Overview

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

Large language models (LLMs) such as ChatGPT that comprise more than 100 billion parameters have ushered in an exciting new era of artificial intelligence (AI). These models exhibit new capabilities, including aspects of general-purpose reasoning, that raise **fundamental scientific questions** that impact not only computer science, but also biology, social science, business, engineering, and education. But LLMs are so large that they cannot be run in inference (i.e., to make predictions) with computing resources available to academics, and it is infeasible to create the needed capabilities at an institutional level. Thus, researchers are hindered in their ability to anticipate, explain, and regulate these systems. The proposed **National Deep Inference Facility (NDIF)** will advance scientific understanding by providing U.S. academic researchers access to a cutting-edge computing service capable of running very large language models while providing transparency to their internal computations—a capability not currently available to academics. NDIF will be developed by a team of researchers at Northeastern University in close collaboration with the NCSA DeltaAI project, which will provide computing resources. **DeltaAI** is the highest-capacity AI-focused cluster in the NSF computing portfolio, combining high-performance GPU nodes and a high-speed storage fabric. Utilizing DeltaAI, the NDIF project will design, develop, and deploy new software infrastructure for LLM inference to address urgent research and societal needs for transparency as a means to advance safe, robust, trustworthy, 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,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 their potential biases, errors, and unknown goals will become part 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.

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 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, without robust 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 deep inference research computing facility for LLMs is necessary due to the unique computational needs of large-model inference research. Performing LLM inference consumes quadrillions of parallel computations in a fraction of a second, requiring both (1) high-performance parallel GPU computing capacity, beyond the scale that is feasible at an institutional level, and (2) software infrastructure to enable scientists to share those high-capacity computing nodes for very brief experiments. Neither existing HPC clusters nor commercial inference services address these needs. HPC clusters do not scale to thousands of simultaneous inference users; and commercial inference services hide internals of the LLMs (Figure 1a), making them unsuitable for research. **The National Deep Inference Facility (NDIF) will enable scientific interrogation of LLM mechanisms by providing a unique scalable transparent deep inference service (Figure 1b), harnessing the high-performance GPU capacity of the NSF DeltaAI project (Figure 1c) to advance urgently needed understanding of *how* LLMs work so industry, government, researchers and the public are able to safely deploy, regulate, use, and study them for the benefit of society.**

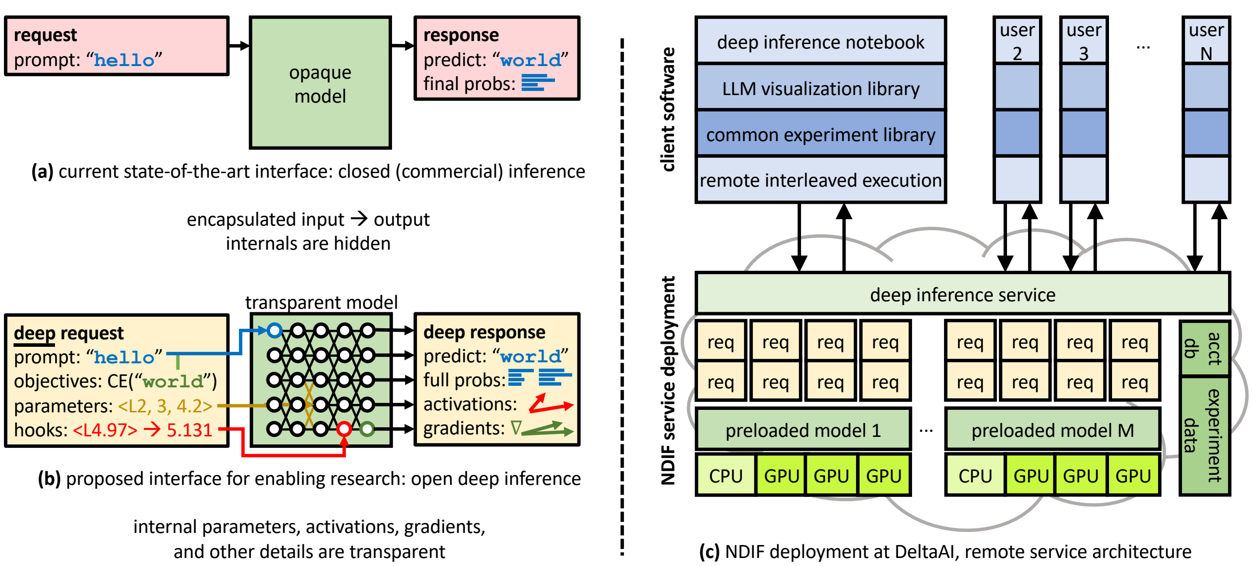

Figure 1: (a) Current services hosting large language models provide very limited interaction functionality (Top). One can send input text in a request, and is then provided an output string (and in some cases scores for predictions). (b) We propose developing infrastructure to provide deep access *to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

NDIF will run on the NSF-supported NCSA DeltaAI project, a recently-funded AI-focused computing resource that consists of a large and uniform pool of compute nodes that provide the high-memory GPU configurations that are necessary for the highly-parameterized massively parallel computations of LLM research. 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 deployment of an online **inference service** hosted at the NSF-supported NCSA DeltaAI cluster, allowing researchers to interrogate and conduct ground-breaking research on the largest available and most scientifically relevant LLMs. (Figure 1c).
2. Development of **open-source server and client software** that will power the service, enabling convenient, transparent and scalable remote execution of experiment protocols by a broad community of researchers, efficiently sharing large models at the online inference service.
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, and inclusive computing at Northeastern University (NU). The team will benefit from the university's well-established organizational structure and advanced facilities. The deployment of the service will be done in close collaboration with the DeltaAI project at National Center for Supercomputing Applications (NCSA) at University of Illinois Urbana-Champaign, who will provision and host the computational resources (see attached letter of collaboration from DeltaAI).

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*." In 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 (from diverse prospectives) highlight the shared urgency for research to explain, audit, evaluate, and manage impacts of large-scale AI.

Meanwhile, **LLMs such as ChatGPT are being adopted more quickly than any previous technology**, with widespread deployment in consumer-facing technologies [12–14], touching every field involving reading, writing, or programming, even as its mechanisms remain unexplained [2, 15]. 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.

Deep inference has unique computational demands. The computing required to support research into LLM inference is not well-supported by either traditional HPC provisioning nor existing commercial inference services. Commercial inference services provide no access to model internals and are therefore unsuited to scientific research. HPC batch allocations, which could provide full access, partition computational resources among users by giving a user exclusive use of capacity for a period of time. This precludes thousands of researchers from performing concurrent experiments on the same models on shared computers. NDIF will solve these critical needs by aggregating together many different researchers' requests to efficiently use existing HPC resources at DeltaAI. Unlike commercial inference services, NDIF will have a unique design, informed by the community of LLM researchers, that will provide the ability to perform inference together with experimental code that can inspect, analyze, and modify every aspect of LLM computations (Figure 2), enabling a wide range of scientific research.

2.1 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 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.

> **NDIF 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 LLMs could transform how such models are used, developed, and regulated. Explaining LLMs is challenging because they are massive artificial neural networks, i.e., computational systems loosely inspired by human neurons [16, 17], with connection strengths determined by an data-driven training process [18, 19]. Because LLMs are not programmed explicitly, the only path to an explicit understanding of their calculations is reverse-engineering their internal computations, which is challenging due to their complexity. The scientist community involved in NDIF design includes experts on methods for understanding both artificial [20, 21] and biological [22] neural networks, part of a growing community investigating the inner-workings of LLMs (e.g., the BlackboxNLP workshop [23]). We have worked with this community to identify capabilities of the NDIF that would empower work on cutting-edge experimental methods, as discussed in Section 2.2.

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, potentially revealing the degree to which models encode bias [24, 25], sentiment [26], linguistic knowledge [27–29], truthfulness [30, 31], and many other kinds of information [32–34]. By providing the unique capability to apply representation analysis to LLM internal states, NDIF will allow scientists to extend 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 [35]. For example, high-stakes settings raise critical evaluation issues in applications such as detecing dementia [36], measuring fairness [37, 38], and studying risks of potentially sensitive personal health training data [39]. 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. In contrast, NDIF will provide capabilities needed for rigorous evaluation, including complete access to output probabilities, internals, and the ability to fine-tune and evaluate models transparently.

d: Creating tools to improve AI safety and manage AI risks. NDIF will enable the development of tools that could be used to mitigate the negative impacts of LLMs, enabling research into

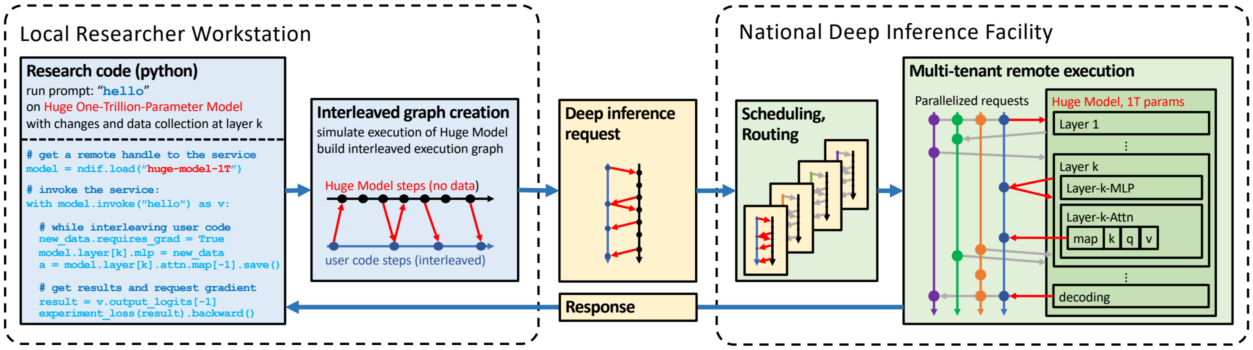

Figure 2: Details of the logical view of a deep inference request. Unlike commercial inference that provides no transparency, the NDIF will allow researchers to execute flexible experiments by interleaving their own code within the internals of the deep network inference process. To maintain safe and efficient co-tenancy, experiments are submitted as computation graphs that enable resource accounting and scheduling.

both short-term risks and long-term risks posed by very large AI models. NDIF will enable research into the detection of machine-generated misinformation [40–42], tools that could detect and mitigate untruths or deception in a model's behavior [30, 31, 43], and methods to erase undesired behavior [44] or improve stability of behavior over long time horizons [45]. Such research requires full access to posteriors and activations, as uniquely provided by NDIF.

e: Enabling research on societal impacts. Understanding social impacts requires studying interactions between LLMs and people. For example, our users include social scientists interested in how people behave differently when talking to a chatbot, and in how persuasive LLMs are. Already, LLMs have been used to judge public opinion [46], to measure political ideology and latent constructs [47], and to analyze text as data [48]. But social scientists tell us that commercial APIs do not provide the transparency that would allow reproducible research. NDIF will provide an environment for ethical human subjects research and will provide both the technical capabilities to support human studies and a process to allow protocols to be overseen by a researcher's IRB.

2.2 Experimental methods uniquely enabled by the NDIF service model

To enable scientists to advance the research agenda in Section 2.1, NDIF will enable a wide range of experimental methods in a **flexible unified framework, allowing researchers to interleave their own experiment code in large models hosted remotely**. This will allow researchers to gather and analyze every aspect of the neural network's internal data (Figure 2). The unified framework will enable critical technical experimental methods that are unavailable via commercial LLM inference services, including *representation probing* [32] and other forms of *representation engineering* [49] such as *causal alignment mapping* [50, 51], *sparse dictionary learning* [52–54], other trained probes of semantics and structure [55–57], and individual neuron studies [26, 58–61]. NDIF will also support *salience mapping* via *gradient methods* [62, 63], as well as access to *attention distributions* [64–66] and *per-token probabilities* [67]. Access to activations will also enable neuroscience-inspired methods such as *representational similarity analysis* [68, 69] and *latent factor analysis* [70–72].

The same framework will also allow direct counterfactual interventions into model internals, which will also enable *causal mediation analysis* at the level of individual neurons [73, 74], representation vectors [75, 76], or subspaces [50, 51]; these methods will also enable *circuit-finding methods* [77] such as *path patching* and *activation patching* [78–80]. The service will also support gradient-based optimization of inserted parameters, to enable *parameter-efficient fine-tuning* (PEFT), which will enable investigations into the intriguing transfer-learning capabilities of LLMs through *soft prompts* and *adapter layers* [81–83]. The same optimization support will also be a tool for creating representation probes [32] and for other training-based experimental methods [50–52, 54].

2.3 NDIF will leverage all 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, after they are trained. NDIF will integrate with every pretrained model that makes its parameters available to academic researchers, adding specialized support for the most popular models. We will collaborate and support ongoing efforts to encourage and deploy new large open models. We will integrate with: (1) **EleutherAI GPT-NeoX** [84], **GPT-J** [85], and **Pythia** [86], 6-to-20 billion parameter LLMs with fully reproducible training data, trained by the EleutherAI nonprofit. (2) **Meta OPT** [87], **Llama** [88], and **Llama 2** [89], sets of models up to 65-to-175-billion parameters trained and licensed by Meta, with parameters available to academic researchers. (3) **AI2 OLMo**, a 70-billion parameter family of multimodal models trained by the Allen Institute for AI planned for 2024. (4) **BigScience Bloom** [90], a 176-billion parameter multilingual model trained by BigScience, a collaboration of European agencies. (5) Model variants fine-tuned with for conversation or multimodal use, including BigScience Bloomz [91], CarperAI [92], Alpaca [93], Vicuna [94], and OpenFlamingo [95]. (6) Ongoing work by the National AI Research Resource (NAIRR) [10], LAION [96], MosaicML [97], and Together [98]. We anticipate many additional publicly available LLMs each year; our configuration committee and scientific advisory board will prioritize model support for maximum scientific impact.

3 Research Infrastucture Development

3.1 Technical readiness

Our team has created a detailed Project Execution Plan (PEP) that delineates requirements, design, and deployment milestones for the NDIF's user model, user-facing and internal software, training, and outreach. We have created a prototype implementation that includes the core interleaved execution framework that will allow the NDIF to efficiently support a wide range of research methods, and we have created an set of online tutorials that demonstrate use of NDIF to reproduce recent scientific results. Testing is underway with the prototype in use by five pilot users conducting research with the system. Please see our attached PEP and Technical Platform Specification for complete details on the development schedule, project management plan, and system design.

3.2 Major deployment milestones

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

Closed pilot, Q1 2025. In the first phase, we will create a web API that can support interleaved-execution research queries observing and intervening in inference for the largest models with open parameters (70b-175b parameters), with both streaming and batch-oriented use patterns, supporting all forward-pass experimental methods. The goal of the project at this stage is to develop the system while validating the system design and user model with a small set of early adopters drawn from our design-participant user community. These early users will work directly with our team and will provide direct design feedback. The team will produce documentation, tooling, and support sufficient to meet pilot user needs. The pilot will culminate with a usage demonstration to teach 100 users to use the system at a major conference tutorial session. Hardware needs: 2 CPU servers and 10 4x A40 GPU nodes, which are already currently deployed at NCSA.

Nationwide open pilot, Q1 2026 (Year 2). In the second phase, we will make the service available openly for early access to all qualified users at any educational institution. Achieving broad usage will require us to provide a stable service with increased robustness, including monitoring and alerting systems, job queuing, and a fairness-oriented scheduler. In this phase we will define Service Level Objectives (SLOs) and measure progress toward meeting them. We will expand functionality to include backpropagation experimental methods and deploy reproducible fine-tuning and optimization functionality. The API will be stable and documented well enough for online self-service, and we will flesh out online tutorials demonstrating all major research methods. We will teach usage of the service at a large local workshop, and demonstrate the service at a large conference tutorial. Hardware capacity needs in year 2 and beyond will be driven by demand; we estimate 4 CPU servers and 10 4x A40 GPU nodes, currently existing, plus 5 4x H100 nodes for large models, growing to 10-20 4x H100 nodes, to be provisioned from DeltaAI based on demand.

National outreach bootcamp and software API full release, Q1 2027 (Year 3). The third phase is concentrated on broadening usage, which will require a high degree of stability, learnability, and functionality. All major user-facing features of the system will meet SLOs, which will require development of refined monitoring and administrative tooling. Support will be added for dedicated allocation for heavy users, to reduce contention on the public queue. User-facing documentation must be complete and tested, and virtual and in-person bootcamp curriculum will be developed and delivered. Over 300 researchers from diverse institutions will be recruited and taught to use the system in a multi-site bootcamp. Capacity in years 3 and beyond will be planned in close coordination with the NSF and NCSA, based on an assessment of community needs.

Operations scale-up, Q1 2028 (Year 4). The fourth phase focuses on scaling the system to support new and larger models to match the state-of-the-art. At this stage, the NDIF will be pivotal in enabling new research on the largest models: its availability should encourage the training of larger-scale academic models, since NDIF will allow researchers to routinely be able to study and reproduce large experiments. We also will continue to expand capacity to meet demand from new users, and we will measure and refine the user experience to improve our ability to efficiently onboard and support new users. We will improve system administration tools to improve issue response and provide a high level of stability. And we will pilot ability to run NDIF workloads on other clusters, to provide a higher level of availability than can be achieved in a single cluster.

4 Metrics and Annual Goals

We will continuously monitor metrics to track the progress of the NDIF project, including (1) Users by amount and diversity (2) Training, outreach, and educational programs delivered (3) Specific research capabilities deployed. The purpose of NDIF is to broadly enable LLM research to enable transformative science and workforce development; therefore the primary goal each year is to enlarge and broaden usage, and to expand research capabilities and outreach to enable that goal. Table 1 summarises high-level goals per-year.

We also summarise the major capability goals for the platform each year. *Year 1:* all major forward-pass inference and intervention methods supported on all open models up to 175 billion parameters. *Year 2:* Parameter-efficient fine-tuning and optimization. Usage accounting and fairness-oriented scheduling. *Year 3:* Support users with heavy usage patterns through approved allocation process. 90% or better continuous uptime, and a set of SLOs are defined, monitored, and met. *Year 4:* Capacity meets current state-of-the-art models and user demand. 90% or better

Metric		Year 1	Year 2	Year 3	Year 4
Users	Active per-month	20	100	1,000	2,000
	States represented	3	20	30	50
Training	Conference tutorial attendees	50	100	100	100
	Week-long bootcamp attendees		100	300, multi-city	100
	Online curriculum	Prototype	Deployed	Revised	Revised

Table 1: Annual goals for high-level metrics. This is a non-exhaustive list of metrics that we will track, review and report. Section 3.9 of the PEP provides a detailed breakdown of milestones.

continuous uptime, and other SLOs are met. The availability of NDIF has enabled major research findings and should be cited in motivation of release of 10^{13}-scale models for academic study.

In addition to the above metrics and goals, throughout the project we we will monitor user diversity, research citations, user feedback, issue response, feature usage, system performance, and system availability. Section 9.11 details our overall project evaluation process.

5 Coordinated Scale-out with NCSA Delta and DeltaAI

We will develop and deploy NDIF utilizing HPC resources operated by the DeltaAI project at the National Center for Supercomputing Applications (NCSA), and we will coordinate deployment with them via biweekly meetings (see the attached letter of collaboration from NCSA director William Gropp). DeltaAI is building computing capacity that is well-suited to NDIF. Together we have estimated an appropriate allocation to this project as part of our proposal. Our initial capacity needs for the first year (estimate: 2 CPU servers and 10 4x A40 GPU nodes) are already available through the Delta cluster. By year 2, we expect to need larger GPU nodes (5-20 4xH100 nodes), which will be built as part of the recent NSF DeltaAI award. We will scale capacity up and down as needed to support demand. For capacity in years 3 and beyond, we will assess demand and current state-of-the-science and discuss with the NSF how best to meet needs of the community, including whether there is a need for additional nodes in DeltaAI or in the NCSA storage environment (for example higher-VRAM nodes or next-generation interconnects). Initial development resources will be allocated through the ACCESS program, and we will work closely with the NSF and NCSA to determine the appropriate allocation method for the service in deployment. Computational capacity will be provided by DeltaAI and is not included in the budget for this project. Each quarter we will track and report utilization metrics to support discussions with DeltaAI and the NSF on how best to meet computational capacity needs of the research community.

6 Outreach, Training and Dissemination Plan

The NDIF will strengthen the US Artificial Intelligence Research & Development ecosystem by providing training and support to the US scientific community to ensure that the infrastructure is accessible and usable, and to address 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. Our training, outreach, and dissemination plan consists of the following:

Nationwide online curriculum We will design training modules to help onboard new researchers and students to NDIF. Modules will cover topics such as: 1. Understanding large foundation models. How to visualize and understand their internal operation using NDIF. 2. How to apply deep inference methodologies such as representation analysis, salience mapping, causal mediation analysis, circuit analysis, and parameter-efficient fine tuning. How to use NDIF to apply these analysis methods. 3. Recent scientific results on auditing, evaluating, and controlling behavior of LLMs with respect to interpretability, bias, safety, robustness, or other task performance. How to reproduce recent research results and extend experimental methods using NDIF.

NDIF week-long bootcamp workshops (Years 2, 3, 4) We will develop and deliver an intensive week-long bootcamp workshop program which will provide graduate students studying in the U.S. with a rapid survey of cutting-edge LLM methods, taking advantage of NDIF infrastrcuture. The purpose of the workshop will be to quickly upskill participants in LLM research methods, taking advantage of NDIF infrastructure. The workshop will begin with LLM fundamentals and culminate with coverage of current research results in LLM interpretability, evaluation, control, and applications. In years two and four, the workshop will be held in Boston, hosting 100 students. In year three, we will conduct a multi-city bootcamp series to reach over 300 students in six geographic regions, leveraging NU's campus network and two university partners, in Oakland, Miami, Washington DC, Dallas, Chicago, and Maine. The cost of attending the bootcamp will be free (leveraging the NU campus network) and will be led primarily by Northeastern PhD students with co-PI Bell and co-PI Guha in attendance. For graduate students whose advisors do not have budget to cover the travel costs we have budgeted a $150k fund to support travel based on need and impact; in awarding these we will prioritize EPSCoR states, MSIs, and PUCs.

Machine learning conference tutorial series We will also 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 one conference each 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.

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. Once developed, we will make materials—which will use the hosted NDIF API—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). We will ensure that we support universities and colleges across the country with a particular focus on outreach to EPSCoR jurisdictions and a diversity of institutions, including SLACs and MSIs.

Democratic and equitable outreach 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 [99, 100]. 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 [101], will lead this effort. We will recruit potential using popular social media channels such as X, 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 already participate in initiatives run by the Center for Inclusive Computing (led by co-PI Brodley).

7 User Engagement and Open-Source Community

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.”

7.1 Gathering user requirements through virtual and open-source communities

Based on strong and broad interest, 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 3), 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 both inside and outside computer science. Their discussions have informed the required research capabilities (Section 2.2) and design of NDIF.

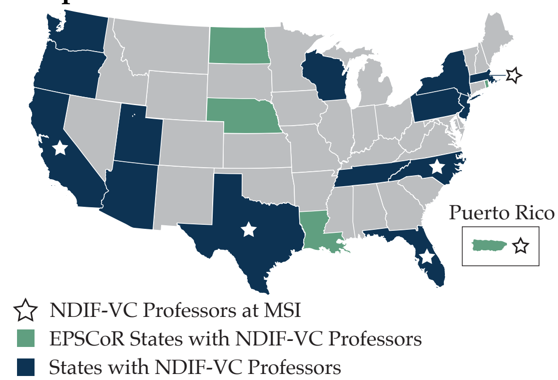

Figure 3: National reach: Our existing virtual community (VC) includes faculty at 34 universities in 19 states (5 EPSCoR jurisdictions) who have proposed specific research to be done with NDIF.

To support this virtual community, we maintain a *Discord* server, bringing together researchers from across the country for both asynchronous and realtime discussions. We also maintain a website with reference materials, tutorials, and examples [102], and share the open-source codebase on GitHub with full code and technical specifications. As NDIF is deployed, we will welcome participants into this community and 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 use the GitHub public issues tool as well as the Discord virtual community to gather, track, and discuss user issues. The open source committee will meet quarterly to review, identify and summarize themes from the virtual and open-source community, and these findings will be reported to NDIF leadership to help inform the technical and policy direction of NDIF.

8 NDIF User Access

The research community will access NDIF in several ways, described below:

NDIF website usage The NDIF service will offer a web interface open to any person 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. Once a user is enrolled, they will have access to a code-free web interface that allows direct chat-style interactions with large models, and that also enables *Logit Lens* visualzations [103–105] of language model internal states during inference.

NDIF API usage Most users of NDIF will use NDIF through the API and python client library. After establishing an account on the website, users can get an API key that can be used together with the NDIF python library, that allows them to write code that performs deep inference on NDIF-hosted models remotely. The library is more flexible and powerful than the web interface alone, because it provides an interleaved-execution context (Figure 2) that allows a researcher to interleave custom experiment code between the steps of a large model, providing full access to model internals and interventions, with a concise and highly reproducible idiom. Programmatic requests can be submitted and fulfilled either serially in best-effort-real-time, or queued in batches of requests whose results are returned asynchronously.

Local API usage The client library will be open source and available to be used freely by all without an NDIF account. As illustrated in Figure 4, the library can be used for local experiments on any pytorch model; it will enable interleaved local model experiments using the same idioms used for huge models hosted by NDIF. The purpose of this design is to provide a smooth "on ramp" where students and researchers can learn and develop their ideas locally, and then use exactly the same code to conduct huge remote model experiments. Since the client library is designed to improve researcher quality-of-life even for small experiments, we aim for and expect broad usage of the library even in client-only mode.

High-volume usage To support high-volume use and to efficiently utilize underlying DeltaAI resources in both high- and low-load conditions, NDIF will design and implement a provisioning technology in collaboration with TAPIS open-source ecosystem, autoscaling Kubernetes containers on Slurm. This will integrate with ACCESS provisioning to accommodate approved users with heavy usage patterns. We will design this system in collaboration with the TAPIS team, meeting with them regularly to consult on the usage of TAPIS in our application. (See letter of collaboration from Joe Stubbs, Development Manager of TAPIS.)

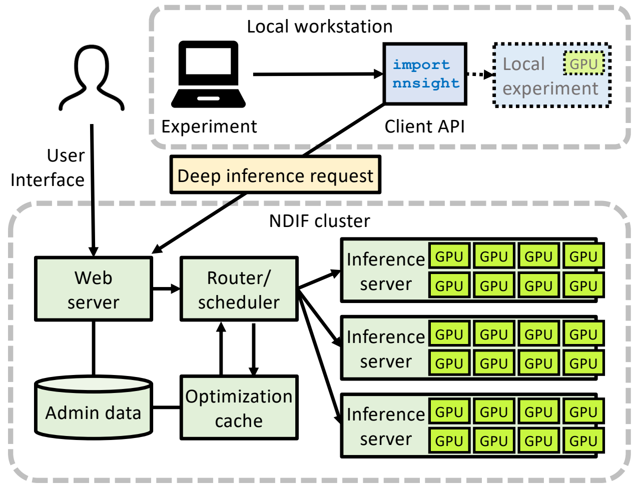

Figure 4: NDIF service architecture, showing request flows between system components.

9 Project Management, Operations and Utilization

9.1 Key Personnel

Figure 5 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 Google. There, he created and managed the Google Talk team and led Boston Google Image Search ranking. 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. Since transitioning to academia, PI Bau has established himself as a leading researcher in interpretability of large neural networks [58, 106–108] and model editing [44, 109–112], and he has been a pioneer in the characterization of causal mechanisms within LLMs [34, 75, 76, 105, 113, 114]. As NDIF Director, Bau will hire and manage project leadership and oversee the overall success of the facility. He will work closely with the project manager and lead software engineer to manage the project and will run monthly meetings of NDIF leadership. 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 1.3) 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

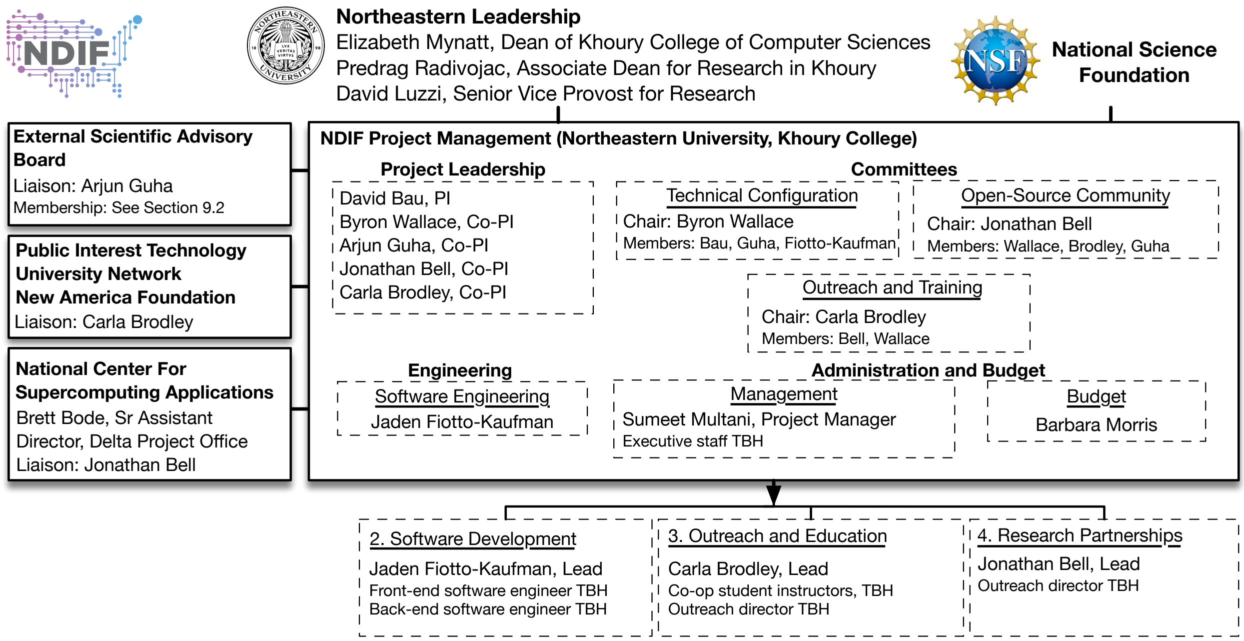

Figure 5: NDIF organization

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 [115–122], as well as their use in biomedical settings [39, 117, 123–125]. 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 [126–129], GPU accelerated domain specific languages [130, 131], and pre-trained models for code generation [132]. 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 Delta AI Liason, Co-PI Bell (Assistant Professor in Khoury College of Computer Sciences, NU) is an expert in software engineering and systems, including architectural design [133], testing and continuous integration [134–137], and analysis [138–140]. As chair of the Open-Source Community Committee, Bell 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. Bell will also serve as liaison to the Delta AI project, ensuring coordination and communication with that team with respect to technical design and capacity. He will also serve on the Outreach and Training Committee.

Project Manager, Multani 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, and all work areas of the project (WBS 1.1-1.4). He will be responsible for management of schedule, budget, scope, and risk. Additionally he will be responsible for all NSF reporting, including quarterly reports, annual reports, and periodic PEP updates.

Lead Software Engineer, Fiotto-Kaufman has spearheaded the technical design and prototype for NDIF. Previously, he served as research scientist and led a wide range of research engineering projects in machine learning and artificial intelligence at Raytheon BBN. As lead engineer, he will be responsible for hiring staff engineers, and for managing the development process (WBS 1.2). He will serve on the Technical Configuration Committee.

9.2 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. One member of the board will represent the Delta AI project. 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.

9.3 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. Agile project mangement methods will be adopted to continuously test the product to iteratively identify and solve problems and to improve the software. The Technical Configuration Committee will conduct an annual review to identify changes in scope and determine where corrections are needed.

9.4 Delta AI / NDIF coordination meeting

The NDIF and Delta AI teams will meet regularly on a biweekly basis, to coordinate technical design and operational and capacity planning. The goal of this meeting will be to coordinate on hardware and cluster configuration and allocation decisions, and also to establish channels of communication for and planning outages and for responding to issues.

9.5 Engaging the public-interest technology research community

The benefits of advances in AI have been realized unequally [141]. To ensure that NDIF enables critical assessment of the potential impact of LLMs on education, policy, privacy, and safety 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 from both technical and social sciences, to provide guidance on the responsible and ethical design, development, deployment, and use of LLMs.

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.

9.6 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 $450,000 (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.

9.6.1 Schedule and schedule contigency

The proposed effort will run from 4/1/2024 to 3/31/2028; please find the schedule in the PEP. In each year of the project we schedule a single major deployment release. Uniformly throghout the schedule, we set aside 5% schedule contingency, to allow time for unanticipated integration, quality assurance, or adjustments in scope. In that case, the project manager will conduct a review to adjust scope, timing, and budget of the project.

9.7 Risk management

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. Risk will be tracked over time, and the effectiveness of the risk-management process will be tracked. Project leadership (the PIs, project management, and lead software engineer) will conduct a comprehensive review of risks annually, reviewing each risk to develop mitigation strategies. As the project proceeds, each risk will remain on the risk register until it is closed. 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.

9.8 Configuration Management

Changes for all project specifications other than software, such as specification of required infrastructure capabilities 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, utilizing software version control through git. Changes will go through code review and will be unit tested before being committed. Integration tests will be conducted before deployments, and deployment version numbers and a change log will be maintained.

9.9 Operations management and governance

After successful deployment of NDIF, the facility will transition to ongoing operations, with the goal of maintaining the service infrastructure to ensure its continued availability to the research community. In the operations phase, NDIF will retain its facility director and external advisory board, and an operating staff that will conduct training, support, and issue management.

9.10 Operating costs and funding sources

The annual operating cost of NDIF is estimated to be about $0.5 million, which includes operations personnel and the cost of ongoing user training and support. This cost will be defrayed through future fundraising for research for which NDIF is a critical resource, and through a backstop funding commitment of $1 million from Northeastern Khoury College of Computer Sciences.

9.11 Evaluation

Project leadership will continuously evaluate the progress of the project towards the specific yearly goals enumerated in Section 3.2. In addition, each committee will continuously collect and monitor fine-grained metrics appropriate to its focus: **The outreach and training committee** gathers and analyzes user information including information on the diversity of users and workshop participants across demographic groups, career stages, geographies, and institutions, as well as research output including citations and major works. **The technical configuration committee** measures experiment throughput and latency, uptime, usage rates of capabilities and models, and error rates. **The open-source committee** monitors and analyzes bug reports, community discussions and pull requests, as well as issue response rates. The committee summarizes and reports on top community issues each quarter. **Other operating metrics** will be developed by the team as part of the service development process. All metrics will be tracked by project management continuously through dashboards and reviewed by the director and the advisory board on a semi-annual basis.

10 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 [35]); 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 [142]. As discussed in Section 2, NDIF will provide the computing resources and training necessary for researchers to characterize benefits and risks of LLMs.

Democratic and equitable access to NDIF: Section 6 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). 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.

11 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) [143]), 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 [144, 145], 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 [146], 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 [147].

12 Divestment

At the end of the lifetime of the facility, NDIF will archive its configuration, documentation, and source code in open-source repositories so that they remain available on GitHub.

13 International Collaborators

Our project does not involve international collaboration. International researchers can participate in our virtual communities, and can use NDIF through allocations with a U.S.-based PI.

14 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, $300,000) 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. This project has not yet resulted in publications.

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 [39, 53, 117, 119, 123–125]. **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 [135], understanding flaky tests [148–150], and making CI builds faster [151]. **Broader Impact:** This project has resulted in significant technology transfers to popular open-source projects Apache Maven [151] and Pitest [149], and creation of educational materials for CI [152, 153].

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 [130, 154–157]. Wasm/k [155] 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] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. "A survey of large language models". In: *arXiv preprint arXiv:2303.18223* (2023).

[14] Demetrios Brinkmann. *MLOps Community LLM Survey Report*. July 2023. URL: `https://mlops.community/surveys/llm/`.

[15] Samuel R Bowman. "Eight things to know about large language models". In: *arXiv preprint arXiv:2304.00612* (2023).

[16] 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.

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

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

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

[20] 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.

[21] 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).

[22] 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.

[23] 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`.

[24] 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.

[25] 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.

[26] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. "Learning to generate reviews and discovering sentiment". In: *arXiv preprint arXiv:1704.01444* (2017).

[27] 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.

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

[29] 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`.

[30] 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`.

[31] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. "Inference-Time Intervention: Eliciting Truthful Answers from a Language Model". In: *arXiv preprint arXiv:2306.03341* (2023).

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

[33] 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`.

[34] 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).

[35] 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).

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

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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).

[43] 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.

[44] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. "Erasing Concepts from Diffusion Models". In: *Proceedings of the 2023 IEEE International Conference on Computer Vision*. 2023.

[45] Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. "Stay on topic with classifier-free guidance". In: *arXiv preprint arXiv:2306.17806* (2023).

[46] 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).

[47] 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).

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

[49] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. "Representation engineering: A top-down approach to ai transparency". In: *arXiv preprint arXiv:2310.01405* (2023).

[50] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D Goodman. "Finding alignments between interpretable causal variables and distributed neural representations". In: *arXiv preprint arXiv:2303.02536* (2023).

[51] Zhengxuan Wu, Atticus Geiger, Christopher Potts, and Noah D Goodman. "Interpretability at scale: Identifying causal mechanisms in alpaca". In: *arXiv preprint arXiv:2305.08809* (2023).

[52] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. *Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023*. URL: https://transformer-circuits.pub/2023/monosemantic-features.

[53] 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.

[54] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. "Efficient sparse coding algorithms". In: *Advances in neural information processing systems* 19 (2006).

[55] 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).

[56] 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.

[57] 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.

[58] 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.

[59] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. "Natural Language Descriptions of Deep Visual Features". In: *International Conference on Learning Representations*. 2021.

[60] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. *Language models can explain neurons in language models*. May 2023. URL: https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

[61] Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. "Explaining black box text modules in natural language with language models". In: *arXiv preprint arXiv:2305.09863* (2023).

[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] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. "Representational similarity analysis-connecting the branches of systems neuroscience". In: *Frontiers in systems neuroscience* (2008), p. 4.

[69] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. "Similarity of neural network representations revisited". In: *International conference on machine learning*. PMLR. 2019, pp. 3519–3529.

[70] Jakob H Macke, Lars Buesing, and Maneesh Sahani. "Estimating state and parameters in state space models of spike trains". In: *Advanced state space methods for neural and clinical data* 137 (2015).

[71] Evan Archer, Il Memming Park, Lars Buesing, John Cunningham, and Liam Paninski. "Black box variational inference for state space models". In: *arXiv preprint arXiv:1511.07367* (2015).

[72] David Sussillo, Rafal Jozefowicz, LF Abbott, and Chethan Pandarinath. "Lfads-latent factor analysis via dynamical systems". In: *arXiv preprint arXiv:1608.06315* (2016).

[73] 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.

[74] Aaron Mueller, Yu Xia, and Tal Linzen. "Causal Analysis of Syntactic Agreement Neurons in Multilingual Language Models". In: *26th Conference on Computational Natural Language Learning, CoNLL 2022 collocated and co-organized with EMNLP 2022*. Association for Computational Linguistics (ACL). 2022, pp. 95–109.

[75] 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.

[76] 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`.

[77] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. "A mathematical framework for transformer circuits". In: *Transformer Circuits Thread* 1 (2021). URL: `https://transformer-circuits.pub/2021/framework/index.html`.

[78] 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`.

[79] Michael Hanna, Ollie Liu, and Alexandre Variengien. "How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model". In: *arXiv preprint arXiv:2305.00586* (2023).

[80] Tom Lieberum, Matthew Rahtz, János Kramár, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. "Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla". In: *arXiv preprint arXiv:2307.09458* (2023).

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

[82] 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`.

[83] 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`.

[84] 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.

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

[86] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. "Pythia: A suite for analyzing large language models across training and scaling". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 2397–2430.

[87] 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).

[88] 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).

[89] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[90] 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, Zden$e$k 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).

[91] 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]`.

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

[93] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. *Stanford Alpaca: An Instruction-following LLaMA model*. `https://github.com/tatsu-lab/stanford_alpaca`. 2023.

[94] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. Mar. 2023. URL: `https://lmsys.org/blog/2023-03-30-vicuna/`.

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

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

[97] *Mosaic ML: MPT Foundation Series*. Oct. 2023. URL: `https://www.mosaicml.com/mpt`.

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

[99] 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/`.

[100] 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/`.

[101] *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`.

[102] *NDIF technical documentation: the nnsight API*. Jan. 2023. URL: https://nnsight.net/.

[103] nostalgebraist. *interpreting GPT: the logit lens*. 2020. URL: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens (visited on 06/23/2023).

[104] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. *Eliciting Latent Predictions from Transformers with the Tuned Lens*. 2023. arXiv: 2303.08112 [cs.LG].

[105] Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. Wallace, and David Bau. "Future Lens: Anticipating Subsequent Tokens from a Single Hidden State". In: *SIGNLL Conference on Computational Natural Language Learning (CoNLL)* (2023).

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. "Unified Concept Editing in Diffusion Models". In: *IEEE/CVF Winter Conference on Applications of Computer Vision* (2024).

[113] Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. "Linearity of Relation Decoding in Transformer Language Models". In: (2023). arXiv: 2308.09124 [cs.CL].

[114] Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. "Function Vectors in Large Language Models". In: (2023). arXiv: 2310.15213 [cs.CL].

[115] 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.

[116] 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`.

[117] 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.

[118] 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.

[119] 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`.

[120] 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.

[121] 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.

[122] 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.

[123] 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`.

[124] 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.

[125] 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.

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

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

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

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] 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.

[134] 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.

[135] 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.

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

[137] 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.

[138] 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.

[139] 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.

[140] 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.

[141] 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.

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

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

[144] 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.

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

[146] 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.

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

[148] 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.

[149] 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.

[150] 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.

[151] 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.

[152] 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.

[153] 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.

[154] 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.

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

[156] 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`.

[157] 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*.