The Returns to Science of Centrally Provisioned GPUs and Optimal Allocation

Connie Xu, Ruby Zhang*

June 17, 2026

Preliminary. Not for quotation without permission of the authors.

Abstract

Artificial intelligence and deep-learning methods require capital investments in the form of Graphical Processing Units (GPUs). GPU clusters are expensive, and universities typically provision them centrally to researchers. For budget-constrained institutions, these clusters are often funded by federal grants, primarily through the NSF Major Research Instrumentation (MRI) and Campus Cyberinfrastructure (CC*) programs. Given the importance of GPU clusters to computational research, where should the NSF allocate infrastructure funding, and how much does it prioritize efficiency versus equity? We construct a novel panel of university HPC capacity and GPU-specific research output for R1 U.S. universities, combining historical snapshots of research-computing websites with bibliometric data from OpenAlex. Leveraging large, discrete jumps in computing capacity we implement an event-study design. We find that while aggregate effects of GPU cluster upgrades on AI-related publications are modest, there is significant heterogeneity: lower-tier institutions exhibit a substantially larger marginal response to capacity upgrades than high-tier institutions. We develop a structural model of the university as a multi-product firm to estimate the production function for GPU-enabled research, derive an allocation rule for capital across fields, and use the estimated elasticities characterize the NSF's revealed preference over efficiency versus equity in infrastructure funding. Our preliminary findings suggest that diminishing returns to computing capital favor a more redistributive allocation of federal cyberinfrastructure grants.

Keywords: Computing, GPU, Multiproduct Production, Grant Allocation, Economics of Science

*(**Corresponding Author**) **Zhang:** Harvard University. Email: rzhang15@g.harvard.edu. **Xu:** Harvard University. Email: conniexu@g.harvard.edu. We gratefully acknowledge funding from the Lab for Economic Applications and Policy Award, the John and Kiendl Gordon Economics Support Fund, the Institute for Humane Studies (IHS020076; Zhang), the National Institute on Aging (T32-AG000186; Xu), and the National Science Foundation (DGE #2140743; Xu). We are grateful to Anton Tarasenko for outstanding research assistance. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

1 Introduction

High performance computing (HPC) has evolved from a specialized tool for physicists to a general-purpose technology fundamental to the scientific production function across numerous fields. The recent explosion in demand for Graphical Processing Units (GPUs), driven by advances in Artificial Intelligence (AI) and deep learning, has exacerbated the capital intensity of modern science (Cockburn et al. 2019). Universities usually maintain on-premise HPC centers, which are characterized by large fixed costs and economies of scale. This may structurally disadvantage smaller or less resourced universities, a phenomenon documented as the "compute divide" in AI research (Ahmed and Wahed 2020, Ahmed et al. 2023, Besiroglu et al. 2024).

What is the return on investment for these expensive GPU clusters? Given the central role of federal funding in building this infrastructure, how should the National Science Foundation (NSF) optimally allocate grants to maximize scientific output?

We answer these questions in three steps. First, we construct a novel panel dataset of GPU cluster capacity across R1 universities in the United States using historical snapshots of research-computing web pages, converting heterogeneous hardware into a common measure of theoretical peak performance. We combine this with bibliometric data from OpenAlex, using text analysis to identify GPU-related research output across disciplines. Second, to identify the causal effect of capital provisioning on scientific output, we rely on the lumpy nature of HPC investments. Some universities upgrade their clusters in discrete, large steps often financed by external grants such as the NSF Major Research Instrumentation (MRI) and Campus Cyberinfrastructure (CC*) programs. We treat these capacity jumps as shocks in an event study framework, following Sun and Abraham (2021) for staggered adoption. We then develop a structural model of the university as a multi-product firm, where each research field is a "product", to derive an allocation rule for how centrally provisioned GPU capital is distributed across fields using first order conditions (in the spirit of De Loecker et al. 2016). We are able to validate our estimated elasticities from the production function estimation with those implied by the subset of universities in our event study framework. Finally, we will use the structurally estimated marginal products of capital (MPK) to simulate counterfactual NSF grant allocations and to recover the NSF's revealed preference over efficiency versus equity.

Our preliminary event-study results provide empirical evidence on the returns to computing capital. We find that while aggregate effects of GPU cluster upgrades on AI-related publications are

modest, there is significant heterogeneity by university endowment and R&D funding. Lower-tier universities, which typically face tighter budget constraints and smaller initial endowments of compute, show a stronger positive response to capacity shocks compared to their high-tier counterparts (all relative to peers). This result is consistent with diminishing returns to computing capital and with researchers at well-resourced universities having outside options such as cloud computing or national facilities. Moreover, the gains at lower-tier institutions appear to be driven primarily by new PIs who arrive after the upgrade rather than by incumbents, suggesting that capital investment attracts talent. This channel is also documented in the broader literature on researcher mobility (Moser et al. 2014, Gofman and Jin 2024). Our next step is to obtain production function estimates which can then be used for counterfactual analysis.

The policy implications follow directly from the heterogeneity in returns. If the objective of federal funding is to maximize total scientific output, then grants should be allocated to equalize marginal returns to capital across universities. Since we mostly observe NSF MRI and CC* grants awarded to mid- and lower-tier universities, this implies that the NSF is already allocating infrastructure funding in an approximately efficient manner. Using our estimates, we plan to simulate the distribution of science under counterfactual allocations which target (1) total output maximization and (2) equity across institutions. Depending on the NSF's objective function, our findings will have important implications for the allocation of federal cyberinfrastructure funding and the subsequent impact on the geography of science.

This paper contributes to several strands of literature. First, a central question in the economics of science is how physical capital affects research output. Waldinger (2016) studies the decline in science output resulting from the loss of physical lab space due to WWII bombings and finds that science recovers within 20 years. Baruffaldi and Gaessler (2026) study the effect of natural disasters on laboratory capital and find persistent productivity declines when specialized capital is lost, suggesting that not all scientific capital is fungible. On the funding side, Jacob and Lefgren (2011) estimate the causal effect of NIH grants on scientific productivity using a regression discontinuity design and find modest average effects, while Azoulay et al. (2019) show that public R&D investments crowd in private-sector patenting.

The closest paper to ours is McKeon and Boudou (2025), who measure the impact of supercomputing allotments on the intensive margin for individual scientists using quasi-random allocations of compute in NSF's national ACCESS program. They find that an increase in computing capacity

affects both output and research novelty. We complement their work in three ways: we focus on the extensive margin of university-level access which serves the broad middle of researchers rather than the right tail served by national facilities; we emphasize cross-university heterogeneity in returns, which is central to policy design; and we connect the estimated returns to an optimal allocation framework for NSF grants.

Second, a growing body of work has also documented increasing concentration of AI research capacity among elite universities and large technology firms. Ahmed and Wahed (2020) use conference publication data to show that non-elite universities have been crowded out of top AI venues since the rise of deep learning in 2012, a pattern they attribute to unequal access to computing power. Ahmed et al. (2023) extend this analysis, documenting that industry increasingly dominates the key inputs to AI research: compute, data, and talent. Besiroglu et al. (2024) show that the compute divide has reduced academic-only teams' representation in compute-intensive research areas such as foundation models. Gofman and Jin (2024) document a parallel "brain drain" of AI faculty from universities to industry, with downstream effects on student entrepreneurship. Our paper causally estimates how centrally provisioned university computing capital can counteract these trends, and develops a framework for evaluating whether federal grants are optimally targeted.

Finally, we model the university as a multi-product firm where each research field constitutes a distinct product. This builds on the tradition of estimating production functions with unobserved input allocations across products (De Loecker et al. 2016), adapted to the setting of research production. In order to allow for flexible substitution between types of capital (CPU versus GPU) and labor (faculty versus graduate students), we model production as a nested CES function and assume an AR(1) process on unobserved productivity. Similar to Rubens et al. (2026), we use a dynamic panel approach (Arellano and Bond 1991, Blundell and Bond 1998) and construct shift-share instruments using GPU prices and federal funding stipends to identify elasticities of substitution. The assumption of nested CES precludes us from using proxy-variable methods (Olley and Pakes 1996, Levinsohn and Petrin 2003, Ackerberg et al. 2015). Our analysis of whether computing capital is misallocated across universities is related to the broader literature on resource misallocation (Hsieh and Klenow 2009, Restuccia and Rogerson 2008), applied here to the context of scientific production rather than manufacturing.

The remainder of this paper proceeds as follows. Section 2 describes the institutional context, and Section 3 describes the data construction. Section 4 presents the event-study framework and

reduced form results. We present our structural model and estimation strategy in Sections 5 and 6. We do not yet have estimation results, but we present our allocation framework and forthcoming counterfactual exercise in Section 7. We conclude in Section 8.

2 High Performance Computing at Universities

High performance computing involves the aggregation of computing power to solve problems that are too large for standard desktop computers. In practice, an HPC cluster will be comprised of many interconnected nodes that perform computationally intensive tasks in parallel. A common measure of compute power is floating-point operations per second (FLOPS), spanning giga-, tera-, and peta-scale systems. We will use Tera-FLOPS (one trillion operations), or TFLOPS, as our unit of computing power throughout this paper.

While traditional supercomputing relied heavily on Central Processing Units (CPUs) for serial processing tasks, the paradigm has shifted towards Graphical Processing Units (GPUs) for parallelizable tasks in deep-learning workloads and traditional simulation. GPUs are composed of thousands of small cores optimized for simple, parallel operations. As of 2025, even a single enterprise-grade GPU (e.g., NVIDIA H200) can retail for upwards of $30,000, making the "democratization" of this capital a significant financial challenge for universities.

Maintaining an on-premise HPC cluster requires data center space, racks of nodes, and power for cooling. It is a large, fixed-cost infrastructure investment. On average, HPC hardware will need to be renewed every 5 to 10 years due to depreciation and warranty expiration. Universities generally adopt one of three models to provide this capital to researchers. The most common is a centrally funded HPC cluster (funded by indirect costs or external grants) which faculty can access for free up to a quota via a fair-share scheduling system. Most universities also layer on a condo model, where faculty can use their own grant money to purchase nodes which are physically hosted and maintained by the university's research computing center. They have priority access to their purchased nodes, and when idle, the nodes are broadly used by the university community. On the other hand, some universities operate their HPC cluster as a service center or core facility by charging their researchers hourly rates for usage to recover costs.[1] This is called a recharge model and reduces the reliance of universities on funding from indirect costs.

[1]The Great Lakes cluster at the University of Michigan is a prominent example.

For many public universities, particularly those outside the elite top-tier, the initial funding for these large capital expenditures comes from federal grants. The NSF Major Research Instrumentation (MRI) program and the Campus Cyberinfrastructure (CC*) program are the primary vehicles for HPC clusters. These grants are discrete and substantial, often funding the purchase of an entire new cluster or a significant partition of GPUs. This institutional feature allows us to observe large, discrete "shocks" to capital stock that are distinct from internal university budgeting decisions. Since 2010, the number of HPC MRI grants related specifically to GPU clusters has increased substantially, as seen in Figure 1.

From the researcher perspective, there are multiple channels to obtain high performance computing. Researchers can use on-premise HPC facilities, purchase their own servers, or pay commercial cloud computing vendors out of their direct costs. Despite the entry of cloud computing, on-premise clusters still persist due to the exorbitant cost of cloud computing, especially for large datasets and GPU-hours. Another channel is national HPC facilities through NSF ACCESS or the DOE. However, national facilities serve only the right tail of computational scientists with very high computational needs. Figure 2 shows that a 2021 survey from the Coalition for Academic Scientific Computing found that on-premise clusters remain a significant computational resource for many universities.

3 Data

We combine several sources on university research capabilities and output to construct a novel panel dataset at the university × field × year level. We restrict our analysis to R1 U.S. universities.

3.1 University Computing Capacity Panel

We manually collect data on the hardware specifications of centralized university high-performance computing using research computing (RC) websites for R1 universities. For each university, we first identify the website link which lists all HPC hardware specifications. Then we extract the latest annual snapshots using Wayback Machine (Internet Archive) to 2010 or the earliest possible year. Note that for many universities, the RC link changes over time and requires manual searching through older university-wide web pages preserved by Wayback Machine.

Once the HTML page is downloaded for each available year, we use a Claude Code LLM agent (Opus 4.6) to extract all relevant hardware specifications in a tabular format. For each university

cluster, we ask the LLM agent to extract the following: total CPU cores, total CPU nodes, GPU models and counts, pricing tier, and funding information. We ask the LLM agent to corroborate the final output with the original HTML page, and find high accuracy based on random spot checks. For a subset of ten universities, we also corroborate the panel by emailing RC centers on the history of major upgrade decisions.

Due to refinement and corroboration of the current LLM algorithm, we will present preliminary results from a subset of 48 universities which are manually verified. We hope to expand this to the full set of 100 R1s in the next month. The subset of R1s is fairly representative across the distribution of endowment and STEM R&D spending.

To compare GPU capacity across different hardware vintages, we convert each GPU model into theoretical peak performance (e.g., FP16 TFLOPS due to deep learning training predominantly using half-precision) and sum across accelerators to obtain total peak GPU compute at university j in year t.

3.2 Scientific Output Measures

Our primary outcome variable is the quantity and quality of GPU-enabled publications from OpenAlex. For the top 100 R1 universities, we download all articles and preprints from 2005 to 2025. This includes information on the title, abstract, journal, authors, institutions, citations, and grants. Furthermore, OpenAlex classifies all publications into primary topics based on their internal text classifier which is trained on the corpus of publication data. Thus, we will rely on their primary field classification.

We identify GPU-relevant publications by searching titles, abstracts, and full text (where open access is available) for keywords related to AI, deep learning, neural networks, and specific GPU hardware models. For non-CS fields, we also track the usage of GPU-accelerated software (e.g., AlphaFold in biology, RELION in cryo-EM). We validate this classification by testing whether the share of GPU-relevant papers is positively correlated with central GPU capacity across universities conditional on year fixed effects in Figure A.1. The share of GPU-relevant papers are positively correlated with GPU capacity (in FP16 TFLOPS) but not CPU capacity (in cores).

We adjust for quality using citation percentiles: normalizing citation counts by publication year and subfield to account for varying citation practices across disciplines. For each paper p, define $\text{CitePct}_p \in [0, 1)$ as the paper's citation percentile normalized within publication year, document type

(article vs. preprint), and subfield. We then allocate paper credit equally across authors and split each author's credit across their institutional affiliations.

Let $A(p)$ denote the set of authors on paper p and let $\mathcal{J}(i)$ denote the set of institutions author i is affiliated with on paper p. Define the contribution of paper p to university j as:

$$y_{p \to j} = \text{CitePct}_p \cdot \sum_{i \in A(p)} \left(\frac{1}{|A(p)|} \cdot \frac{\mathbf{1}\{j \in \mathcal{J}(i)\}}{|\mathcal{J}(i)|} \right). \tag{3.1}$$

Aggregating across papers in year t yields university-year output:

$$Y_{jt} = \sum_{p \in \mathcal{P}_{jt}} y_{p \to j}. \tag{3.2}$$

We also consider threshold outcomes (e.g., counts of papers in the top 95th citation percentile) and alternative weighting schemes such as first and last authors only and share of authors with any affiliation to university j.

3.3 Labour Measure

For the structural model, we will require a count of both quasi-fixed computational labour (i.e. PIs) and flexible labour (i.e. PhD students and postdocs) by field. Since OpenAlex provides primary field designations, we construct a measure of "computational" publication which we then use to infer author fields. We create an extensive list of keyword searches in the abstract for CPU- and GPU-based publications. For every author $\times$ year, we construct both binary threshold-based and share-based computational labour measures. As a proxy for whether the author is quasi-fixed or flexible labour, we use authorship position: last authors are classified as quasi-fixed and middle authors as flexible. This convention is well-established in the life sciences, where last authorship typically signals the supervising PI, but conventions differ substantially across fields. For example, computer science tends to list authors by contribution or alphabetically. We are currently refining this classification using supplementary signals such as grant PI status from OpenAlex and career length from first publication date.

3.4 University Characteristics

We supplement our panel with university-specific characteristics to use as control variables. We obtain institutional covariates such as total number of graduate students, faculty count, and R&D expenditure by field from the Integrated Postsecondary Education Data System (IPEDS) and the NSF Higher Education Research and Development (HERD) Survey. We obtain annual university

endowment and donation data from the Center for Measuring University Performance. Finally, we obtain annual counts of doctoral and post-doctoral students by field and funding source from the NSF Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS) for all R1 universities. We will use the counts of faculty and students as a robustness check to our labour measures.

4 Event Study

We present preliminary event-study evidence from the subset of 48 manually verified universities. Figures 3 and 4 first establish that our measure of GPU capital is meaningfully tied to university resources: within-university variation in GPU capacity exhibits a high elasticity with respect to lagged endowment and lagged STEM R&D spending, conditional on university fixed effects. This confirms that GPU investment is not idiosyncratic noise but tracks the financial capacity of the institution.

4.1 Empirical Strategy

To estimate the returns to computing capital, we exploit variation in the timing and magnitude of HPC cluster upgrades. We define an "event" as a year in which a university's GPU capacity increases by more than 5,000 FP16 TFLOPS in a single step. The control group consists of universities that experienced only small or incremental capacity changes during the same window. We construct treatment and control sets within three tiers of universities, defined by baseline endowment and STEM R&D spending, to reduce confounding from persistent level differences across institution types. Figures A.2 and A.3 present the trajectory of GPU capacity over time for the 48 universities, split by tier.

The key identifying assumption is that, conditional on university and year fixed effects and within-tier comparisons, the *exact timing* of a large capacity jump is not driven by contemporaneous shocks to research productivity. While upgrade decisions are clearly non-random since better-resourced universities invest more in compute, the lumpy, grant-funded nature of these investments introduces plausible quasi-exogenous variation in timing: NSF grant review cycles, procurement lead times, and data center construction schedules all generate delays between a university's decision to invest and the actual availability of new hardware. We assess this assumption via pre-trend tests.

The estimation specification is:

$$Y_{jt} = \gamma_j + \lambda_t + \sum_{k=-3}^{3} \beta_k \mathbf{1}\{K_{jt} = k\} + \epsilon_{jt} \tag{4.1}$$

where Y_{jt} is the outcome of interest for university j in year t, γ_j and λ_t are university and year fixed effects, K_{jt} denotes time relative to the capacity expansion event, and β_k estimates the dynamic treatment effect k years from the upgrade. Since adoption is staggered across universities, we follow the interaction-weighted estimator of Sun and Abraham (2021) to avoid bias from heterogeneous treatment effects. We cluster standard errors at the university level. Outcomes include (i) citation-weighted GPU-related publications, (ii) counts of papers above the 95th citation percentile, and (iii) placebo outcomes using non-GPU-related publications.

A natural concern is that GPU upgrades coincide with other investments, such as faculty hiring or new grant funding, which independently raise productivity. We address this in two ways. First, the incumbent-only analysis in Section 4.3 holds the set of researchers fixed, removing compositional changes as a confound. Second, the computer science placebo test in Section 4.4 shows that non-GPU publications at treated universities do not increase, which would be inconsistent with a story in which general departmental investments drive the results.

4.2 Aggregate and Tier-Level Results

Figure 5 shows limited average effects of GPU cluster upgrades on aggregate AI-related publication outcomes. However, the average effect masks substantial heterogeneity across institution types. Figures 6 and 7 show that low-tier and medium-tier universities exhibit economically and statistically significant increases in AI-related output following an upgrade, with the effect appearing within one to two years and persisting through the end of our event window. In contrast, Figure 8 shows no detectable increase for high-tier institutions. Pre-trends are flat for all three tiers, supporting the identifying assumption.

The pattern is consistent with diminishing marginal returns to computing capital. High-tier universities already possess substantial GPU resources; an additional cluster represents an incremental expansion of capacity, and substitutes such as commercial cloud computing and national facilities are readily accessible to well-funded researchers. For lower-tier universities, by contrast, a new GPU cluster may represent a discrete shift from being compute-constrained to having meaningful capacity.

4.3 Decomposition: Incumbent vs. Entrant PIs

The aggregate results could in principle reflect either increased productivity of existing researchers or the arrival of new researchers attracted by improved infrastructure. To distinguish these channels, we re-estimate Equation (4.1) restricting the outcome to publications by *incumbent* PIs, defined as authors with at least one publication at the university prior to and after the upgrade event.

Subsetting to incumbents, the positive effects at low- and medium-tier universities largely disappear. This suggests that the output gains are primarily driven by *entrant* PIs—new researchers who arrive at the university after the capacity expansion. This is consistent with GPU infrastructure serving as an attraction mechanism for computational researchers, complementing evidence from the broader literature on how capital investment shapes researcher mobility (Moser et al. 2014, Gofman and Jin 2024). We plan to investigate this channel further.

4.4 Returns to HPC for Computer Science

As a more targeted test, we restrict the sample to computer science publications and split them into GPU-related and non-GPU-related. If the event-study effects reflect a genuine return to GPU capital rather than a general improvement in departmental resources, we should see effects concentrated in GPU-related CS publications with no corresponding increase in non-GPU CS output.

Figures 9 and 10 confirm this pattern: low-tier and medium-tier universities show significant increases in citation-weighted GPU-related CS publications following an upgrade, with no corresponding movement in non-GPU CS publications. Figure 11 shows no relative increase for high-tier universities in either category. The absence of effects on non-GPU CS output serves as a within-department placebo test, reinforcing the interpretation that the reduced-form effects are not driven by unobserved departmental shocks.

5 Structural Model

Despite purchasing and system installation lags, the standard simultaneity problem persists: universities endogenously choose research inputs (such as GPU capital) in response to the same unobserved productivity term which affects output. Since universities endogenously invest in computing capital, event study analysis is only centered around sufficiently large capital increases. Not only does this discard many panel observations and valuable data variation, but also implies imperfect treatment

and control groups. We address the simultaneity problem with a structural production function and model unobserved productivity as an AR(1) process.

We model the university as a multi-product firm where each research field is a distinct "product," in the spirit of De Loecker et al. (2016). We do not take a normative stance on how universities should aggregate research outputs across fields. Thus, we estimate field-level production functions where each university × field has a different unobserved productivity term evolving via an AR(1) process. However, we do not observe capital at the field-level, only the university level. We model the university's objective and derive intra-university field-level capital allocations from the first-order conditions.

5.1 Setup

Let fields be indexed by $f \in \mathcal{F}$ and let $g(f) \in \mathcal{G}$ denote the broad field group containing field f. We define the following groups: computer science, life sciences, physical sciences, engineering, math & statistics, social sciences, and psychology. Humanities are excluded since GPU capital does not plausibly enter the production function.

University j enters period t with four types of state variables: CPU capital (in cores) denoted by C_{jt}, GPU capital (in FP16 TFLOPS) denoted by G_{jt}, faculty counts across fields denoted by $\mathbf{L}^F_{jt} = \{L^F_{jft}\}_{f\in\mathcal{F}}$, and unobserved productivity across fields denoted by $\boldsymbol{\omega}_{jt} = \{\omega_{jft}\}_{f\in\mathcal{F}}$. Capital is at the university level while faculty count and productivity are at the university × field level. Faculty is treated as a quasi-fixed input that cannot be adjusted within a period.

Production technology. For a field f in group $g = g(f)$, output is produced according to a Cobb–Douglas technology with nested CES for capital and labor inputs:

$$Y_{jft} = \Omega_{jft}\, \mathcal{K}^{\alpha^K_g}_{jft}\, \mathcal{L}^{\alpha^L_g}_{jft}, \qquad \nu_g \equiv \alpha^K_g + \alpha^L_g. \tag{5.1}$$

$$\mathcal{K}_{jft} = \left[\gamma_g\, C^{\rho_K}_{jft} + (1-\gamma_g)\left(G_{jft}\right)^{\rho_K}\right]^{1/\rho_K}, \qquad \rho_K \equiv \frac{\sigma_K - 1}{\sigma_K}. \tag{5.2}$$

$$\mathcal{L}_{jft} = \left[\phi_g\, (L^V_{jft})^{\rho_L} + (1-\phi_g)\, (L^F_{jft})^{\rho_L}\right]^{1/\rho_L}, \qquad \rho_L \equiv \frac{\sigma_L - 1}{\sigma_L}. \tag{5.3}$$

where $\Omega_{jft} = \exp(\omega_{jft})$ is Hicks-neutral productivity, L^V_{jft} is flexible labor (postdocs and PhD students), L^F_{jft} is quasi-fixed labor (faculty), and C_{jft} and G_{jft} are field-level CPU and GPU capital. Parameters $(\alpha_{Kg}, \alpha_{Lg}, \gamma_g, \phi_g)$ are group-specific, allowing the returns to inputs and distribution share

within nests to differ across computer science, life sciences, physical sciences, engineering, math & statistics, social sciences, and psychology. Within-nest capital and labor substitution elasticities (σ_K, σ_L) are not group-specific, implying that the rate of substitution between CPU and GPU capital are common across groups. Since investment events are at the university-level, there is insufficient variation to separately identify group-specific substitution elasticities. A similar assumption is made for the substitution elasticity of quasi-fixed versus flexible labor.

Compared to a Cobb-Douglas production function in all four inputs, allowing for capital and labor nests relaxes the assumption that CPU, GPU, faculty, and graduate students substitute equally for each other. Instead, CPU and GPU capital have their own elasticity of substitution for scientific workflows and likewise for faculty versus graduate students. A graduate student cannot substitute for the type of work that a CPU or GPU can do (and vice versa).

5.2 The University's Optimization Problem

Since we do not observe field-level capital allocations, we model the university's problem to derive an allocation rule. The university maximizes total weighted publications across multiple "products" (i.e. fields) over time subject to discount factor β. Let $\{\lambda_{jf}\}_{f\in\mathcal{F}}$ denote time-invariant prestige weights that university j places on output across fields f. The university has two categories of choice variables.

Static (intratemporal) choices. At the beginning of each time period t, given the current stock of capital and faculty, the university chooses flexible labor (i.e. graduate students and post-doctorates) L^V_{jft} and the allocation of aggregate capital across fields, $\{C_{jft}, G_{jft}\}_{f\in\mathcal{F}}$, to maximize current period publication output.

Dynamic (intertemporal) choices. At the end of each time period t, the university decides total CPU investment I^C_{jt}, total GPU investment I^G_{jt}, and faculty hiring $\{H_{jft}\}_{f\in\mathcal{F}}$ across fields. This will increase next period's stock of capital and faculty.

The Bellman equation is:

$$V(C_{jt}, G_{jt}, \mathbf{L}^F_{jt}, \boldsymbol{\omega}_{jt}) = \max_{\substack{I^C_{jt}, I^G_{jt}, \{H_{jft}\}_f, \\ \{L^V_{jft}\}_f, \{C_{jft}\}_f, \{G_{jft}\}_f}} \left\{ \sum_f \lambda_{jf}\, Y_{jft} + \beta\, \mathbb{E}\big[V\big(C_{jt+1}, G_{jt+1}, \mathbf{L}^F_{jt+1}, \boldsymbol{\omega}_{jt+1}\big) \,\big|\, \boldsymbol{\omega}_{jt}\big] \right\} \tag{5.4}$$

subject to the laws of motion for the state variables:

$$C_{jt+1} = (1 - \delta_C)\, C_{jt} + I^C_{jt} \tag{5.5}$$

$$G_{jt+1} = (1 - \delta_G)\, G_{jt} + I^G_{jt} \tag{5.6}$$

$$L^F_{jft+1} = (1 - \delta_F)\, L^F_{jft} + H_{jft} \tag{5.7}$$

$$\omega_{jft+1} = \varrho_g \omega_{jft} + \xi_{jft+1} \tag{5.8}$$

where δ_C, δ_G, and δ_F are depreciation rates for CPU capital, GPU capital, and faculty, respectively. Productivity evolves according to (5.8), where the autoregressive parameter differs by broad group.

The university also faces three static constraints. Total CPU and GPU capital allocated across fields cannot exceed the aggregate stocks:

$$\sum_f C_{jft} \leq C_{jt} \tag{5.9}$$

$$\sum_f G_{jft} \leq G_{jt} \tag{5.10}$$

and total expenditure on flexible inputs and investment cannot exceed the university's budget:

$$\sum_f W^V_{jft}\, L^V_{jft} + \sum_f W^F_{jft}\, L^F_{jft} + P^C_t\, I^C_{jt} + P^G_t\, I^G_{jt} \leq B_{jt} \tag{5.11}$$

where W^V_{jft} and W^F_{jft} are wage rates for flexible and quasi-fixed labor, P^C_t and P^G_t are capital prices, and B_{jt} is total available funding (e.g., HERD R&D expenditures and endowment payouts). We assume that faculty hiring is costless, and the only cost is maintaining the faculty stock.

5.3 Deriving the Allocation Rules

The static first-order conditions for the university's intratemporal problem yield allocation rules that map aggregate capital stocks into field-level inputs. This allows us to estimate the production function at the field level even though we do not observe field-level capital directly. Since we are not estimating the full dynamic model, the field-level prestige weights $\{\lambda_{jf}\}_f$ drop out and we do not need data on wages, prices, and the budget.

Let μ^C_{jt}, μ^G_{jt}, and η_{jt} denote the Lagrange multipliers on the CPU constraint (5.9), the GPU constraint (5.10), and the budget constraint (5.11), respectively. The period-t Lagrangian for university j is:

$$\mathcal{L} = \sum_f \lambda_{jf}\, Y_{jft} + \beta\, \mathbb{E}[V(\cdot) \mid \boldsymbol{\omega}_{jt}]$$
$$- \mu^C_{jt}\left(\sum_f C_{jft} - C_{jt}\right) - \mu^G_{jt}\left(\sum_f G_{jft} - G_{jt}\right)$$
$$- \eta_{jt}\left(\sum_f W^V_{jft}\, L^V_{jft} + \sum_f W^F_{jft}\, L^F_{jft} + P^C_t\, I^C_{jt} + P^G_t\, I^G_{jt} - B_{jt}\right) \tag{5.12}$$

Since the continuation value $\beta\,\mathbb{E}[V(\cdot)]$ depends only on next-period states (which are functions of the dynamic choices $(I^C_{jt}, I^G_{jt}, H_{jft})$ and current states), it does not depend on the static choices $(C_{jft}, G_{jft}, L^V_{jft})$. Therefore, the continuation value drops out of all static first-order conditions.

Differentiating $\mathcal{L}$ with respect to the three static choice variables and multiplying both sides by the corresponding choice variable yields:

$$\mu^C_{jt}\, C_{jft} = \alpha^K_g\; s^{\mathcal{K}}_C(jft) \cdot \lambda_{jf} Y_{jft}, \tag{5.13}$$

$$\mu^G_{jt}\, G_{jft} = \alpha^K_g\; s^{\mathcal{K}}_G(jft) \cdot \lambda_{jf} Y_{jft}, \tag{5.14}$$

$$\eta_{jt}\, W^V_{jft}\, L^V_{jft} = \alpha^L_g\; s^{\mathcal{L}}_V(jft) \cdot \lambda_{jf} Y_{jft} \tag{5.15}$$

with within-nest value shares

$$s^{\mathcal{K}}_C(jft) = \gamma_g \Big(\frac{C_{jft}}{\mathcal{K}_{jft}}\Big)^{\rho_K}, \qquad s^{\mathcal{K}}_G(jft) = (1-\gamma_g)\Big(\frac{e^{\delta} G_{jft}}{\mathcal{K}_{jft}}\Big)^{\rho_K} = 1 - s^{\mathcal{K}}_C(jft), \tag{5.16}$$

$$s^{\mathcal{L}}_V(jft) = \phi_g \Big(\frac{L^V_{jft}}{\mathcal{L}_{jft}}\Big)^{\rho_L}, \qquad s^{\mathcal{L}}_F(jft) = 1 - s^{\mathcal{L}}_V(jft) \tag{5.17}$$

From (5.15), $\lambda_{jf} Y_{jft} = \eta_{jt} W^V_{jft} L^V_{jft} / [\alpha^L_g\; s^{\mathcal{L}}_V(jft)]$ which we substitute into (5.13) and (5.14) to cancel out field-level prestige weights and output. We then divide the capital FOCs by the summed version across fields to cancel out Lagrange multipliers μ^C_{jt}, μ^G_{jt}, and η_{jt} and yield the following allocation rules:

$$C_{jft} = C_{jt} \cdot \frac{\psi^C_{jft}\, W^V_{jft}\, L^V_{jft}}{\sum_h \psi^C_{jht}\, W^V_{jht}\, L^V_{jht}}, \qquad \psi^C_{jft} \equiv \frac{\alpha^K_g}{\alpha^L_g} \cdot \frac{s^{\mathcal{K}}_C(jft)}{s^{\mathcal{L}}_V(jft)} \tag{5.18}$$

$$G_{jft} = G_{jt} \cdot \frac{\psi^G_{jft}\, W^V_{jft}\, L^V_{jft}}{\sum_h \psi^G_{jht}\, W^V_{jht}\, L^V_{jht}}, \qquad \psi^G_{jft} \equiv \frac{\alpha^K_g}{\alpha^L_g} \cdot \frac{s^{\mathcal{K}}_G(jft)}{s^{\mathcal{L}}_V(jft)} \tag{5.19}$$

Equations (5.18)–(5.19) are not closed-form because ψ^C_{jft} and ψ^G_{jft} depend on the within-nest shares $s^{\mathcal{K}}_C(jft)$ and $s^{\mathcal{L}}_V(jft)$, which in turn depend on the field-level allocations. Given parameter vector $\boldsymbol{\theta}$ and observables, allocations solve the fixed-point system formed by (5.18)–(5.19) together with the share definitions. Cobb–Douglas is the special case $\sigma_K = \sigma_L = 1$, where shares are constants $(s^{\mathcal{K}}_C = \gamma_g,\ s^{\mathcal{L}}_V = \phi_g)$ and the fixed point collapses to a closed form.

Common Wage Assumption. Since we do not have wage data or labor spending share but rather labor count, we assume that flexible-labour wages are constant across fields within a university-year:

$$W^V_{jft} = W^V_{jt} \quad \forall f \tag{5.20}$$

Then wages cancel and the allocation rule simplifies to the following:

$$\boxed{C_{jft} = C_{jt} \cdot \frac{\psi^C_{jft} L^V_{jft}}{\sum_h \psi^C_{jht} L^V_{jht}}} \tag{5.21}$$

$$\boxed{G_{jft} = G_{jt} \cdot \frac{\psi^G_{jft} L^V_{jft}}{\sum_h \psi^G_{jht} L^V_{jht}}} \tag{5.22}$$

Within a broad field group, fields with more flexible labor receive proportionally more compute. Across groups, capital tilts toward groups with larger compute-intensity ratios ψ_{jft}.

Utilization. The allocation rules require the capital constraints to bind. In practice, HPC clusters do not run at 100% instantaneous utilization, but fairshare schedulers (e.g., SLURM) allocate the full stock of compute-hours across research groups over annual or quarterly cycles. Anecdotally, GPU resources are commonly at 100% capacity while CPU resources tend to hover stably at around 80% utilization to account for demand spikes (see Figure A.4 for Harvard's utilization). Furthermore, allocation shares are not affected if utilization is stable across fields. Let $u_{jt} \in (0, 1]$ denote the university-year utilization rate, so that effective capital is $u_{jt}C_{jt}$. If u_{jt} does not vary systematically across fields within a university-year, it cancels from the share equations: $C_{jft}/C_{jt} = u_{jt}C_{jft}/(u_{jt}C_{jt})$. The allocation *shares* are therefore invariant to the utilization rate. Idle time affects only the scale of effective field-level capital, which is absorbed into the iid error ε_{jft} in the production function.

Testable implications. For the subset of universities that publish HPC usage reports by department (e.g., University of Arizona, Boston University), the model-implied field shares can be compared directly to observed compute allocations. Additionally, the allocation rules imply short-run crowd-out: holding total university capacity fixed, a positive shock to flexible labor in one field group mechanically reduces the capital share available to other groups. Given that most universities operate on a fairshare system to allow different researchers equal access, this implication seems reasonable.

6 Estimation

The production function estimation is complicated by the fact that we do not have a closed-form allocation rule. Since unobserved productivity is assumed to be AR(1), we will take ϱ_g-differences and use GMM to estimate the outer nest. For each guess of the parameter vector $\boldsymbol{\theta}$, we will compute field-level capital allocations via fixed point iteration. We do not yet have estimation results, but we plan to test the necessity of nested CES under the null hypothesis that $\sigma_K = \sigma_L = 1$. We also plan to present elasticity measures assuming plain Cobb-Douglas using a dynamic panel approach.

6.1 Production Function Estimation

Taking logs of the aggregate production function we have

$$y_{jft} = \alpha_g^K \log \mathcal{K}_{jft}(\boldsymbol{\theta}) + \alpha_g^L \log \mathcal{L}_{jft} + \omega_{jft} + \varepsilon_{jft} \tag{6.1}$$

where ε_{jft} is an iid error term, labor nest $\mathcal{L}_{jft}$ is computed from observed (L^V_{jft}, L^F_{jft}), and capital nest $\mathcal{K}_{jft}(\boldsymbol{\theta})$ is the model-implied capital aggregator evaluated at the fixed-point allocations. Apply the AR(1) productivity process by taking ϱ_g-differences:

$$y_{jft} - \varrho_g\, y_{jf,t-1} = \alpha_g^K \left[\log \mathcal{K}_{jft}(\boldsymbol{\theta}) - \varrho_g \log \mathcal{K}_{jft-1}(\boldsymbol{\theta})\right] + \alpha_g^L \left[\log \mathcal{L}_{jft} - \varrho_g \log \mathcal{L}_{jft-1}\right] + u_{jft} \tag{6.2}$$

$$u_{jft} \equiv \xi_{jft} + \varepsilon_{jft} - \varrho_g\, \varepsilon_{jft-1} \tag{6.3}$$

u_{jft} is the composite iid error, and thus orthogonal to predetermined inputs and any instrument observed by $t-2$. The estimation procedure will have a GMM outer loop based on iid error u_{jft}, and an inner loop which computes $\mathcal{K}_{jft}(\boldsymbol{\theta})$ for each outer loop estimate.

Inner Loop. Given a guess of the full parameter vector $\boldsymbol{\theta} \equiv \left(\{\alpha_g^K, \alpha_g^L, \gamma_g, \phi_g, \varrho_g\}_{g\in\mathcal{G}}, \sigma_K, \sigma_L\right)$ and pair (j,t), we calculate field-level capital $\{C_{jft}, G_{jft}\}_{f\in\mathcal{F}}$ using the following fixed-point procedure:

1. Initialize $\hat{C}^{(0)}_{jft}, \hat{G}^{(0)}_{jft}$ at the Cobb–Douglas allocation ($\sigma_K = \sigma_L = 1$) under common wage assumption (5.20):
$$\hat{C}^{(0)}_{jft} = \frac{C_{jt} \cdot (\alpha_g^K \gamma_g / \alpha_g^L \phi_g)\, L^V_{jft}}{\sum_h (\alpha^K_{g(h)} \gamma_{g(h)} / \alpha^L_{g(h)} \phi_{g(h)}) L^V_{jht}}$$
and analogously for $\hat{G}^{(0)}_{jft}$.
2. Compute $\mathcal{K}^{(k)}_{jft}$ from (5.2) using $\hat{C}^{(k)}_{jft}, \hat{G}^{(k)}_{jft}$, and $\mathcal{L}_{jft}$ from (5.3) using observed labor.
3. Compute shares $s_C^{\mathcal{K},(k)}(jft), s_G^{\mathcal{K},(k)}(jft), s_V^{\mathcal{L}}(jft)$.
4. Update $\psi^{C,(k)}_{jft}$ and $\psi^{G,(k)}_{jft}$.

5. Update allocations from (5.21)–(5.22).

6. Stop when $\left\|\hat{C}_{jft}^{(k+1)} - \hat{C}_{jft}^{(k)}\right\| < 10^{-8}$ and similarly for $\{G_{jft}\}_f$. We use a tight tolerance of 10^{-8} since this is an inner loop optimization.

Outer Loop. Given fixed-point allocations $\hat{C}_{jft}(\boldsymbol{\theta}), \hat{G}_{jft}(\boldsymbol{\theta})$, construct $\log \mathcal{K}_{jft}(\boldsymbol{\theta})$ and form iid residuals $u_{jft}(\boldsymbol{\theta})$ from (6.2). Let Z_{jft} be the instrument vector, we define a set of moments for each broad group $g \in \mathcal{G}$:

$$\hat{m}_g(\boldsymbol{\theta}) \;=\; \frac{1}{N_g} \sum_{jt} \sum_{f:\, g(f)=g} u_{jft}(\boldsymbol{\theta}) \otimes Z_{jft}, \qquad N_g = \#\{(j,f,t) : g(f) = g\} \tag{6.4}$$

Stacking the blocks gives the full moment vector

$$\hat{m}(\boldsymbol{\theta}) \;=\; \left[\hat{m}_{g_1}(\boldsymbol{\theta})', \ldots, \hat{m}_{g_{|\mathcal{G}|}}(\boldsymbol{\theta})'\right]' \tag{6.5}$$

and the GMM objective stacks all group-specific moments into one estimator:

$$\hat{\boldsymbol{\theta}} \;=\; \arg\min_{\boldsymbol{\theta}} \; \hat{m}(\boldsymbol{\theta})' \, W \, \hat{m}(\boldsymbol{\theta}) \tag{6.6}$$

We use the efficient weighting matrix $W = \Omega^{-1}$ by doing a two-step GMM. In the first step, we use $W = (Z'Z)^{-1}$. Then we compute the first-step residuals $\hat{u}_{jft}^{(1)}$ and use $W = \hat{\Omega}^{-1}$ in the second step where we cluster on university to account for serial correlation:

$$\hat{\varphi}_j \;=\; \left[\, \textstyle\sum_t \sum_{f:g(f)=g_1} \hat{u}_{jft}^{(1)} Z_{jft}^{g_1}, \;\ldots,\; \sum_t \sum_{f:g(f)=g_{|\mathcal{G}|}} \hat{u}_{jft}^{(1)} Z_{jft}^{g_{|\mathcal{G}|}} \,\right]', \tag{6.7}$$

$$\hat{\Omega} \;=\; \frac{1}{N} \sum_j \hat{\varphi}_j \, \hat{\varphi}_j' \tag{6.8}$$

The off-diagonal blocks of $\hat{\Omega}$ capture the covariance between groups' moments induced by the common parameters (σ_K, σ_L). Standard errors are computed by block-bootstrap on universities ($B = 200$ draws), re-estimating the inner fixed point and outer GMM on each draw so that all estimation error propagates. The outer optimization uses the derivative-free Nelder–Mead method, which avoids finite-difference gradients of an objective that embeds an inner fixed-point loop.

6.2 Instruments

The set of instruments Z_{jft} must be independent of u_{jft}, which is a function of $(\xi_{jft}, \varepsilon_{jft}, \varepsilon_{jft-1})$. Since ε_{jft} is iid error on the final output, then we only need that instruments are independent of the innovation to unobserved productivity. Current period capital and quasi-fixed labor are chosen before observing ξ_{jft} while flexible labor is chosen after observing ξ_{jft}. Thus we motivate our first set of instruments which are all lagged inputs and current quasi-fixed inputs:

$$Z_{jft}^{\text{inputs}} = \left\{ \log L_{jft-1}^V, \log L_{jft}^F, \log L_{jft-1}^F, \log C_{jt}, \log C_{jt-1}, \log G_{jt}, \log G_{jt-1} \right\} \tag{6.9}$$

Since CPU and GPU capital measures are aggregated to the university level and thus does not vary across fields, we supplement the instrument set with lagged and current capitalized equipment expenditure by field from the NSF HERD survey:

$$Z_{jft}^{\text{equip}} = \left\{ \log E_{jft}, \log E_{jft-1} \right\} \tag{6.10}$$

However, in order to identify substitution elasticities within the nests (σ_K, σ_L), we must have instruments which shift the *relative price* for CPU versus GPU capital and for quasi-fixed versus flexible labor. We construct shift-share instruments to identify the capital and labor nest elasticities of substitution (σ_K, σ_L).

Capital substitution σ_K. During the main analysis time frame of 2015 to 2025, Moore's Law slowed down substantially for CPUs. However, new GPU vintages that doubled or tripled computing capacity were released during this time period. The primary shift will be NVIDIA major architecture transition years and the share will be pre-period university GPU vintages. Let $\mathcal{T}^{\text{NVDA}} = \{t_{\text{V100}\to\text{A100}}, t_{\text{A100}\to\text{H100}}, \ldots\}$ denote the set of NVIDIA architecture-transition years. For each $t \in \mathcal{T}^{\text{NVDA}}$, define the log change in price per FP16 TFLOP of newly purchased GPUs:

$$\Delta p_t^G \equiv \log P_t^G - \log P_{t-1}^G \tag{6.11}$$

where P_t^G is the year-t price per FP16 TFLOP for the contemporary generation (from NVIDIA vendor list prices). For $t \notin \mathcal{T}^{\text{NVDA}}$, set $\Delta p_t^G = 0$.

For each university j, we compute the predetermined share of its installed GPU TFLOPS that belongs to an *outgoing* generation just prior to transition year t. We define *outgoing* generation as vintages which are at least three generations behind the transition year, indexed by $\mathcal{V}^{\text{prev}}(t)$.

$$\text{share}_{jt-1}^{\text{prev}} \equiv \frac{\sum_{v \in \mathcal{V}^{\text{prev}}(t)} \text{TFLOPS}_{jt-1}^{v}}{G_{jt-1}} \tag{6.12}$$

We construct the following instrument which varies at the university $\times$ year level.

$$Z_{jt}^{\text{NVDA}} \equiv \Delta p_t^G \cdot \text{share}_{jt-1}^{\text{prev}} \tag{6.13}$$

Since we have one overall σ_K parameter, we do not need field-level variation. The shock is national and exogenous to university-level productivity innovations and the predetermined vintage share which is set by prior procurement decisions uncorrelated with year-t productivity shocks.

Labour substitution σ_L. We use changes in the federal stipend rate for doctoral and post-doctoral students, interacted with the share supported by the funding agency. From the NSF GSS survey, we can see the number of doctoral and post-doctoral students supported by the various federal

agencies, namely NSF and NIH, by university × field × year. For doctoral students, we use yearly stipend changes from the NSF Graduate Research Fellowship Program (GRFP). For post-doctoral students, we use yearly stipend changes from the NIH National Research Service Award Fellowship (NRSA). These changes are tracked from annual solicitation announcements. For each program $p \in \{\text{GRFP}, \text{NRSA}\}$, define the log change in annual stipend:

$$\Delta s_t^p \equiv \log \text{stipend}_t^p - \log \text{stipend}_{t-1}^p \tag{6.14}$$

For each university-field (j, f), we compute the lagged share of L_{jft-1}^V supported by the federal agency in charge of program p (NSF for GRFP and NIH for NRSA):

$$\text{share}_{jft-1}^p \equiv \frac{L_{jft-1}^{V,p}}{L_{jft-1}^V} \tag{6.15}$$

Note that we are only counting doctoral students for NSF GRFP and post-doctoral students for NIH NRSA. Then for each program p and each (jft), we have the instrument:

$$Z_{jft}^{\text{funding}} \equiv \Delta s_t^p \cdot \text{share}_{jft-1}^p \tag{6.16}$$

Our overall instrument set generates $|\mathcal{G}|$ moments per instrument, leading to overidentification:

$$Z_{jft} = Z_{jft}^{\text{inputs}} \cup Z_{jft}^{\text{equip}} \cup Z_{jt}^{\text{NVDA}} \cup Z_{jft}^{\text{funding}}$$

7 NSF Allocation and Counterfactuals

The heterogeneity in returns documented in Section 4 has direct implications for federal cyberinfrastructure funding. Under the nested CES model in Section 5, universities differ in baseline output, capital stocks, field composition, and labor shares. In this section, we formalize the planner's efficiency–equity tradeoff, describe counterfactual allocation exercises, and how to estimate the NSF's revealed preference over efficiency versus equity.

7.1 The Planner's Problem

We model the NSF as a social planner allocating a fixed cyberinfrastructure budget $\bar{B}_{\text{NSF}}$ across universities. Since MRI and CC* grants fund local GPU clusters rather than a full re-optimization of every university input, we treat CPU stocks, quasi-fixed labor, and field composition as part of each university's baseline state. The policy margin is GPU support: the planner chooses additional GPU capacity $\Delta G_j \geq 0$ for each university j.

The planner's objective is an Atkinson-style social welfare function over predicted scientific

output:

$$W(\rho) = \sum_{j=1}^{J} \frac{\mathcal{Y}_j(\Delta G_j;\theta)^{1-\rho}}{1-\rho} \tag{7.1}$$

where

$$\mathcal{Y}_j(\Delta G_j;\theta) \equiv \widehat{Y}_{jt}(G_{jt} + \Delta G_j;\theta) \tag{7.2}$$

is university j's total predicted output under the counterfactual GPU stock, and $\rho \geq 0$ is the inequality-aversion parameter. The parameter ρ governs the planner's willingness to sacrifice aggregate output in order to raise output at lower-producing institutions:

- $\rho = 0$ (pure efficiency): $W = \sum_j \mathcal{Y}_j$. The planner maximizes total national scientific output regardless of its distribution.
- $\rho = 1$ (moderate equity): $W = \sum_j \ln \mathcal{Y}_j$. The planner values proportional improvements equally across universities.
- $\rho \to \infty$ (Rawlsian): $W = \min_j \mathcal{Y}_j$. The planner maximizes the output of the lowest-producing university.

The planner solves:

$$\begin{aligned} \max_{\{\Delta G_j\}_{j=1}^{J}} \quad & \sum_{j=1}^{J} \frac{\mathcal{Y}_j(\Delta G_j;\theta)^{1-\rho}}{1-\rho} \\ \text{s.t.} \quad & \sum_{j=1}^{J} p_G \, \Delta G_j \leq \bar{B}_{\text{NSF}}, \qquad \Delta G_j \geq 0 \ \ \forall \, j \end{aligned} \tag{7.3}$$

The first-order condition equates equity-weighted marginal products of GPU capital (MPG) across universities:

$$\mathcal{Y}_j(\Delta G_j;\theta)^{-\rho} \cdot \frac{1}{p_G} \, MPG_j^{\text{post}}(G_{jt} + \Delta G_j;\theta) = \lambda \tag{7.4}$$

where λ is the shadow value of the planner's budget. When $\rho = 0$, this reduces to the standard efficiency condition: equalize marginal products of GPU capital across all funded universities. When $\rho > 0$, the planner tolerates higher marginal products at high-output universities—effectively directing more resources toward universities with lower baseline output.

7.2 Counterfactual Exercises

We propose four exercises that use the estimated production function and the planner's framework to evaluate alternative allocations of NSF GPU funding.

Exercise 1: Pure efficiency benchmark. Set $\rho = 0$ in (7.3) and compute the allocation that maximizes total predicted scientific output. The implementation proceeds in two steps. First, rank all universities by their return per dollar: MPG_j^{post}/p_G. Second, iteratively allocate the budget to whichever university currently offers the highest return, updating marginal products after each increment (since MPG^{post} is decreasing in G_{jt} under diminishing returns). This yields the maximum attainable aggregate output, which serves as the efficiency benchmark against which all other allocations are measured.

Exercise 2: Evaluating the status quo. Using the observed distribution of NSF MRI and CC* GPU-related grants, compute predicted scientific output under the actual allocation. The gap between this output and the pure-efficiency benchmark measures the scientific cost—if any—of the NSF's current allocation choices. Because the production function allows returns to differ across field groups, this exercise also reveals which scientific areas gain or lose under the status quo relative to the efficient benchmark. For instance, if the NSF disproportionately funds universities with large life-science departments, the status quo may overweight GPU-enabled biology relative to the efficiency optimum.

Exercise 3: Efficiency–equity frontier. Solve (7.3) for a grid of values $\rho \in \{0, 0.5, 1, 2, 5, 10\}$. This traces out the frontier showing how much aggregate output the planner sacrifices to direct more compute toward lower-output institutions. At each point on the frontier, we report the implied allocation across university tiers, the distribution of output gains across field groups, and whether the allocation induces new adoptions at previously GPU-absent universities. The grouped model makes these exercises especially informative: a redistribution toward a given university can disproportionately benefit different field groups depending on that university's flexible-labor composition.

Exercise 4: Revealed preference estimation. Finally, we estimate where the NSF currently sits on the efficiency–equity frontier. Taking logs of the interior FOC (7.4), the optimality condition for post-adoption universities that received NSF funding becomes:

$$\ln\left(MPG_{jt}^{\text{post}}\right) = \alpha_t + \rho \ln\left(\widehat{Y}_{jt}^{\text{base}}\right) + u_{jt}, \tag{7.5}$$

where α_t absorbs the shadow value λ_t and the common GPU price p_G in year t, and $\widehat{Y}_{jt}^{\text{base}}$ is baseline predicted output prior to the marginal grant. The coefficient on baseline output recovers the planner's revealed inequality aversion $\hat{\rho}$. If $\hat{\rho} \approx 0$, the NSF acts as a pure efficiency maximizer, equalizing

marginal products across funded universities. If $\hat{\rho} > 0$, the NSF is deliberately funding universities with lower marginal products because it values a more equitable distribution of scientific capacity.

Note that the framework serves as a revealed-preference exercise over the distribution of *scientific output generated by compute grants*, not a complete welfare analysis of NSF funding. The NSF may value geographic dispersion (e.g., EPSCoR), institutional capacity building, training of future researchers, or face political constraints that are orthogonal to scientific output. The estimated $\hat{\rho}$ should therefore be read as a reduced-form summary of how the agency trades off output against concentration in the data, not as a literal structural welfare weight.

8 Conclusion

In this paper, we compile a novel dataset of university GPU computing capacity and estimate the returns to centrally provisioned computing capital for scientific research. Using an event-study design that exploits discrete capacity expansions at U.S. R1 universities, we find that the aggregate effect of GPU cluster upgrades on AI-related publications is modest, but that this average masks substantial heterogeneity: lower-tier universities exhibit significantly larger marginal responses to capacity shocks than their well-resourced counterparts. The gains at less-funded institutions appear to be driven primarily by the arrival of new researchers rather than increased productivity among incumbents, suggesting that computing infrastructure serves as a mechanism for attracting computational talent.

These findings have direct implications for federal science policy. The concave nature of returns to computing capital implies that the current concentration of GPU resources at elite institutions may be suboptimal for maximizing aggregate scientific output. We develop a structural framework—modeling the university as a multi-product firm—that connects the reduced-form estimates to an optimal allocation rule for NSF cyberinfrastructure grants and allows us to recover the NSF's revealed preference over efficiency versus equity in infrastructure funding.

8.1 Next Steps

This paper is preliminary. Several extensions are in progress. First, we are expanding the university computing capacity panel from 48 to 100 R1 universities, which will substantially improve the precision of our event-study estimates and enable richer heterogeneity analysis. Second, we are completing the production function estimation using the estimation procedure outlined in Section 6, which will

provide structural elasticities that can be compared to the reduced-form estimates and used for counterfactual simulations. Third, we plan to implement the counterfactual allocation exercises described in Section 7, quantifying the output gains from alternative NSF grant distributions and estimating where the NSF currently sits on the efficiency-equity frontier. As the capital intensity of AI research continues to rise, understanding where federal infrastructure dollars generate the highest returns to science is increasingly important for policy design.

References

Ackerberg DA, Caves K, Frazer G (2015) Identification properties of recent production function estimators. *Econometrica* 83(6):2411–2451.

Ahmed N, Wahed M (2020) The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv:2010.15581* .

Ahmed N, Wahed M, Thompson NC (2023) The growing influence of industry in AI research. *Science* 379(6635):884–886.

Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58(2):277–297.

Azoulay P, Graff Zivin JS, Li D, Sampat BN (2019) Public R&D investments and private-sector patenting: Evidence from NIH funding rules. *The Review of Economic Studies* 86(1):117–152.

Baruffaldi S, Gaessler F (2026) The returns to physical capital in knowledge production: Evidence from lab disasters. *American Economic Journal: Applied Economics* 18(1):248–85.

Besiroglu T, Bergerson SA, Michael A, Heim L, Luo X, Thompson N (2024) The compute divide in machine learning: A threat to academic contribution and scrutiny? *arXiv:2401.02452* .

Blundell R, Bond S (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87(1):115–143.

Cockburn IM, Henderson R, Stern S (2019) The impact of artificial intelligence on innovation: An exploratory analysis. Agrawal A, Gans J, Goldfarb A, eds., *The Economics of Artificial Intelligence: An Agenda*, 115–146 (University of Chicago Press).

De Loecker J, Goldberg PK, Khandelwal AK, Pavcnik N (2016) Prices, markups, and trade reform. *Econometrica* 84(2):445–510.

Gofman M, Jin Z (2024) Artificial intelligence, education, and entrepreneurship. *The Journal of Finance* 79(1):631–667.

Hsieh CT, Klenow PJ (2009) Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics* 124(4):1403–1448.

Jacob BA, Lefgren L (2011) The impact of research grant funding on scientific productivity. *Journal of Public Economics* 95(9–10):1168–1177.

Levinsohn J, Petrin A (2003) Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies* 70(2):317–341.

McKeon J, Boudou J (2025) Innovation under resource constraints: Supercomputing in scientific research. *SSRN Electronic Journal* .

Moser P, Voena A, Waldinger F (2014) German jewish émigrés and US invention. *American Economic Review* 104(10):3222–3255.

Olley GS, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64(6):1263–1297.

Restuccia D, Rogerson R (2008) Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics* 11(4):707–720.

Rubens M, Wu Y, Xu MJ (2026) Exploiting or augmenting labor? *American Economic Review: Insights* 8(1):72–89.

Sun L, Abraham S (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* 225(2):175–199.

Waldinger F (2016) Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge. *The Review of Economics and Statistics* 98(5):811–831.

Figures

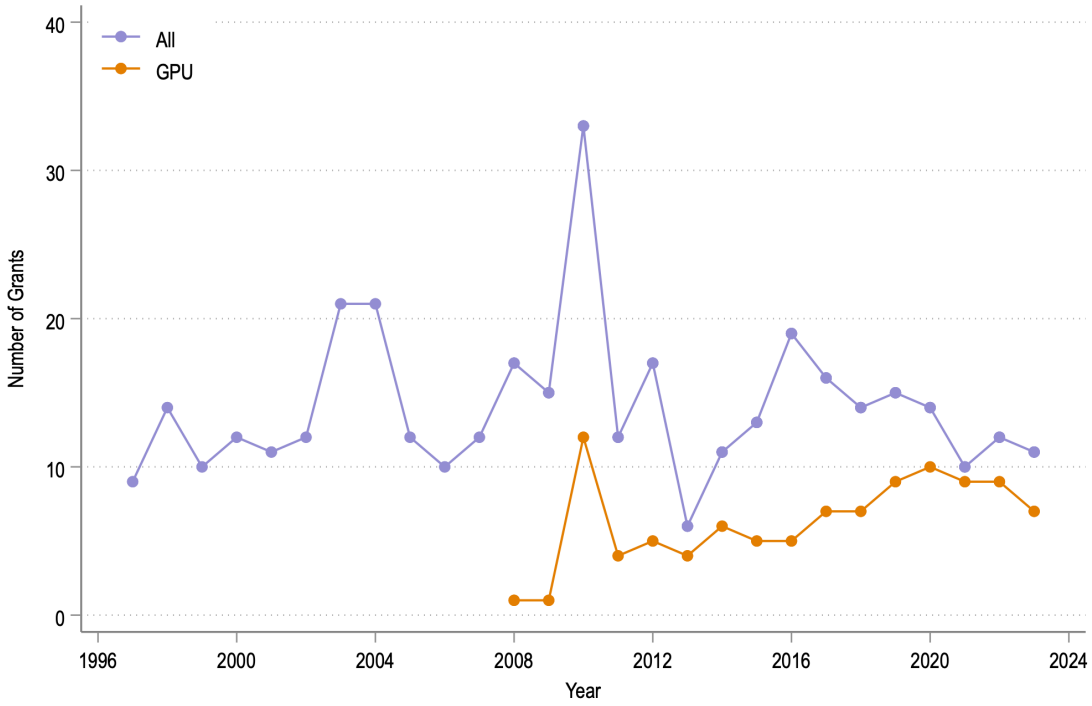


Figure 1: NSF Major Research Instrumentation Grants for High Performance Computing Clusters

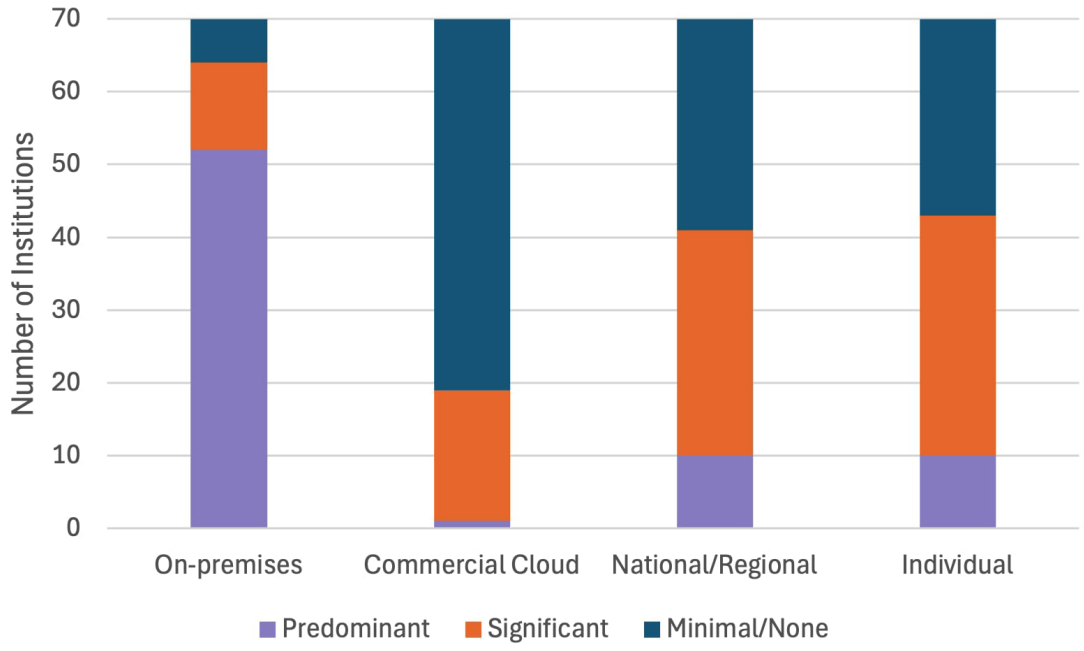


Figure 2: Survey from the Coalition for Academic Scientific Computing on Most Significant Computational Resource (2021).

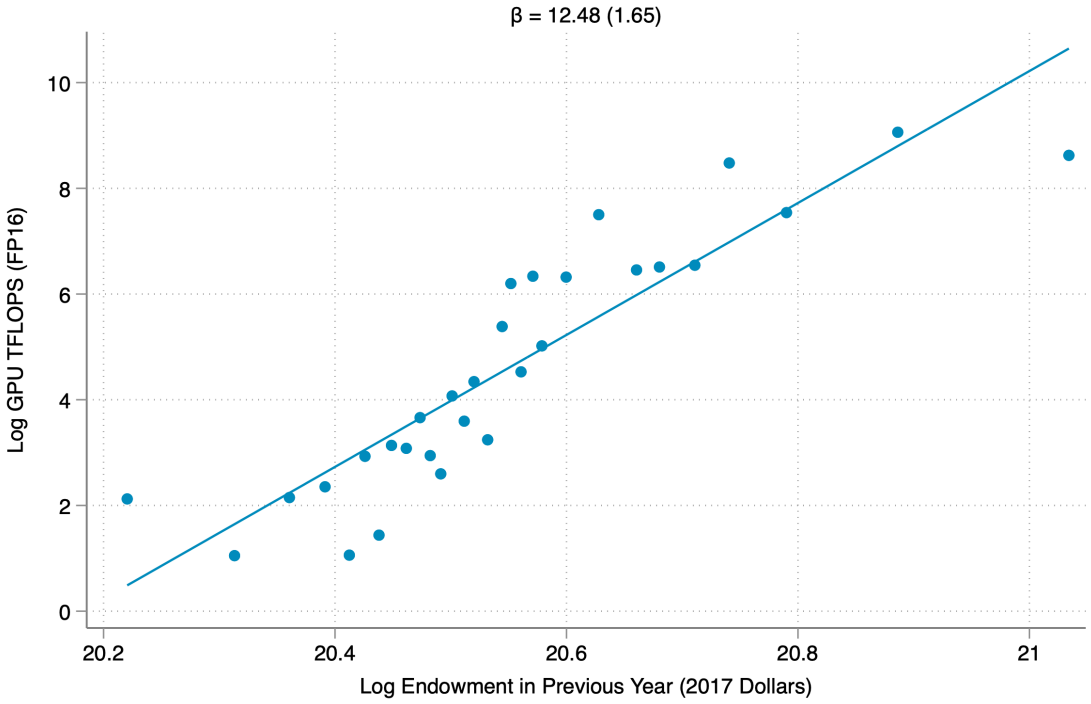


Figure 3: Binscatter of University Log GPU Capacity against Lagged Log Endowment Over Time (Controlling for University Fixed Effects).

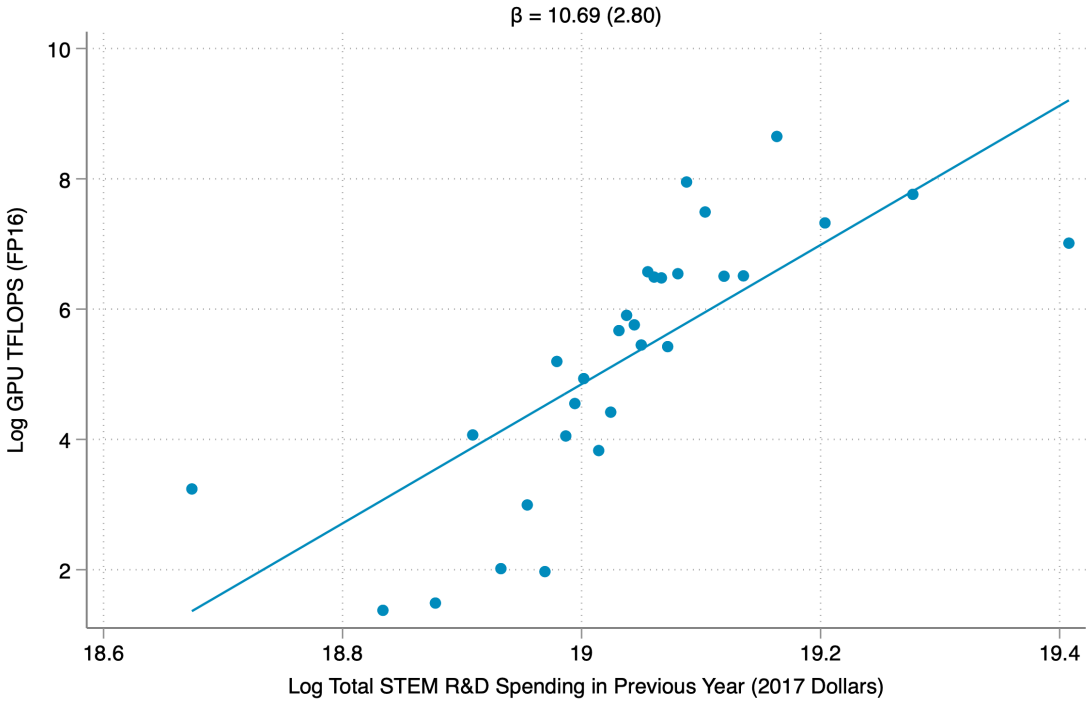


Figure 4: Binscatter of University Log GPU Capacity against Lagged Log STEM R&D Spending Over Time (Controlling for University Fixed Effects).

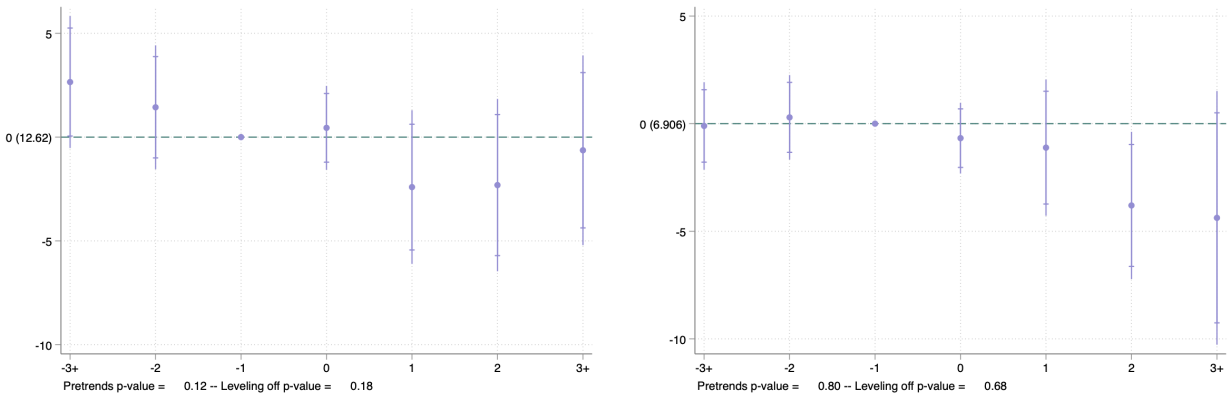


Figure 5: Event Study of Total AI Publications for All and Incumbents

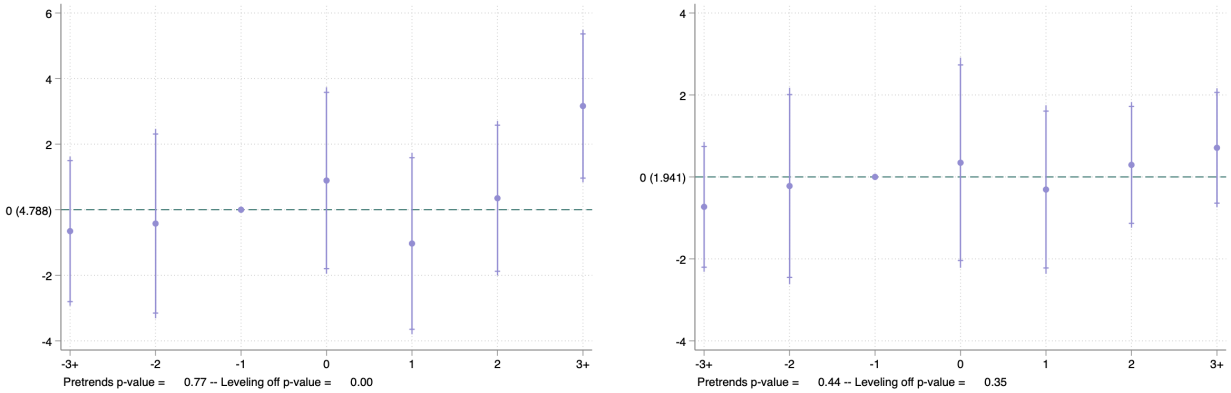


Figure 6: Event Study of Low-Tier AI Publications for All and Incumbents

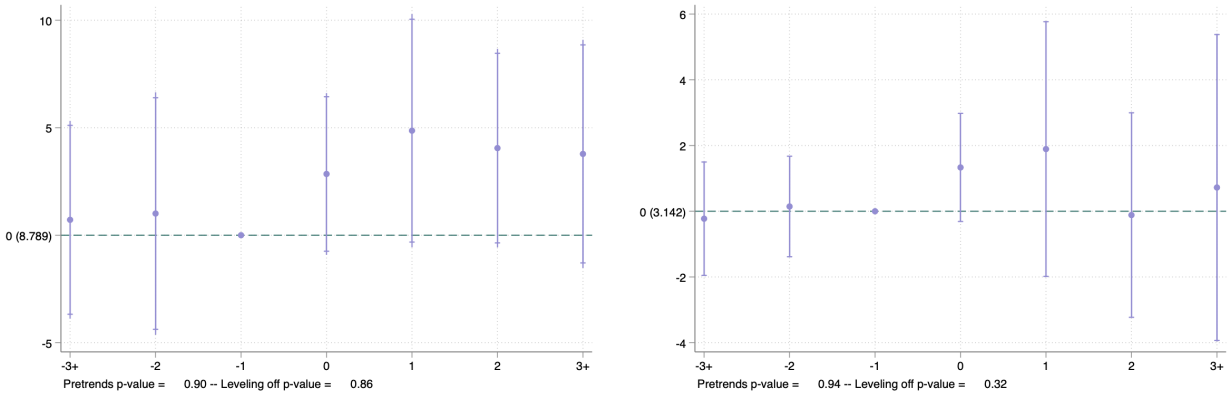


Figure 7: Event Study of Medium-Tier AI Publications for All and Incumbents

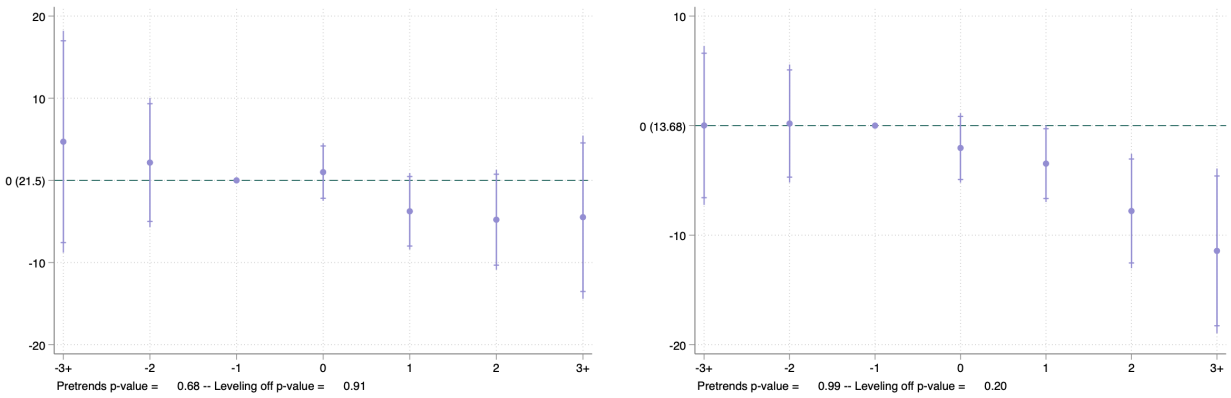


Figure 8: Event Study of High-Tier AI Publications for All and Incumbents

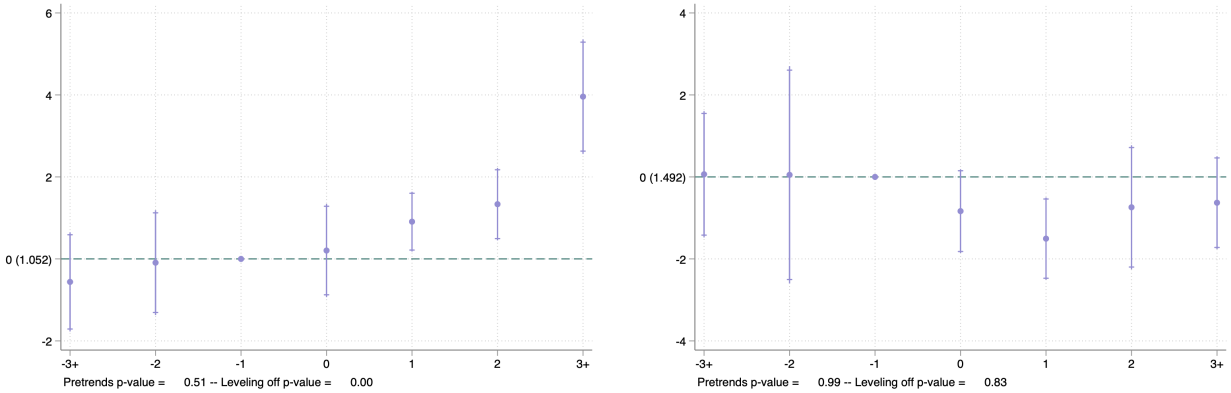


Figure 9: Event Study of GPU and Non-GPU Computer Science Publications for Low-Tier

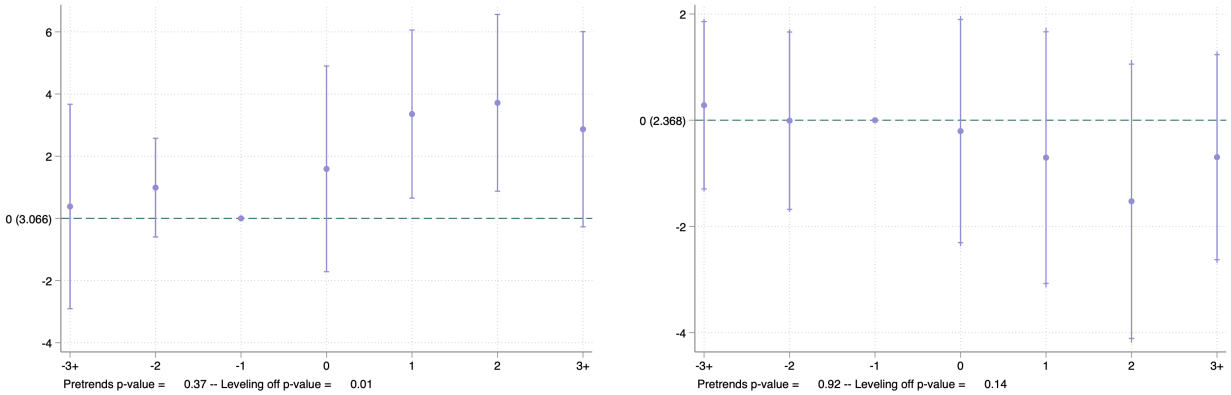


Figure 10: Event Study of GPU and Non-GPU Computer Science Publications for Medium-Tier

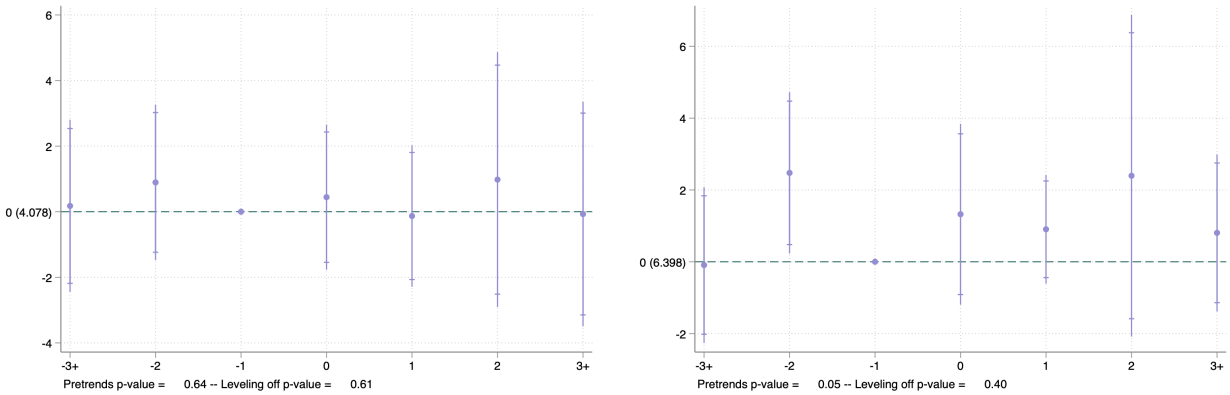


Figure 11: Event Study of GPU and Non-GPU Computer Science Publications for High-Tier

Appendix Figures

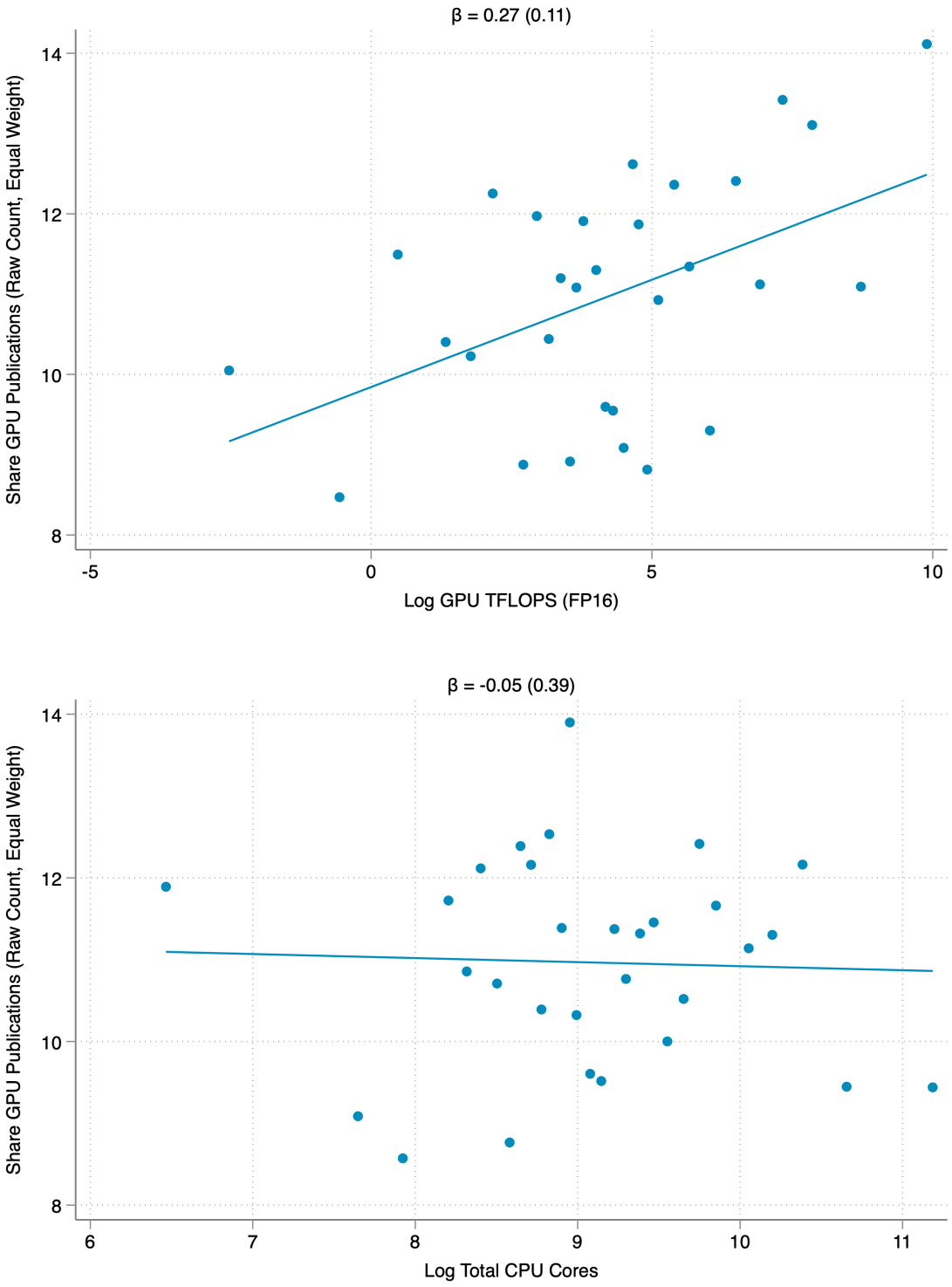


Figure A.1: Binscatter of Share GPU-Relevant Publications Against Log GPU and CPU Capacity (Across Institutions)

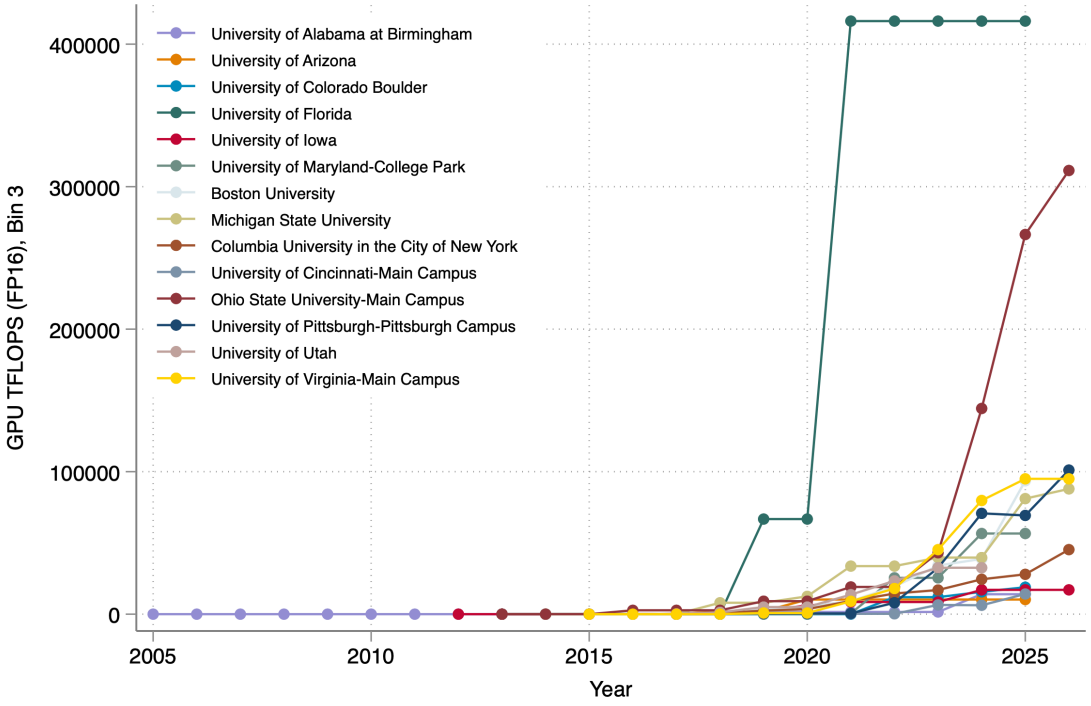
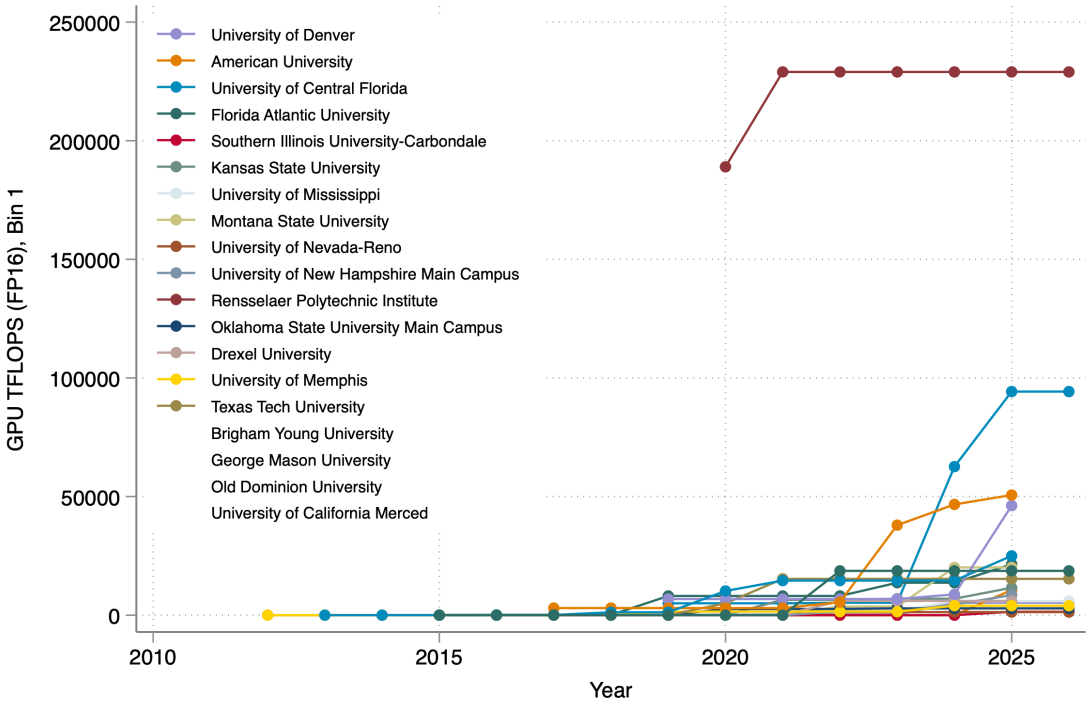


Figure A.2: GPU TFLOPS (FP16) Over Time (High Tier Universities)



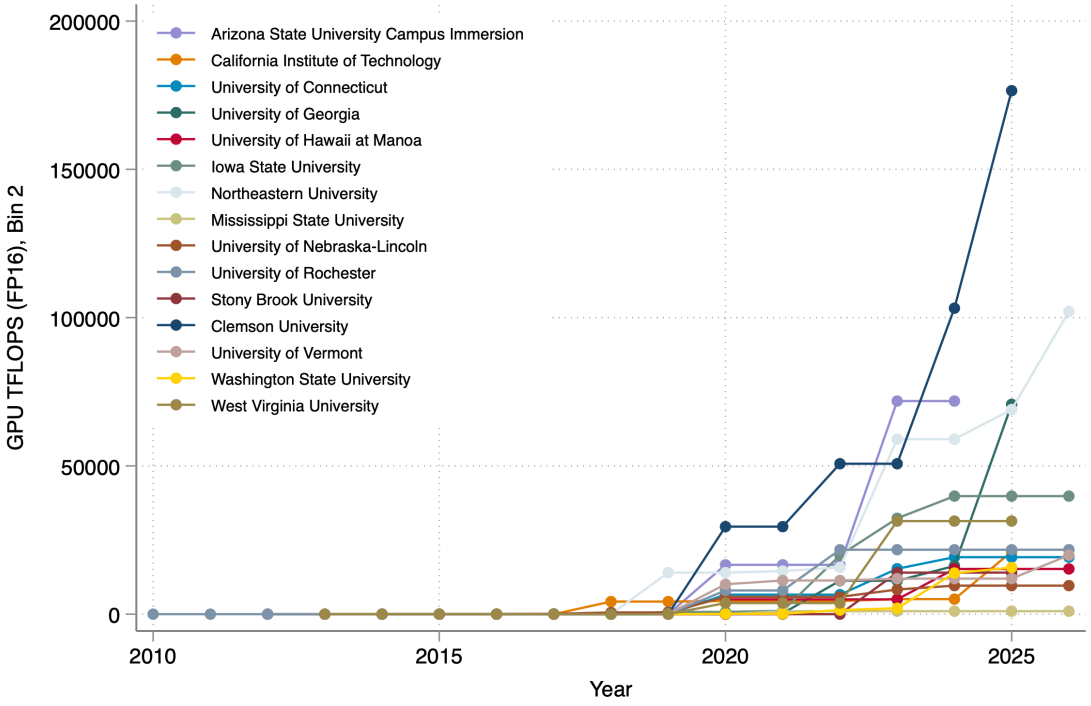


Figure A.3: GPU TFLOPS (FP16) Over Time (Low and Medium Tier Universities)

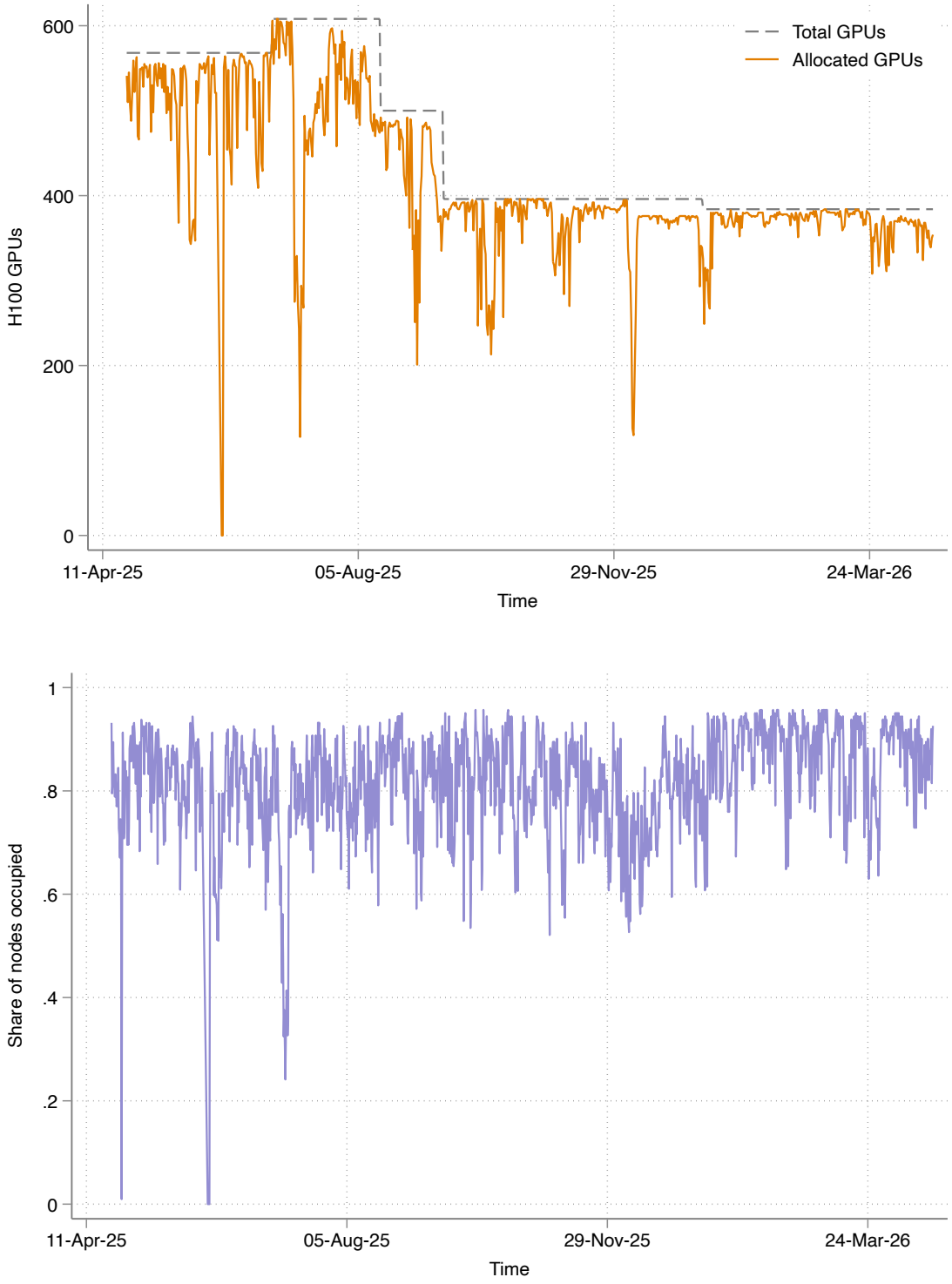


Figure A.4: Harvard Research Computing Total Utilization for NVIDIA H100 GPUs and All Nodes