The Returns to Science of Centrally Provisioned GPUs and Optimal Allocation

Connie Xu, Ruby Zhang*

March 26, 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—often funded by NSF grants—we implement an event-study design following Sun and Abraham (2021). 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 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: Innovation, Computing, GPU, 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.

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).

In this paper, we ask: what is the return on investment for these expensive GPU clusters? Furthermore, 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 will estimate the returns to GPU clusters using a proxy variable approach (Ackerberg et al. 2015) and the derived allocation rule for a subset of public universities where we observe cloud computing expenditure (from purchase order data). 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—a channel 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.

Returns to capital in science production. 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.

The compute divide and AI as a general-purpose technology. A growing body of work documents 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.

Production function estimation and capital misallocation. 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. We use proxy-variable methods from the production function literature (Olley and Pakes 1996, Levinsohn and Petrin 2003, Ackerberg et al. 2015) to address the simultaneity of capital investment and unobserved productivity shocks. 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 Section 5. We do not yet have estimation results, but we present our allocation framework and forthcoming counterfactual exercise in Section 6. We conclude in Section 7.

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.

For many public universities, particularly those outside the elite top-tier, the initial funding for

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

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 $\times$ field $\times$ year level. We restrict our analysis to the top 100 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 the top 100 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 25 universities which are manually collected and 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, presented in Figure 5.

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.

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 counts of post-doctoral students by field from the NSF Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). We will use the counts of faculty and students as a robustness check to our labour measures.

3.5 University Procurement and Spending

We observe all university purchase orders from GovSpend, a commercial data vendor that aggregates procurement records from public institutions. These transaction-level data contain item descriptions, prices, quantities, spending, purchase dates, supplier identifiers, and purchasing institutions. Unfortunately, we cannot distinguish hardware purchases for individual usage from that of the centralized cluster. However, we are able to observe total cloud computing purchases, which we will use as a proxy variable in a robustness check.

4 Event Study

We present preliminary event-study evidence from the subset of 25 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 ($\beta = 4.75$, s.e. $= 1.09$) and lagged STEM R&D spending ($\beta = 2.96$, s.e. $= 0.92$), 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 25 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—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:

$$Y_{jt} = \gamma_j + \lambda_t + \sum_{k=-4}^{4} \beta_k \mathbf{1}\{K_{jt} = k\} + \epsilon_{jt} \tag{4.1}$$

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—that 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 6 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 7 and 8 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 9 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 10 and 11 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 12 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 Empirical Model

Despite covariates, many universities invest continuously in computing capital over time, which runs into the fundamental simultaneity problem in productivity estimation: universities may adjust computing inputs in response to unobserved productivity shocks that also affect output. To address this, we model the university's optimization decision as a multi-product firm where each research field is a distinct "product" (in the spirit of De Loecker et al. 2016). The model serves two purposes: first, it delivers an allocation rule that maps aggregate university-level GPU capacity to field-level capital, which we cannot directly observe; second, it provides the timing assumptions needed for a proxy-variable production function estimator (Ackerberg et al. 2015) that addresses simultaneity.

5.1 Setup

Since the focus is on research production rather than revenue generation, we assume that university j maximizes an unknown weighted sum of publications across fields. University j enters period t with the following state variables: total GPU capacity K_{jt} (determined by last period's investment), a vector of quasi-fixed faculty across fields $\mathbf{L}^F_{jt}$ (determined last period), and a vector of (log) realized productivity shocks $\boldsymbol{\omega}_{jt}$ which follow a first-order Markov process.

Let λ_{jf} denote a time-invariant prestige weight that university j places on field f. The university makes two types of choices. *Dynamic (intertemporal) choices* include investment I_{jt} and faculty hiring $\mathbf{H}_{jt}$, which incur costs today but yield benefits tomorrow. *Static (intratemporal) choices* include flexible labor $\mathbf{L}^V_{jt}$ (postdocs and PhD students) and the allocation of total GPU capital across fields $\mathbf{K}_{jt}$, which are chosen to maximize the current period's payoff given the state.

The Bellman equation is:

$$V(K_{jt}, \mathbf{L}^F_{jt}, \boldsymbol{\omega}_{jt}) = \max_{I_{jt}, \mathbf{H}_{jt}, \mathbf{L}^V_{jt}, \mathbf{K}_{jt}} \left\{ \sum_f \lambda_{jf} Y_{jft} + \beta \mathbb{E}\left[V(K_{jt+1}, \mathbf{L}^F_{jt+1}, \boldsymbol{\omega}_{jt+1}) \mid \boldsymbol{\omega}_{jt} \right] \right\} \tag{5.1}$$

subject to the laws of motion:

$$K_{jt+1} = (1-\delta) K_{jt} + I_{jt} \tag{5.2}$$

$$L^F_{jft+1} = (1-\delta_F) L^F_{jft} + H_{jft} \tag{5.3}$$

$$\omega_{jft+1} = g(\omega_{jft}) + \xi_{jft+1} \tag{5.4}$$

and the following constraints:

$$Y_{jft} = \Omega_{jft}(L^V_{jft})^{\alpha_V}(L^F_{jft})^{\alpha_F}K^{\alpha_K}_{jft} \tag{5.5}$$

$$K_{jt} \geq \sum_f K_{jft} \tag{5.6}$$

$$B_{jt} \geq \sum_f W^V_{jt}L^V_{jft} + \sum_f W^F_{jt}L^F_{jft} + P^I_t I_{jt} \tag{5.7}$$

where $P^I_t I_{jt}$ captures the cost of capital investment and B_{jt} represents total available funding (e.g., R&D expenditures and endowment payouts).

5.2 Allocation Rule

The key advantage of the multi-product firm structure is that the static first-order conditions yield a closed-form rule for how GPU capital is allocated across fields—without requiring us to observe field-level capital directly.

We set up the period-t Lagrangian, letting μ_{jt} denote the multiplier on the capital constraint (5.6) and η_{jt} the multiplier on the budget constraint (5.7). The first-order condition for field-level capital K_{jft} is:

$$\lambda_{jf}\alpha_K \frac{Y_{jft}}{K_{jft}} = \mu_{jt} \tag{5.8}$$

and the first-order condition for flexible labor L^V_{jft} is:

$$\lambda_{jf}\alpha_V \frac{Y_{jft}}{L^V_{jft}} = \eta_{jt}W^V_{jt}. \tag{5.9}$$

Note that the continuation value $\beta\mathbb{E}[V(\cdot)]$ drops out of both conditions because it does not depend on the static choice variables.

Rearranging (5.9) gives $\lambda_{jf}Y_{jft} = \eta_{jt}W^V_{jt}L^V_{jft}/\alpha_V$. Substituting into (5.8):

$$\mu_{jt}K_{jft} = \frac{\alpha_K}{\alpha_V}\eta_{jt}W^V_{jt}L^V_{jft}. \tag{5.10}$$

Summing across all fields and using the binding capital constraint ($\sum_f K_{jft} = K_{jt}$):

$$\mu_{jt}K_{jt} = \frac{\alpha_K}{\alpha_V}\eta_{jt}W^V_{jt}\sum_f L^V_{jft}. \tag{5.11}$$

Taking the ratio of the field-specific equation to the aggregate eliminates μ_{jt}, η_{jt}, W^V_{jt}, and the production function parameters, yielding the allocation rule:

$$\boxed{K_{jft} = K_{jt} \cdot \frac{L^V_{jft}}{\sum_f L^V_{jft}}} \tag{5.12}$$

That is, each field's share of GPU capital equals its share of flexible labor. Importantly, this result does not depend on the unobserved prestige weights λ_{jf}, field-level output Y_{jft}, or the production function elasticities—it holds for any values of $(\alpha_V, \alpha_F, \alpha_K)$ under the Cobb-Douglas specification.

Validating the Allocation Rule Equation (5.12) is a testable prediction. We plan three validation exercises.

First, for a small sub-sample of universities where field-level compute allocations are directly observed—through published HPC usage reports or data from the NSF ACCESS/XSEDE program—we can regress the actual field-share of compute on the flexible labor share from OpenAlex. A coefficient close to one would provide direct evidence that the structural assumption holds empirically.

Second, the allocation rule implies short-run crowd-out across fields: holding total K_{jt} fixed, a positive shock to flexible labor in one field mechanically reduces the capital share available to

other fields. We can test this by regressing field f's output on flexible labor in other fields $f' \neq f$, controlling for university fixed effects and total capacity. A negative coefficient would be consistent with the allocation rule binding in practice.

Third, once production function estimates are obtained, we can use the structural parameters together with the allocation rule to predict the field-level output response to a university-level capacity shock, and compare this prediction to the reduced-form event-study estimates from Section 4. Agreement between the two would serve as a joint validation of the model.

5.3 Production Function Estimation

We estimate the production function at the university × field × year level using the ACF procedure (Ackerberg et al. 2015). Letting $\tilde{k}_{jft}$ denote field-level allocated capital from Equation (5.12), the log-linear production function is:

$$y_{jft} = \alpha_V l^V_{jft} + \alpha_F l^F_{jft} + \alpha_K \tilde{k}_{jft} + \omega_{jft} + \epsilon_{jft} \tag{5.13}$$

where y_{jft} is log citation-weighted output, l^V_{jft} and l^F_{jft} are log flexible and quasi-fixed labor respectively, ω_{jft} is unobserved productivity, and ϵ_{jft} is measurement error.

Identification relies on the following timing assumptions. Quasi-fixed labor (l^F_{jft}) and capital ($\tilde{k}_{jft}$) are determined at $t-1$: tenured faculty take time to hire, and total GPU capacity is installed before the period begins.[2] Flexible labor (l^V_{jft}) is chosen at t after observing the current productivity shock ω_{jft}. Productivity follows a first-order Markov process, $\omega_{jft} = g(\omega_{jft-1}) + \xi_{jft}$, where ξ_{jft} is the innovation.

We require an intermediate proxy variable that is flexible and monotonically increasing in ω_{jft} conditional on state variables. Candidate proxies include cloud computing expenditure from GovSpend (available for the subset of public universities), HERD non-personnel operational expenditures, or field-level grant volume.

Stage 1. We assume proxy demand is strictly monotonic in productivity: $m_{jft} = m(l^F_{jft}, \tilde{k}_{jft}, \omega_{jft})$. Inverting and substituting into (5.13), all inputs are absorbed into a non-parametric function:

$$y_{jft} = \Phi(l^V_{jft}, l^F_{jft}, \tilde{k}_{jft}, m_{jft}) + \epsilon_{jft}. \tag{5.14}$$

We approximate $\Phi(\cdot)$ with a second- or third-order polynomial and save the predicted values $\hat{\Phi}_{jft}$, which purge measurement error but do not identify any structural parameters.

Stage 2. For a candidate parameter vector $\theta = (\alpha_V, \alpha_F, \alpha_K)$, we compute implied productivity:

$$\omega_{jft}(\theta) = \hat{\Phi}_{jft} - \alpha_V l^V_{jft} - \alpha_F l^F_{jft} - \alpha_K \tilde{k}_{jft} \tag{5.15}$$

and recover the innovation $\xi_{jft}(\theta)$ as the residual from a non-parametric regression of $\omega_{jft}(\theta)$ on $\omega_{jft-1}(\theta)$. The structural parameters are identified by the moment conditions:

$$\mathbb{E}\left[\xi_{jft}(\theta) \begin{pmatrix} l^F_{jft} \\ \tilde{k}_{jft} \\ l^V_{jft-1} \\ e_{jft} \end{pmatrix}\right] = \mathbf{0} \tag{5.16}$$

where e_{jft} denotes field-specific HERD capitalized equipment expenditure, included as an additional instrument to address measurement error in scraped GPU capacity.[3] The first two moments exploit

[2] The internal allocation of capital across fields occurs at t via the allocation rule, but the total capacity on which it operates is fixed at $t-1$.

[3] HERD equipment expenditure is determined at $t-1$ by budget cycles and thus satisfies the orthogonality condition with respect to the period-t innovation ξ_{jft}.

the fact that quasi-fixed labor and capital are predetermined; the third instruments for flexible labor with its own lag; and the fourth provides an external instrument for capital.

6 NSF Allocation and Counterfactuals

The heterogeneity in returns documented in Section 4 has direct implications for federal infrastructure funding. If less-resourced universities exhibit higher marginal returns to GPU capital, then reallocating infrastructure grants toward these institutions could increase aggregate scientific output. In this section, we outline a framework for (i) recovering university-level marginal products of computing capital, (ii) simulating counterfactual grant allocations, and (iii) estimating the NSF's revealed preference over efficiency versus equity. We do not yet have production function estimates, so we present the framework and defer the quantitative results to a subsequent draft.

6.1 Recovering Heterogeneous Marginal Products

The goal is to recover university-specific marginal products of computing capital, $MPK_j = \partial Y_j / \partial K_j$. We pursue two approaches.

First, using the event-study estimates from Section 4, we can recover university-level elasticities for the subset of universities that experience large capacity shocks. Let $\hat{\gamma}_j$ denote the estimated percentage effect on output and $\hat{\gamma}_j^K$ the percentage change in GPU capacity at the event. The implied elasticity is:

$$\hat{\beta}_j^K \approx \frac{\hat{\gamma}_j}{\hat{\gamma}_j^K} \tag{6.1}$$

and the marginal product is $MPK_j = \hat{\beta}_j^K \times Y_j / K_j$, where baseline output and capital are measured prior to the event.

Second, the structurally estimated production function from Section 5.3 yields $MPK_j = \hat{\alpha}_K \cdot \hat{Y}_j / K_j$ for all universities in the sample, not just those with discrete upgrade events. We can then examine the dispersion in MPK_j across universities and test whether it varies systematically with baseline resources.

To characterize returns to scale, we estimate the following cross-sectional regression:

$$\hat{\gamma}_j = \alpha_0 + \alpha_1 \log(K_j) + \alpha_2 \log(\Delta K_j) + \theta X_j + \epsilon_j \tag{6.2}$$

where X_j includes observables such as total funding, faculty size, and field composition, and ΔK_j controls for dosage. The coefficient α_1 is the key parameter: $\alpha_1 < 0$ implies diminishing returns to baseline capital, meaning that less-endowed universities have higher marginal returns and there is no equity-efficiency tradeoff. If $\alpha_1 > 0$, returns are increasing and the planner faces a genuine tradeoff between concentrating resources at productive institutions and distributing them broadly.

6.2 Efficiency vs. Equity

We model the NSF as a social planner distributing a fixed infrastructure budget $\bar{K}_{NSF}$ across universities. The planner maximizes an Atkinson-style social welfare function parameterized by an inequality aversion parameter $\rho \geq 0$:

$$W = \sum_{j=1}^{J} \frac{Y_j^{1-\rho}}{1-\rho} \tag{6.3}$$

When $\rho = 0$, the objective reduces to $W = \sum_j Y_j$ (pure efficiency: maximize total output regardless of distribution). As $\rho \to 1$, the function approaches $W = \sum_j \ln(Y_j)$, placing increasing weight on proportional equity. In the limit $\rho \to \infty$, the objective becomes Rawlsian: $W = \min_j Y_j$.

The first-order condition for the optimal allocation equates equity-weighted marginal products across universities:

$$Y_j^{-\rho} \cdot MPK_j = \lambda \quad \forall j \tag{6.4}$$

where λ is the shadow value of the budget constraint. When $\rho = 0$, this reduces to the standard condition that MPK_j is equalized. When $\rho > 0$, universities with lower baseline output receive a higher effective weight, and the planner tolerates unequal marginal products in order to distribute scientific capacity more broadly.

6.3 Planned Counterfactual Exercises

Using the estimated MPK_j values, we plan three exercises.

Pure efficiency counterfactual. We take the observed total NSF GPU/equipment budget and reallocate it to equalize MPK_j across universities, iteratively assigning the marginal dollar to the university with the highest current MPK_j and updating capital stocks (which lowers MPK_j due to diminishing returns when $\alpha_K < 1$). This yields the maximum feasible total output and serves as the efficiency benchmark.

Evaluating the status quo. We compute the total scientific output implied by the actual allocation of NSF MRI and CC* grants in our data and compare it to the efficiency benchmark. The gap between the two represents the output cost—if any—of the NSF's current allocation, which may reflect equity considerations such as funding institutions in EPSCoR states to build geographic research capacity.

Revealed preference estimation. Taking logs of the planner's first-order condition (6.4):

$$\ln(MPK_j) = \ln(\lambda) + \rho \ln(Y_j) \tag{6.5}$$

For the subset of universities that received NSF infrastructure funding in a given year, we regress the structurally estimated $\ln(MPK_j)$ on baseline $\ln(Y_j)$. The coefficient on log-output is a direct estimate of $\hat{\rho}$, the NSF's implicit inequality aversion. If $\hat{\rho} \approx 0$, the NSF is allocating as a pure efficiency maximizer. If $\hat{\rho}$ is positive and statistically significant, the NSF is deliberately directing funds toward universities with lower marginal products—accepting some aggregate output loss in exchange for a more equitable distribution of scientific capacity.

7 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.

7.1 Next Steps

This paper is preliminary. Several extensions are in progress. First, we are expanding the university computing capacity panel from 25 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 proxy-variable approach outlined in Section 5.3, 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 6, 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.

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* .

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.

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

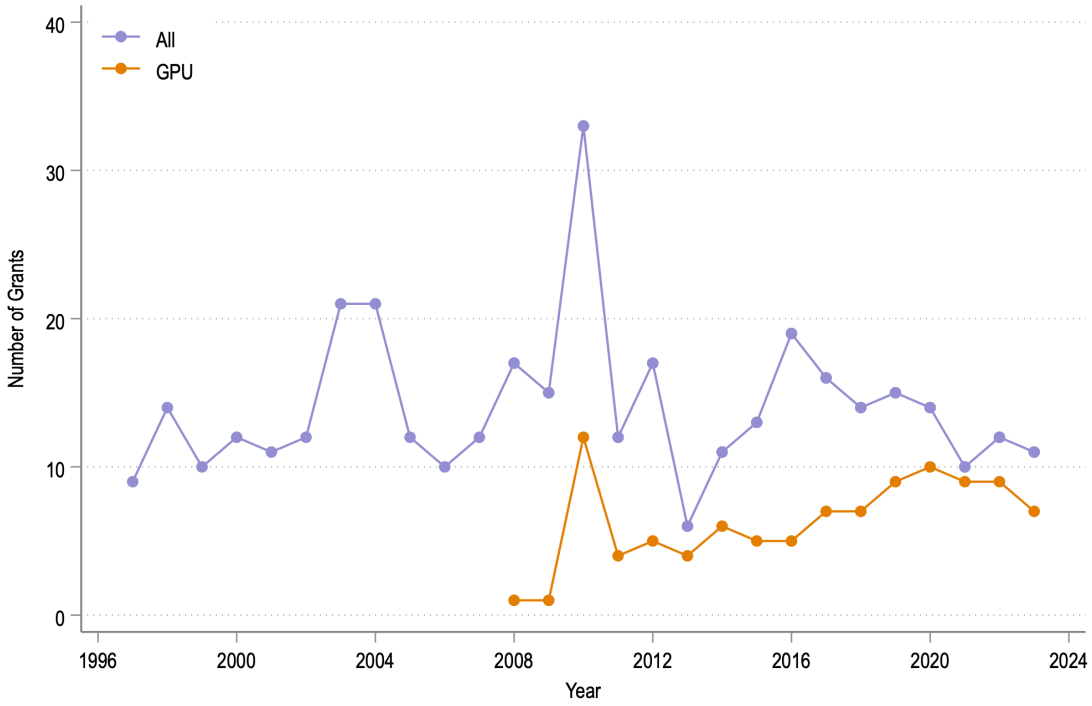


Figure 1: NSF Major Research Instrumentation Grants for High Performance Computing Clusters

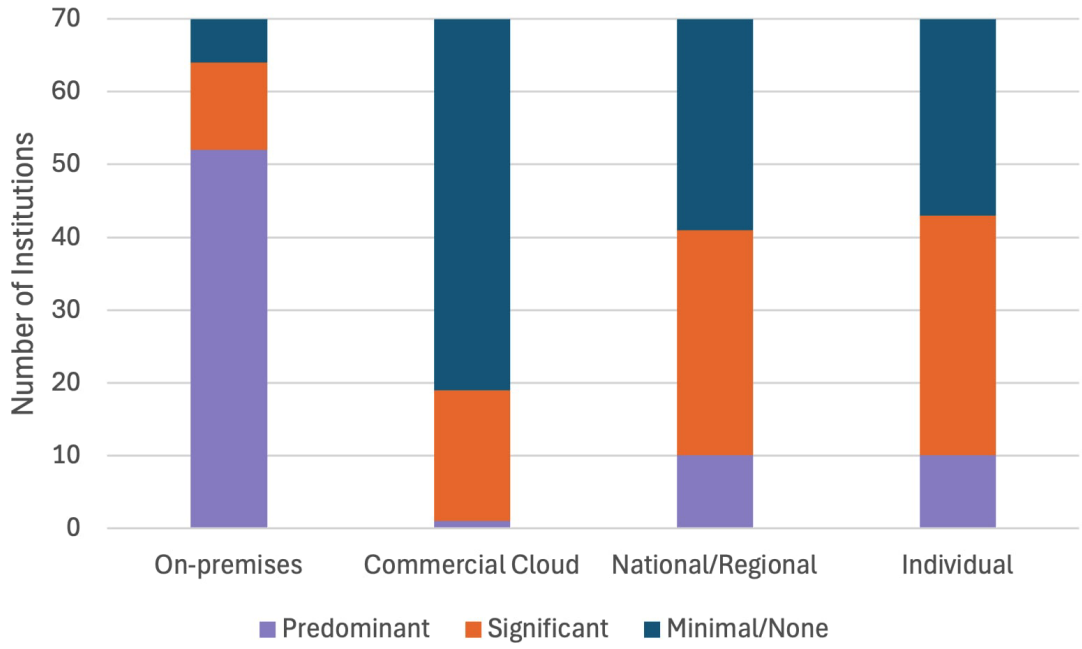


Figure 2: Survey from the Coalition for Academic Scientific Computing on Most Significant Computational Resource (2021).

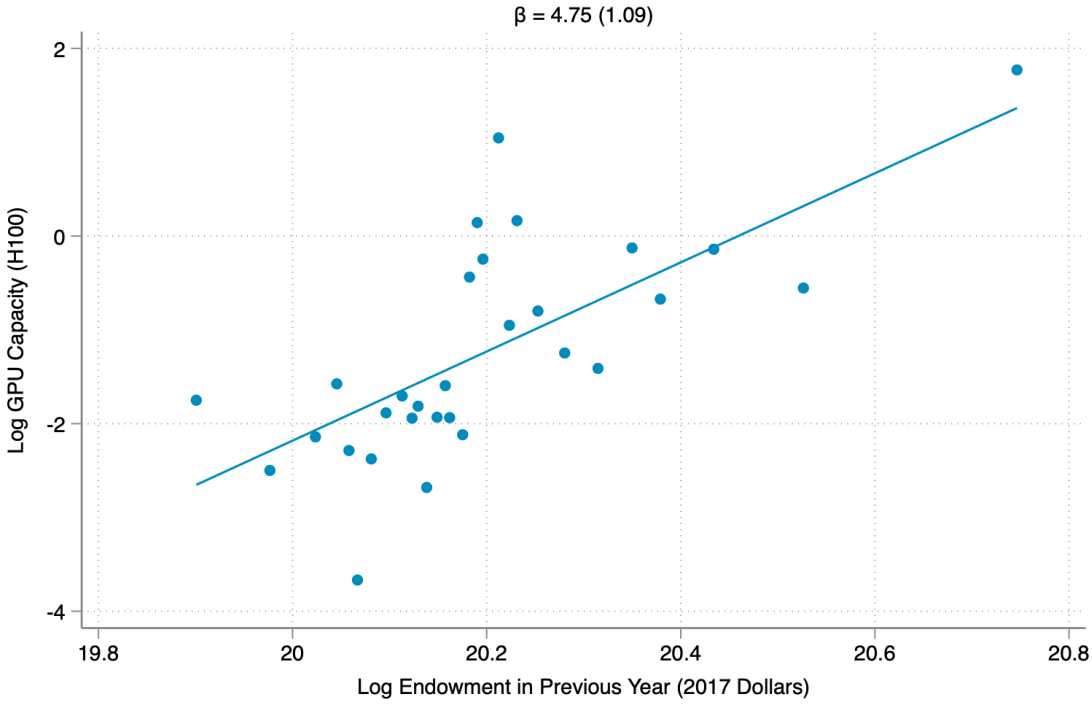


Figure 3: Binscatter of University Log GPU Capacity against Lagged Log Endowment Over Time (Controlling for University Fixed Effects).

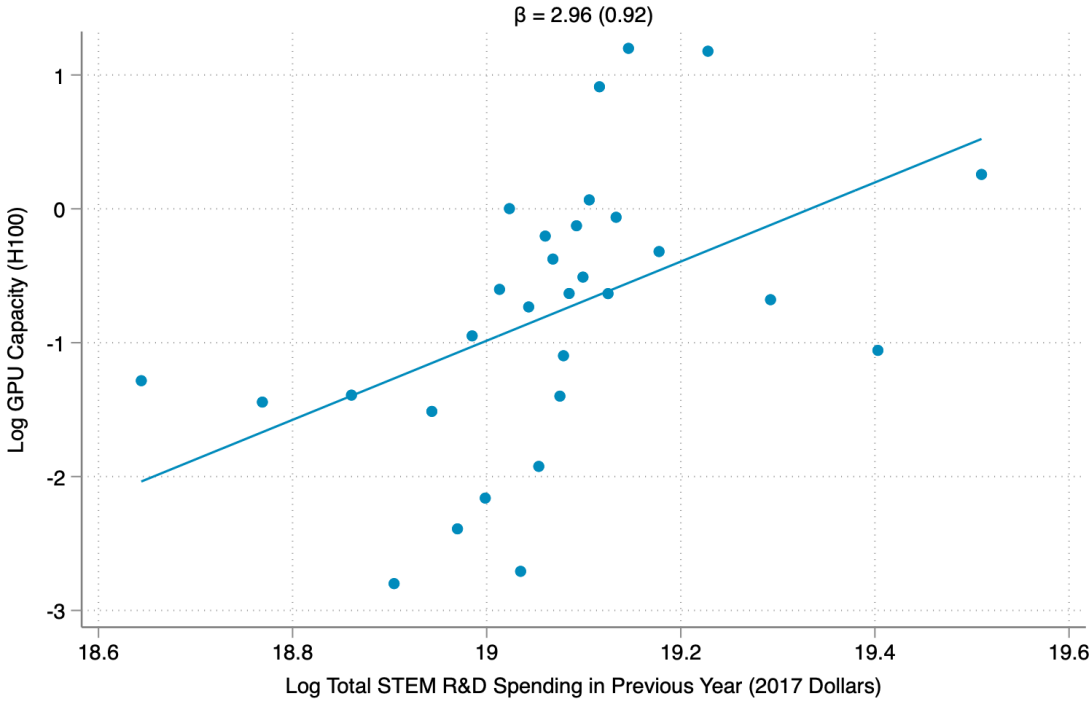


Figure 4: Binscatter of University Log GPU Capacity against Lagged Log STEM R&D Spending Over Time (Controlling for University Fixed Effects).

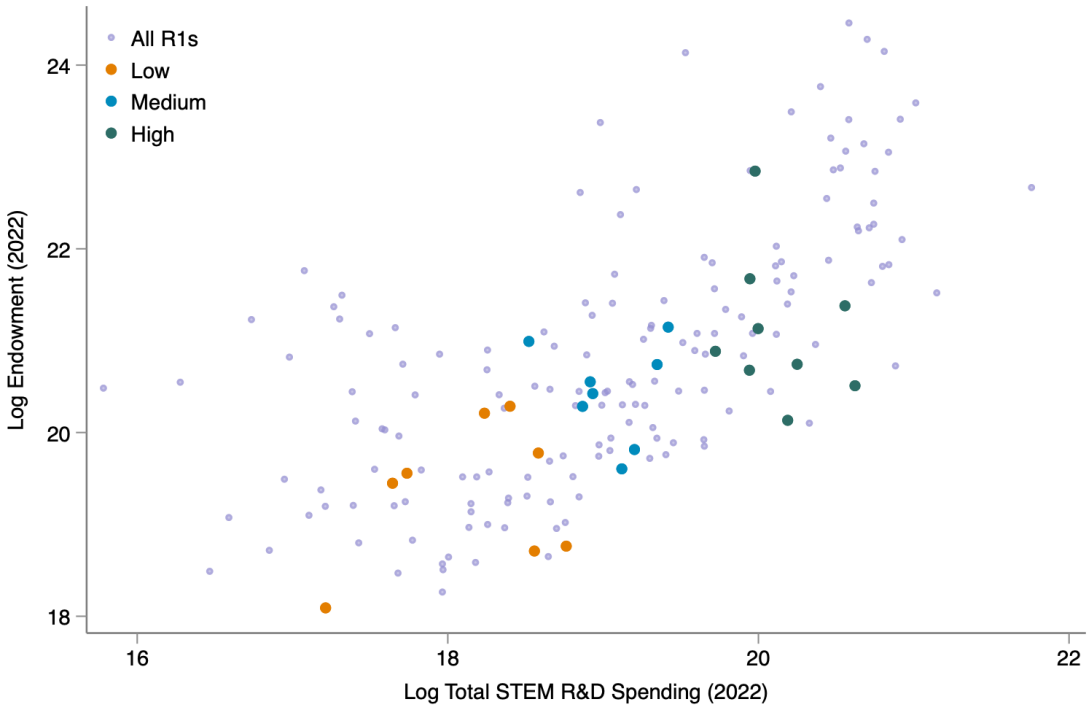


Figure 5: Endowment vs. STEM R&D Spending of All R1 Universities (2022)

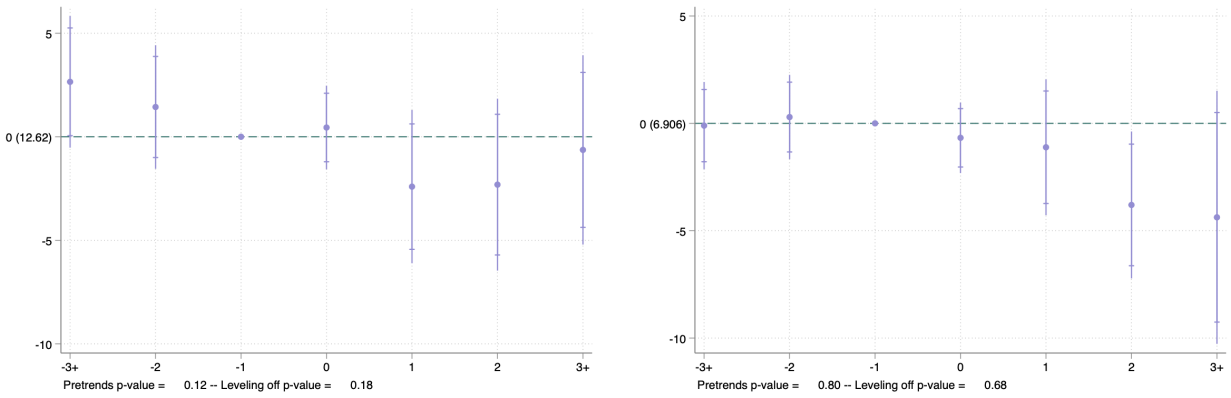


Figure 6: Event Study of Total AI Publications for All and Incumbents

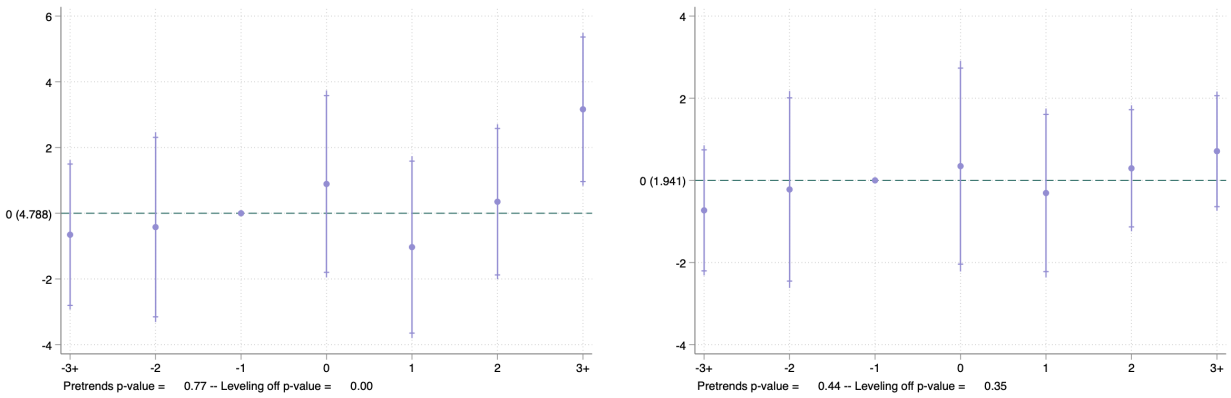


Figure 7: Event Study of Low-Tier AI Publications for All and Incumbents

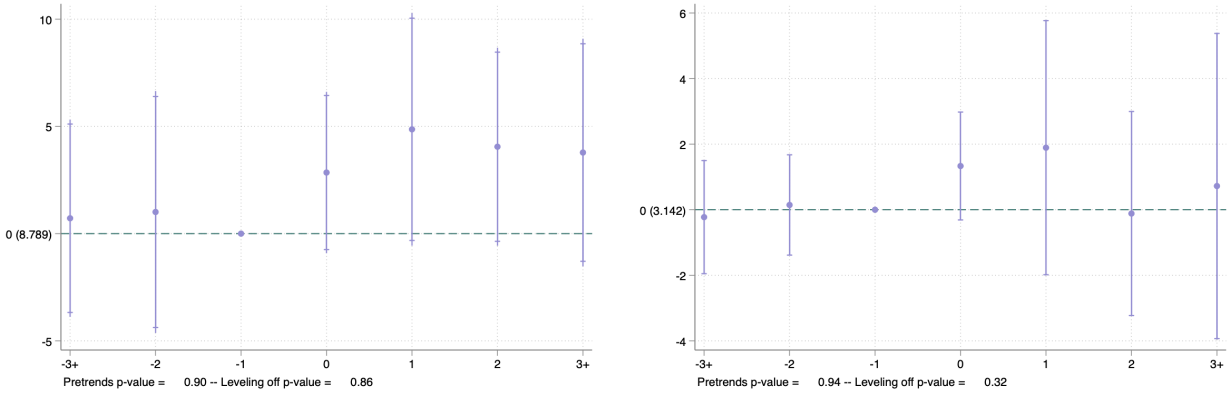


Figure 8: Event Study of Medium-Tier AI Publications for All and Incumbents

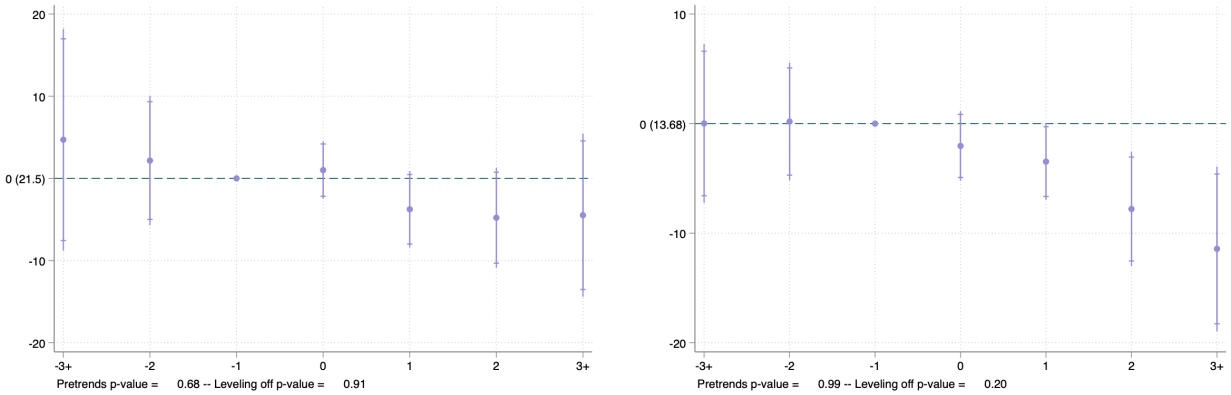


Figure 9: Event Study of High-Tier AI Publications for All and Incumbents

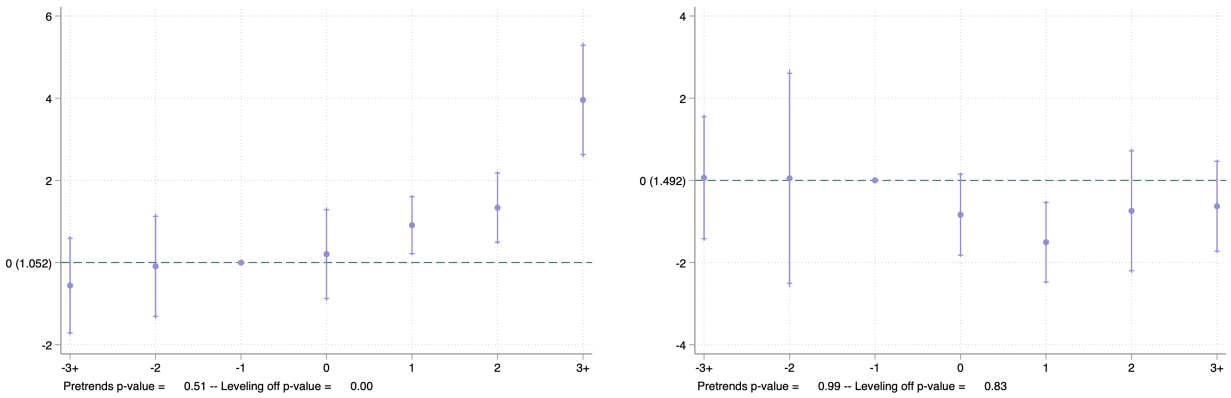


Figure 10: Event Study of GPU and Non-GPU Computer Science Publications for Low-Tier

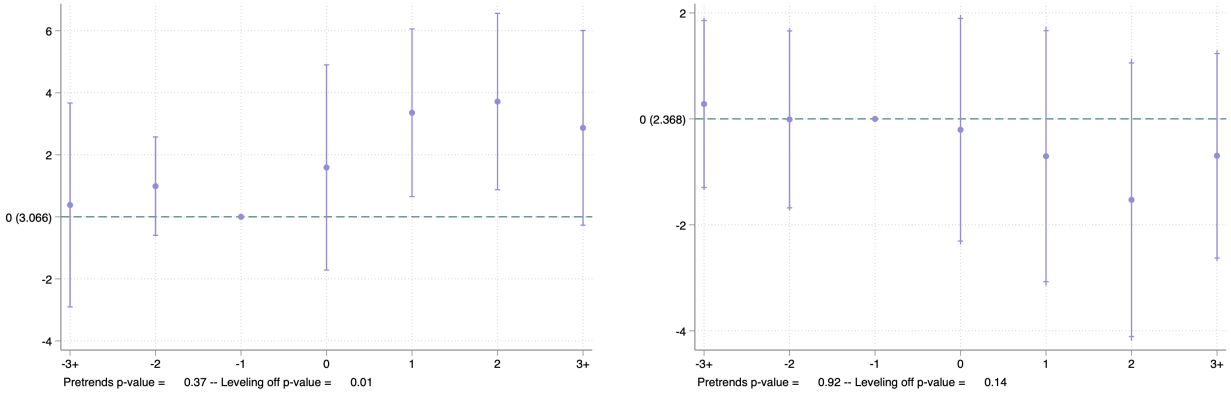


Figure 11: Event Study of GPU and Non-GPU Computer Science Publications for Medium-Tier

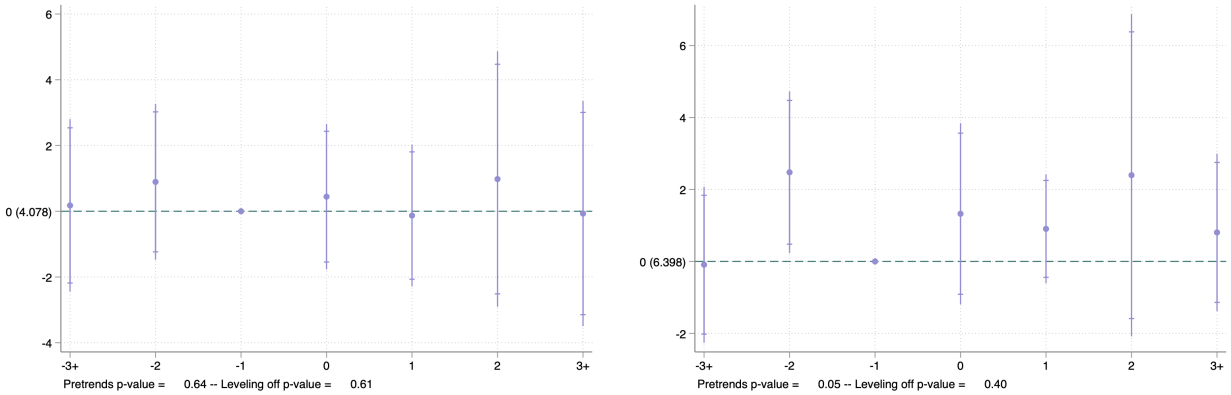


Figure 12: Event Study of GPU and Non-GPU Computer Science Publications for High-Tier

Appendix Figures

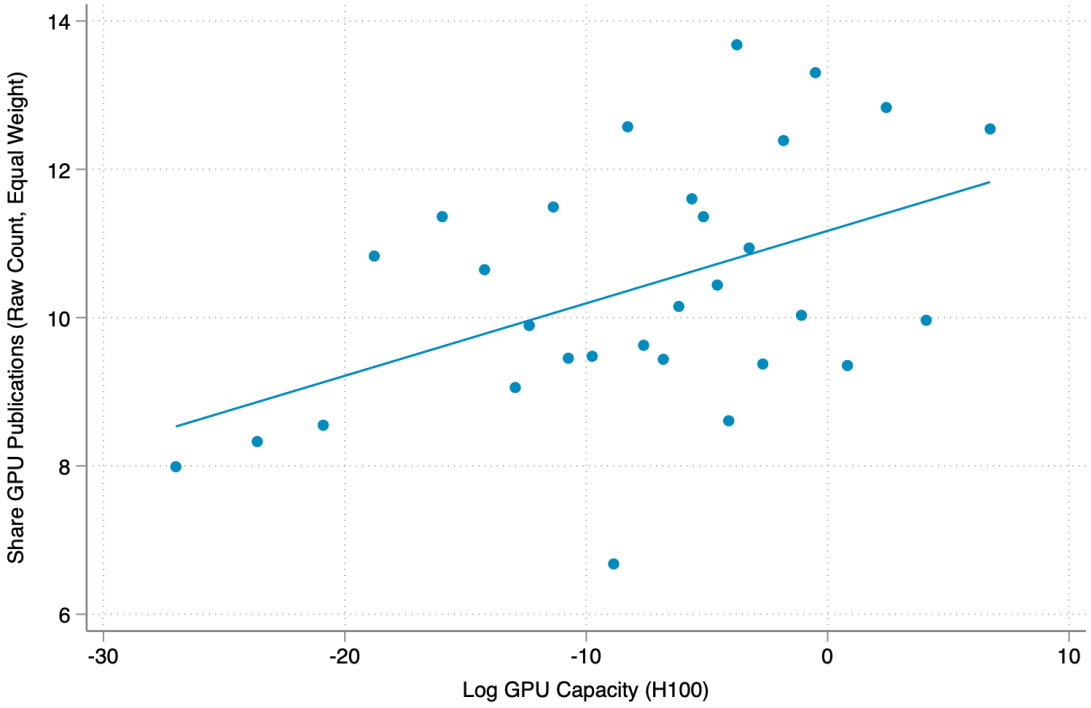


Figure A.1: Binscatter of Share GPU-Relevant Publications Against Log GPU Capacity (Across Institutions)

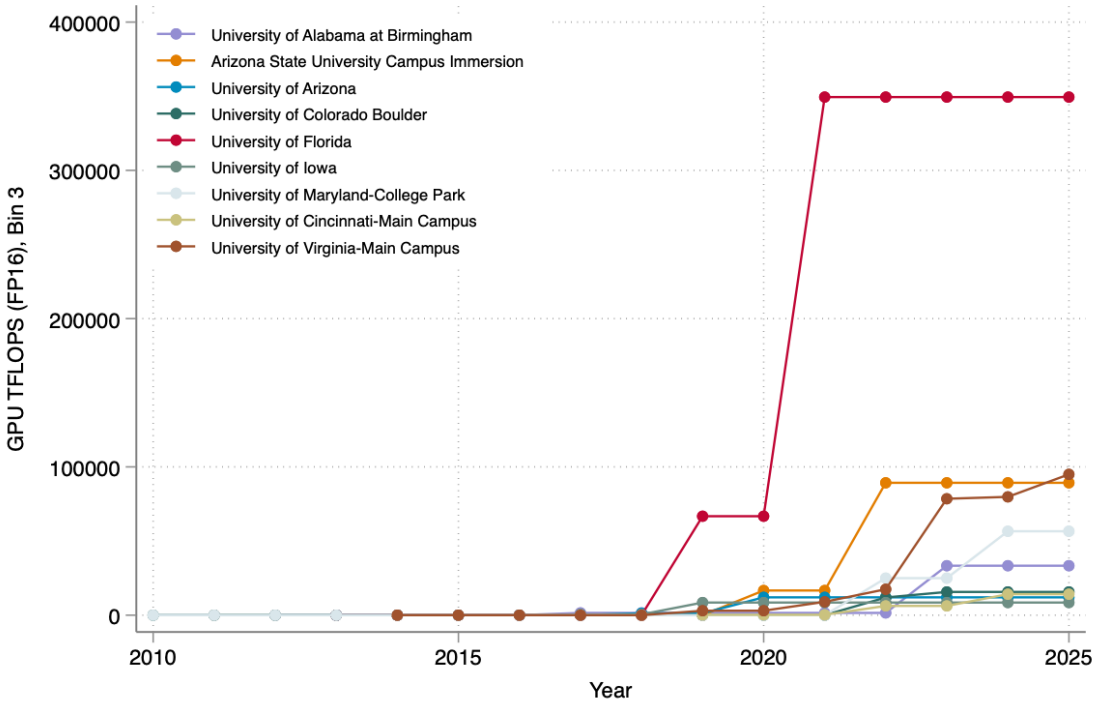


Figure A.2: GPU TFLOPS (FP16) Over Time (High Tier Universities)

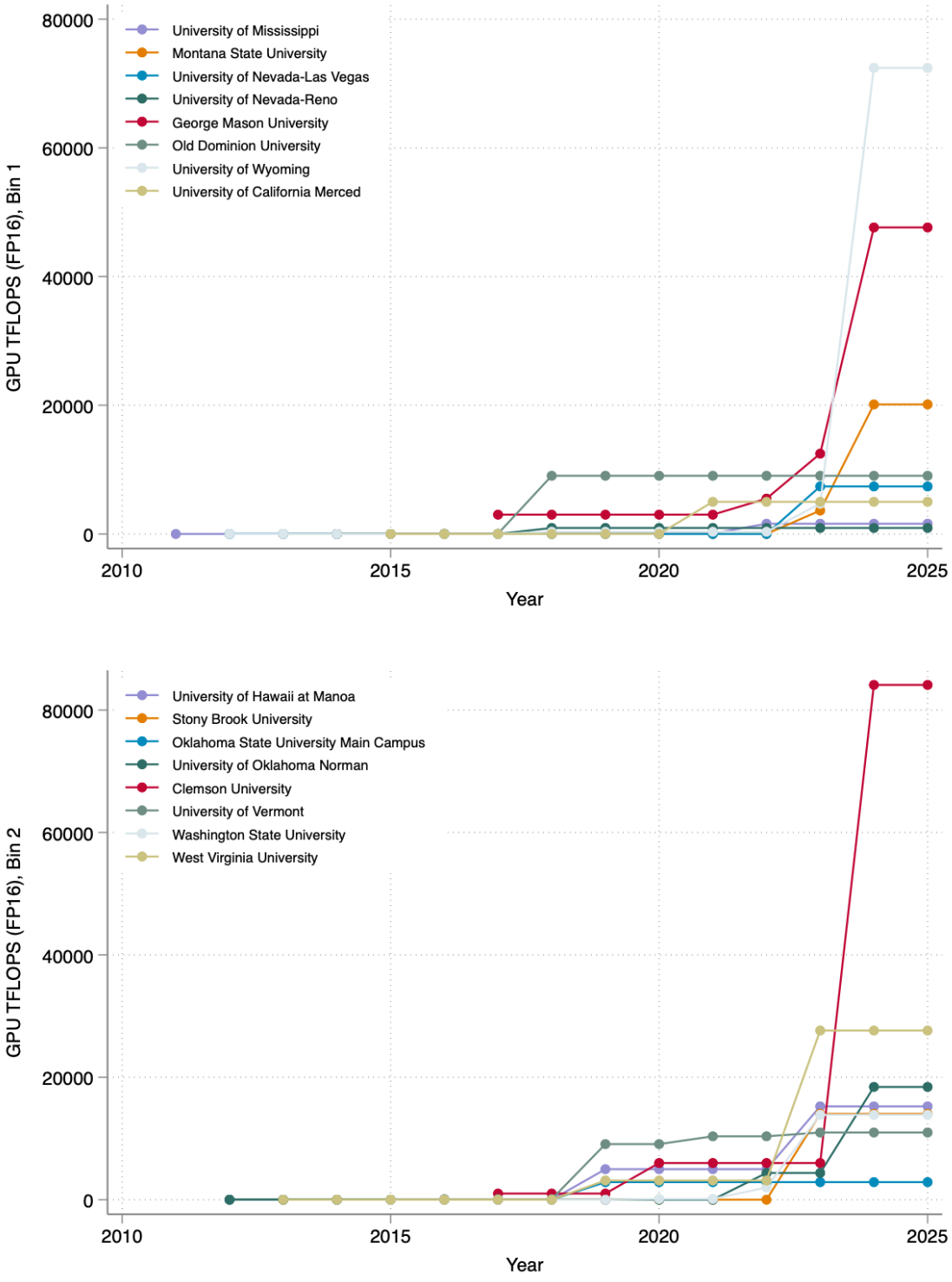


Figure A.3: GPU TFLOPS (FP16) Over Time (Low and Medium Tier Universities)