Overview

The most relevant directorate for review of this proposal is CISE, and divisions: CNS, IIS and OAC.

Large language models (LLMs) such as ChatGPT that surpass 100 billion parameters have ushered in an exciting new era of artificial intelligence (AI). State-of-the-art LLMs exhibit new capabilities, including some aspects of general-purpose reasoning, that raise **fundamental scientific questions** that impact not only computer science, but also **biology, the social sciences, business, engineering, and education.** But the computational requirements needed to run such models have made it infeasible for academic researchers to conduct research into how they work: LLMs are so large that they cannot be run in inference (i.e., to make predictions) with computational resources available to academics, and it is infeasible to create the needed capabilities at an institutional level. Thus, researchers are currently hindered in their ability to anticipate, explain, and regulate these systems. The proposed **National Deep Inference Facility (NDIF)**, led by a team of researchers at Northeastern University, will advance scientific understanding by providing U.S. academic researchers with access to a cutting-edge computing service capable of running very large language models while giving complete transparency to their internal computations—a capability not currently accessible to academics. Therefore, by designing, building, and deploying both computing hardware and software infrastructure, NDIF addresses urgent research and societal needs for transparency as a means to advancement of safe, robust, trustworthy, and explainable AI.

Intellectual Merit

Deep inference describes 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 provides the computational capacity, instrumentation, transparency, broad access and training 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.

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 that could potentially transform the way LLMs are explained, applied, trusted, and regulated by potentially establishing a foundational understanding of their internal mechanisms.

Broader Impacts

Highly-capable LLMs will increasingly be deployed into use with widespread implications for society, because when they are broadly applied to read and write text, they have the potential to insert AI predictions that may contain biases, misinformation, and unknown goals into a wide variety of intellectual work worldwide. *But scientists cannot explain the predictions of such models.* Academics are well-positioned to critically scrutinize the inner-workings of very large AI models, but the infrastructure required to perform such research is out of reach for most academic labs. NDIF will enable U.S.-based academics to conduct critical research into LLMs that is currently not feasible, spurring advances exemplified by our community of researchers in computing, medicine, neuroscience, linguistics, social sciences and humanities. To ensure that these models are deployed ethically and in a socially responsible way, we will engage public interest technology groups as we design, build, and operationalize the facility and as we directly train hundreds of student-users.

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.

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.** If we continue to deploy LLMs without the ability to explain them, society will enter this new era of AI blindfolded and without tools for anticipating, auditing, or regulating the mechanisms of these large-scale systems, even as they begin to impact every aspect of society.

A national-scale infrastructure to explain LLMs is necessary due to the demanding computational requirements for conducting research-oriented inference on very large models. Existing computation clusters partition resources across users to serve batch jobs. By contrast, a deep inference service must share a small set of large models on relatively few servers and make them accessible to many users. **Creating this infrastructure at the institutional level is not feasible** because running the first deep inference experiment on a trillion-parameter model would require a multi-million-dollar investment in unique hardware and software. While companies like OpenAI offer commercial inference services such as ChatGPT that spread costs over many users (Figure 1a), those services do not expose the internals of the LLMs, making it impossible to study their mechanisms. By providing a transparent deep inference service (Figure 1b), **NDIF will enable scientific interrogation of LLM mechanisms, advancing urgently needed understanding of *how* they work**

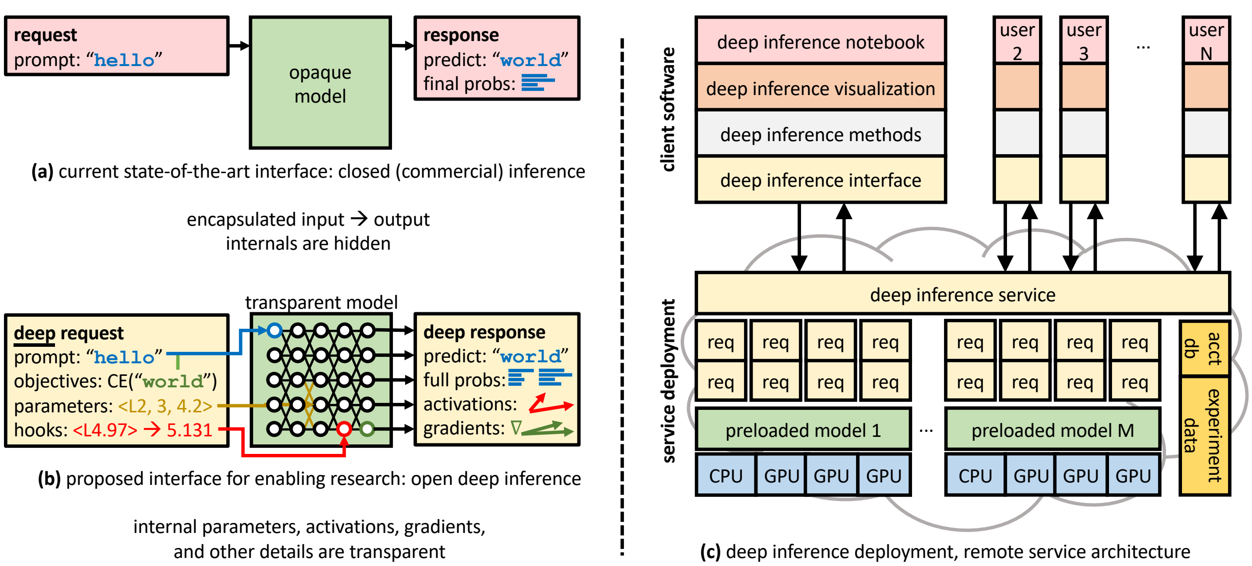


Figure 1: (a) Current services hosting large language models provide very limited interaction functionality (Top). One can send input text in a request, and is then provided an output string (and scores associated with the final predictions). (b) We propose developing infrastructure to provide deep access *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.*

so industry, government, researchers and the public are able to safely deploy, regulate, use, and study LLMs for the benefit of society.

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. NDIF consists of three complementary components:

1. Creation and testing of an online inference service with the support staff and infrastructure necessary to provide researchers with the ability to interrogate and conduct ground-breaking 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 expanding deep inference research capabilities by deploying on national high-performance computing (HPC) clusters or commercial cloud providers.
3. Outreach and training for students and researchers in every region of the country to use NDIF to advance understanding of large neural network models, developing a highly skilled workforce of scientists and engineers to lead the world in ethical use of state-of-the-art LLMs.

NDIF will be developed under the leadership of a unique team of experts in machine learning, software engineering, deep network interpretability, language modeling, high-performance computing, and inclusive computing at Northeastern University (NU). The team will benefit from the university's well-established organizational structure and advanced facilities. The hardware cluster will be deployed at the Massachusetts Green High Performance Computing Center, a shared computation facility in which NU is one of five university partners.

2 Intellectual Merit

Explaining AI systems is a national and global priority: In October 2022, the White House Office of Science and Technology Policy 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.*" In January, 2023, the National AI Research Resource Task Force [10] identified one of the four critical opportunities for strengthening the U.S. AI R&D ecosystem as the development of 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*." Two months later (March 2023), the Future of Life Institute published a "Pause Giant AI" open letter [9] which has since garnered more than 25,000 signatories, including many national leaders in AI research, recommending "*a significant increase in public funding for technical AI safety research in the areas of alignment, robustness and assurance, and explainability and interpretability*" [11]. These three documents published in the last six months alone, highlight the urgency of research to explain, audit, evaluate, and manage impacts of LLMs.

Meanwhile, **LLMs such as ChatGPT are being adopted more quickly than any previous technology**, with widespread deployment in consumer-facing technologies [12], touching every field involving reading, writing, or programming, 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 of developing trustworthy AI**.

Academic researchers are ideally-suited to investigate the mechanisms of LLMs, but are unable to conduct this critical research due to the lack of large-scale LLM research infrastructure – a new need that stems from from the unprecedented scale of state-of-the-art LLMs. NDIF will directly address this need through a robust investment in a shared hardware and software platform.

2.1 The challenge: the scale of LLMs has created a new crisis of transparency

While the emergence of LLMs such as GPT-3 [13] has energized the Natural Language Processing (NLP) and broader Machine Learning (ML) research communities, the scale of those models has also presented the research community with a crisis of transparency that is qualitatively different from the previous generation of "large-scale" AI.

When the AlexNet [14] 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 laboratories to be able to reproduce, validate, modify, retrain, and study the model using a desktop workstation outfitted with a consumer GPU. Similarly, when the first successful pre-trained models for NLP—e.g., ELMO [15] and BERT [16]—emerged, these were small enough for academic researchers to run, interrogate, and tinker with locally, enabling important research into their capabilities and limitations [17]. 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 [18].

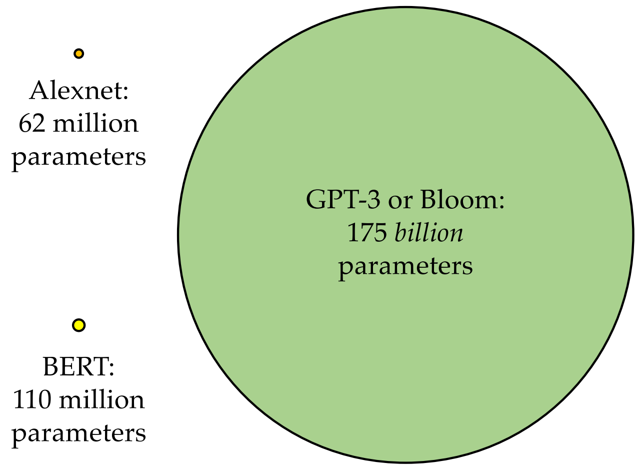


Figure 2: LLMs such as GPT-3 or Bloom are so much larger than the previous generation of deep networks (such as Alexnet or BERT) that investigating LLM inference requires specialized infrastructure. NDIF will provide this.

The current advancement made possible by GPT-3 [13] and similar very large language models (LLMs) is qualitatively different. The 175-billion parameter GPT-3 model is huge and private. Alternative, comparably sized LLMs (such as OPT [19], Bloom [20], and NEO-X [21]) are technically available to researchers, but often *de facto* inaccessible because merely running them requires specialized engineering and expensive data-center equipment. Meanwhile, a recent survey of established benchmarks [22] catalogued over 175 different capabilities that emerge in LLMs but that do not appear in smaller models. These include the ability to perform multi-digit arithmetic, unscramble words, correctly select truthful answers when baited by commonly-stated misconceptions, and multi-step reasoning "chain-of-thought" reasoning [23, 24]. Smaller language models do not exhibit this range of capabilities. Yet most academic researchers do not have sufficient resources to run LLMs, and are consequently unable to probe these phenomena 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. That 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, where AI advances lack the competitive scrutiny provided by independent academics.

Moreover, opaque models are a significant barrier to developing trustworthy AI. Our ignorance about the mechanisms that give rise to human-level LLM capabilities creates a troublesome dilemma between performance and transparency [25], making it difficult to anticipate how models will behave when deployed in the real world [26, 27]. Critically, our lack of understanding renders it impossible to regulate these systems and ensure safety, especially given the speed with which these technologies are being deployed [28, 29].

By enabling the diverse academic research community to study and explain how such models work, NDIF will empower important research into the potential risks of LLMs that are beyond the purview of industry. The disciplinary diversity of academic researchers in our user community (Section 2.6) 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 that operate them commercially. 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.

2.2 Scientific justification

The unique LLM deep inference capabilities provided by NDIF will enable scientists to advance **five critical research priorities identified by the NAIRR Task Force** [10], specifically by enabling scientists to advance a) understanding of AI decisions; b) AI auditing capabilities; c) AI testing and evaluation; d) tools to manage AI risks, and e) research on societal implications of AI.

Enabling Critical Research Priorities:

Understanding AI decisions
AI auditing capabilities
AI evaluation and testing
Tools to manage AI risks
Studying societal impacts of AI

a: Improving understanding of AI decisions. Mechanistic understanding of LLM decisions could transform how such models are used, developed, and regulated. Explaining LLMs is challenging in part because they are massive artificial neural networks, i.e., computational systems loosely inspired by human neurons [30, 31], with connection strengths determined by a training process which aims to match "training" examples [32, 33]. Because LLMs are not programmed explicitly, the only way to develop an explicit understanding of their decisions is to examine their internal calculations. This is challenging due to the complexity of the networks. The community of scientists involved in NDIF design discussions includes experts on methods for understanding neural networks, both the artificial network [34, 35] and biological [36] variety. They are part of a fast-growing community investigating the mechanistic inner-workings of LLMs, perhaps best represented by the BlackboxNLP workshop [37], which has been held annually at Association of Computational Linguistics (ACL) conferences for the past four years. These experts have worked with us to identify capabilities of the NDIF that would empower advancement of their cutting-edge research by enabling experimental methods in LLMs such as *casual mediation analysis*, *saliency mapping*, *and representation similarity analysis*. We describe key experimental methods on LLMs that will be enabled by NDIF in Section 2.4.

b: Improving AI auditing capabilities. Auditing LLMs would allow users to identify the knowledge contained within a network. This capability could be transformative by redefining the way that humans interact with LLMs. NLP researchers and linguists in our user research community are keen to understand the linguistic knowledge implicitly encoded in LLMs [38–40], including the degree to which such models encode bias [41, 42], and they have demonstrated such measurements on smaller networks [43–45]. Similarly, human-computer interaction and data visualization experts want to study the effects of allowing users to see and interact model internals to facilitate interactions. Reading out the internals of an LLM as it operates may reveal implicit model biases and better allow users (i.e., humans) to appropriately calibrate their trust in model outputs. By providing the unique capability to apply experimental methods such as *representation probing* to LLMs (see Section 2.4), NDIF will allow our community of scientists to develop and extend such auditing capabilities for modern LLMs.

c: Developing AI evaluation and testing methods. Rigorous evaluation of LLMs is essential, especially when they are applied in high-stakes application areas such as bio-medicine [46]. Many of our community members are performing such research with opaque model access to GPT-3/4, where they do not have complete control over the evaluation setting. We have been working with them to ensure that NDIF provides them with capabilities they need for rigorous evaluation, including complete access to posterior probabilities, access to model internal activations, and the ability to fine-tune and evaluate models transparently for precise application domains. For example, high-stakes settings raise critical issues related to learned representations and fairness [47, 48] as well as risks of training LLMs on potentially sensitive personal health data [49]. When our community members investigate the use of LLMs to detect dementia from patient-elicited speech [50], or when they study how (health-related) domain knowledge is stored in LLMs, they will also require transparent access to LLM representations which NDIF will uniquely provide and that is unavailable from commercial services.

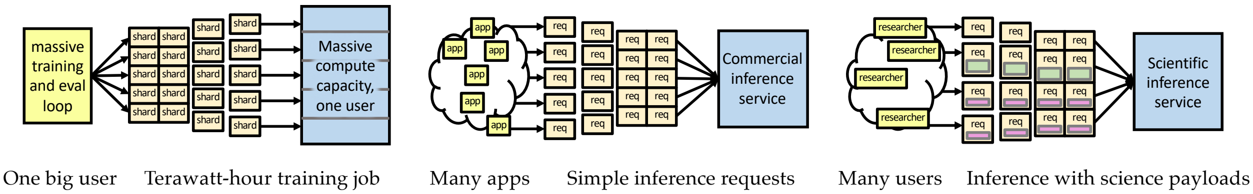
d: Creating tools to manage AI risks. NDIF will enable the development of tools that could be used to mitigate the negative impacts of LLMs, for example by detecting machine-generated misinformation [51–53] or tools that could detect possible untruths or deception in a model's behavior [54, 55]. Our community members have advised that applying such methods in the era of LLMs requires us to use LLMs with full access to posteriors and activations, a capability that will be uniquely provided by NDIF and currently unavailable from commercial inference providers.

e: Enabling research on societal impacts. Understanding societal impacts requires studying interactions between LLMs and people. For example, our community of researchers includes social scientists interested in studying whether and to what extent people will behave differently when they are aware that they are talking to a chatbot. Or the extent to which an LLM is persuasive in human conversation. Already, social scientists have begun using LLMs as tools for judging public opinion in ways that would be impractical to scale using other means [56], as well as using them to measure political ideology and other latent constructs from texts [57], and applying LLMs to various "text-as-data" tasks to permit subsequent analysis [58]. While some of these research settings may, on the surface, seem suitable for commercial inference services, researchers have told us that the commercial APIs limit experiment designs that they can use in their research, and they do not provide the transparency that would allow reproducible research. Unlike commercial services, NDIF will provide an environment suitable for ethical conduct of human subjects research, that will provide both the technical capabilities to support interactive human studies and a process to allow protocols to be overseen by a researcher's IRB.

2.3 Four current barriers to deep inference research

Deep inference research on LLMs is hindered by four factors. The lack of (1) available **computational resources** for researchers, (2) open **inference software for research**, and (3) **transparency** with respect to the training data, architecture, computations, and parameters of the models. Also (4) the choice by companies to maintain **closed models** as proprietary secrets. NDIF addresses the first three of these needs and relies on collaborations with open-model training efforts to address the fourth issue (Section 2.5), because computational demands of training are substantially different from those of inference. Further, there are numerous ongoing efforts to *pre-train* large open-models, but to our knowledge efforts focused on inference for research do not yet exist.

Deep inference has unique computational demands. The compute power required to support research into LLM *inference* is not well-supported by traditional HPC clusters. HPC clusters partition computational resources among users and give users exclusive use of a portion of capacity for a period of time. Because of this design, HPC clusters are well-suited to handling longer jobs, such as LLM pre-training jobs that may run for an extended training loop, with one user utilizing



(a) HPC applied to LLM training (b) Commercial LLM inference service (c) **NDIF**: scientific LLM inference facility

Figure 3: The computational workload of (a) training LLMs on traditional HPC clusters, where a single user runs a massive job for months, differs from (b) commercial inference services, which serve many apps' small (e.g. sub-second) requests on a concentrated sever; both also differ from (c) NDIF, which adds diverse scientific payloads that serve many different types of experiments. NDIF infrastructure will meet a need that is currently not met.

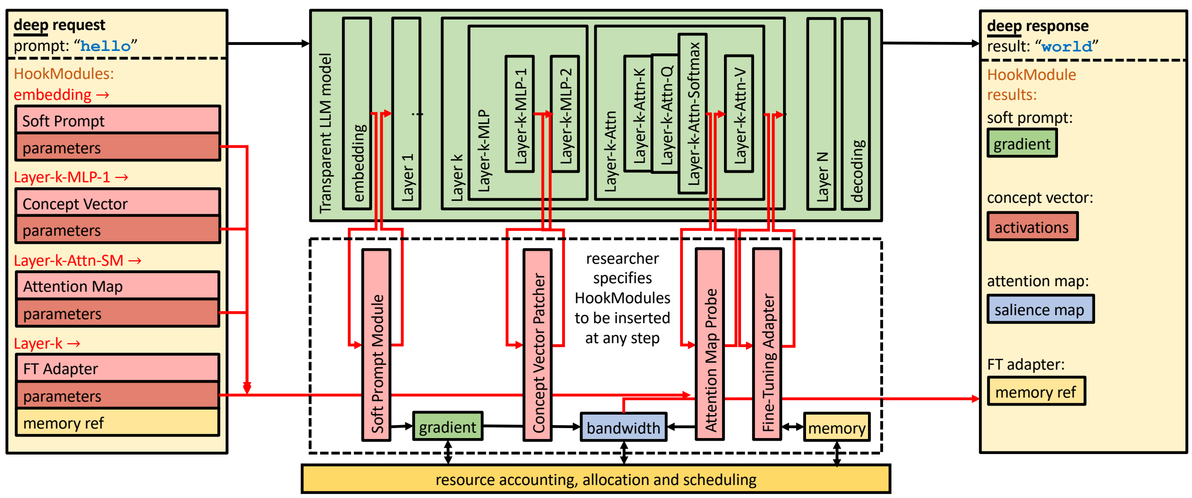


Figure 4: Details of the logical view of a deep inference request. Unlike commercial inference that provides no transparency, with the NDIF, researchers can execute flexible experiments by inserting computations in the internals of the deep network inference process. To maintain safe and efficient co-tenancy,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

hundreds of GPU devices for months (Figure 3a). 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 on a shared set of computers. Unlike commercial inference services (Figure 3b), NDIF will provide the ability to probe, inspect, and modify details of the pre-trained model to support the range of experimental methods, which means its requests will carry diverse scientific payloads (Figures 3c,4).

Unlike other ML infrastructure efforts such as XSEDE/ACCESS or commercial inference APIs, NDIF is singularly focused on the infrastructure needed for efficiently running LLMs to enable research *after* they are trained. NDIF will provide open-source tools, allowing large-scale HPC clusters to support deep inference research if they choose to configure their clusters to support inference service workloads.

2.4 Experimental methods uniquely enabled by NDIF

To enable scientists to advance the research agenda discussed in Section 2.2, NDIF will enable critical technical experimental methods (Figure 5) that are unavailable via commercial LLM inference services.

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* [59]. This is illustrative of a body of emerging research probing internal representations for implicit linguistic structure (e.g., [60, 61]). Work on LLMs for healthcare has shown that neural representations of health records implicitly encode patient race, which has fairness implications [47]. 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 [54]; 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) [62, 63]. Alternatively, one can analyze model *attention* distributions. In small models such analysis has revealed how simple dependencies are processed [64–66], including the discovery of very explicit copying circuits in transformer models [35]. Analyzing per-token model probabilities can reveal model self-knowledge [67] and differences between human and AI-generated text [51]. Extending these lines of inquiry to large models requires transparent access to model internals.

	Compute profile			Transparency needs				
Deep inference research methods on LLMs enabled by the NDIF	Interactive	Batch	Optimization	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 5: Deep inference research methods enabled by NDIF.

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 the specific computations within a language model that cause gender bias in language models [68]; that cause indirect object identification in sentences that name multiple subjects [69]; and that recall world knowledge within LLMs, such as knowledge of the relationships, associations and properties of real-world entities [70, 71]. Using such methods to interrogate larger models requires direct access to internal states.

Parameter-efficient fine-tuning. One of the most compelling properties of LLMs is their ability to be quickly fine-tuned to a specific task using a small amount of data [72]. NDIF will advance investigation of such capabilities by supporting parameter-efficient fine-tuning methods such as *adapter layers* [73, 74], 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" [72], which similarly introduce a small set of tunable parameters, albeit in this case they are viewed as pseudo input token embeddings.

2.5 NDIF will leverage existing and future open LLMs

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. There are several currently-available open LLM models that NDIF will integrate with, and we will collaborate and support ongoing efforts to create and deploy new large models. We have already begun efforts to integrate **EleutherAI**'s 20-billion parameter GPT-NeoX [21] and 6-billion parameter GPT-J [75] (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** [20], a 176-billion parameter multilingual model trained by BigScience, a collaboration of European agencies, the Huggingface company, and many others. (2) **Meta OPT** [19] and Llama [76], 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 [77], CarperAI [78] and OpenFlamingo [79]. (5) Ongoing work by the National AI Research Resource (NAIRR) [10], Large-Scale Artificial Intelligence Open Network (LAION) [80], and Together Computer [81]. 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.

2.6 The NDIF research user community and illustrative application areas

Support for the NDIF project is strong in the US research community — **over 400 researchers indicated that their research goals were blocked in a twitter community survey.**

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 Zoltan Majdik (North Dakota State) laid out the benefits: "Interpretability would easily be my number-one target. On multiple levels: for academic LLM experts, but also ... make interpretability interpretable for social scientists, non-computer-science." 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."

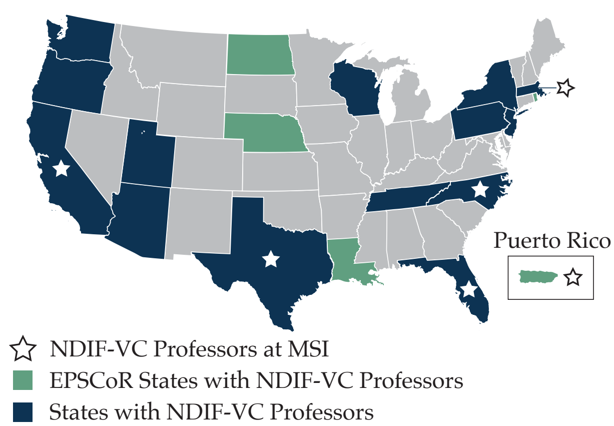


Figure 6: National reach: Our existing virtual community (VC) includes faculty at 34 universities in 19 states (5 EPSCoR jurisdictions) who have proposed specific research to be done with NDIF.

Based on our Twitter community survey, we have established a virtual research user community (NDIF-VC) that includes 40 professors from 34 different universities in 19 states including 5 EPSCoR jurisdictions (see Figure 6), and 8 minority-serving institutions (including an HBCU) who have suggested specific research projects that will benefit from NDIF. The researchers, who will directly use NDIF as early adopters, span a broad range of disciplines, including researchers who have discussed plans for projects in computational linguistics, NLP, human-computer-interaction, data visualization from CISE as well users whose research lie in other fields including network science, robotics, and hardware description language research from ENG, neuroscience and biomedicine from BIO, and political discourse, psycholinguistics, and narratology from SBE. The breadth of experiments that this diverse community wishes to conduct has defined our scientific requirements and has informed our design for NDIF.

3 Preliminary Activities

We have accomplished three key preliminary tasks to ensure that immediate development of the infrastructure toward the first pilot deployment can begin as soon as funds are awarded.

3.1 Identification of goals together with the user community

As a facility intended to serve the needs of the wider researcher community 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, we established **an open design process** involving outreach to the relevant research community through an online Discord server. This platform enables real-time and asynchronous text communication organized by channel, and provides audio and video chat as well. This is 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 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 LLMs** 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 LLMs in practice. Our user community has made significant progress toward 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 5). 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 with potential NDIF users online, we have hosted an in-person outreach event at the **International Conference on Learning Representations (ICLR) 2023, held on May 2, 2023.** Our event brought together over 100 members of the ML research community to discuss challenges faced by researchers studying LLMs on academic budgets. During the event, the community emphasized the need to make scientific tools for LLMs broadly accessible to researchers at low cost, with priorities on inclusiveness and openness. Through a poll and a live discussion, the community identified a range of specific research and infrastructure priorities and challenges that will need to be addressed by both the NDIF and other open-science efforts in LLMs; these are consistent with the findings in our online virtual 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 University of California, San Diego, has led software development projects at Google, the Broad Institute, and General Dynamics, among others. Brockman has been participating in our design process and helping to develop our technical specifications. If NDIF is funded, he 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 NU, 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 technical development planning

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 laboratory. These prototypes inform an architectural specification and technical development plan for NDIF. An overview of the plan is given in Section 4.1, and full details can be found in our Project Execution Plan (PEP).

4 Implementation

4.1 Technical readiness

Our team has created a detailed technical development plan that delineates requirements, design, and deployment milestones for the NDIF's user model, user-facing and internal software, hardware, training, and outreach. Please see our Project Execution Plan (PEP) for complete details.

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 the 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. Establish preliminary client library with five 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 the NDIF team. Increased robustness, including monitoring and alerting, improved job queuing, and a fairness-oriented scheduler. Define Service Level Objectives (SLOs) and measure progress toward 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 (see Section 6.1).

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 ten A100 GPUs to be located in the same physical sever (eight is 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. If high-density hardware is unavailable, we will deploy low-density nodes with interconnects. 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 four 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 (i.e., three 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 four 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 [10] public AI training resource.

4.1.3 Software stack and service architecture.

The software layer is critical 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 NDIF 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 [81–84] and/or nonprofit cluster [80, 85] and the future NAIRR [10].

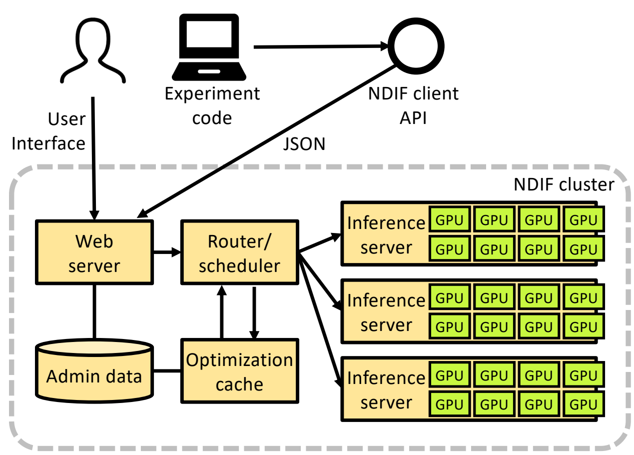


Figure 7: NDIF service architecture, showing request flows between system components.

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

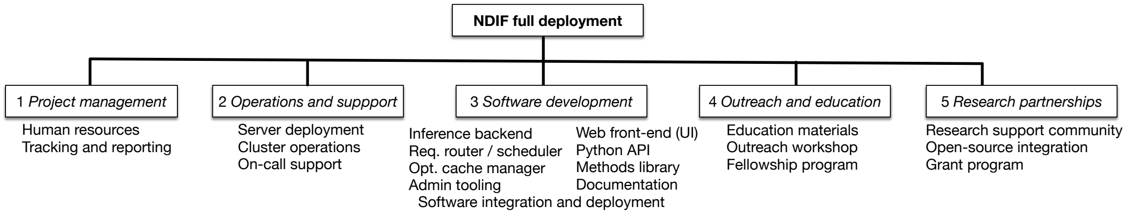


Figure 8: The work breakdown structure of NDIF construction.

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 [86] 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.

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 management of 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, including logging and monitoring software, and tools and scripts for administrative tasks, including user administration and moderation, quota management, cluster model allocation, health monitoring, load monitoring, and debugging tools for 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` [87] 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

Figure 9 shows the organization chart. 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.

PI Bau - NDIF Director (Assistant Professor, Khoury College of Computer Sciences, NU) brings a unique skillset as a late-career academic who worked in industry for 20+ years, 12 of them at

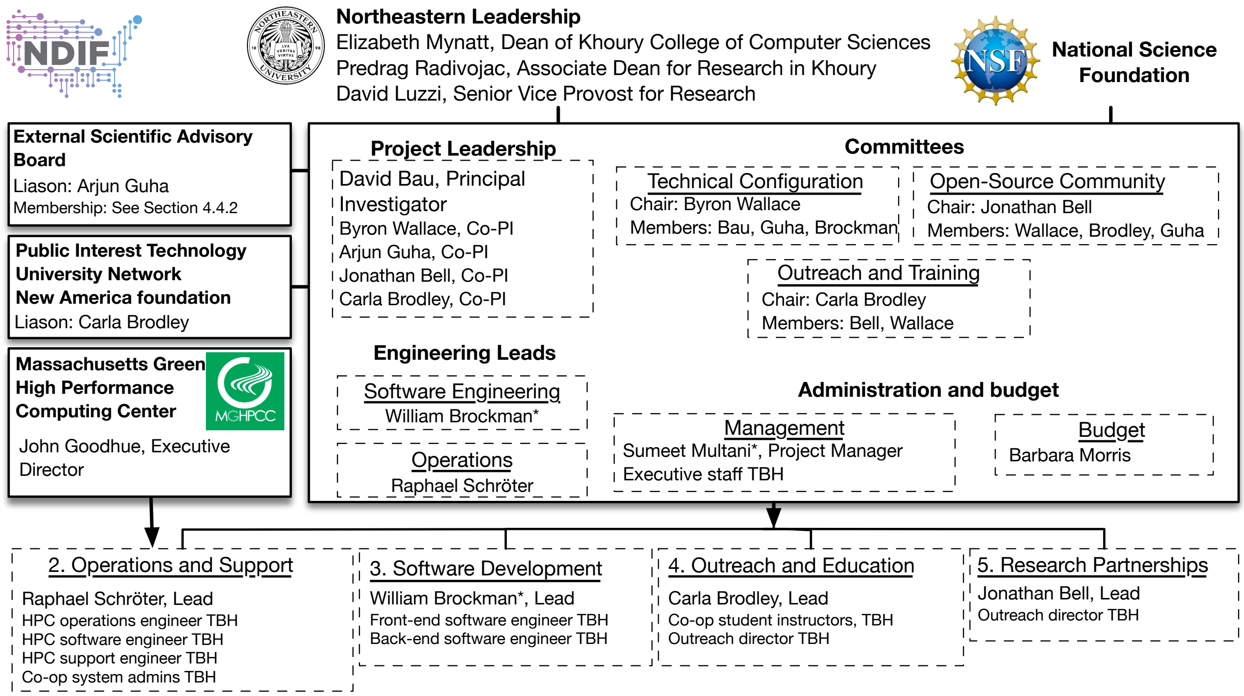


Figure 9: NDIF organization. Asterisks (*) indicate tentative hires.

Google. There, he created and managed the Google Talk and Hangouts team and led the Boston Google Image Search ranking team. He has the track record to implement an infrastructure of this scale, having successfully managed 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 [88–91] and editing of large models [92–94], and he has been a pioneer in the explicit characterization of causal computational mechanisms within language models [45, 70, 71]. As NDIF Director, Bau will hire and manage project leadership, as well as oversee the direction, development, and overall success of the facility. He will work closely with the project manager and lead software engineer to oversee the development of the project, and 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, Founding Executive Director of the Center for Inclusive Computing, NU; 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 ML research has advanced computer science 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.

Co-PI Wallace - Technical Configuration Committee Chair (Sy and Laurie Sternberg Interdisciplinary Associate Professor in Khoury College of Computer Sciences, NU) has extensive research expertise in NLP and interpretability of such models [95–102], as well as their use in biomedical settings [49, 97, 103–105]. Wallace will chair the Technical Configuration Committee, which will meet at least quarterly 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 NLP 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 College of Computer Sciences, NU) brings deep experience in programming languages, including language-based security [106–109], GPU accelerated domain specific languages [110, 111], and pre-trained models for code generation [112]. 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 College of Computer Sciences, NU) is an expert in software engineering and systems, including architectural design [113], testing and continuous integration [114–117], and analysis [118–120]. Bell will chair the Open-Source Community Committee. He will be responsible for incubating the open-source community and overseeing open-source activities, 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 NU) organizes strategic planning for research computing resources at NU; 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 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 at Google, TripAdvisor, and Akamai Technology. As project manager, Multani will work closely with the PI, as well as all four work 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 or project management expert or 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. The advisory board will meet twice per year, once virtually and once in person.

4.2.3 Engaging the public-interest technology research community

The benefits of advances in AI have been realized unequally [121]. To ensure that NDIF enables 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 NDIF by establishing a Public Interest Technology (PIT) Advisory Group comprised of 10-15 interdisciplinary experts to provide guidance on the responsible and ethical design, development, deployment, and use of LLMs. The Advisory Group will include experts from both technical and social sciences.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 related 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 decisions will be made in consultation with the External Scientific Advisory Board and 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 budget) 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 proposed effort 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 capabilities 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 successful deployment of NDIF, the facility will transition to ongoing operations, with the aim will of maintaining the HPC infrastructure to 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 its External Scientific Avisory Board; 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 continues to address their needs. 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 $2.1 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 queues. This program will help to ensure the long-term sustainability of NDIF.

5.2.1 Access and utilization plan

NDIF will be open to all individuals 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 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 ten 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, 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 this 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 that researchers with high computing needs but limited funds are able to enroll as paid partners, NDIF will award "NDIF computing grants" that will provide free access to high-bandwidth queues. We will broadly advertise these compute grants, particularly to early career faculty, faculty in EPSCOR states, and MSI faculty. The outreach group will review and choose grantees based on need and scientific merit. Periodically we will review usage with respect to the proposed policy: if many more users need high capacity than anticipated we will rethink our policies with the goal of providing access to less well-resourced institutions.

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 towards the four goals, and 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 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 interactions 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: LLMs are already being rapidly integrated into consumer products, and are impacting fields outside of CS (e.g., medicine [46]); their impact on society will continue to grow. LLMs have advanced so rapidly that some have called for a temporary pause on LLM development until academic research can catch-up [122]. NDIF will provide the hardware, software, and training necessary for researchers to characterize benefits and risks 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.

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. Northeastern is well-known for experiential learning — every student completes at least one six-month full-time Co-Op — and will build on our existing efforts to recruiting students to develop software.

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 NDIF, 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 [123, 124]. Thus throughout all outreach we will ensure that we are reaching a diverse set of institutions, researchers and students, with a focus on reaching early-career faculty, and professors in EPSCOR states, MSIs, PUIs, and CCs. Co-PI Brodley, who is a nationally recognized expert in broadening participation in computing [125], 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 Center for Inclusive Computing (led by co-PI Brodley).

Developing national expertise We will design training modules to help onboard new researchers 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. with hands-on access to the experts who build and maintain NDIF. In year three we will expand this bootcamp to reach over 300 students in six different geographic regions, leveraging NU's campus network (NU offers programs at nine global campuses) and two university partners, with a focus on cities with a major airport hub; we will run six bootcamps during the summer of 2026 in Oakland, Miami, Washington DC, Dallas, Chicago, and Maine. 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 and impact; in awarding these we will prioritize EPSCoR states, MSIs, and PUCs. Additionally we will offer one-day workshops using the in-person tutorials co-organized with major machine learning conferences (e.g., 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 After in-person events, training and mentoring will continue virtually. This virtual community will provide space for researchers to learn more about NDIF, and to showcase and discuss ongoing research on NDIF. While preparing this proposal, we created the NDIF virtual community using the *Discord* platform. In its first month of operation, this platform brought together 43 researchers from across the country to discuss the design and use-cases for NDIF. We will organize an Annual NDIF 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—we will use the public issues tool to gather and track user issues. We will integrate this virtual community with our in-person training to broaden the availability of NDIF and reduce barriers to its adoption.

6.2 Undergraduate education

We are committed to ensuring that undergraduates in the U.S. benefit from NDIF. 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 NU (e.g., Machine Learning I and II, NLP, and Neural Networks). PI Bau and Co-PI 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 materials—which will use the hosted NDIF API developed under this project—publicly available, and we will support their use by to faculty at other institutions, scaling the impact by enabling U.S. undergraduates in CS to gain hands-on experience analyzing and working with the internals of LLMs. This is not currently possible at the vast majority of institutions given the resources required to run such models (and the limited access to model internals that commercial APIs provide, as discussed above). As discussed in Section 6.1 we will ensure that we support universities/colleges across the country with a particular focus on outreach to EPSCoR jurisdictions and a diversity of institutions, including SLACs and MSIs.

7 Institutional Commitment to Inclusion

Khoury College of Computer Sciences is a leader in broadening participation in CS. Khoury is home to the Center for Inclusive Computing (CIC) [126]), which aims to increase the representation of women of all races and ethnicities majoring in CS across the U.S. The CIC works with 100+ domestic institutions to remove institutional barriers to students discovering and excelling in computing. Under co-PI Brodley's leadership, Khoury piloted and scaled the Align Master's (MS) in Computer Science program [127, 128], 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 were 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 [129], a network of 23 institutions now offering the MS in CS for non-majors. Khoury also has a verified college-wide broadening participation in computing plan [130].

8 Divestment

NU 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 [49, 97, 99, 103–105, 131]. **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 [115], understanding flaky tests [132–134], and making CI builds faster [135]. **Broader Impact:** This project has resulted in significant technology transfers to popular open-source projects Apache Maven [135] and Pitest [133], and creation of educational materials for CI [136, 137].

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 [110, 138–141]. Wasm/k [139] 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, Hanaa Albanna, Mousa Ahmad Albashrawi, Indranil Bose, Lawrence Brooks, Buhalis Dimitrios, Lemuria Carter, Soumyadeb Chowdhury, Tom Crick, Scott W. Cunningham, Gareth H. Davies, Robert M. Davison, Rahul Dé, Denis Dennehy, Yanqing Duan, Rameshwar Dubey, Rohita Dwivedi, Marijn Janssen, Paul Jones, Iris Junglas, Sangeeta Khorana, Sascha Krause, Kai R. Larsen, Paul Latreille, Sven Laumer, F. Tegwen Malik, Abbas Mardani, Marcello Mariani, Sunil Mithas, Emmanuel Mogaji, Jeretta Horn Nord, Siobhan O'Connor, Fevzi Okumus, Margherita Pagani, Neeraj Pandey, Savvas Papagiannidis, Ilias O. Papas, Jan Pathak Nishith Pries-Heje, Ramakrishnan Raman, Nripendra P. Rana, Sven-Volker Rehm, Samuel Ribeiro-Navarrete, Alexander Richter, Franz Rowe, Suprateek Sarker, Bernd Carsten Stahl, Manoj Kumar Tiwari, Wil van der Aalst, Viswanath Venkatesh, Giampaolo Viglia, Michael Wade, Paul Walton, Wirtz Jochen, and Ryan Wright. ""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, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. "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, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe P Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H Guss,

Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. "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] Future of Life Institute. *Pause Giant AI Experiments: an Open Letter*. `https://futureoflife.org/open-letter/`. 2023.

[10] *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`.

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

[12] Kevin Roose. "Why A Conversation With Bing's Chatbot Left Me Deeply Unsettled". In: *New York Times* (2023). URL: `https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html`.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[14] 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.

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

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

[17] 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.

[18] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault. "The AI index 2021 annual report". In: (2022). URL: `https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf`.

[19] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. "OPT: Open pre-trained transformer language models". In: *arXiv preprint arXiv:2205.01068* (2022).

[20] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Sasko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton

Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. "Bloom: A 176b-parameter open-access multilingual language model". In: *arXiv preprint arXiv:2211.05100* (2022).

[21] Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. "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.

[22] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. "Emergent Abilities of Large Language Models". In: *Transactions on Machine Learning Research* (2022). Survey Certification. ISSN: 2835-8856. URL: `https://openreview.net/forum?id=yzkSU5zdwD`.

[23] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. "Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 346–361.

[24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. "Chain of thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems*. 2022.

[25] 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, and Niels Peek. "Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries". In: *Journal of the American Medical Informatics Association* 28.10 (Aug. 2021), pp. 2128–2138. ISSN: 1527-974X. DOI: `10.1093/jamia/ocab127`. eprint: `https://academic.oup.com/jamia/article-pdf/28/10/2128/40408843/ocab127\_supplementary\_data.pdf`. URL: `https://doi.org/10.1093/jamia/ocab127`.

[26] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Elhage, Nelson, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Catherine Olsson Kamal Ndousse, Daniela Amodei, Dario Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Chris Olah, and Jack Clark. "Predictability and surprise in large generative models". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1747–1764.

[27] Prasann Singhal, Jarad Forristal, Xi Ye, and Greg Durrett. "Assessing Out-of-Domain Language Model Performance from Few Examples". In: *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2385–2397. URL: `https://aclanthology.org/2023.eacl-main.175`.

[28] 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.

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

[30] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.

[31] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386.

[32] Jeffrey L Elman. "Finding structure in time". In: *Cognitive science* 14.2 (1990), pp. 179–211.

[33] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. "A neural probabilistic language model". In: *Advances in neural information processing systems* 13 (2000).

[34] Ruth C Fong and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3429–3437.

[35] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. "In-context learning and induction heads". In: *arXiv preprint arXiv:2209.11895* (2022).

[36] Grace W Lindsay, Daniel B Rubin, and Kenneth D Miller. "A unified circuit model of attention: neural and behavioral effects". In: *bioRxiv* (2019), pp. 2019–12.

[37] 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.

[38] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. "Probing Pretrained Language Models for Lexical Semantics". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7222–7240.

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

[40] John Hewitt and Percy Liang. "Designing and Interpreting Probes with Control Tasks". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2733–2743. DOI: 10.18653/v1/D19-1275. URL: https://aclanthology.org/D19-1275.

[41] 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.

[42] 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.

[43] Yonatan Belinkov. "Probing classifiers: Promises, shortcomings, and advances". In: *Computational Linguistics* 48.1 (2022), pp. 207–219.

[44] Tiwalayo Eisape, Vineet Gangireddy, Roger Levy, and Yoon Kim. "Probing for Incremental Parse States in Autoregressive Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2801–2813. URL: https://aclanthology.org/2022.findings-emnlp.203.

[45] 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).

[46] 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).

[47] 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.

[48] 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.

[49] 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.

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

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

[52] Leon Fröhling and Arkaitz Zubiaga. "Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover". In: *PeerJ Computer Science* 7 (2021), e443.

[53] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. "Detectgpt: Zero-shot machine-generated text detection using probability curvature". In: *arXiv preprint arXiv:2301.11305* (2023).

[54] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. "Discovering Latent Knowledge in Language Models Without Supervision". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=ETKGuby0hcs.

[55] Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei Wei, David Wu, Hugh Zhang, and Markus Zijlstra. "Human-level play in the game of Diplomacy by combining language models with strategic reasoning". In: *Science* 378.6624 (2022), pp. 1067–1074.

[56] 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).

[57] 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).

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

[59] 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).

[60] John Hewitt and Christopher D Manning. "A structural probe for finding syntax in word representations". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2019, pp. 4129–4138.

[61] 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 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1073–1094. DOI: 10.18653/v1/N19-1112. URL: https://aclanthology.org/N19-1112.

[62] 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).

[63] 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.

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

[65] Jesse Vig and Yonatan Belinkov. “Analyzing the Structure of Attention in a Transformer Language Model”. In: *Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 63–76. DOI: `10.18653/v1/W19-4808`. URL: `https://aclanthology.org/W19-4808`.

[66] Samira Abnar and Willem Zuidema. “Quantifying Attention Flow in Transformers”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Online: Association for Computational Linguistics, July 2020, pp. 4190–4197. DOI: `10.18653/v1/2020.acl-main.385`. URL: `https://aclanthology.org/2020.acl-main.385`.

[67] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. “Language models (mostly) know what they know”. In: *arXiv preprint arXiv:2207.05221* (2022).

[68] 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.

[69] Kevin Ro 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: *The Eleventh International Conference on Learning Representations*. 2023. URL: `https://openreview.net/forum?id=NpsVSN6o4ul`.

[70] 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.

[71] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. “Mass-Editing Memory in a Transformer”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: `https://openreview.net/forum?id=MkbcAHIYgyS`.

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

[73] 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 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`.

[74] 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`.

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

[76] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[77] 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]`.

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

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

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

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

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

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

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

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

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

[87] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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`.

[96] 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`.

[97] 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.

[98] 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.

[99] 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`.

[100] 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.

[101] 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.

[102] 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.

[103] 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`.

[104] 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.

[105] 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.

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

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

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

[109] 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.

[110] 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.

[111] 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.

[112] 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`.

[113] 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`.

[114] 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`.

[115] 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.

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

[117] 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.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

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

[123] 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/.

[124] 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/.

[125] *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.

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

[127] 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 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.

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

[129] 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.

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

[131] 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.

[132] 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`.

[133] 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`.

[134] 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`.

[135] 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`.

[136] 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.

[137] 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.

[138] 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.

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

[140] 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`.

[141] 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*.