Overview

Large language models (LLMs) such as ChatGPT that surpass 100 billion parameters have ushered in an exciting new era of artificial intelligence (AI). But their massive size has created a crisis of transparency; the computational requirements to run such models have made it infeasible for academic researchers to conduct research into how they work. To address this critical research need, we propose to design, build, and deploy computing infrastructure for **deep inference** in LLMs. By deep inference we mean the instrumentation and study of the behavior, mechanisms, and impact of an AI model when it is used to perform tasks *after* it has been trained.

The **National Deep Inference Facility (NDIF)** will advance scientific understanding by providing U.S. academic researchers with efficient, reproducible, and fully transparent access to the complete computations of pre-trained LLMs beyond 100 billion parameters—a capability not currently accessible to academics today. This facility will launch a new phase of AI innovation in the U.S., improve understanding of AI models, unlock cutting-edge advances and enable widespread training of a highly skilled workforce to lead the world in the ethical use of state-of-the-art LLMs.

Intellectual Merit

The National Artificial Intelligence Research Resource Task Force has identified the development of **trustworthy AI** as one of the critical priorities for strengthening U.S. Artificial Intelligence (AI) R&D. The NDIF directly addresses this national priority by providing the computational capacity, instrumentation, transparency, broad access and training that is necessary to enable research on LLMs to advance trust, including investigations of societal implications, auditing of internal mechanisms, reproducible testing and evaluation, and studies of AI safety.

State-of-the-art machine language models trained to predict language use on large text data sets have exhibited new capabilities when models are scaled over 100 billion parameters, including some aspects of general-purpose reasoning. The emergence of these capabilities has posed a wide range of **fundamental scientific questions that impact almost every discipline**. However, the ability to train LLMs does not directly lead to an understanding of how they are able to achieve such feats at inference time (i.e., when they are run). That hinders our ability to anticipate, explain, and regulate these systems. The NDIF addresses an urgent need for transparency.

Working with a community of dozens of scientists nationwide and under the leadership of a unique team of experts in machine learning, deep network interpretability, language modeling, software engineering, high-performance computing, and inclusive computing, the proposed project will yield **open-source software, tools, and a broadly-available national computation resource for transparent LLM inference** to enable the U.S. academic community to conduct cutting-edge research into the mechanisms, limits, and impacts of state-of-the-art LLMs.

Broader Impacts

Highly-capable LLMs will increasingly be deployed into use, with potentially widespread implications for society. *But scientists cannot explain the predictions of such models*. Although academics are well-suited to critically scrutinize the inner-workings of large AI models, the infrastructure required to perform such research is out of reach for most academic labs. The NDIF will enable U.S.-based academics to conduct critical research into LLMs that is currently not feasible, spurring broad advances exemplified by our collaborating researchers in **computing, medicine, neuroscience, linguistics, social sciences and humanities**.

The inference service and outreach will directly support the research agendas of graduate students in AI, thereby playing **a central role in training the next generation of researchers**. Moreover, we will develop undergraduate and graduate-level course materials and, through workshops and fellowships targeting PUIs and MSIs, make these resources broadly available across the nation.

Mid-scale RI-1 (M1:IP): The National Deep Inference Facility (NDIF) for Hundred-Billion-Parameter Language Models

This Mid-scale RI-1 implementation project has no anticipated environmental or cultural impacts.

1 A Computational Microscope for Large Language Models

Powerful large language models (LLMs) such as ChatGPT [1] herald a new era of artificial intelligence (AI) that is poised to reshape society [2], but *scientists cannot explain their predictions*. LLMs are able to write cogently about real-world topics [3], follow human instructions [4], and even pass legal [5], medical [6], and computer programming [7] exams. Both policymakers [8] and researchers [9] have stressed the urgency of explaining how they perform such tasks. Because we know how to *create* LLMs, we can now clearly envision the instrumentation necessary to open up their black box calculations and *explain* them. **Just as physicists characterize particles using atom smashers and biologists catalog genes using DNA sequencers, researchers will explain machine intelligence by running LLMs under a computational microscope.**

The need for national-scale instrumentation to explain LLMs arises due to the demanding computational requirements for conducting research-oriented inference on these very large models. *Training* LLMs requires massive computational resources using graphical processing units (GPUs) to analyze huge bodies of text. Once trained, LLMs are used for *inference*—i.e., the models are run with learned parameters. Inference still requires significant resources, as the trained model cannot fit onto a single GPU. While private companies like OpenAI offer commercial inference services (Figure 1a) such as ChatGPT, those services do not expose their internals, making it impossible to study their mechanisms. By providing a transparent inference service (Figure 1b), NDIF will enable rigorous study of LLMs, increasing critical scientific understanding and supporting research on the impact of LLMs on society. The NDIF consists of three complementary investments:

1. Creation and testing of an online inference service with the support staff and infrastructure necessary to provide academics with the ability to interrogate and conduct groundbreaking

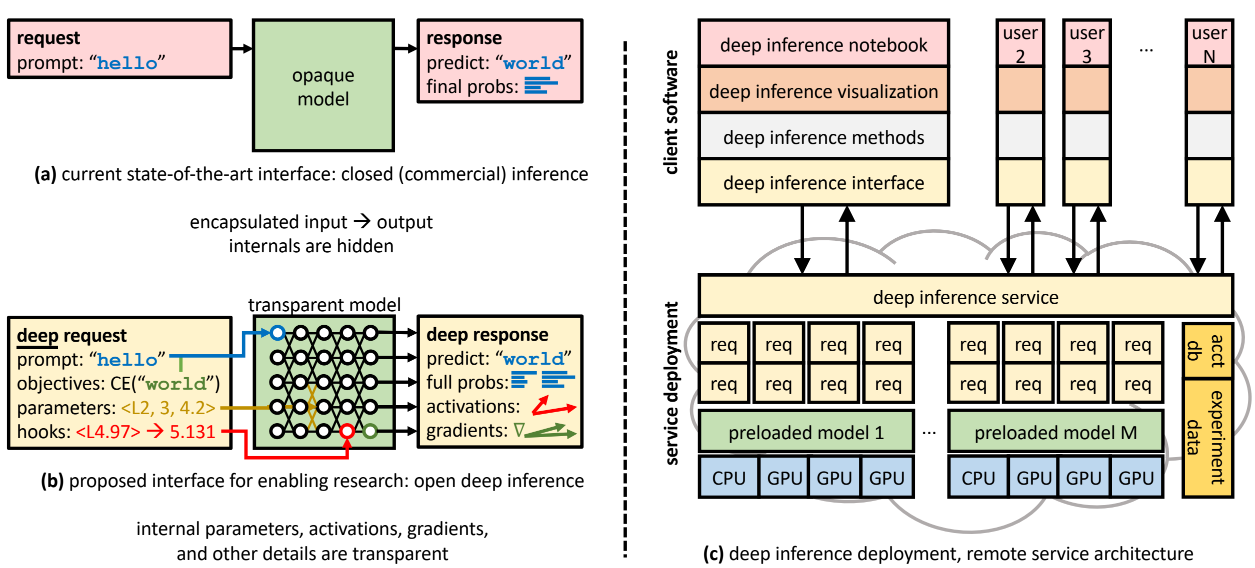


Figure 1: (a) Current services hosting large language models provide very limited interaction functionality (Top). One can send input text in a request, and is then provided an output string (and scores associated with the final predictions). (b) We propose developing infrastructure to provide deep access *to hosted language model instances (bottom), which will permit critical research without necessitating researchers hosting such models themselves. (c) The infrastructure consists of new software libraries and a deployed distributed service to be shared by researchers nationwide.*

research on the largest available and most scientifically relevant LLMs. (Figure 1c).
2. Development of an open-source server and client software stack that will power the service, suitable for deployment on other large clusters.
3. Outreach and training for PhD students and U.S. researchers to use this facility to advance understanding of large neural network models and their impacts.

NDIF is complementary to other projects—detailed in Section 2.2—that have the goal of *pre-training* open LLMs. Those efforts aim to produce LLMs with open parameters such as Bloom [10] and OPT [11], but they do not provide the inference service infrastructure that would be needed by academics to study them in detail. NDIF will provide academic researchers with that infrastructure. By "deep inference" we denote the instrumentation and study of the behavior, mechanisms, and impact of an AI model when it is used to perform tasks *after* it has been trained.

NDIF will be developed at Northeastern University, building on our existing organizational structure, facilities, and experience in research computing. The hardware cluster will be deployed at the Massachusetts Green High Performance Computing Center, a shared computation facility, in which Northeastern is one of the five university partners.

Why not a traditional HPC cluster? U.S. academics can access accelerated computing hardware via several DoE and NSF funded clusters, such as XSEDE/ACCESS. Those focus on batch jobs and partition resources across users, whereas NDIF will be optimized for many users to share access to a few large models persistently loaded on relatively few servers. By concentrating usage, this approach greatly reduces operating costs for this workload, as described in Section 2.1.3.

2 Intellectual Merit

Explaining AI systems is a national and global priority: the National AI Research Resource Task Force [12] identified that one of the four critical opportunities for strengthening the U.S. AI R&D ecosystem is to develop trustworthy AI by "supporting research on AI's societal implications, developing testing and evaluation approaches, improving auditing capabilities, and developing best practices for responsible AI R&D can help improve understanding and yield tools to manage AI risks." Similarly, the Future of Life Institute, whose "Pause Giant AI" open letter [9] has more than 25,000 signatories including many other leaders in AI, has recommended "a significant increase in public funding for technical AI safety research" in the areas of alignment, robustness and assurance, and explainability and interpretability [13]. And the White House Office of Science Technology Policy has released a Blueprint for an AI Bill of Rights [8] delineating a consumer's right to AI systems that "provide explanations that are technically valid, meaningful and useful."

Meanwhile, LLMs such as ChatGPT are being adopted more quickly than any previous technology, with widespread deployment in consumer-facing technologies [14], touching intellectual work in almost every field even as its mechanisms remain unexplained [2]. Because we do not understand how LLMs make their predictions, we find ourselves in a situation where the most impactful class of AI model today is inscrutable: the opacity of LLMs has become a foundational challenge to our national goal to develop trustworthy AI.

Academic researchers are well-suited to investigate the mechanisms of LLMs, but for the reasons we shall detail in Section 2.1.3—primarily a lack of infrastructure—they are unable to do this critical research. The barriers to research are a new problem, largely stemming from the unprecedented scale of state-of-the-art LLMs. NDIF will directly address these needs through a robust investment of a shared hardware and software platform. NDIF will enable a diverse community of researchers to conduct rigorous analyses of LLMs to understand their inner-workings and failure modes.

2.1 Scientific justification

Since the emergence of computing as a discipline, researchers have pursued the creation of generalized or "human-like" machine intelligence, and debated how to measure it [15, 16]. Previous generations of AI systems have surpassed human-level capabilities in narrow domains such as playing chess [17], answering quiz-show questions [18], playing Go [19], and classifying images [20]. Capable as they are, those systems are qualitatively different from the emerging technology of LLMs, because LLMs are *generalists*. LLMs are able to perform a variety of tasks without explicit supervision [21], which makes them a uniquely interesting subject of study.

The flexible capabilities of LLMs have emerged from pre-training with the simple language-modeling objective of predicting the next word in a sequence, given the preceding words.[1] Despite this simplicity, LLMs such as GPT-4 [23] and OPT [11] are capable—to varying degrees—of answering questions about the world [24], translating between natural languages [21], performing mathematical reasoning [25], obeying descriptive requests to perform a variety of tasks [21], applying theory-of-mind reasoning about the knowledge of people [3], and learning to perform a new task given a small set of input examples without further training [21].

The purpose of the NDIF is to provide the broad academic research community with a platform to conduct experiments that explain the mechanisms and impacts of such emergent phenomena.

2.1.1 Large language models have created a new crisis of transparency

While the emergence of very large models such as GPT-3 [21] has energized the Natural Language Processing (NLP) and broader Machine Learning (ML) research communities, the dominant success of such models has also presented the research community with a crisis of transparency that is very different from the previous generation of "large-scale" AI.

When the AlexNet [26] model shocked the computer vision community in 2012 by winning the ImageNet Visual Recognition Challenge, it comprised 62 million learned parameters. That was large for the time, but sufficiently small for academic labs to be able to reproduce, validate, modify, retrain, and study the model using relatively cheap hardware. Similarly, when the first successful pre-trained models for NLP—e.g., ELMO [27] and BERT [28]—emerged, these were small enough for academic researchers to run, interrogate, and tinker with locally, enabling important research into their capabilities and limitations [29]. That accessibility led to an explosion of creativity and innovation, with a doubling of AI papers published annually from 2011 to 2021, and a 30-fold increase in the annual number of AI-related patents filed [30].

The current advance represented by GPT-3 [21] and similar very large language models (LLMs) is qualitatively different. The 175-billion parameter GPT-3 model is private. Alternative, comparably sized LLMs (such as OPT [11], Bloom [10], and NEO-X [31]) are technically available to researchers, but often *de facto* inaccessible due to their size: Most academic researchers do not have sufficient resources to run such models, and so they are unable to probe them in depth. Much academic work on analyzing LLMs therefore relies on the paid Application Programming Interfaces (APIs) that OpenAI or other vendors make available for integrating with other commercial products. Inference API services obviate the need for one to run (very large) models locally to interact with them. But this approach comes with a critical trade-off: Commercial inference APIs provide only limited access to model outputs (Figure 1a), in part to ensure that model weights remain proprietary. This precludes researchers from characterizing the internal mechanisms that models have learned from data, and that in turn threatens to slow the pace of innovation, shielding new developments behind the cloak of private ownership such that advances in AI cannot be subject to the kind of competitive scrutiny provided by independent academics.

Moreover, opaque models directly conflict with the aim of developing trustworthy AI. Our

[1]GPT-3 and GPT-4 are understood to have also been "fine-tuned" using human feedback, but OpenAI provides practically no details on this [22], which is illustrative of their opaque modus operandi. The degree to which such explicit supervision is required to realize the impressive performance of LLMs is yet another open question which the NDIF will help researchers investigate.

ignorance about the mechanisms that give rise to surprising LLM capabilities creates a troublesome dilemma between performance and transparency [32], making it difficult to anticipate how models will behave when deployed in the real world [33, 34]. Furthermore, our lack of understanding hinders our ability to regulate these systems and ensure safety [35, 36].

By enabling the diverse academic research community to scrutinize how such models work, NDIF will empower important research into the potential risks of LLMs that are beyond the purview of industry labs. The broad range of academic researchers in our user community (Section 2.3) demonstrates that NDIF will enable research not only in computer science, but also in biomedical science, neuroscience, bioinformatics, and social sciences. We should not leave critical research on LLMs—their capabilities, biases, functioning, and shortcomings—only to companies with clear self-interest in their positive reception. Such research should be conducted by academic groups in a transparent manner that emphasizes reproducibility and ethical conduct of research, and should be subject to the rigors of peer-review.

2.1.2 Research methods for understanding LLMs

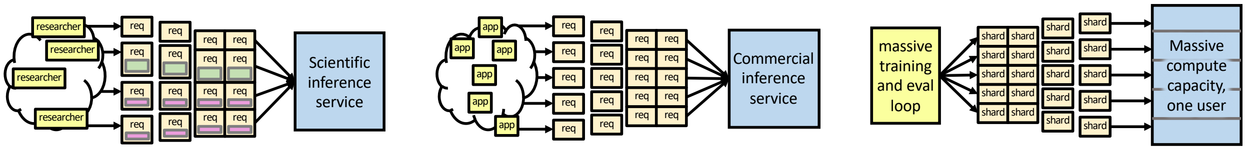
The emergent capabilities of large models pose a fundamental question: *How do they work*? When a large model makes a surprising decision, what information contributed to this? What rules did the model apply to make its choice? What information has the model learned to store, and where does it put it? Understanding such mechanisms is important to distinguish profound computational capabilities from the mere *appearance* of them. The NDIF will enable several classes of experimental methods to enable researchers to characterize LLM mechanisms; we enumerate some in Figure 2, and we detail several of these methods below (we do not review them all in detail owing to space constraints).

Deep inference research methods on LLMs enabled by the NDIF	Compute profile: Interactive	Batch	Optimization	Transparency needs: Activations	Gradients	Interventions	Parameters	Training data
Human subject studies	✓							
Representation probing	✓			✓				
Interactive visualization		✓		✓				
Saliency mapping		✓	✓	✓	✓			
Causal mediation analysis		✓		✓		✓		
Input synthesis methods			✓	✓	✓			
Parameter efficient fine-tuning			✓		✓	✓	✓	
Direct model editing			✓	✓		✓	✓	
Influence functions			✓		✓			✓
Representation similarity analysis		✓		✓				
Latent factor modeling		✓		✓				
Neuron response analysis		✓		✓		✓		✓
Memorization analysis		✓						✓

Figure 2: Deep inference research methods enabled by NDIF.

Representation probing. One major line of inquiry asks: What information does the network encode? For instance, computational linguistics might want to know whether and to what degree LLMs encode varieties of *semantics* [37]. This is illustrative of a body of emerging research probing internal representations for implicit linguistic structure (e.g., [38, 39]). Elsewhere, work on LLMs for healthcare has shown that neural representations of health records implicitly encode patient race, which has fairness implications [40]. As another intriguing example, recent work found that even when a language model is conditioned to output falsehoods, it may contain a hidden state that represents the true answer internally [41]; this discovery is only possible with access to model internals.

Salience mapping. A model can also be better understood by asking: What parts of the input are most affecting its response? Saliency techniques aim to answer this question. These can be based on gradients, which can directly capture the magnitude of change expected in the output distribution as a result of small perturbations to inputs (or intermediate parameters) [42, 43]. Alternatively, one can analyze model *attention* distributions. In small models such analysis has revealed how simple dependencies are processed [44–46], including the discovery of very explicit copying circuits in transformer models [47]. Analyzing per-token model probabilities can reveal



Many users Inference with science payloads — (a) **NDIF**: scientific LLM inference facility

Many apps Simple inference requests — (b) Commercial LLM inference service

One big user Terawatt-hour training job — (c) HPC applied to LLM training

Figure 3: The computational workload of (a) NDIF consists of scientific inference requests which are small (e.g., sub-second) tasks with diverse scientific payloads that serve many different experiments. These differ from (b) commercial inference, which serves apps with no scientific payloads and also very different from (c) training LLMs on traditional HPC clusters, where a single user runs a massive job for months.

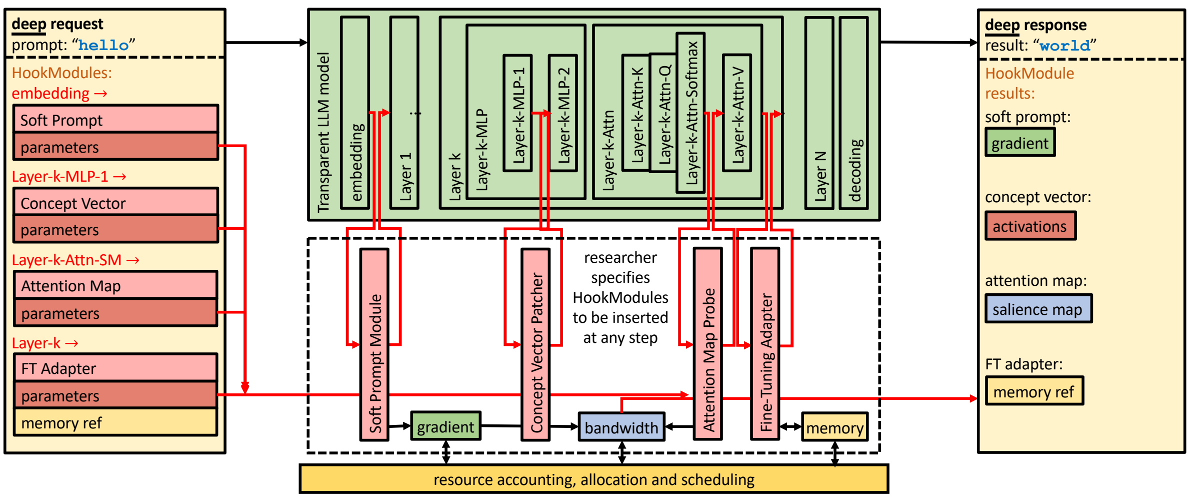


Figure 4: Details of the logical view of a deep inference request. Unlike commercial inference that provides no transparency, with the NDIF, researchers can execute flexible experiments by inserting computations in the internals of the deep network inference process. To maintain safe and efficient co-tenancy, experiment computations are packaged as `HookModules` *that enable resource accounting and scheduling.*

model self-knowledge [48] and differences between human and AI-generated text [49]. Extending these lines of inquiry to large models requires transparent access to model internals.

Causal mediation analysis. Another way emergent learned algorithms can be understood is through measuring the impact of modifying individual computational steps within a model. Such *causal* analysis has been applied to identify attention heads that mediate gender bias in language models [50]; indirect object identification in sentences that name multiple subjects [51]; and the recall of world knowledge within large language models, such as knowledge of the relationships, associations and properties of real-world entities [52, 53]. Using such methods to interrogate larger models requires direct access to internal states.

Parameter-efficient fine-tuning. One of the most compelling properties of large language models is their ability be quickly fine-tuned to a specific task using a small amount of data [54]. The NDIF will support investigation of such capabilities by supporting parameter-efficient fine-tuning methods such as *adapter layers* [55, 56], which are free parameters inserted into the network and then fine-tuned for a specific task while other network parameters remain fixed. NDIF will also enable methods such as "soft prompts" [54], which similarly introduce a small set of tunable parameters, albeit in this case they are viewed as pseudo input token embeddings.

2.1.3 Computational profile of research methods for inference

LLM interpretability methods are diverse and have unique computational requirements with access patterns that are distinct from standard inference needs and traditional high-performance computing workloads (Figure 3); this demands novel infrastructure development.

Deep inference involves small, heterogeneous requests with scientific payloads. The access pattern created by research on inference is characterized by running small inference requests together with transparent access to various aspects of model computations, including activations, gradients, parameters, interventions, and original training data (see Figure 2). Because each experimental method can be decomposed into a small number of types of needed access, the workload lends itself to a design in which users can submit streams of inference requests, each containing a modular specification of scientific payloads to attach to each request (Figure 4a). While the scientific payload has a modular structure, the variety of different ways to apply and combine the modules creates a heterogeneous workload, where many requests may be inexpensive, but others may consume more memory, bandwidth, or computation.

This unique access pattern can also be divided into three major use cases: (1) Simple generative access that includes only lightweight scientific payloads for which it is possible maintain low latency; (2) Batch access, where the scientific payloads may be larger but real-time response is not needed; and (3) Optimization usage, where iterated calculations are done to calculate a statistic or optimize an objective. Support for these successively more complex categories of usage correspond to phases of the software development and deployment of NDIF, as we discuss in Section 4.1.1.

Commercial inference servers support only uniform requests. The workload of NDIF differs from existing non-research-oriented commercial inference services, although they resemble NDIF in some ways (Figure 4b): they are structured with a service API architecture, serving a stream of small requests by running inference on a preloaded model. Like NDIF, these services amortize the cost of loading a model, and they achieve economies of scale by batching requests from many users together. However, unlike NDIF, they have no scientific payloads, so all the requests are similar and small. Commercially available inference APIs include: The OpenAI inference API providing access to OpenAI's GPT-3 and other large models; The Azure inference API, which offers several Microsoft-proprietary models; the Huggingface inference API; and the Cohere inference API, and the Amazon bedrock service. Because **the goal of our proposed service is to support fundamental research**, our inference service must provide additional functionality that is not available commercially, particularly the ability to manage scientific payloads safely and efficiently.

High-Performance Computing (HPC) facilities are suited to pre-training rather than inference. The compute required to support research into LLM *inference* is different from what *pre-training* such models demands. An inference research service must support multiple users making small (often less than one second!), diverse requests to a shared pre-loaded model. In contrast, when pre-training a large model, a single job may run for many weeks or even months, executing the same training procedure in parallel across thousands of GPU devices (Figure 4c). One job from a single user assumes complete ownership over a large model which it modifies on every iteration. Therefore, the compute infrastructure for pre-training need only support a small number of users accessing many nodes uniformly; this is a use-case well-served by existing HPC facilities.

In contrast, deep inference workloads demand fine-grained flexibility, including the ability to accept and respond to a stream of very small requests from research users accessing the same models, and the ability to probe, inspect, and modify details of the pre-trained model to support the range of experimental methods as discussed above in Section 2.1.2. Unlike other machine learning infrastructure efforts, the NDIF is singularly focused on the infrastructure needed for *efficiently running* large-scale models to enable research on them *after* they are trained. The NDIF also provides open-source tools that will allow large-scale training clusters to support deep inference research if they choose to schedule inference service jobs rather than large-model training workloads.

2.2 NDIF will leverage open large language models

As described in the prior section, existing high-performance computing facilities have enabled researchers to *train* large language models, but are not suitable for performing inference on them. Efforts to train open LLMs are complementary to NDIF, as NDIF focuses on addressing barriers to research at the *inference* stage — on those open LLMs. We have already begun efforts to integrate **EleutherAI**'s 20-billion parameter GPT-NeoX [31] and 6-billion parameter GPT-J [57] (see attached letter of collaboration from Stella Biderman, Executive Director of EleutherAI). We are also following EleutherAI's efforts to train an even larger, 150-200-billion parameter LLM. Other related efforts that we will engage with include: (1) **BigScience Bloom** [10], a 176-billion parameter multilingual model trained by BigScience, a collaboration of European agencies, the Huggingface company, and many others. (2) **Meta OPT** [11] and Llama [58], sets of models based on commercially licensed language models trained by Meta, with parameters that are made available to academic researchers. The OPT family includes a 175-billion parameter model and the largest Llama variant has 65 billion parameters. (3) **Tsinghua GLM** is a 130-billion-parameter Chinese-English model supported by Zhipu.AI. (4) Variants of these models are fine-tuned with human feedback, including BigScience Bloomz [59], CarperAI [60] and OpenFlamingo [61]. (5) Ongoing work by the National AI Research Resource (NAIRR) [12], Large-Scale Artificial Intelligence Open Network (LAION) [62], and Together Computer [63]. This is a (very) fast-moving area, and we anticipate many additional publicly available LLMs to be available within the coming years; our configuration committee and scientific advisory board will work with the community to identify new models to add to NDIF to maximize scientific impact.

2.3 Our research user community and illustrative application areas

Support for the NDIF project is strong in the US research community. Beginning with outreach via Twitter in December 2022, over 400 researchers indicated that their research goals were blocked due to the present impracticality of deep inference on LLMs. Many emphasized the strong need for infrastructure given the practical difficulties of investigating models whose parameters do not fit into the memory of typical research computing nodes. Professor Boaz Barak (Harvard) observed, "Any model that doesn't fit on one GPU starts to be complicated for researchers to use even if they do have enough GPUs to fit... A central engineering resource that all academics can share would be a game changer." Professor Tom Dietterich (Oregon State) said, "I strongly support a public National Deep Inference service.... We will want to support many different things: fine tuning, access to the training data, access to external resources." Professor Ana Marasović (University of Utah) noted, "Having academic access ... would enable not only machine learning academics, but also academics without expertise in training models, to study large language models."

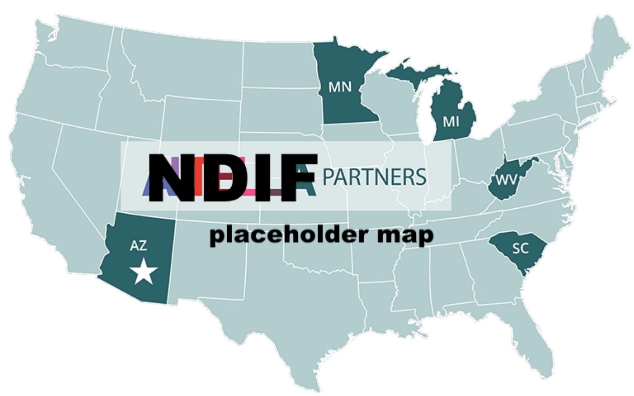


Figure 5: National reach: Our community includes faculty at 34 universities in 19 states (6 EPSCor states) who have proposed specific research to be done with NDIF.

Based on this strong national interest expressed via Twitter, we have established an initial virtual research community that includes 40 professors from 34 different universities in 19 states including 6 EPSCoR states, (see Figure 5) and 8 minority-serving institutions (including an HBCU) who have suggested specific research projects that will benefit from NDIF. These projects span a broad range of topic areas. Here we provide illustrative, concrete examples of how a diverse set of academic researchers will directly benefit from the NDIF infrastructure.

Computational linguists and natural language processing (NLP) researchers are keen to understand the linguistic knowledge implicitly encoded in LLMs [64–66]. This sub-community is also interested in the degree to which such models encode bias [67, 68], especially in sensitive domains such as healthcare [40, 69]. There is also a quickly growing sub-community investigating the mechanistic inner-workings of LLMs, perhaps best represented by the BlackboxNLP workshop [70], which has been held annually at ACL conferences for the past four years.

Computational linguistics and NLP researchers form a core constituency of enthusiastic would-be NDIF users, and representatives of this group have been actively taking part in design discussions accordingly. However, we stress that the NDIF will benefit researchers from a diverse set of academic communities, especially as LLMs begin to impact these fields—a few examples follow.

Social and political scientists we have spoken with are interested in the use of LLMs as tools for judging public opinion in ways that would be impractical to scale using other means [71], as well as using them to measure political ideology and other latent constructs from texts [72], and applying LLMs to various "text-as-data" tasks to permit subsequent analysis [73]. While some of these research settings may, on the surface, seem suitable for black-box inference services (e.g. APIs provided by OpenAPI), researchers noted that in practice, the lack of reproducibility makes it impossible to investigate critical questions, such as to understand implicit biases of such models.

Neuroscientists would like to use the NDIF to analyze artificial neural networks using tools from neuroscience [74, 75]. This research will have implications for both AI and the brain. LLMs have the advantage of being decidedly easier to open, inspect, probe, and manipulate than brains, but only if infrastructure like the NDIF exists to permit such experimentation.

Researchers in healthcare are interested in a range of applications of LLMs [76]. For example, community members we have spoken to are investigating the use of LLMs to detect dementia from patient-elicited speech [77]. Other researchers we have spoken to are interested in studying how (health-related) domain knowledge is stored in LLMs, which will require deep inference. And then there are the critical issues related to learned representations and fairness [40] as well as risks of training LLMs on potentially sensitive personal health data [78]. Investigating such issues requires access to model activations and parameters.

These are only a small and illustrative subset of the potential use cases across fields that we have discussed with researchers in this formative period. As LLM technology continues to impact new fields, so too will the need for rigorous LLM science which is not, in general, possible to conduct using commercial inference APIs.

3 Preliminary Activities

We have accomplished several preliminary tasks including recruiting prospective leadership staff, forming an open design process, outreach to relevant research community members, identification of goals, and preliminary development of technical plans.

3.1 Identification of goals together with the user community

As a facility intended to serve the needs of the wider community of researchers interested in LLMs, it is essential for the development team to have an intimate understanding of the likely needs of users of the facility, who will be pursuing a diverse range of research programs. Toward this end, have established on online Discord server devoted to design discussions, currently involving 43 active researchers from universities across the nation pursuing LLM research across various fields including computer science, neuroscience, political science, and biomedical sciences. This forum is open to all prospective users of the NDIF and will serve as a direct conduit between our development team and the wider LLM research user community. We use the forum to conduct regular design discussions, enabling us to set priorities, gather detailed requirements, and validate development plans. Working directly with our user community, we have identified three core goals that will drive the detailed design of NDIF: (1) The facility will **enable research into the most capable state-of-the-art open large language models** as well as large multimodal models after they are trained. (2) It will provide **full transparency and reproducibility**, including access to model internals such as activations, weights, overrides, gradients, and the ability to control random seeds. (3) It will **prioritize community support**, with a focus on enabling academic researchers studying the mechanisms and impact of large language models in practice. Our user community has made significant progress towards identifying capabilities that should be enabled by the facility after full deployment. Specifically, we have identified a list of experimental methods applied to neural systems that are a priority for our community (Figure 2). Furthermore, the community has begun to characterize key unknowns and available resources and has started the process of prioritizing detailed research capabilities to enable in the first phase of development.

In addition to reaching out to and connecting wtih potential NDIF users online, we have hosted an in-person outreach event at the International Conference on Learning Representations (ICLR) 2023. Our event brought together over 100 members of the machine learning research community to discuss challenges faced by researchers studying LLMs on academic budgets. During the event we identified a range of research priorities that will need to be addressed by the NDIF, and we recruited researchers to participate directly in our user community.

3.2 Leadership recruitment

We acknowledge the challenge of attracting highly-qualified professionals in this field, and we are fortunate to have recruited two exceptional individuals who could be available to join our leadership team pending the funding of the project. William Brockman, PhD in Mathematics from UC San Diego, has led software development projects at top organizations such as Google, the Broad Institute, and General Dynamics. Brockman has been participating in our design process and helping to develop our technical specifications. If the NDIF is funded, could be available to serve as our lead software engineer. Sumeet Multani, PMP, holds a Master's Degree in Computer Systems Networking and Telecommunications from Northeastern University, and has served as a technical program manager at Google, TripAdvisor, and Akamai Technology. If the project is funded, Multani could be available to serve as our project manager. Both Brockman and Multani are based in Boston. Their roles in NDIF are described in Section 4.2.1.

3.3 Prototype and specification development

Working with the community, we have developed several small-scale prototypes that implement aspects of the NDIF service model. These include a software package used for instrumenting single-GPU neural networks that we have validated and used for several published research works, as well as a prototype web service to run research-oriented inference on a multi-GPU language model at a scale suitable for use by a single lab. We have used what we have learned from these prototypes to inform an architectural specification for NDIF. An overview of this preliminary specification is given in Section 4.1, and the specification is included in supplementary materials.

4 Implementation

4.1 Technical Readiness

Our team has developed a detailed technical specification and development plan that delineates requirements, design, and deployment milestones for the NDIF's user model, user-facing and internal software, and hardware.

4.1.1 Major Deployment Milestones

Development and deployment of NDIF will proceed in several phases, each one increasing NDIF capabilities, robustness, usability, user support, outreach, and user base. The plan is designed to deliver value to researchers as early as possible while maximizing opportunities to respond to user feedback and outside events.

Pre-funding pilot, Summer 2023. Develop single-server cluster that can serve user requests and streaming interactions on medium-sized models. Preliminary client library with 5 local users.

Closed pilot, Q1 2024 (Year 1). First phase hardware (see Section 4.1.2) serving sustained research queries observing and modifying the largest models of interest, with both streaming and batch-oriented use patterns. Documentation sufficient for 20 selected early adopters drawn from our design-participant user community, working directly with our team and supported by our engineers and researchers.

Open pilot, Q1 2025 (Year 2). Second phase hardware deployment enables opening early access to qualified users at any educational institution, with limited support from NDIF team. Increased robustness, including monitoring and alerting, improved job queuing, and a fairness-oriented scheduler. Define Service Level Objectives (SLO)s and measure progress towards meeting them. Preliminary optimization and gradient functionality. Documentation is complete enough for early adopters; draft tutorials are prepared.

Software API full release, Q1 2026 (Year 3). Robust support for optimization and gradient methods. Initial support for user-defined aggregation on-cluster. All major user-facing features of the system meeting SLOs. Documentation is complete and undergoes user testing and improvement. We will teach 100+ researchers how to use the system via a multi-site bootcamp.

Operations scale-up, Q2 2027 (Year 4). Refresh hardware to support new models, larger models, more models, and more users. Continued refinement of ability to onboard new users. Robust user-defined on-cluster computation. Refine system administration tools to improve issue response and stability. Pilot ability to run NDIF on other clusters and to route traffic to other HPC clusters.

Cluster self-hosting, Q2 2028 (Year 5). Administrative tooling is complete and robust enough to support distributing NDIF software to other HPC clusters. Hire permanent director, release major code version, and prepare for sustained operations.

4.1.2 Hardware Design and Scale-Out

NDIF will consist of a high-density cluster of GPUs, along with an open-source software platform to enable the efficient utilization of that hardware for deep inference. Present technology allows for a maximum of eight A100 GPUs to be located in the same physical sever (enough to support current LLMs, of the size of GPT-3), but we anticipate that over the span of this five-year project, GPU density and performance will increase. Hence, we will phase the hardware roll-out of NDIF, so that we can begin constructing the software and supporting present-day LLMs immediately, and continuously improve the hardware resource as vendors release newer, higher density hardware. Our budget estimates the cost of higher-density nodes by including quotes for larger sets of lower-density nodes to reach the same amount of GPU vRAM. The software stack (Section 4.1.3) will support parallelizing workloads across multiple nodes if necessary, and will support a heterogeneous cluster where different nodes may have different capabilities.

175-Billion Parameter Capacity, Q4 2023 (Year 1). The first phase is sized to match the current state-of-the-art: there are two open-parameter models at the 175-billion parameter size (similar to GPT-3), and we anticipate one more soon. Thus we plan 10 nodes, each with 640 GB of VRAM via 8x Nvidia 80GB A100 GPUs. This phase suffices to run three different models of this size, with three inference servers each, and one spare to reduce downtime.

500-Billion Parameter Capacity, Q4 2024 (Year 2). The second phase adds 6 nodes, each containing 1.2 TB VRAM, for example, through 16x Nvidia 80GB A100 GPUs. This will provide enough capacity to serve one 500-billion parameter model (five inference servers, with one spare node to reduce downtime). Currently the only models at this scale are proprietary, but we anticipate the availability of open models in this timeframe.

Trillion-Parameter Capacity, Q4 2026 (Year 4). The third phase adds 4 nodes, each with 2.5 TB of VRAM, e.g., through 32x Nvidia 80GB A100 GPUs. This will provide enough capacity to serve a 1000-billion parameter model (three inference servers plus one spare node), matching the goals of the NAIRR [12] public AI training resource.

4.1.3 Software Stack and Service Architecture.

The software layer is equally important to the success of NDIF, in two distinct ways. It needs to promote efficient utilization of the hardware, but it also needs to provide smooth on-ramps and highly productive steady-state usage patterns for new and experienced researchers. For transparency and reuse, all our software will be open-source and developed in public repositories with the active engagement of the user community and open-source contributors. The system architecture will rest on widely-adopted open-source platforms to enable its use in a variety of contexts, for instance to handle spikes in demand. We will test deployment in at least one commercial cloud provider [63, 79–81] and/or nonprofit cluster [62, 82] and the future NAIRR [12].

Inference backend. The workhorse of NDIF is the single-node multi-gpu inference backend, which aggregates the stream of incoming inference requests into batches and executes the instrumented models, including all scientific payloads. It must track the association between individual experiments and batched data, orchestrate data flow including pipelining, manage the calculation of gradients, and track resources used. The inference server will be built on the open-source Nvidia Triton [83] inference server framework, with a new backend for LLM scientific payloads.

Request router and scheduler. As inference requests arrive, the router is responsible for queueing, ordering, and routing those requests based on availability and prioritization. In the initial pilot, a naive (FIFO) scheduling algorithm is implemented, but in subsequent milestones, an adaptive scheduler will be developed to sort and group different classes of usage to improve utilization, and to maintain fair resource allocation across the cluster.

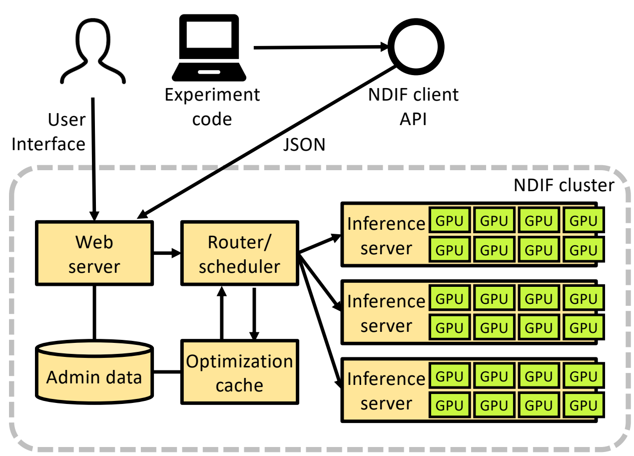


Figure 6: NDIF service architecture, showing request flows between system components.

Optimization cache manager. To reduce the bandwidth consumed by stateless operation of common operations, NDIF will support data caching on each node. The optimization cache manager will manage all temporary storage and caching of user data, including, managing queues and cached intermediate results. The cache manager tracks cached data that may be present on any node, and it is able to orchestrate movement of data between nodes when needed. The optimization cache is essential for speeding up operations like gradient descent.

Administrative tooling. Administrators and engineers will develop a set of tools for maintaining a robust facility, inlucing logging and monitoring software, and tools and scripts for administrative tasks, including user administration and moderation, quota management, cluster model allocation,

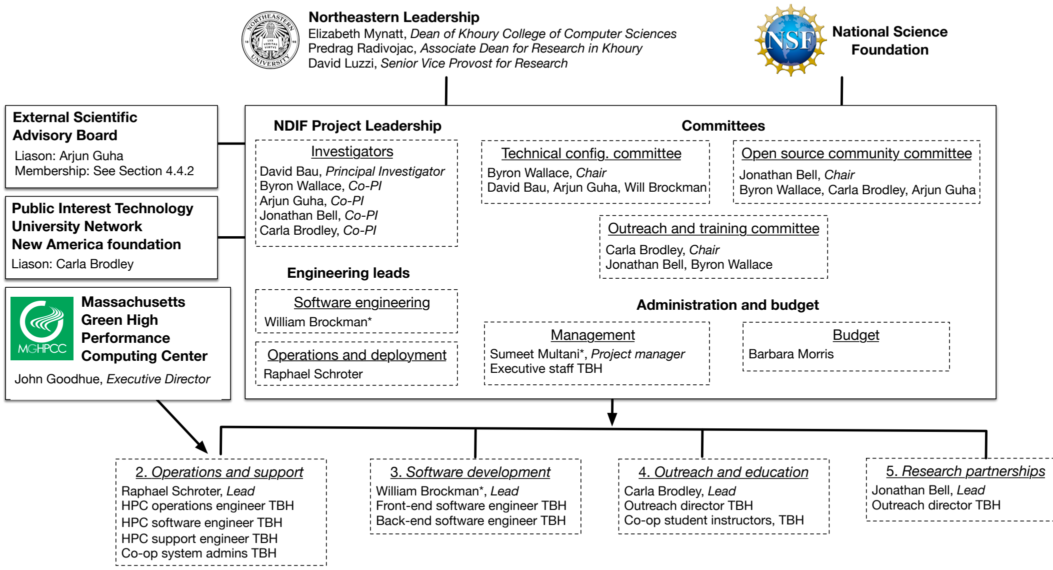


Figure 7: NDIF organization. Asterisks (*) indicate tentative hires.

health monitoring, load monitoring, and debugging tools for inspecting activity on the cluster.

User interface frontend. This webserver forms the boundary to the user-facing aspects of NDIF. It includes the end-user visible views of the system, including signup, login, experiment console pages, as well as an interactive interface for directly conducting experiments with a model. It will support an HTTPs JSON API for submitting inference experiment requests and receiving results, to enable the community to integrate other systems using any language or framework.

Client-side Python API. The primary way researchers will conduct experiments will be through an open-source python library built on the `PyTorch` [84] deep learning framework that runs on the user's workstation. This library will provide a modular way to conduct LLM experiments, as shown in Figure 4, while supporting remote inference on NDIF models. The design priority is to provide a practical and accessible "on-ramp" for researchers to do research on LLMs.

Experiment methods library. Built on top of the core python API, we will provide modules that implement higher-level algorithms, interactions, analyses, and visualizations to implement the important experimental methods for various lines of LLM research.

4.2 Planned Project Management

4.2.1 Key Personnel

The organization chart is in Figure 7. This project brings together an interdisciplinary group with deep expertise in ML/NLP, programming languages, software engineering, and large-scale computing, as well as experience in development and operation of large-scale computing systems and the creation and administration of multi-institution research programs.

NDIF Director and PI Bau (Assistant Professor in Khoury College of Computer Sciences at Northeastern) brings a unique skillset as a late-career academic who worked in industry for 20+ years, 12 of them at Google at which he created and managed the Google Talk and Hangouts team and led the Boston Google Image Search ranking team. He has a track record managing projects to develop large-scale online platforms with global reach and real-world impact, processing exabytes of data and answering billions of user queries each day. Since transitioning to academia, PI Bau has established himself as a leading researcher in interpretability of large neural networks [85–88] and editing of large models [89–91], and he has been a pioneer in the explicit characterization of causal computational mechanisms within language models [52, 53, 92]. Bau will serve as Principal

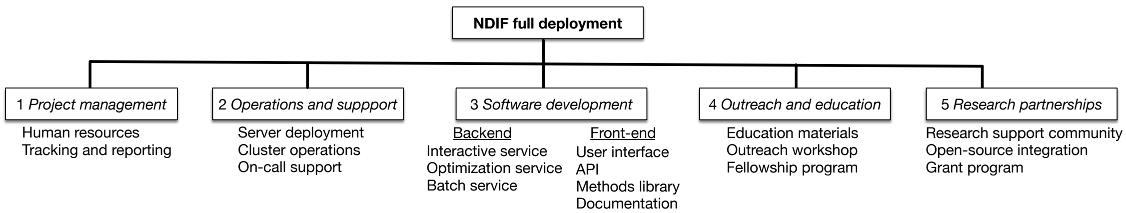


Figure 8: The work breakdown structure of NDIF construction.

Investigator of NDIF. He will hire and manage project leadership, as well as oversee the direction, development, and overall success of the facility. Bau will work closely with the project manager and lead software engineer to oversee the development of the project, and he will run monthly meetings of the NDIF leadership team. He will also serve on the technical configuration committee.

Outreach Lead and Co-PI Brodley (Dean of Inclusive Computing at Northeastern and Founding Executive Director of the Center for Inclusive Computing at Northeastern University; former Dean of Khoury College) is a fellow of the Association for Computing Machinery (ACM), the Association for the Advancement of Artificial Intelligence (AAAI) and the American Association for the Advancement of Science (AAAS). Her interdisciplinary machine learning research has advanced CS as well as remote sensing, neuroscience, digital libraries, astrophysics, image retrieval, computational biology, chemistry, and evidence-based medicine. Brodley will lead outreach efforts (WBS 1.4) and ensure broad participation in setting research priorities and the educational mission of NDIF. She will chair the outreach and training committee. She will serve as liaison to the public interest technology university network, and will serve on the open-source community committee.

Technical Configuration Committee Chair and Co-PI Wallace (Sy and Laurie Sternberg Interdisciplinary Associate Professor in Khoury) has extensive research expertise in NLP and interpretability of such models [93–100], as well as the use of NLP in biomedical settings [78, 101–104]. Wallace will chair the technical configuration committee, which will meet quarterly (and more frequently as needed) to review the design of the service and ensure that its design meets research aims. Wallace will be responsible for establishing academic priorities for the facility, and for conducting outreach to the Natural Language Processing and Biomedical research communities. He will serve on the open-source community committee and the outreach and training committee.

External Advisory Board Liaison and Co-PI Guha (Associate Professor in Khoury) brings deep experience in programming languages, including language-based security [105–108], GPU accelerated domain specific languages [109, 110], and pre-trained models for code generation [111]. Guha will serve as liaison for the External Scientific Advisory board and will be responsible for engaging and recruiting leaders of the academic community to the board to ensure that long-term needs of the academic community are met. Guha will also serve on the technical configuration committee and the open-source community committee.

Open-Source Chair and Co-PI Bell (Assistant Professor in Khoury) is an expert in software engineering and systems, including architectural design [112], testing and continuous integration [113–116], and analysis [117–119]. Bell will chair the open-source community committee. He will be responsible for incubating the open-source community and overseeing open-source activities, and providing academic oversight on open source policies, technical contributions, and quality assurance. He will also serve on the outreach and training committee.

HPC Operations Lead Schröter (Director of Research Computing at Northeastern University) organizes strategic planning for research computing resources at Northeastern; he works with university researchers across all disciplines, to achieve research goals using HPC infrastructure.

Schröter will manage deployment and operations for NDIF (WBS 1.2) and supervise the staff of HPC engineers. He will be responsible for managing the physical colocation of the facility, as well as day-to-day operations of the computing service.

Project Manager Multani (tentative) has extensive experience leading the definition, planning, and execution of both user-facing and infrastructure projects and has served as a technical program manager Google, TripAdvisor, and Akamai Technology. As project manager, Multani will work closely with the PI, as well as all fourork areas of the project (WBS 1.1-1.5). He will be responsible for schedule management, budget management, scope management, and risk management. Additionally he will be responsible for all NSF reporting, including monthly reports, quarterly reports, annual reports, and periodic PEP updates.

Lead Software Engineer Brockman (tentative) has a wealth of experience leading projects in high performance computing, data science, and mathematical modeling, and has led software development projects at Google, the Broad Institute, and General Dynamics. As lead engineer, he will be responsible for hiring staff engineers, and for managing the development process (WBS 1.3). Brockman will serve on the technical configuration committee.

4.2.2 External Scientific Advisory Board

We will establish an External Scientific Advisory Board to provide input into key aspects of the project. The board will consist of 5-10 members, each of whom will be a subject area expert or a representative of a relevant constituency such as a university administrator. Several prominent members of our academic user community have offered to serve as initial board members.

4.2.3 Engaging the Public-Interest Technology Research Community

The benefits of advances in AI have been realized unequally [120]. To ensure critical assessment of the potential impact of LLMs on education, policy, privacy, and safety we will engage with academics who work in the public-interest technology sector. We will work with the New America Foundation's Public Interest Technology University Network (PIT-UN) to bring both AI and non-AI faculty to workshops with AI researchers/students to discuss issues of interest and promote the public good (see attached letter of collaboration). PIT-UN has a membership of 63 universities and colleges, 19 of which are Minority Serving Institutions (MSIs). PIT-UN will support the National Deep Inference Facility by establishing a Public Interest Technology (PIT) Advisory Group to provide guidance on the responsible and ethical design, development, deployment, and use of LLMs with an interdisciplinary approach that includes experts from both technical and social sciences.

PIT-UN will establish a PIT Advisory Group comprised of 10-15 interdisciplinary experts from PIT-UN to provide guidance on the responsible and ethical design, development, deployment, and use of LLMs. PIT-UN seeks to align PIT with both informal and formal STEM learning through the capacity development of its 63 member universities. By undertaking the LLM project, PIT-UN aims to meet the following goals: (1) Engage conversations regarding LLMs to be community–driven; (2) Promote equitable and broad participation in the emerging field of AI through LLMs; (3) Advance the knowledge base of LLM learning by advancing PIT with formal reports out from New America reflecting feedback from semi-annual roundtables of its Advisory Group; (4) Develop formal learning experiences and environments through strategic activities in collaboration with other New America teams such as Open Technology Institute and the Ranking Digital Rights program as needed to support and represent possible frameworks; (5) Develop professional capacity within member universities themselves to deliver informal AI learning using the LLMs within a PIT framework; (6) Host an annual webinar to distill key findings and build a base of new AI learners through exposure to PIT and its applications in relation to LLMs.

4.2.4 Scope Control

The technical configuration committee and the PIs will define the experimental capabilities that will be enabled by NDIF during each phase of deployment. These choices will be made in consultation with the external Scientific Advisory board and the open source community. After each phase is released for usage in phase 2 and beyond, agile project management methods will be adopted to continuously test the product to identify and solve problems, to iteratively improve the software and infrastructure. We will monitor customer-reported issues, cluster efficiency, and open-source contributions. The technical configuration committee will conduct an annual review to identify changes in scope and determine where corrections are needed.

4.2.5 Budget and Budget Contingency

We begin with a baseline budget and budget justification included with this proposal. Throughout each phase of the project, the project manager will update the budget and provide NSF with updated cost estimate for both capital and soft costs. We set $900,000 (5% of total direct cost) as the budget contingency, which covers any extra costs, including risk, increases in scope, and unknown tasks. This money is not allocated to any area of work and will only be used as needed.

4.2.6 Schedule and Schedule Contigency

The project will run from 9/1/2024 to 9/1/2029; please find the schedule in the PEP. In each year of the project we schedule a single major deployment release to increase the experimental capabilties of NDIF. After each of these releases, we set aside two months schedule contingency. If the schedule is followed, we will use these two months to collect customer feedback and focus on design review for the next phase. If not, the contingency allows time for unanticipated integration, performance tuning, quality assurance, or adjustments in scope. In that case, the project manager will conduct a program review to adjust scope, timing and budget of the project.

4.2.7 Risk Management

Our project will use the following process for managing risks:

Risk Identification. We will identify project risk through structured brainstorming sessions with stakeholders through the duration of the project, utilizing SWOT analysis, cause and effect diagramming, assumptions analysis, and risk breakdown structures. The project manager will conduct sessions focused on risk categories, e.g., scope risks, financial risks, and quality risks.

Risk Analysis. The project manager will analyze risks identified through structured brainstorming sessions using a variety of qualitative and quantitative methods, including SWIFT analysis, interviewing experts, analysis of expected monetary value, and sensitivity analysis. This analysis will be used to establish the appropriate risk tracking and control procedures.

Risk Tracking, Control, and Monitoring. The risk-mitigation process will guide us in selecting appropriate mitigation strategies for each risk, given the value impact and probability of each option. Possible mitigation include risk avoidance, risk transfer, risk reduction, and risk acceptance. Risk will be tracked over time, and the effectiveness of the risk-management process will be tracked.

Project leadership (the PIs, project management, operations lead, and lead software engineer) will conduct a comprehensive review of risks annually. As the project proceeds, each risk will remain on the risk register until it is closed. Additionally, the team will review each risk to develop mitigation strategies. We have conducted an initial risk assessment (refer to the PEP for the risk register). Some of the major risks include hardware failure, user adoption risk, and software technical performance risk. Our project plan mitigates these risks where possible.

4.2.8 Configuration Management

Changes for all project specifications other than software, such as specification of required infrastructure capabilities, hardware specifications, or changes in policies or legal agreements, will be managed through a formal change control process. Staff and leadership will propose changes with input from external stakeholders, and the changes will be reviewed by the Technical Configuration Committee. Updates to specifications will be communicated at the required time, for example during contract renewal. Change control for software will be narrowly controlled by the software engineering team. They will utilize software version control through git, and all changes will go through code review. Unit testing will be conducted before changes are committed, and integration testing will be conducted before any changes are deployed to customer-facing services. All deployments will be given a release number, and a change log file will be maintained.

5 Operations and Utilization

After the successful deployment of NDIF, the organization will transition to ongoing operations. The primary focus will be to maintain the high performance computing infrastructure and ensure its continued availability to the research community. To achieve this, several key personnel and operational changes will be implemented.

5.1 Operations Management and Governance

In the final phase of the grant we will hire a full-time facility director to oversee scientific operations and to ensure that NDIF continues to provide cutting-edge computational resources to researchers. NDIF will also maintain an external advisory board consisting of members of the scientific community and policy community from outside organizations; the board will advise on allocation priorities, ethical issues, and strategic direction of the facility. Operating staff will include an outreach director who will continue to update educational materials and run workshops, and engage with the research community to ensure that NDIF remains accessible and user-friendly. It will also include software engineers, who will maintain software, respond to open-source contributions, and implement updates to keep up with the latest science. NDIF will fully staff system administration and operations. As a mature application, NDIF application-level system administration staff will report into Northeastern Research Computing alongside HPC operations at MGHPCC. The HPC staff will maintain a high level of service for NDIF, updating software and making hardware repairs, and responding to operational issues.

5.2 Operating Costs and Funding Sources

The annual operating cost of NDIF is estimated to be about $2 million, which includes personnel, the cost of conducting outreach, and the cost of maintaining the hardware at the deployed level of computation. This estimate does not include any investments in increases in capacity. Our plan is to ask the NSF to cover the annual operating costs of NDIF, while defraying some costs through a research partnership program; researchers can contribute funds to the facility in exchange for access to priority scheduling queues. This program will help to ensure the long-term sustainability of NDIF.

5.2.1 Access and Utilization Plan

The NDIF will be open to all people with an educational affiliation to use, free-of-charge (using the NSF and DOE-supported "CILogon" Service for authentication), after agreeing to a service agreement and submitting a brief statement of intended use. To allocate scarce resources when oversubscribed, we will implement an online adaptive scheduling algorithm that estimates and monitors heterogeneous resource use in order to fairly distribute computation, bandwidth, and memory. Based on our estimates of computing capacity of a state-of-the-art software implementation using our hardware configuration, we estimate a user will be able to get 10 tokens-per-second latency under light load (with 30 simultaneous users per node, when scientific payloads do not require smaller batch sizes). When usage is heavier or when users are placing sustained scientific load on the service, latency will naturally rise as capacity saturates, and heavy users will have their requests queued and throttled so that overall capacity is distributed equitably.

Paid Partnerships: the NDIF will also offer a paid "NDIF Partnership" program to allow researchers to subsidize capacity that they can allocate for sustained high-bandwidth usage for their research. For example, partnership fees can be paid for by researchers' grants, and they will give partners access to their own allocated queue where they can be assured of a level of throughput that is independent of baseline load on the public scheduling queue.

Need-Based Resource Grants: to ensure the broadest possible access to researchers who may have high computing needs without a source of sufficient funds enroll as a paid partner, NDIF will also award "NDIF computing grants" that will provide free-of-cost access to allocated high-bandwidth queues. We will broadly advertise these computing grants, particularly to new PIs, to PIs in EPSCOR states and to PIs at minority serving institutions. A committee appointed by the NDIF leadership will review and choose grantees based on need and scientific merit.

5.3 Evaluation

We will continuously evaluate the project, both at the component-level (e.g. latency of individual APIs), and at the full facility-level. Our project is driven by four measurable goals: (1) **Advance scientific understanding** of large language models. (2) Provide **broad access** to researchers and students for inference not served elsewhere. (3) Enable **efficient use** of scarce computational resources. (4) **Train students** on LLMs, to build the next generation of AI engineers and researchers.

These goals correspond to metrics that we will track. To measure our progress in realizing **impact** by providing **broad access** and **efficiency**, we will track and aim to increase: (1) **Sustained server utilization** in the deployed service, a core measure of efficiency. Our aim is to size the facility to maintain an overall utilization of 50% or more. (2) **Experiment response latency** which quantifies the technical accessibility of the facility to researchers. The goal will be for latency to be low enough to enable interactive human studies with real-time interaactions with large models. (3) **Number of monthly academic users** of the deployed service, a core measure of reach, along with metrics of the diversity of those users. (4) **Number of peer-reviewed research works** that use our service or software in experiments. (5) **The number of deployments** of our software stack on clusters beyond the initial service. Other operating metrics will be developed by the team as part of the service development process. These metrics will be tracked by the project continuously through dashboards and reviewed by the director and the advisory board on a semi-annual basis.

6 Broader Impacts

Understanding the impact of AI across society: Large language models (LLMs) are already being rapidly integrated into consumer products, and are impacting fields outside of computer science from mathematics to medicine, and their impact on society is likely to continue to grow. However: LLMs have advanced so rapidly that civic leaders have called for a temporary pause on LLM development until academic research can catch-up [121]. NDIF will provide the hardware, software, and training necessary for researchers to better characterize the possibilities and potential of LLMs. For example: our collaborators in psychology plan to analyze AI using tools from neuroscience, and collaborators in linguistics will analyze how aspects of knowledge are captured by LLMs. Without this vital research, it will be difficult-to-impossible for policymakers to design regulations to ensure that state-of-the-art AI systems are safe, transparent and robust.

Building national research capacity in AI: This project will have a transformational impact on the U.S. workforce, by training an intellectually diverse group of scholars to understand the potential and mitigate the harms of powerful new AI capabilities. Our outreach and education programs will help train the next generation of researchers to ask and answer critical questions about the capabilities and limitations of large language models, and their impact on society.

Democratic and equitable access to NDIF: Section 6.1 describes our outreach, training and support plan, which will ensure democratized access to NDIF. Our outreach plan is structured to build upon our established partnerships with Northeastern's Center for Inclusive Computing (led by co-PI Brodley), supplemented by collaborations with the Computing Research Association (see attached letter of collaboration from Tracy Camp). Need-based resource grants will ensure that access to the NDIF is not simply prioritized to the largest research institutions. Our training and support plan will build a scalable network of experts across the country that can further promote the NDIF and help us understand the local needs of the different sites that we serve.

Workforce development: This project will directly contribute to the training of undergraduate, masters, and doctoral students who will be engaged in the development, operations, and evaluation of the NDIF. Building on our experiences designing project-based software engineering education, we will create course projects that engage students in NDIF development. We will make a special effort to engage students in Northeastern's "Align" masters program, which provides a direct pathway into computing for students without a CS background. We will deploy our training curricula throughout our institution's network of campuses in Virgina, North Carolina, Florida, Maine, Massachusetts, California, and Washington. Northeastern is well-known for its experiential learning — every student completes at least one six-month full-time Co-Op — we have had success in the past recruiting students to develop software, and will continue this approach for this project.

6.1 Strengthening National AI: Outreach, Training, Virtual Community and Support

The NDIF will strengthen the US Artificial Intelligence Research & Development ecosystem. As such, we are committed to ensuring that we provide training and support to the US scientific community to ensure that the infrastructure is accessible and usable. Beyond offering "open" access to NDFI, our goal is to provide democratized and equitable access to the facility by addressing knowledge, technical, and social barriers that could limit adoption. Core to our outreach plan is an effort to build a scalable network of experts who can respond to local needs.

Democratic and Equitable Access It is critical that we ensure that the NDIF does not further widen the gap between AI researchers from majority groups and those from groups historically marginalized in tech [122, 123]. Thus throughout all outreach we will ensure that we are reaching a diverse set of institutions, researchers and students, with a focus on reaching new PIs, and PIs in EPSCOR states and at minority serving institutions. Co-PI Brodley, who is a nationally recognized expert in broadening participation in computing [124], will lead this effort. We will recruit potential users in several ways: using popular social media channels such as twitter, through the CRA (see attached letter from CRA Exec Director Tracy Camp), by running workshops at AI/ML conferences, and by utilizing the deep network of 100+ (R1 and non-R1) institutions that participate in initiatives run by the CIC (led by co-PI Brodley).

Developing National Expertise We will design training modules to help onboard new PIs and students to the NDIF. Modules will cover topics such as: 1. How to perform reproducible inference experiments on the NDIF. 2. How to apply deep inference methodologies such as representation probing, attention mapping, causal mediation analysis and parameter-efficient fine tuning. How to perform these experiments on NDIF. 3. How to deploy NDIF on your own GPU infrastructure. In the second year we will pilot an intensive in-person "bootcamp" in Boston, which will provide graduate students studying in the U.S. hands-on access to the experts who build and maintain NDIF. In year three we will expand this bootcamp to six different geographic regions, leveraging Northeastern's campus network (Northeastern offers programs at nine global campuses) and two universities partners, with a focus on cities with a major airport hub; we will run five bootcamps during the summer of 2026 in Oakland, Miami, Washington DC, Dallas and Chicago. The cost of attending the bootcamp will be free (leveraging our campus network) and will be led primarily by

Northeastern PhD students with co-PI Bell and co-PI Gupta in attendance. For graduate students whose advisors do not have budget to cover the travel costs we have budgeted a $50k fund to support travel based on need. Additionally we will offer one-day workshops using the in-person tutorials co-organized with major machine learning conferences, such as NeurIPS, ICML, ICLR, ACL, EMNLP, AAAI. We will select two conferences per year with the goal of maximizing the diversity of locations in the U.S. The students who participate in the bootcamps and tutorials will become part of a network of experts, providing embedded expertise within their own institutions, and helping us to provide support that is responsive to local needs across the nation.

Nurturing a Virtual Community Training and mentoring will continue virtually beyond in-person events. The goal of the virtual community is to provide space for researchers to learn more about the NDIF, and to showcase and discuss ongoing research on the NDIF. In preparation for this proposal, we created the NDIF virtual community using the *Discord* platform. This platform enables real-time and asynchronous text communication organized by channel and thread, and provides audio and video chat as well. In its first month of operations, this platform brought together 43 researchers from across the country to discuss the design and use-cases for NDIF. We will organize an annual virtual conference, providing students and researchers with a space to showcase their ongoing work and to have "ask me anything" interactions with the project team. We will also maintain a website with reference materials, tutorials, and examples, as well as an open-source codebase on github and we will use the public issues database as a conduit for gathering and tracking user issues. We will integrate this virtual community with our in-person training, with the goal of broadening the availability of NDIF and limiting potential barriers to its adoption.

6.2 Undergraduate Education

We are committed to ensuring that undergraduates in the U.S. benefit from the proposed infrastructure. To this end we will develop materials—lectures, exercises, and assignments—that cover analysis of large language models. We will pilot and refine these materials in relevant courses at Northeastern (e.g., Machine Learning I and II, NLP, and Practical Neural Networks). Bau and Wallace regularly lead these offerings. Further, Wallace is Director of the Bachelors in Data Science program (and serves on the undergraduate curriculum committee), so is well-positioned to ensure that developed materials are incorporated into course curricula.

Importantly, once developed, we will make these materials—which will use the hosted API developed and instantiated under this project—available to faculty at other institutions, scaling the impact by enabling undergraduates in CS across the U.S. to gain hands-on experience analyzing and working with the internals of massive language models, which increasingly dominate the AI landscape. Note that such exercises are not currently possible at the vast majority of institutions given the resources required to run such models. And even if students are willing and able to pay for access via a commercial API, they would not be able to access model internals in a granular way, in turn severely limiting the kinds of analysis possible. As discussed in Section 6.1 we will ensure that we are supporting universities/colleges across the country with a particular focus on outreach to minority-serving institutions, women's colleges, HSIs, and HBCUs.

7 Institutional Commitment to Inclusion

Khoury College of Computer Sciences is a leader in broadening participation in Computer Science (CS). Khoury is home to the Center for Inclusive Computing (CIC) [125]), which aims to increase the representation of women of all races and ethnicities majoring in CS across the U.S. The CIC works with over 100+ institutions across the country to remove institutional barriers to students discovering and persisting in computing. Under co-PI Brodley's leadership, Khoury piloted and scaled the Align MS in CS program [126, 127], which provides a pathway to an MS in CS for students without CS backgrounds. This unique program attracts a notably diverse student body;

in 2022 more than half of the incoming class are women and 20% of the domestic students identify as Hispanic, Latino, African-American, Native American, or Pacific Islander. In 2019, the CIC brought this innovation to other universities and established the MS Pathways Consortium [128], a networked community of 23 institutions who are now offering the MS in CS for non-majors. Khoury also has a verified college-wide broadening participation in computing plan [129].

8 Divestment

Northeastern Research Computing will maintain NDIF hardware for its lifetime at no direct cost (including costs of safe disposal as needed). Should the project become insolvent, Research Computing will erase all disk storage on NDIF and repurpose the hardware for other research, at no direct cost. Source code would remain available on GitHub.

9 International Collaborators

Our project does not involve international collaboration. No-fee usage of the facility will be provided to US educational users only. When the service is established, international researchers can apply to join the facility as paid NDIF partners.

10 Results of Prior NSF Support

PI Bau has no prior NSF support.

Co-PI Brodley is PI/Co-PI on four current NSF grants, all of which share the same **Broader Impact:** to increase the representation of populations historically minoritized in tech in the undergraduate and graduate computing populations. The award most relevant to this proposal is #2137907: BPC-DP: Distributed Research Apprenticeships for Master's (DREAM), (2021-2023) supports MS students in the MS Pathways Consortium universities to participate in research. **Intellectual Merit:** The diverse demographics of the Consortium programs provide a unique opportunity to recruit Ph.D. students from a previously untapped population of students.

Co-PI Wallace is PI on multiple active NSF awards; most relevant to this proposal is "RI: Medium: Learning Disentangled Representations for Text to Aid Interpretability and Transfer" (NSF 1901117, $999,990.00, 2019-2023). **Intellectual Merit:** The aim is to develop neural networks that yield *disentangled* representations, i.e., which factorize into interpretable sub-components. Such representations can afford *interpretability* by being explicit about what aspects of a text they encode. The project has yielded several publications describing progress toward these ends [78, 95, 97, 102–104, 130]. **Broader Impact:** The technical focus of this project—interpretabile neural networks via disentanglement—has clear implications with respect to fairness, as it provides mechanisms to inspect *what* models encode. The project has also supported undergraduate research.

Co-PI Bell's most relevant recent award is CCF-2100037 "SHF: Medium: Collaborative Research: Enhancing Continuous Integration Testing for the Open-Source Ecosystem" ($400K, 2018–2023). **Intellectual Merit:** This project addresses the problem of regression testing in the new setting of continuous integration (CI), and has focused on detecting flaky tests [114], understanding flaky tests [131–133], and making CI builds faster [134]. **Broader Impact:** This project has resulted in significant technology transfers to popular open-source projects Apache Maven [134] and Pitest [132], and creation of educational materials for CI [135, 136].

Co-PI Guha is PI on NSF Award "SHF: Small: A Language-based Approach to Faster and Safer Serverless Computing (SHF-2102288, $441,149, 2020-2022). **Intellectual Merit:** This project aims to develop new programming abstractions and tools for serverless computing. The project has produced several papers [109, 137–140]. Wasm/k [138] implements continuations for WebAssembly, a growing platform for serverless computing. **Broader Impact**: PI Guha is standardizing WebAssembly effect, informed by Wasm/k.

References

[1] John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright, Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, and Nick Ryder. *ChatGPT: Optimizing Language Models for Dialogue*. `https://openai.com/blog/chatgpt/`. Dec. 2022.

[2] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. "“So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy". In: *International Journal of Information Management* 71 (2023), p. 102642.

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. "Sparks of artificial general intelligence: Early experiments with gpt-4". In: *arXiv preprint arXiv:2303.12712* (2023).

[4] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep reinforcement learning from human preferences". In: *Advances in neural information processing systems* 30 (2017).

[5] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. "Gpt-4 passes the bar exam". In: *Available at SSRN 4389233* (2023).

[6] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. "Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations". In: *arXiv preprint arXiv:2303.18027* (2023).

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. "Evaluating large language models trained on code". In: *arXiv preprint arXiv:2107.03374* (2021).

[8] White House Office of Science and Technology Policy. *Blueprint for an AI bill of rights*. `https://www.whitehouse.gov/ostp/ai-bill-of-rights/`. Oct. 2022.

[9] et al. Yoshua Bengio. *Pause Giant AI Experiments: an Open Letter*. `https://futureoflife.org/open-letter/`. 2023.

[10] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. "Bloom: A 176b-parameter open-access multilingual language model". In: *arXiv preprint arXiv:2211.05100* (2022).

[11] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. "OPT: Open pre-trained transformer language models". In: *arXiv preprint arXiv:2205.01068* (2022).

[12] *Strengthening and Democratizing theU.S. Artificial Intelligence Innovation Ecosystem, An Implementation Plan for a National Artificial Intelligence Research Resource*. Jan. 2023. URL: `https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf`.

[13] Future of Life Institute. *Policymaking in the Pause*. https://futureoflife.org/open-letter/. 2023.

[14] Kevin Roose. "A Conversation With Bing's Chatbot Left Me Deeply Unsettled". In: (2023).

[15] Alan M Turing. "Computing Machinery and Intelligence". In: *Mind* 59.236 (1950), pp. 433–460.

[16] John McCarthy. "What is artificial intelligence". In: (2007).

[17] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. "Deep blue". In: *Artificial intelligence* 134.1-2 (2002), pp. 57–83.

[18] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. "Building Watson: An overview of the DeepQA project". In: *AI magazine* 31.3 (2010), pp. 59–79.

[19] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. "Mastering the game of go without human knowledge". In: *nature* 550.7676 (2017), pp. 354–359.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[22] Steve Mollman. "OpenAI is getting trolled for its name after refusing to be open about its A.I." In: *Fortune* (Mar. 2023). URL: https://fortune.com/2023/03/17/sam-altman-rivals-rip-openai-name-not-open-artificial-intelligence-gpt-4/.

[23] OpenAI. "GPT-4 Technical Report". In: *arXiv preprint arXiv:2303.08774* (2023).

[24] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. "Language Models as Knowledge Bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2463–2473. DOI: 10.18653/v1/D19-1250. URL: https://aclanthology.org/D19-1250.

[25] Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, et al. "A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level". In: *Proceedings of the National Academy of Sciences* 119.32 (2022), e2123433119.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[27] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. "Semi-supervised sequence tagging with bidirectional language models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1756–1765. DOI: 10.18653/v1/P17-1161. URL: https://aclanthology.org/P17-1161.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`. URL: `https://aclanthology.org/N19-1423`.

[29] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A primer in BERTology: What we know about how bert works". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 842–866.

[30] Daniel Zhang, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Nestor Maslej, Andre Barbe, Helen Ngo, Latisha Harry, Ellie Sakhaee, Benjamin Bronkema-Bekker, et al. "The AI index 2021 annual report". In: (2022). URL: `https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf`.

[31] Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. "GPT-NeoX-20B: An Open-Source Autoregressive Language Model". In: *Proceedings of BigScience Episode\# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. 2022, pp. 95–136.

[32] Sabine N van der Veer, Lisa Riste, Sudeh Cheraghi-Sohi, Denham L Phipps, Mary P Tully, Kyle Bozentko, Sarah Atwood, Alex Hubbard, Carl Wiper, Malcolm Oswald, et al. "Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries". In: *Journal of the American Medical Informatics Association* 28.10 (2021), pp. 2128–2138.

[33] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. "Predictability and surprise in large generative models". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1747–1764.

[34] Prasann Singhal, Jarad Forristal, Xi Ye, and Greg Durrett. "Assessing Out-of-Domain Language Model Performance from Few Examples". In: *arXiv preprint arXiv:2210.06725* (2022).

[35] Chris Reed. "How should we regulate artificial intelligence?" In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128 (2018), p. 20170360.

[36] Kay Firth-Butterfield. "Artificial Intelligence and the Law: More Questions than Answers?" In: *Scitech Lawyer* 14.1 (2017), pp. 28–31.

[37] Zhaofeng Wu, William Merrill, Hao Peng, Iz Beltagy, and Noah A Smith. "Transparency Helps Reveal When Language Models Learn Meaning". In: *arXiv preprint arXiv:2210.07468* (2022).

[38] John Hewitt and Christopher D Manning. "A structural probe for finding syntax in word representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4129–4138.

[39] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. "Linguistic Knowledge and Transferability of Contextual Representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1073–1094. DOI: 10.18653/v1/N19-1112. URL: https://aclanthology.org/N19-1112.

[40] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. "Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. CHIL '20. Toronto, Ontario, Canada: Association for Computing Machinery, 2020, pp. 110–120. ISBN: 9781450370462. DOI: 10.1145/3368555.3384448. URL: https://doi.org/10.1145/3368555.3384448.

[41] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. "Discovering Latent Knowledge in Language Models Without Supervision". In: *arXiv preprint arXiv:2212.03827* (2022).

[42] Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. ""Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification". In: *arXiv preprint arXiv:2111.07367* (2021).

[43] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. "Gradient-based attribution methods". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (2019), pp. 169–191.

[44] Jesse Vig. "A multiscale visualization of attention in the transformer model". In: *arXiv preprint arXiv:1906.05714* (2019).

[45] Jesse Vig and Yonatan Belinkov. "Analyzing the structure of attention in a transformer language model". In: *arXiv preprint arXiv:1906.04284* (2019).

[46] Samira Abnar and Willem H Zuidema. "Quantifying Attention Flow in Transformers". In: *ACL*. 2020.

[47] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. "In-context learning and induction heads". In: *arXiv preprint arXiv:2209.11895* (2022).

[48] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. "Language models (mostly) know what they know". In: *arXiv preprint arXiv:2207.05221* (2022).

[49] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. "Gltr: Statistical detection and visualization of generated text". In: *arXiv preprint arXiv:1906.04043* (2019).

[50] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M Shieber. "Investigating Gender Bias in Language Models Using Causal Mediation Analysis". In: *NeurIPS*. 2020.

[51] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. "Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small". In: *arXiv preprint arXiv:2211.00593* (2022).

[52] Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. "Locating and editing factual associations in gpt". In: *Advances in Neural Information Processing Systems*. 2022.

[53] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. "Mass-editing memory in a transformer". In: *arXiv preprint arXiv:2210.07229* (2022).

[54] Brian Lester, Rami Al-Rfou, and Noah Constant. "The power of scale for parameter-efficient prompt tuning". In: *arXiv preprint arXiv:2104.08691* (2021).

[55] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. "Parameter-Efficient Transfer Learning for NLP". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 2790–2799. URL: `https://proceedings.mlr.press/v97/houlsby19a.html`.

[56] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. "On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 2208–2222. DOI: `10.18653/v1/2021.acl-long.172`. URL: `https://aclanthology.org/2021.acl-long.172`.

[57] Ben Wang and Aran Komatsuzaki. *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*. `https://github.com/kingoflolz/mesh-transformer-jax`. May 2021.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[59] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. *Crosslingual Generalization through Multitask Finetuning*. 2022. arXiv: `2211.01786 [cs.CL]`.

[60] *Carper AI*. Apr. 2023. URL: `https://carper.ai/`.

[61] *Announcing OpenFlamingo: an open-source framework for training vision-language models with in-context learning*. Apr. 2023. URL: `https://laion.ai/blog/open-flamingo/`.

[62] *Large-scale AI Open Network*. Apr. 2023. URL: `https://laion.ai/`.

[63] *Together Computing*. Apr. 2023. URL: `https://together.xyz/`.

[64] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. "Probing pretrained language models for lexical semantics". In: *arXiv preprint arXiv:2010.05731* (2020).

[65] Zining Zhu and Frank Rudzicz. "An information theoretic view on selecting linguistic probes". In: *arXiv preprint arXiv:2009.07364* (2020).

[66] John Hewitt and Percy Liang. "Designing and interpreting probes with control tasks". In: *arXiv preprint arXiv:1909.03368* (2019).

[67] Li Lucy and David Bamman. "Gender and representation bias in GPT-3 generated stories". In: *Proceedings of the Third Workshop on Narrative Understanding*. 2021, pp. 48–55.

[68] Abubakar Abid, Maheen Farooqi, and James Zou. "Persistent anti-muslim bias in large language models". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 298–306.

[69] Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. "Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22. Oxford, United Kingdom: Association for Computing Machinery, 2022, pp. 7–21. ISBN: 9781450392471. DOI: 10.1145/3514094.3534203. URL: https://doi.org/10.1145/3514094.3534203.

[70] Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe, eds. *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022. URL: https://aclanthology.org/2022.blackboxnlp-1.0.

[71] Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. "Language models trained on media diets can predict public opinion". In: *arXiv preprint arXiv:2303.16779* (2023).

[72] Patrick Y Wu, Joshua A Tucker, Jonathan Nagler, and Solomon Messing. "Large Language Models Can Be Used to Estimate the Ideologies of Politicians in a Zero-Shot Learning Setting". In: *arXiv preprint arXiv:2303.12057* (2023).

[73] Joseph T Ornstein, Elise N Blasingame, and Jake S Truscott. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts". In: (2022).

[74] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. "A deep learning framework for neuroscience". In: *Nature neuroscience* 22.11 (2019), pp. 1761–1770.

[75] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. "Dissociating language and thought in large language models: a cognitive perspective". In: *arXiv preprint arXiv:2301.06627* (2023).

[76] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. "Capabilities of gpt-4 on medical challenge problems". In: *arXiv preprint arXiv:2303.13375* (2023).

[77] Felix Agbavor and Hualou Liang. "Predicting dementia from spontaneous speech using large language models". In: *PLOS Digital Health* 1.12 (2022), e0000168.

[78] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. "Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?" In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2021, pp. 946–959.

[79] *Microsoft Azure*. Apr. 2023. URL: https://azure.microsoft.com/.

[80] *Amazon AWS*. Apr. 2023. URL: https://aws.amazon.com/.

[81] *Google Cloud*. Apr. 2023. URL: https://cloud.google.com/.

[82] *BigScience Petals*. Apr. 2023. URL: https://petals.ml/.

[83] Nvidia. *Nivida Triton Inference Server: open-source inference serving software.* https://developer.nvidia.com/nvidia-triton-inference-server. Apr. 2023.

[84] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[85] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. "Understanding the role of individual units in a deep neural network". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30071–30078.

[86] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. "Network dissection: Quantifying interpretability of deep visual representations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6541–6549.

[87] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[88] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks". In: *International Conference on Learning Representations*. 2018.

[89] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. "Rewriting a deep generative model". In: *European conference on computer vision*. Springer. 2020, pp. 351–369.

[90] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. "Rewriting geometric rules of a gan". In: *ACM Transactions on Graphics (TOG)* 41.4 (2022), pp. 1–16.

[91] Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. "Editing a classifier by rewriting its prediction rules". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 23359–23373.

[92] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. "Emergent world representations: Exploring a sequence model trained on a synthetic task". In: *arXiv preprint arXiv:2210.13382* (2022).

[93] Pouya Pezeshkpour, Sarthak Jain, Byron Wallace, and Sameer Singh. "An Empirical Comparison of Instance Attribution Methods for NLP". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Online: Association for Computational Linguistics, June 2021, pp. 967–975. DOI: `10.18653/v1/2021.naacl-main.75`. URL: `https://aclanthology.org/2021.naacl-main.75`.

[94] Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. "Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Online: Association for Computational Linguistics, July 2020, pp. 5553–5563. DOI: `10.18653/v1/2020.acl-main.492`. URL: `https://aclanthology.org/2020.acl-main.492`.

[95] Jered McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron C. Wallace. "That's the Wrong Lung! Evaluating and Improving the Interpretability of Unsupervised Multimodal Encoders for Medical Data". In: *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*. 2022.

[96] Sarthak Jain, Varun Manjunatha, Byron C. Wallace, and Ani Nenkova. "Influence Functions for Sequence Tagging Models". In: *Proceedings of the Findings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*. 2022.

[97] Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace. "Disentangling Representations of Text by Masking Transformers". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 778–791. URL: https://aclanthology.org/2021.emnlp-main.60.

[98] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*. 2020, pp. 4443–4458.

[99] Sarthak Jain and Byron C. Wallace. "Attention is not Explanation". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2019, pp. 3543–3556.

[100] Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. "Learning to Faithfully Rationalize by Construction". In: *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*. 2020, pp. 4459–4473.

[101] Jered McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron Wallace. "That's the Wrong Lung! Evaluating and Improving the Interpretability of Unsupervised Multimodal Encoders for Medical Data". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3626–3648. URL: https://aclanthology.org/2022.emnlp-main.238.

[102] Diego Garcia-Olano, Yasumasa Onoe, Ioana Baldini, Joydeep Ghosh, Byron Wallace, and Kush Varshney. "Biomedical Interpretable Entity Representations". In: *Proceedings of the Findings of the Association for Computational Linguistics (ACL)*. Online: Association for Computational Linguistics, 2021, pp. 3547–3561. DOI: 10.18653/v1/2021.findings-acl.311. URL: https://aclanthology.org/2021.findings-acl.311.

[103] Sanjana Ramprasad, Denis Jered McInerney, Iain J. Marshall, and Byron C. Wallace. "Automatically Summarizing Evidence from Clinical Trials: A Prototype Highlighting Current Challenges". In: *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL), System Demonstrations*. 2023.

[104] Silvio Amir, Jan-Willem van de Meent, and Byron C. Wallace. "On the Impact of Random Seeds on the Fairness of Clinical Classifiers". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2021, pp. 3808–3823.

[105] Arjun Guha, Mark Reitblatt, and Nate Foster. "Machine Verified Network Controllers". In: *ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*. 2013.

[106] Arjun Guha, Matthew Fredrikson, Benjamin Livshits, and Nikhil Swamy. "Verified Security for Browser Extensions". In: *IEEE Security and Privacy (Oakland)*. 2011.

[107] Arjun Guha, Shriram Krishnamurthi, and Trevor Jim. "Using Static Analysis for Ajax Intrusion Detection". In: *World Wide Web Conference (WWW)*. 2009.

[108] Rian Shambaugh, Aaron Weiss, and Arjun Guha. "Rehearsal: A Configuration Verification Tool for Puppet". In: *ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*. 2016.

[109] Abhinav Jangda, Sandeep Polisetty, Arjun Guha, and Marco Serafini. "Accelerating Graph Sampling for Graph Machine Learning Using GPUs". In: *European Conference on Computer Systems (EuroSys)*. 2021.

[110] Abhinav Jangda and Arjun Guha. "Model-Based Warp-Level Tiling for Image Processing Programs on GPUs". In: *International Conference on Parallel Architectures and Compilation Techniques (PACT)*. 2020.

[111] Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. *MultiPL-E: A Scalable and Extensible Approach to Benchmarking Neural Code Generation*. 2022. DOI: 10.48550/ARXIV.2208.08227.

[112] Nicolas Viennot, Mathias Lécuyer, Jonathan Bell, Roxana Geambasu, and Jason Nieh. "Synapse: A Microservices Architecture for Heterogeneous-Database Web Applications". In: *Proceedings of the Tenth European Conference on Computer Systems*. EuroSys '15. Bordeaux, France: Association for Computing Machinery, 2015. ISBN: 9781450332385. DOI: 10.1145/2741948.2741975. URL: https://doi.org/10.1145/2741948.2741975.

[113] Jonathan Bell, Owolabi Legunsen, Michael Hilton, Lamyaa Eloussi, Tifany Yung, and Darko Marinov. "DeFlaker: Automatically Detecting Flaky Tests". In: *Proceedings of the 2018 International Conference on Software Engineering*. ICSE 2018. 2018. URL: http://jonbell.net/publications/deflaker.

[114] Abdulrahman Alshammari, Christopher Morris, Michael Hilton, and Jonathan Bell. "FlakeFlagger: Predicting Flakiness Without Rerunning Tests". In: *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. 2021, pp. 1572–1584. DOI: 10.1109/ICSE43902.2021.00140.

[115] Jonathan Bell and Gail Kaiser. "Unit Test Virtualization with VMVM". In: *ICSE*. 2014.

[116] Jonathan Bell, Eric Melski, Gail Kaiser, and Mohan Dattatreya. "Accelerating Maven by Delaying Test Dependencies". In: *3rd International Workshop on Release Engineering*. RELENG '15. Florence, Italy: IEEE Press, May 2015, p. 28. URL: http://dl.acm.org/citation.cfm?id=2820690.2820703.

[117] Jonathan Bell and Gail Kaiser. "Phosphor: Illuminating Dynamic Data Flow in Commodity JVMs". In: *ACM International Conference on Object Oriented Programming Systems Languages & Applications*. OOPSLA '14. Portland, Oregon, USA: ACM, Oct. 2014, pp. 83–101. ISBN: 978-1-4503-2585-1. DOI: 10.1145/2660193.2660212. URL: http://doi.acm.org/10.1145/2660193.2660212.

[118] Jonathan Bell and Luís Pina. "CROCHET: Checkpoint and Rollback via Lightweight Heap Traversal on Stock JVMs". In: *Proceedings of the 2018 European Conference on Object-Oriented Programming*. ECOOP 2018. 2018.

[119] Katherine Hough and Jonathan Bell. "A Practical Approach for Dynamic Taint Tracking with Control-Flow Relationships". In: *ACM Trans. Softw. Eng. Methodol.* 31.2 (Dec. 2021). ISSN: 1049-331X. DOI: 10.1145/3485464. URL: https://doi.org/10.1145/3485464.

[120] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35.

[121] *Pause Giant AI Experiments: An Open Letter*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. 2023.

[122] Ayanna Howard and Charles Isbell. "Diversity in AI: The Invisible Men and Women". In: *MIT Sloan Management Review* (Sept. 2020), pp. 20–22. URL: https://sloanreview.mit.edu/article/diversity-in-ai-the-invisible-men-and-women/.

[123] Gabriel Ramos. "Why we must act now to close the gender gap in AI". In: *World Economic Forum* (Aug. 2022), pp. 20–22. URL: https://www.weforum.org/agenda/2022/08/why-we-must-act-now-to-close-the-gender-gap-in-ai/.

[124] *Carla Brodley recieves the 2021 ACM Francis E. Allen Award for Ourstanding Mentoring*. Apr. 2022. URL: https://www.acm.org/articles/bulletins/2022/april/allen-award-2021-brodley.

[125] *Center for Inclusive Computing at Northeastern University*. Jan. 2023. URL: https://cic.northeastern.edu/.

[126] Carla Brodley, Megan Barry, Aidan Connell, Catherine Gill, Ian Gorton, Benjamin Hescott, Bryan Lackaye, Cynthia LuBien, Leena Razzaq, Amit Shesh, Tiffani Williams, and Andrea Danyluk. "An MS in CS for non-CS Majors: Moving to increase diversity of thought and demographics in CS". In: *Proceedings of the 51th ACM Technical Symposium on Computer Science Education*. SIGCSE '20. Portland, OR, USA: Association for Computing Machinery, 2020, pp. 1248–1254. ISBN: 9781450367936. DOI: 10.1145/3328778.3366802. URL: https://doi.org/10.1145/3328778.3366802.

[127] *Align MS in Computer Science at Northeastern*. Jan. 2023. URL: https://www.khoury.northeastern.edu/programs/align-masters-of-science-in-computer-science/.

[128] Carla Brodley and Jan Cuny. "The MSCS New Pathways Consortium-a National Invitation". In: *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*. Vol. 1. IEEE. 2020, pp. 1–2.

[129] *Verified Departmental BPC Plans*. Jan. 2023. URL: https://bpcnet.org/verified-departmental-bpc-plans/.

[130] Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh, and Byron C. Wallace. "Intermediate Entity-based Sparse Interpretable Representation Learning". In: *Proceedings of the BlackboxNLP Workshop at EMNLP*. 2022.

[131] Wing Lam, Stefan Winter, Anjiang Wei, Tao Xie, Darko Marinov, and Jonathan Bell. "A Large-Scale Longitudinal Study of Flaky Tests". In: *Proc. ACM Program. Lang.* 4.OOPSLA (Nov. 2020). DOI: 10.1145/3428270. URL: https://doi.org/10.1145/3428270.

[132] August Shi, Jonathan Bell, and Darko Marinov. "Mitigating the Effects of Flaky Tests on Mutation Testing". In: *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 112–122. ISBN: 9781450362245. URL: https://doi.org/10.1145/3293882.3330568.

[133] Alessio Gambi, Jonathan Bell, and Andreas Zeller. "Practical Test Dependency Detection". In: *Proceedings of the 2018 IEEE Conference on Software Testing, Validation and Verification*. ICST 2018. 2018. URL: http://jonbell.net/publications/pradet.

[134] Pengyu Nie, Ahmet Celik, Matthew Coley, Aleksandar Milicevic, Jonathan Bell, and Milos Gligoric. "Debugging the Performance of Maven's Test Isolation: Experience Report". In: *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. ISSTA 2020. Virtual Event, USA: Association for Computing Machinery, 2020, pp. 249–259. ISBN: 9781450380089. DOI: `10.1145/3395363.3397381`. URL: `https://doi.org/10.1145/3395363.3397381`.

[135] Jonathan Bell, Adeel Bhutta, Ferdinand Vesely, and Mitch Wand. *CS4530, Spring 2022: Fundamentals of Software Engineering*. `https://neu-se.github.io/CS4530-Spring-2022/`. 2022.

[136] Jonathan Bell, Adeel Bhutta, Ferdinand Vesely, and Mitch Wand. *CS4530, Spring 2022: Fundamentals of Software Engineering Source Materials*. `https://github.com/neu-se/CS4530-Spring-2022`. 2022.

[137] Donald Pinckney, Federico Cassano, Arjun Guha, Jonathan Bell, Massimiliano Culpo, and Todd Gamblin. "Flexible and Optimal Dependency Management via Max-SMT". In: *IEEE/ACM International Conference on Software Engineering (ICSE)*. 2023.

[138] Donald Pinckney, Yuriy Brun, and Arjun Guha. "Wasm/k: Delimited Continuations for WebAssembly". In: *Dynamic Languages Symposium (DLS)*. 2020. DOI: `10.1145/3426422.3426978`.

[139] Luna Phipps-Costin, Carolyn Jane Anderson, Michael Greenberg, and Arjun Guha. "Solver-based Gradual Type Migration". In: *Proceedings of the ACM on Programming Languages (PACMPL)* 5.OOPSLA (2021). DOI: `https://doi.org/10.1145/3485488`.

[140] James Perretta, Andrew DeOrio, Arjun Guha, and Jonathan Bell. "On the use of mutation analysis for evaluating student test suite quality". In: *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*.