transparency, faces its own set of challenges. Once a model is released, it cannot be "recalled," which can lead to uncontrolled misuse. This problem arose in Stability AI's release of the parameters of Stable Diffusion, wherein the model's capabilities enabled the proliferation of deep fake porn child sexual abuse material[3]. Because the model has been broadly copied, the company has no practical way to reverse the ongoing damage caused by its capabilities.

Both approaches struggle with long-term resilience. Closed models concentrate power and knowledge in a few companies, while open models lead to a lack of accountability for damaging effects. Neither approach adequately balances the needs for innovation, transparency, and responsible development, and neither approach puts society on a resilient path for safe deployment of AI in the long run.

NDIF and NNsight: A Balanced Approach

The National Deep Inference Fabric (NDIF)[4] is an inference service that will host large models using a structured access API[5] called NNsight[6]. NNsight is a highly transparent programming interface that provides much more visibility and control than existing commercial AI inference APIs. It extends pytorch with tracing context managers that allow researchers to specify code that modifies the execution of remotely-hosted large AI systems without taking custody of the underlying model. Figure 1 illustrates the architecture of NNsight and NDIF.

Together, these technologies provide a middle ground between open and closed approaches. Researchers can probe and modify model internals without requiring their own copy of the weights, exploiting transparent access to model activations, gradients, fine-tuning, and interventions through remote customization. This setup not only addresses key challenges in current AI research but also provides

[1] Here we examine resilience as sustained adaptability in the face of unanticipated challenges, as distinguished from reliability and robustness. See *Four concepts for resilience and the implications for the future of resilience engineering*, Woods, 2015.

[2] A survey of categories of research methods on AI safety that are hindered by black-box commercial API access can be found in *Black-Box Access is Insufficient for Rigorous AI Audits*, Casper, et al, 2024.

[3] For a detailed analysis of how the open-source release of Stable Diffusion has led to the proliferation of child sexual abuse material, see *Generative ML and CSAM: Implications and Mitigations*, David Thiel, et al, 2023.

[4] The National Deep Inference Fabric project is described at `https://ndif.us`

[5] The proposal for structured access APIs was made by Bucknall, et al. in their analysis of recent AI research methods; see *Structured Access for Third-Party Research on Fronter AI Models*, Bucknall and Trager, 2023.

[6] Technical documentation for the NNsight API can be found at `https://nnsight.net`

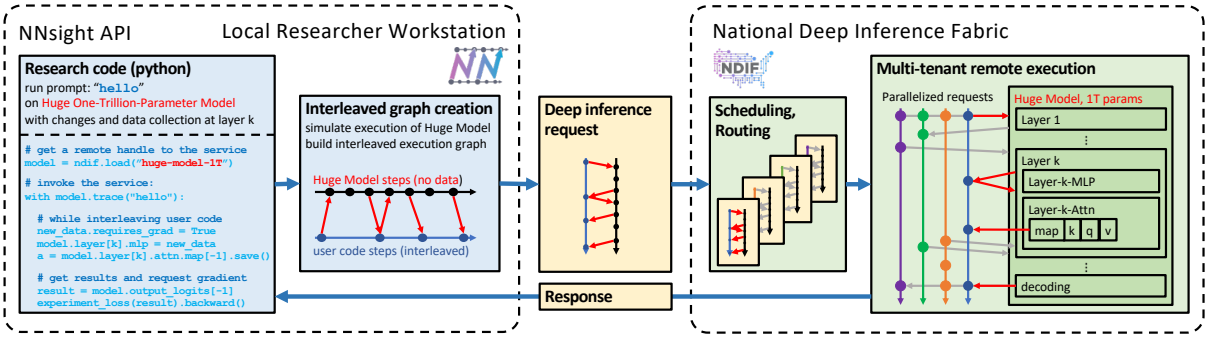

Figure 1: NNsight is a structured access API that allows researchers and application developers to probe and customize a large model without taking custody of the weights. The National Deep Inference Fabric (NDIF) is an NSF-hosted infrastructure that will host large models efficiently using the NNsight API.

a framework for enhancing AI safety. Unlike current approaches that focus primarily on evaluations of anticipated safety concerns, this ecosystem approach creates a structure resilient to unanticipated problems and changes in technology and usage.

Building Resilience through Standardization

The NDIF and NNsight framework offers several key features that promote resilience:

- **Openness in Practice:** Shared computational resources through NDIF lower the barrier to entry for analysis, testing, and customization, giving small-scale scientific efforts and application developers access to the same types of technical capabilities as larger-scale competitors[7].
- **Standards-Enabled Innovation:** A uniform method of interacting with AI models facilitates the development of a diverse ecosystem of model developers and customizers, leading to faster development of common tools, methods, and collaborative innovation.
- **Collaborative Safety Research:** By enabling diverse researchers to share tools and models, the standards-based ecosystem also facilitates collaborative efforts in identifying and addressing potential safety issues before they become critical problems.
- **Balanced Transparency:** The technical approach allows for scrutiny of AI models without propagating copies of the models themselves, enabling the development of invasive audits and scientific advances without the risks posed by irreversible and uncontrolled model releases.
- **Rapid Safety Interventions:** Standardized access through a shared fabric allows for quick implementation of safety measures such as automatic model updates when new risks are identified, enhancing overall ecosystem safety.

Models deployed using using the NDIF and NNsight framework will equip the ecosystem to handle unexpected changes or problems by leveraging a diverse, interconnected community with standardized tools and shared resources.

AI Safety and Stakeholder Concerns

This framework addresses a critical gap in current AI safety approaches. While most other AI safety efforts focus on monitoring and testing to improve AI reliability and robustness against classes of problems that can be anticipated today, NDIF and NNsight create a resilient societal structure capable of responding to unforeseen challenges[8]. Empowered by transparent tools, third-party audits will not be static gates; instead they will come from a living ecosystem of scientists and industry participants who

[7] Efficient deployment of a trillion-parameter model at half-precision requires multiple terabytes of GPU RAM, which implies conducting research on a high-bandwidth multiple-node GPU cluster, a steep leap in engineering cost and complexity beyond single-GPU experiments that researchers can perform on smaller models.

[8] Such as dangerous new modes of use or novel emergent behaviors that are seen in usage only after deployment in society.

continuously improve their understanding of AI and monitor its impacts on society. Transparency and community involvement rely on open standards, which are crucial for long-term AI safety, as they allow the ecosystem to evolve its safety measures in response to emerging risks.

Both open- and closed-model stakeholders will have legitimate concerns about intellectual property and secrecy when they choose how to deploy their models. To address these, NDIF envisions developing a fabric that stitches together multiple service providers, including providers who offer technical safeguards against weight exfiltration for closed models[9], and providers who can be trusted as disinterested third parties, protecting the intellectual property interests of both model creators and application developers[10] (initially the National Science Foundation service will be a neutral host).

Both open and closed stakeholders can benefit from the NDIF framework: for the open-model community, the shared fabric offers a solution to practical resource constraints, lowering costs and increasing the ability of model users to work with much larger models than are currently practical. The closed-model community stands to benefit from the creative efforts of a more productive community of application developers who will be willing to make larger investments due to greater safety, transparency, and control over the models that they build upon.

The Initial Mission of NDIF: Enabling Large-Model Research

The National Deep Inference Fabric has been funded by the National Science Foundation to enable scientific research on large-scale AI. While enabling the advancement of AI research is only a portion of the long-term vision for NDIF, it is the first achievable goal for the framework. As part of its NSF charter, the NDIF will create the standards and services that will allow academic researchers to conduct invasive experiments on very large-scale open-parameter models. In the future, those same standards will form the basis for AI safety audits and other industry collaborations.

Under the NSF, NDIF will serve large open models such as LLama3-405b. Although models with published parameters are theoretically available for academics to study without any special framework, in reality when those models grow beyond about 100 billion parameters, the hardware and engineering challenges that must be overcome to conduct research on the models put research out of reach of most academics in practice. Therefore the initial goal of NDIF will be to create a shared computing fabric that solves the difficult engineering problems as a common resource, and to enable academic researchers to share the fabric for efficiently probing and customizing the largest open models in a cost-effective way.

The NSF-funded NDIF can be seen as a pilot deployment for a general framework of standards that can serve as a model for commercial or regulatory deployments in the future.

Conclusion

AI is undergoing a period of rapid change. We cannot anticipate all the complex problems that will arise as AI is deployed in society, but we can respond to this uncertainty by creating a *resilient AI ecosystem*.

NDIF and NNsight offer a path towards a more resilient AI ecosystem by defining a technical approach that balances the needs for openness, innovation, and safety. By providing standardized transparent access and shared resources, these technologies will help the AI community respond to both anticipated and unanticipated challenges. The approach complements existing AI safety efforts by creating a flexible, responsive framework capable of adapting to new safety concerns as they emerge.

By working towards standardization of responsible transparency in AI, we can create an AI landscape that is more adaptable, collaborative, and ultimately safer and more beneficial to society.

---

[9] Besides contractual obligations forbidding weight extraction, NDIF will enable technical measures protecting against exfiltration of very large closed-parameter models including limits on egress bandwidth, made feasible by NNsight programming capabilities that allow users to conduct the full range of legitimate work within NDIF using only minimal data downloads. The ratio of model size to available bandwidth can be raised so that an exfiltration effort would be highly detectable by a vigilant service provider, requiring the aggregate bandwidth of all users of the API for many months or years.

[10] Application developers are understandably reluctant to share their precious fine-tuning data and innovations with large-model developers who are their potential competitors; this points to the need for neutral NDIF service providers who have no incentives to compete with suppliers or developers.

Thanks to Alex Loftus and Jonathan Bell for discussions and suggestions on drafts of this memo.