Returns to Science of Centrally Provisioned GPUs and Optimal Allocation

Connie Xu, Ruby Zhang (Job Market Paper)*

September 6, 2026

Abstract

Should federal infrastructure funding go to more productive universities or to universities where external funding is most likely to relieve financing constraints? We construct a novel panel of CPU and GPU capacity at U.S. R1 universities from historical research-computing websites and link it to publication and researcher histories from OpenAlex. Exploiting large, discrete capacity expansions in a staggered difference-in-differences design, we find substantial heterogeneity in the impact of upgrades on computer science paper publications. Less-resourced universities experience persistent gains, while effects at well-resourced universities are short-lived. We model the university as a multi-product organization, estimate field-level production elasticities with respect to computing capital and researchers, and use the implied marginal products of capital to evaluate counterfactual funding allocations. In a counterfactual holding fixed all inputs except GPU capital, a purely publication-maximizing allocation concentrates capital funding more heavily at relatively high-output institutions than the observed portfolio from the National Science Foundation (NSF), consistent with the agency valuing a broader distribution of scientific capacity.

Keywords: Computing, GPU, Multi-Product Production, Grant Allocation

*(**Corresponding Author**) **Zhang:** Harvard University. Email: rzhang15@g.harvard.edu. **Xu:** Harvard University. Email: conniexu@g.harvard.edu. We are deeply grateful to Shane Greenstein, Robin Lee, Ariel Pakes, and Jesse Shapiro for their guidance and support throughout this project. 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

Graphics processing units (GPUs) have become essential inputs in not only frontier artificial intelligence (AI) and deep-learning research, but also an expanding set of computational workflows in biology, chemistry, engineering, and the physical sciences. This shift has made scientific production more capital-intensive (Cockburn et al. 2019, Suh 2026), and therefore universities increasingly operate high fixed-cost, on-premise high-performance computing (HPC) centers with GPU clusters. Unequal access to this infrastructure may in turn reinforce the "compute divide" between research institutions with and without the financial capacity to fund computing capital (Ahmed and Wahed 2020, Ahmed et al. 2023, Besiroglu et al. 2024).

Federal programs such as the National Science Foundation's (NSF) Major Research Instrumentation (MRI) and Campus Cyberinfrastructure (CC*) programs finance many of these discrete, university-level investments, particularly outside the set of universities able to fund large clusters internally.[1] However, it is not clear ex ante where scarce funding should be allocated. A well-resourced university has stronger complementary inputs such as productive researchers, but also has better access to outside options—cloud computing and national facilities—and can self-finance HPC investments. A less-resourced university may not be as productive at baseline, but also has a tighter financing constraint preventing investments in talent and capital, which increases the marginal value of an external grant. Policymakers must consider the total return to an external grant after accounting for the university's endogenous private investment and adjustment of complementary research labor. There are dynamic considerations even beyond total research output—building up distributed scientific capacity generates local economic spillovers (Hausman 2022), trains the future scientific workforce (Gofman and Jin 2024), and increases the diversity of research directions (Hofstra et al. 2020).

We ask two related questions. First, how much does an expansion in university GPU capacity change scientific output, and how do those returns vary with institutional resources? Second, given the returns to GPU capacity, what allocation of a fixed federal cyberinfrastructure budget would maximize aggregate output, and how does the observed NSF allocation pattern reveal the agency's willingness to trade off aggregate output against equal distribution of scientific capacity?

[1] The NSF is a dominant federal funder in this space: its Directorate for Computer and Information Science and Engineering (CISE) estimates that it provides 87% of federal funding for fundamental computer science research at U.S. academic institutions (National Science Foundation 2021).

We answer these questions by constructing a novel panel of university computing capacity and combining estimates from an event study and a scientific production function. Using historical Wayback Machine snapshots of research-computing websites for 53 U.S. R1 universities, we recover the CPU and GPU hardware installed in each year. Heterogeneous GPU models and vintages are converted into a common measure of theoretical peak performance: FP16 TFLOPS.[2] We then combine the HPC capacity panel with publication and author histories from OpenAlex, using text analysis to identify computational research across computer science, engineering, geosciences, the physical sciences, and the life sciences.

Our first empirical design exploits the lumpy timing of major capacity expansions. We define an event as a within-year increase exceeding five FP16 TFLOPS per research-personnel headcount and estimate interaction-weighted event studies for staggered adoption following Sun and Abraham (2021).[3] We compare universities within baseline endowment tiers and decompose the publication response into incumbent researchers, who were already present before the upgrade, and entrants. The difference between the all-researcher and incumbent outcomes is informative about the importance of adjusting complementary assets, namely talent (Teece 1986, Bresnahan et al. 2002).

Our second approach estimates a scientific production function. We model the university as a multi-product organization in which fields are separate research products and centrally provisioned CPU and GPU stocks must be allocated across them. Field-level output is produced with an outer Cobb–Douglas technology and nested CES aggregates of CPU versus GPU capital and flexible versus quasi-fixed labor. Flexible labor is measured as counts of PhDs and postdocs and assumed to be adjustable within each period. Quasi-fixed labor is measured as counts of faculty (inferred from publication data). Because field-level compute allocations are not observed, we use the university's static first-order conditions to derive allocation rules that map aggregate capital into field-level inputs. We estimate the labor nest from a labor first-order condition and estimate the outer output elasticities using dynamic-panel IV–GMM moments, allowing for persistent shocks to productivity in each field within each university. Finally, we develop a framework that combines the estimated marginal products with a reduced-form investment policy function to study counterfactual NSF allocations and the

[2]FP16 TFLOPS is trillions of 16-bit ("half-precision") floating-point operations per second, the arithmetic standard for deep-learning workloads. Theoretical peak is the vendor-rated maximum rather than realized performance. Because every GPU model reports this specification, it provides a consistent benchmark for aggregating hardware across vendors and vintages. Section 4 describes the harmonization in detail.

[3]The event threshold translates to at least one NVIDIA A100 for every 60 graduate students, postdocs, and staff scientists, or about 30 A100s at the median university. These are large discrete increases: only the top third of year-over-year capacity increases are above the threshold.

agency's revealed efficiency–equity tradeoff.

Results from our first empirical design show modest average effects but substantial heterogeneity across tiers of endowment. Across all fields, large GPU upgrades produce limited changes in aggregate AI-related publications. The response is sharper in computer science, where publication lags are shorter. Low-tier universities experience a persistent increase of roughly 50% in computer science publications after an upgrade. At high-tier universities, publications initially rise by about 30%, but the effect dissipates within a year. Medium-tier universities display an economically similar but imprecisely estimated increase. Roughly half of the low-tier response is produced by incumbent researchers, while the medium-tier response is driven primarily by entrants. This pattern is consistent with both a direct productivity return to existing researchers and a dynamic response of complementary talent (Moser et al. 2014, Gofman and Jin 2024).

Turning to the production function, we find that a 10% expansion of a university's computing stock raises predicted computational output by roughly 2.6% and 1% in computer science and life sciences, respectively. The capital elasticity is smaller and statistically indistinguishable from zero in engineering and the physical sciences. Using the university's first-order condition for flexible labor, we find that flexible labor (PhD students and postdocs) and quasi-fixed labor (faculty) are strong substitutes in producing computational papers, and the estimator rejects Cobb–Douglas. The field-level capital elasticities are similar under various levels of labor substitution elasticities. The capital substitution elasticity is not identified at the current level of aggregation, so we calibrate CPU and GPU capital to be complements, and the current specification treats persistent productivity as a fixed effect. Because the capital elasticities are well below one, the marginal product of GPU capital diminishes quickly in the installed stock. Thus, an output-maximizing planner does not simply fund the most productive universities, but equalizes marginal products across universities—favoring those with strong baseline output and relatively small existing GPU stocks.

We then reallocate the observed FY2021–2025 MRI and CC* computing budget among universities in the estimation panel. In the current counterfactual, labor and CPU stocks are fixed, and each grant dollar translates directly into installed GPU capacity, assuming no endogenous investment response. At the 2021 baseline, the observed allocation raises predicted computational output by 242 papers relative to a no-grant baseline of roughly 7,300 papers, while an output-maximizing allocation of the same $30.5 million budget raises output by 544 papers. This translates to a gain of 3.3% versus 7.5% of baseline output, or 7.9 versus 17.8 papers per million dollars. The observed

allocation therefore captures about 45% of the attainable output gain, and the corresponding share at the 2020 baseline is 49%. The efficient allocation concentrates funding among universities with high baseline computational output but relatively small installed GPU stocks. Computer science experiences the largest proportional increase, while the life sciences account for the largest level increase. A revealed-preference exercise based on the planner's first-order condition implies an inequality-aversion parameter between 0.40 and 0.98 across the two baselines. A parameter value of zero maps to an efficiency-only objective, which maximizes aggregate output, while a parameter value of one maps to an objective that maximizes *proportional* output increases across universities. Given NSF's many objectives, we interpret this parameter as a reduced-form summary of the agency's allocation rule.

This paper contributes to three strands of literature. First, it contributes to work on physical capital and the production of scientific knowledge. Existing research studies losses of laboratory space and specialized equipment, the rising capital intensity of science, and the productivity effects of research grants (Waldinger 2016, Baruffaldi and Gaessler 2026, Suh 2026, Jacob and Lefgren 2011, Azoulay et al. 2019). The closest paper is McKeon and Boudou (2025), who use quasi-random allocations in NSF's national ACCESS program to estimate the intensive-margin effect of supercomputing allotments on individual scientists. We instead study local university infrastructure, which determines broad access for researchers who may not use national facilities, and emphasize heterogeneity across institutions.

Second, we contribute to the literature on the concentration of AI research capacity. Non-elite universities and academic-only teams have become less represented in compute-intensive research as industry and elite institutions increasingly control compute, data, and talent (Ahmed and Wahed 2020, Ahmed et al. 2023, Besiroglu et al. 2024). Furthermore, the general movement of AI faculty toward industry has downstream consequences for training, entrepreneurship, and research direction (Gofman and Jin 2024, Yue 2024). We provide causal evidence on whether centrally provisioned university infrastructure can offset these disparities and show why the answer depends on both baseline productivity and the financing response of the institution. A related strand of literature shows that the cost of and access to shared research inputs shape who participates in science and which directions advance: institutions that lower the cost of biological materials, data, and research tools disproportionately broaden participation beyond elite teams (Furman and Stern 2011, Nagaraj et al. 2020, Teodoridis 2018). We study computing capital, an input whose expense and lumpiness make institutional provision, and therefore institutional inequality, central to these dynamics. Finally, we

adapt multi-product production function methods to the university. As in De Loecker et al. (2016), aggregate inputs must be allocated across products that are separately observed in output. We combine static first-order conditions with dynamic-panel moments (Arellano and Bond 1991, Blundell and Bond 1998, Rubens et al. 2026) to recover field-level capital allocations and production elasticities in a setting with nested capital and labor inputs. This framework connects the economics of knowledge production to the broader literature on allocative efficiency (Hsieh and Klenow 2009, Restuccia and Rogerson 2008).

The remainder of this paper proceeds as follows. Section 2 describes the institutional context, and Section 3 presents a stylized model. Section 4 describes the data construction. Section 5 presents the event-study framework and reduced-form results. Section 6 develops the structural model of university production and investment, and Sections 7 and 8 describe estimation and results. Section 9 presents the allocation framework and counterfactual exercises. Section 10 concludes.

2 High-Performance Computing at Universities

High-performance computing (HPC) involves the aggregation of computing power to solve problems that are too large for standard desktop computers. In practice, an HPC cluster will be composed 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 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 toward Graphics 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 five years due to depreciation and warranty expiration, implying an average depreciation rate of 20%. Universities generally adopt one of three models to provide this capital to

researchers. The most common model is a central HPC cluster—funded by indirect costs or external grants—which faculty can access for free up to a quota. The second model is a condo model, where faculty use their own grant money to purchase nodes that are then 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. Most universities combine the first two models. The last model is a recharge model—universities operate HPC clusters as a service center or core facility and charge their researchers hourly rates to recover costs.[4] The prices are set with the objective of cost recovery, not profit maximization, which reduces the reliance on funding from indirect costs.

From the researcher's 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 own funding. 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.[5] Another channel is national HPC facilities through NSF ACCESS or the Department of Energy (DOE). However, national facilities serve only the right tail of computational scientists with very high computational needs. A 2021 survey from the Coalition for Academic Scientific Computing found that on-premise clusters remain the most significant computational resource for many universities (Figure 1).

2.1 External Funding for Research Computing

Federal support for academic research computing flows primarily through the NSF. Within the NSF's Directorate for Computer and Information Science and Engineering (CISE)—which estimates that it provides 87% of federal funding for fundamental computer science research at U.S. academic institutions—the Office of Advanced Cyberinfrastructure (OAC) is the unit provisioning advanced computing, data, software, and networking resources for the national research community (National Science Foundation 2021). A large share of OAC's investment funds national, leadership-class systems that are part of the ACCESS program, which allocates time on shared national resources. Our focus

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

[5]In 2011, the Department of Energy's Magellan project concluded that its computing centers were typically 3 to 7 times less expensive than equivalent commercial cloud capacity (U.S. Department of Energy, Office of Science 2011). The report highlighted two conditions under which commercial cloud would be cheaper: uncertain or sporadic demand, and facility capacity constraints. In 2019, Emeras et al. (2019) still reach a similar conclusion comparing Amazon EC2 against an in-house HPC cluster.

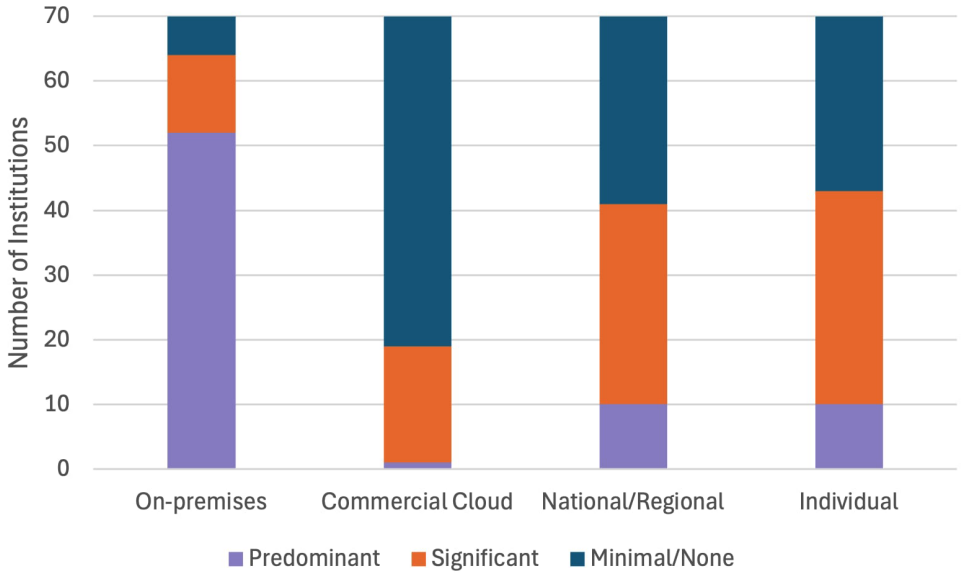


Figure 1: Survey from the Coalition for Academic Scientific Computing on Most Significant Computational Resource (2021). For each type of computational resource—on-premise clusters, commercial cloud, national/regional facilities, and individually owned hardware—bars report the number of surveyed institutions rating it a Predominant, Significant, or Minimal/None part of their computing. On-premise clusters are the only resource a majority of institutions rate as predominant.

is instead on the university-level layer of this ecosystem: on-premise clusters owned and operated by individual institutions.

Broadening access to research computing is an explicitly stated goal of the agency. NSF's cyberinfrastructure blueprint calls for "democratizing availability of and access to" cyberinfrastructure (National Science Foundation, Office of Advanced Cyberinfrastructure 2021), the ACCESS program's stated goal is "ensuring democratized and equitable access" to the national ecosystem (National Science Foundation 2022), and in January 2024 the NSF launched the National Artificial Intelligence Research Resource (NAIRR) pilot to democratize access to AI-scale computing, data, and models (NAIRR Task Force 2023, National Science Foundation 2024b).

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. MRI describes itself as a capacity-building program that increases access to shared, multi-user instrumentation across diverse institution types; awards range from $100,000 to $4 million, and the program's long-standing 30% institutional cost-share requirement was waived for five years beginning with the FY2023 solicitation under the CHIPS and Science Act (National Science Foundation 2023). CC* funds coordinated campus-level computing and networking driven by science

needs and explicitly encourages proposals that overcome geographic disparities in cyber-connectivity (National Science Foundation 2024a). These grants are discrete and substantial, often funding the purchase of an entire new cluster or a significant partition of GPUs. Since 2010, the number of HPC MRI grants related specifically to GPU clusters has increased substantially, as seen in Figure 2.

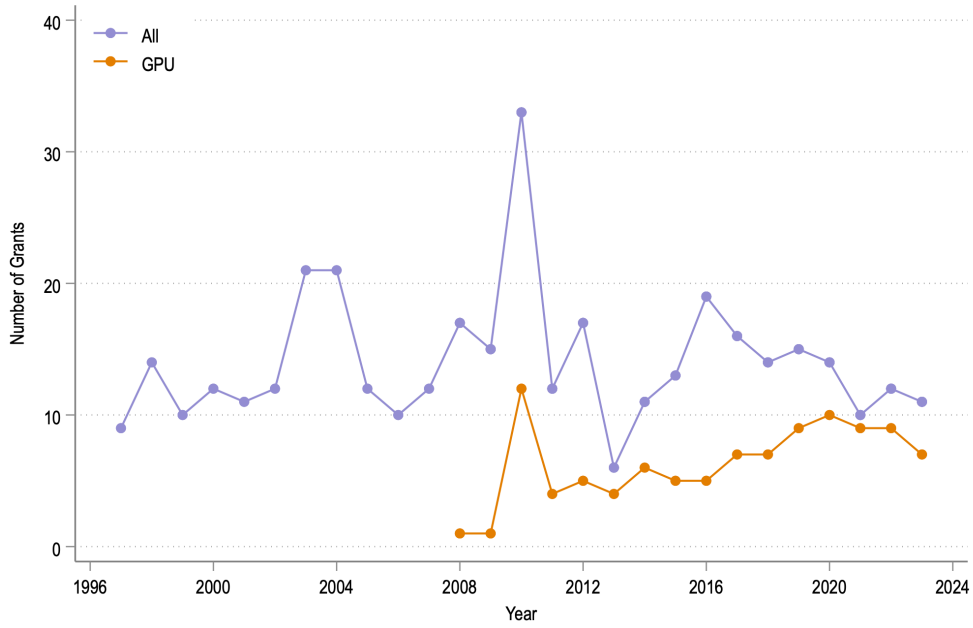


Figure 2: NSF Major Research Instrumentation Grants for High-Performance Computing Clusters. Annual counts of NSF Major Research Instrumentation (MRI) awards for HPC clusters, 1997–2023. The *All* series counts all HPC-related MRI awards; the *GPU* series counts the subset that specifically fund GPU clusters, identified from award-abstract text. The large spike in 2010 is due to increased funding from the American Recovery and Reinvestment Act (ARRA).

3 Stylized Model

We will now present a simple static model that organizes the subsequent empirical analysis. The model highlights two opposing consequences of institutional resources. Well-resourced universities have higher Hicks-neutral productivity, so a given bundle of GPU capital and labor produces more scientific output. At the same time, well-resourced universities have larger internal budgets, allowing them to finance more inputs and purchase GPUs on their own. Higher productivity raises the technical return to compute, while larger baseline input stocks and greater self-financing ability can reduce the return to an external GPU grant. The goal of a central funder such as the NSF is to maximize scientific output compared to the counterfactual defined as the absence of *external* funding.

3.1 Setup

Consider a one-period university indexed by j. Universities differ in internal budgets B_j and Hicks-neutral productivity $\Omega_j \equiv \exp(\omega_j)$. We refer to universities with larger B_j as better resourced and consider the empirically relevant case in which higher-B_j universities also have higher ω_j. The budget B_j determines how much input the university can finance, while ω_j determines output per unit of chosen inputs.

The university produces scientific output using GPU capital G_j and labor L_j:

$$Y_j = \Omega_j G_j^{\alpha_G} L_j^{\alpha_L}, \qquad \alpha_G > 0, \quad \alpha_L > 0, \quad s \equiv \alpha_G + \alpha_L < 1 \tag{3.1}$$

Cobb–Douglas is used for illustrative purposes. GPU capital and labor are complements, and science production exhibits diminishing returns to scale. Let N_j denote centrally provided GPU capacity and let I_j denote GPU capacity financed by the university. Total GPU capital is

$$G_j = N_j + I_j \tag{3.2}$$

One unit of privately financed GPU capital costs p_G, and one unit of labor costs w. The university chooses private GPU investment and labor to maximize scientific output:

$$\max_{I_j, L_j \geq 0} \quad \Omega_j (N_j + I_j)^{\alpha_G} L_j^{\alpha_L} - p_G I_j - w L_j \tag{3.3}$$

subject to the financing constraint

$$p_G I_j + w L_j \leq B_j \tag{3.4}$$

Let $\eta_j \geq 0$ denote the Lagrange multiplier on (3.4). At an interior solution, the first-order conditions are

$$\frac{\partial Y_j}{\partial G_j} = p_G(1 + \eta_j) \tag{3.5}$$

$$\frac{\partial Y_j}{\partial L_j} = w(1 + \eta_j) \tag{3.6}$$

When $\eta_j = 0$, the university can finance its privately optimal bundle of GPU capital and labor. When $\eta_j > 0$, the financing constraint binds, and the marginal scientific return to both inputs exceeds their marginal cost. These conditions also describe the university's input tradeoff. Dividing (3.5) by (3.6) gives

$$\frac{L_j}{G_j} = \frac{\alpha_L p_G}{\alpha_G w} \tag{3.7}$$

Under Cobb–Douglas production, all universities choose the same input ratio at an interior solution. A larger budget allows a university to operate at a larger scale, while higher Hicks-neutral

productivity raises the output and marginal products generated by a given input bundle.

3.2 Resources and the Marginal Product of Compute

Assume the financing constraint binds and set external grants $N_j = 0$. Then the university spends its entire budget, and optimal choice variables satisfy:

$$G_j^c = \frac{\alpha_G}{s}\frac{B_j}{p_G} \tag{3.8}$$

$$L_j^c = \frac{\alpha_L}{s}\frac{B_j}{w} \tag{3.9}$$

Substituting these choices into the marginal product of GPU capital gives

$$\left.\frac{\partial Y_j}{\partial G_j}\right|_{\eta_j>0} = K\Omega_j B_j^{s-1} = p_G(1+\eta_j) \tag{3.10}$$

where $K > 0$ is a constant as a function of parameters $(\alpha_G, \alpha_L, p_G, w)$. Because B_j and Ω_j are primitives of the university's problem, while η_j is the endogenous shadow value of the financing constraint, Equation (3.10) can also be written as

$$1 + \eta_j = \frac{K}{p_G}\Omega_j B_j^{s-1}. \tag{3.11}$$

Thus, holding productivity fixed, a larger budget relaxes the financing constraint:

$$\frac{\partial \log(1+\eta_j)}{\partial \log B_j} = s - 1 < 0. \tag{3.12}$$

Holding the budget fixed, higher productivity raises the shadow value of an additional dollar of internal resources one-for-one in logs. Equation (3.10) therefore captures the central tradeoff. Higher Hicks-neutral productivity Ω_j raises the marginal product of GPU capital. A larger budget B_j allows the university to accumulate more GPU capital and labor, but, since $s < 1$, this larger scale lowers the marginal return to an additional GPU.

To see this directly, consider a high-resource university H and a low-resource university L, with $B_H > B_L$ and $\Omega_H > \Omega_L$. Among financially constrained universities,

$$\frac{MP_H^G}{MP_L^G} = \frac{\Omega_H}{\Omega_L}\left(\frac{B_H}{B_L}\right)^{s-1}, \qquad MP_j^G \equiv \frac{\partial Y_j}{\partial G_j}. \tag{3.13}$$

The high-resource university has the larger technical return to compute if and only if

$$\frac{\Omega_H}{\Omega_L} > \left(\frac{B_H}{B_L}\right)^{1-s}. \tag{3.14}$$

Thus, higher productivity favors allocating compute to the high-resource university, while

diminishing returns from its larger existing input bundle favor the low-resource university. The model does not impose which force dominates. At sufficiently high values of B_j, the financing constraint becomes slack. Then $\eta_j = 0$, and the university privately invests until the marginal products equal input prices. A larger B_j makes this outcome more likely, although a larger Ω_j also raises the university's privately desired scale.

3.3 Externally Funded GPUs

Now consider an external grant of N_j units of GPU capital. Suppose first that the university remains financially constrained and continues to purchase a positive amount of GPU capital privately. The binding constraint can be rewritten as

$$p_G G_j + wL_j = B_j + p_G N_j \tag{3.15}$$

The constrained choices are

$$G_j^c(N_j) = \frac{\alpha_G}{s}\frac{B_j + p_G N_j}{p_G} \tag{3.16}$$

$$L_j^c(N_j) = \frac{\alpha_L}{s}\frac{B_j + p_G N_j}{w} \tag{3.17}$$

$$I_j^c(N_j) = \frac{\alpha_G B_j - \alpha_L p_G N_j}{p_G s} \tag{3.18}$$

For a sufficiently small grant,

$$\frac{\partial I_j^c}{\partial N_j} = -\frac{\alpha_L}{s} \tag{3.19}$$

$$\frac{\partial G_j^c}{\partial N_j} = \frac{\alpha_G}{s} \tag{3.20}$$

$$\frac{\partial L_j^c}{\partial N_j} = \frac{\alpha_L p_G}{sw} \tag{3.21}$$

There are thus three effects of the grant. It directly provides GPU capital, partially displaces private GPU investment, and releases internal resources that the constrained university uses to hire complementary labor. A GPU grant can then affect scientific output partly through the entry of new researchers rather than solely through the productivity of incumbent researchers.

For completeness, the response is different when the financing constraint is slack. Provided the grant is smaller than the university's desired GPU stock, the university reduces private investment one-for-one:

$$\frac{\partial I_j}{\partial N_j} = -1, \qquad \frac{\partial G_j}{\partial N_j} = 0, \qquad \frac{\partial L_j}{\partial N_j} = 0 \tag{3.22}$$

The grant changes who pays for the GPUs but does not change the university's research-input bundle. So an external GPU grant is less additional at a well-resourced university that could have financed the equipment itself. The reduced-form output response to centrally provided GPUs can be decomposed as:

$$\frac{dY_j}{dN_j} = \underbrace{\frac{\partial Y_j}{\partial G_j}}_{\text{technical return to compute}} \times \underbrace{\left(1 + \frac{\partial I_j}{\partial N_j}\right)}_{\text{behavioral response}} + \underbrace{\frac{\partial Y_j}{\partial L_j}\frac{\partial L_j}{\partial N_j}}_{\text{complementary-input response}} \tag{3.23}$$

For a university that remains financially constrained, the decomposition simplifies further. Combining the first-order conditions (3.5)–(3.6) with (3.20)–(3.21) gives

$$\frac{dY_j}{dN_j} = p_G(1 + \eta_j(N_j)) = MP_j^G(N_j), \tag{3.24}$$

where

$$1 + \eta_j(N_j) = \frac{K}{p_G}\Omega_j(B_j + p_G N_j)^{s-1}. \tag{3.25}$$

An external GPU grant therefore has a larger marginal return when it is assigned to a university with a larger financing wedge. At the same time, since $s < 1$, the marginal return falls as the grant itself relaxes that wedge. This means that the identity of the university with the highest marginal return can change with the size of the grant.

Equation (3.23) is the main implication of the stylized model. Well-resourced universities may have a high technical return to compute because their Hicks-neutral productivity Ω_j is high. However, their larger budgets also allow them to begin with more GPU capital and labor and make them more likely to finance GPUs on their own. Less-resourced universities have lower Hicks-neutral productivity, but smaller existing input bundles and a larger financing wedge. The return to an external GPU grant is therefore ambiguous ex ante: it depends on the productivity advantage of the high-resource university, diminishing returns from its larger input stock, and the extent of private-investment crowd-out.

3.4 Illustrative Simulation

We illustrate these tradeoffs with a simple two-university calibration. We set $\alpha_G = 0.25$, $\alpha_L = 0.45$, and normalize $p_G = w = 1$. The low-resource university has $(B_L, \Omega_L) = (1, 3)$, while the high-resource university has $(B_H, \Omega_H) = (4, 4.5)$. Both universities remain financially constrained over the range of

grants considered below, so the simulation focuses on the case with $\eta_j > 0$. The grant N_j is measured in units of GPU capital, so under the normalization $p_G = 1$, a grant of $N_j = 0.1$ is therefore worth 0.1 normalized budget units.

Figure 3 plots the equilibrium marginal product of GPU capital over combinations of (B_j, Ω_j) for which the financing constraint binds. Moving upward in the figure raises Hicks-neutral productivity and therefore raises MP_j^G. Moving to the right raises the university's internal budget, which allows it to operate at a larger scale and lowers MP_j^G through diminishing returns. The contour lines report the implied financing multiplier η_j. Despite having four times the internal budget, the high-resource university's higher productivity offsets most of the decline in the financing wedge, and the two universities are similarly constrained.

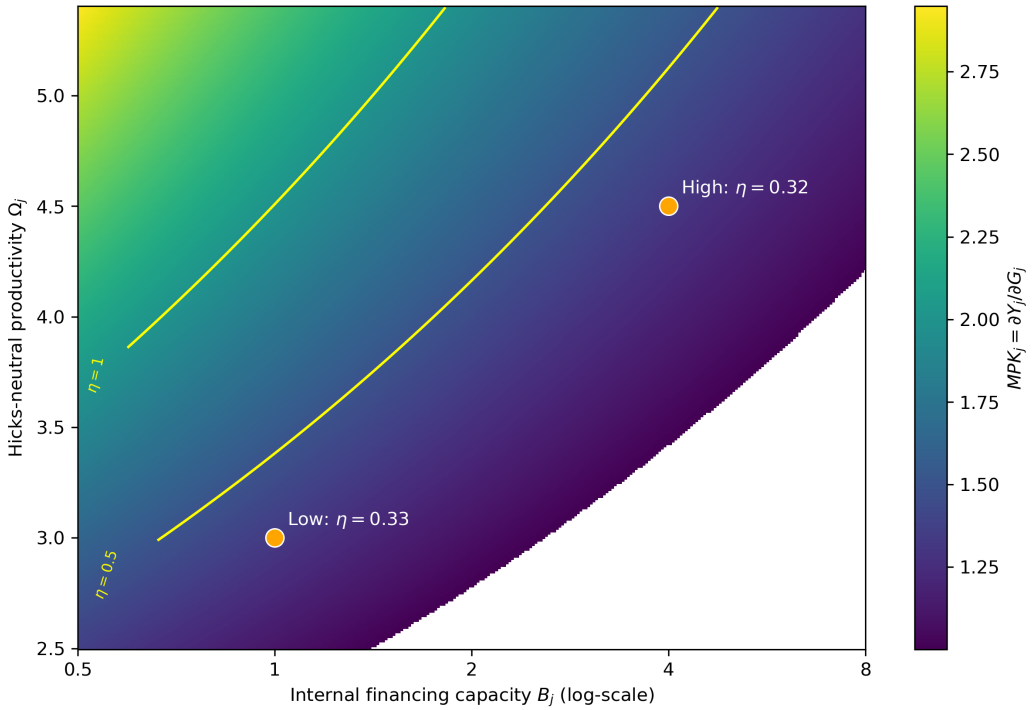


Figure 3: Marginal Product of GPU Capital over Budget and Productivity. The color scale reports MP_j^G for financially constrained universities. Contours report the implied financing multiplier η_j. The points show the low-resource calibration $(B_L, \Omega_L) = (1, 3)$ and the high-resource calibration $(B_H, \Omega_H) = (4, 4.5)$. All quantities are in normalized units.

We can see the comparison clearly in Figure 4, which plots the relationship between financing capacity B_j and the financing wedge η_j for high- and low-resource universities holding productivity Ω_j fixed. At any common value of B_j, the higher-productivity university has a larger shadow value because an additional unit of resources generates more scientific output. At the calibrated points, the low-resource university has $\eta_L = 0.331$, and the high-resource university has $\eta_H = 0.317$. Thus,

the low-resource university is slightly more financially constrained in equilibrium even though the high-resource university is substantially more productive.

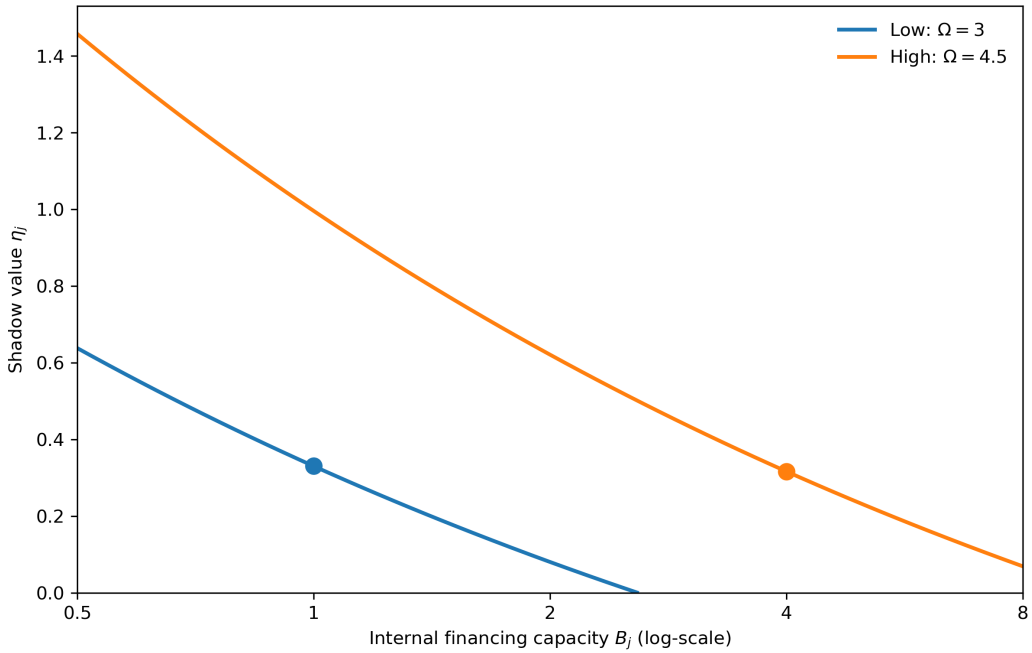


Figure 4: Internal Budget and the Shadow Value of the Financing Constraint. Each curve holds Hicks-neutral productivity fixed and varies the university's internal budget. Larger budgets reduce η_j, while higher productivity shifts the relationship upward. Markers indicate the two calibrated university types.

Figure 5 assigns GPU capital to one university at a time and plots the marginal scientific return dY_j/dN_j over a small grant range $N_j \in [0, 0.1]$. At zero external funding, the low-resource university's smaller budget and tighter financing constraint initially outweigh the high-resource university's productivity advantage. However, the same grant is large relative to the low-resource university's internal budget. As N_j rises, its financing wedge is relieved more quickly, and its marginal return falls more steeply. The two returns cross at $N_j \simeq 0.048$. Beyond this point, the high-resource university's Hicks-neutral productivity advantage dominates. Thus, it is an empirical question where an additional unit of external GPU capital has the highest marginal return since (B_j, Ω_j) are two opposing forces.

4 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.

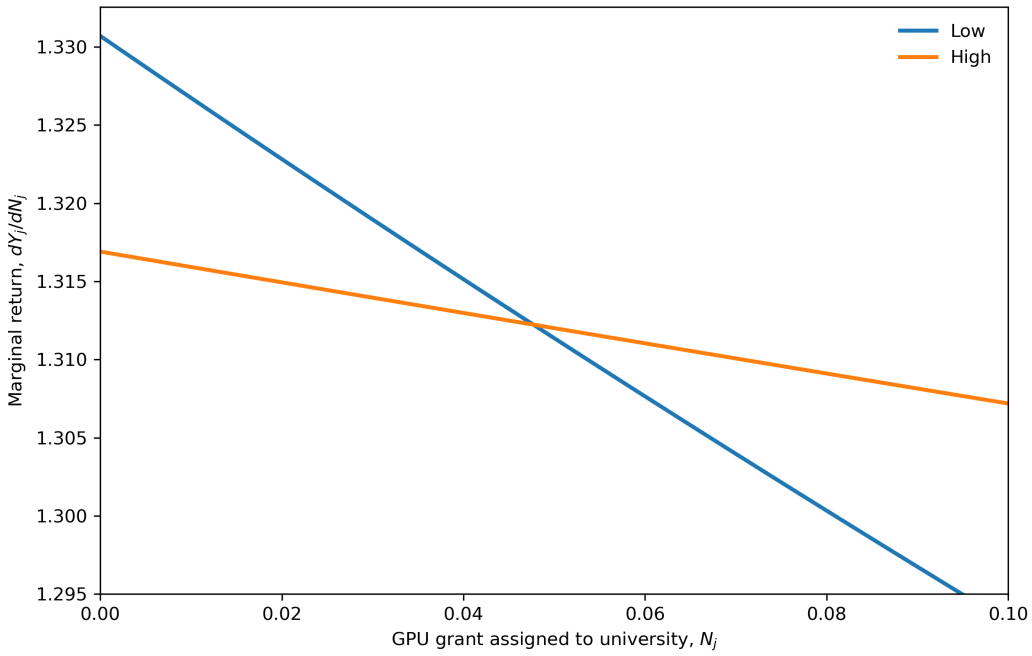


Figure 5: Marginal Return to an External GPU Grant. The figure plots dY_j/dN_j as the same marginal grant is assigned to the low- or high-resource university. The low-resource university has the larger return for very small grants because its financing constraint is tighter. As its constraint is relaxed, its marginal return declines more rapidly, and the high-resource university overtakes at approximately $N_j = 0.048$. N_j is in the same units as budget B_j.

4.1 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 that 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 prelimi-

nary results from a subset of 53 universities that are manually verified. We hope to expand this to the full set of 100 R1s. The subset of R1s is fairly representative across the distribution of endowment and STEM R&D spending. Figures 6 and 7 show that our measure of GPU capital is meaningfully tied to university resources: *within-university* variation in GPU capacity exhibits a high correlation 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. We now describe how we measure GPU capital based on GPU model counts.

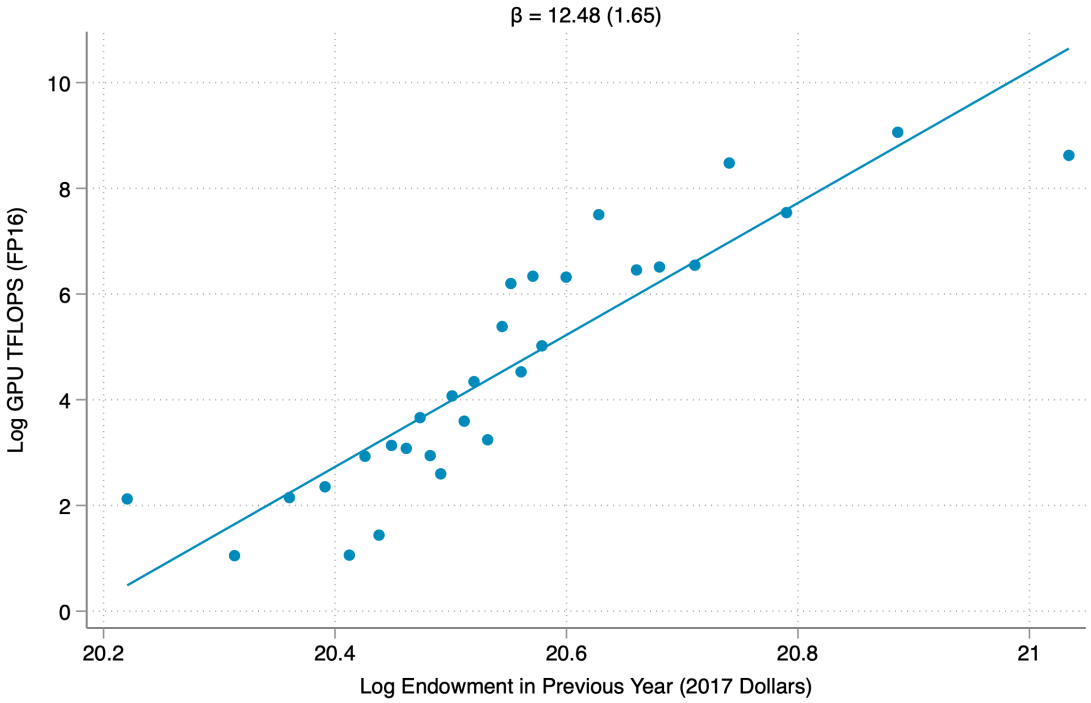


Figure 6: Binscatter of University Log GPU Capacity against Lagged Log Endowment Over Time (Controlling for University Fixed Effects). Binned scatter of log GPU capacity (FP16 TFLOPS) on once-lagged log university endowment (2017 dollars) after partialling out university fixed effects; each marker is a bin of the within-university residuals. The reported slope and standard error are from the underlying fixed-effects regression. The positive within-university relationship indicates that GPU capacity tracks an institution's financial resources rather than idiosyncratic timing.

Harmonizing GPU Capacity. To compare GPU capacity across different hardware vintages, we convert each GPU model into its theoretical peak throughput in floating-point operations per second (FLOPS) and sum across all GPU models to obtain total peak GPU compute at university j in year t. FLOPS measures how many decimal calculations a computer processor can perform in one second. Floating-point numbers are stored as 16, 32, and 64 bits, and while accuracy increases with

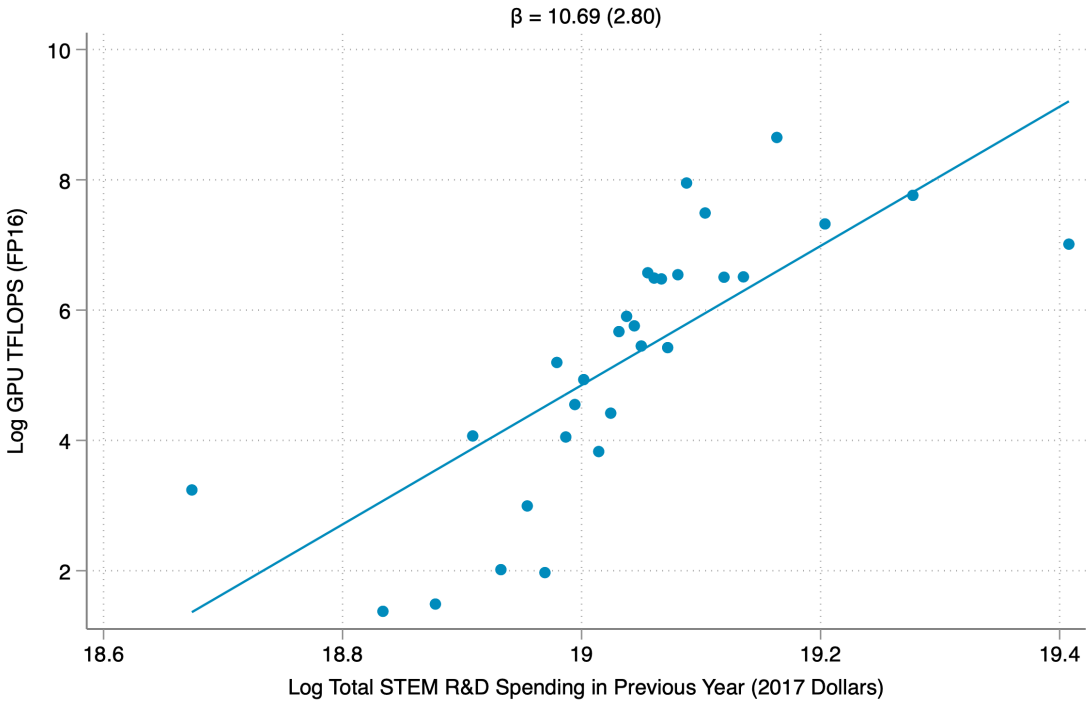


Figure 7: Binscatter of University Log GPU Capacity against Lagged Log STEM R&D Spending Over Time (Controlling for University Fixed Effects). As in Figure 6, with once-lagged log total STEM R&D spending (2017 dollars, NSF HERD) as the covariate. Markers are bins of within-university residuals after absorbing university fixed effects; the slope and standard error are from the corresponding regression. GPU capacity rises with lagged STEM research expenditure within institutions.

the number of bits, memory and speed decrease. For example, scientific simulation requires high accuracy and thus 64-bit arithmetic, while matrix multiplication in neural network training tolerates 16-bit arithmetic. Our headline measure is FP16 (tensor-dense) throughput since that's the rated performance of dedicated matrix multiplication units (i.e. "tensor cores"), which dominate deep-learning workflows. GPUs released prior to 2017 do not have tensor cores and contribute zero, and so we also construct FP32 TFLOPS as a robustness measure. FP32 is the standard general precision rate, which is available for every GPU model in our dataset. We use specifications from the Epoch AI Machine Learning Hardware (Epoch AI 2026) and TechPowerUp GPU databases (TechPowerUp 2026), cross-checked against manufacturer datasheets (NVIDIA Corporation 2017, 2021, Advanced Micro Devices 2021) when unsure.

4.2 Scientific Output Measures

Our primary outcome variable is the quantity and quality of GPU-enabled publications from OpenAlex. For all U.S. 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 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 is 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:

$$y_{p\to j} = \text{CitePct}_p \cdot \sum_{i\in A(p)} \left(\frac{1}{|A(p)|} \cdot \frac{\mathbf{1}\{j \in \mathcal{J}(i)\}}{|\mathcal{J}(i)|} \right). \tag{4.1}$$

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

$$Y_{jt} = \sum_{p\in\mathcal{P}_{jt}} y_{p\to j}. \tag{4.2}$$

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.

4.3 Labor Measure

For the structural model, we require a count of both quasi-fixed labor (i.e. PIs) and flexible labor (i.e. PhD students and postdocs) by field. We construct flexible labor using external survey data from the NSF Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). However, GSS only reports doctoral enrollment from 2017 onward, so we extend the series backward using the ProQuest Dissertations and Theses panel of Shvadron et al. (2025). They report PhD graduates by university, field, and year. We construct quasi-fixed labor using OpenAlex and authorship position, along with career length from first publication date. 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.

4.4 University Characteristics and Grants

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. Furthermore, we download all NSF MRI and CC* awards that are publicly available and classify computing grants based on the abstract text.

5 Event Study

We present preliminary quasi-experimental evidence that university-provisioned GPU capital affects research output from the subset of 53 manually verified universities. We exploit within-year large capacity upgrades in a staggered event-study framework. First, we evaluate whether total versus GPU-relevant research output changes relative to peers by tiers of baseline university financial resources, and decompose whether the change is driven by incumbent versus new researchers. We then isolate our analysis to computer science, where the gains are expected to be greatest and outcomes suffer the least from publication lag. Finally, we also investigate whether isolating to only NSF-funded upgrades generates changes in research output.

5.1 Empirical Strategy

To estimate the returns to computing capital, we exploit variation in the timing of HPC cluster upgrades. We define an "event" as a year in which a university's GPU capacity increases by more than five FP16 TFLOPS per research personnel. The treated sample consists of 30 universities with event years spanning 2019 to 2025. We use the NSF GSS to define a time-invariant measure of research personnel as the annual mean over 2015 to 2019 of full-time graduate students, postdocs, and non-faculty researchers in STEM. We normalize by the headcount to evaluate comparable changes in computing intensity at large and small universities as opposed to a fixed number.

The control group then consists of both universities that experienced only small or incremental capacity changes and universities that are not yet treated. We construct treatment and control sets within three tiers of financial capacity: low, medium, and high baseline endowment (from 2022).

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 not random since better-resourced and more productive 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, often up to 18 to 24 months.

The estimation specification is:

$$Y_{jt} = \gamma_j + \lambda_t + \sum_{k=-4}^{4} \beta_k \mathbf{1}\{K_{jt} = k\} + \epsilon_{jt} \tag{5.1}$$

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. We present equal-author-weighted raw paper counts, and will extend future versions to counts of high-quality papers (defined as above the 95th citation percentile within subfield × publication year).

A natural concern is that GPU upgrades coincide with other investments in complementary inputs such as labor. For each figure, we present an incumbent-only event study. We define an incumbent as a quasi-fixed researcher who joined the main institution at least one year prior to the upgrade event. Since there is attrition of short-term labor such as students, we focus the analysis on quasi-fixed labor.

5.2 Aggregate and Tier-Level Results

Figure 8 shows limited average effects of GPU cluster upgrades on aggregate AI-related publication outcomes, with no differences restricting to incumbent only (except for levels). Figure 9 shows a slight increase in total AI-related publications at high-tier universities, which fades out by period $t + 2$, with a corresponding increase in incumbent publications. Figures 10 and 11 show no AI publication increase for medium- and low-tier universities. There is a slight increase for incumbent PIs at low-tier universities. Pre-trends are flat for all three tiers, supporting the identifying assumption.

5.3 Returns to HPC for Computer Science

As a more targeted test, we restrict the sample to computer science publications. Figure 14 shows a steady and significant 50% increase over time in computer science publications for low-tier universities. Half the effect is driven by incumbent PIs. Medium-tier universities also experience a 50% increase over time, though it is not statistically distinguishable from zero in Figure 13. Unlike low-tier universities, the increase is driven by entrants. Figure 12 shows a significant 30% increase in CS publications in the initial period, which immediately becomes insignificant. Similarly, slightly less than half of the initial increase is due to incumbent PIs.

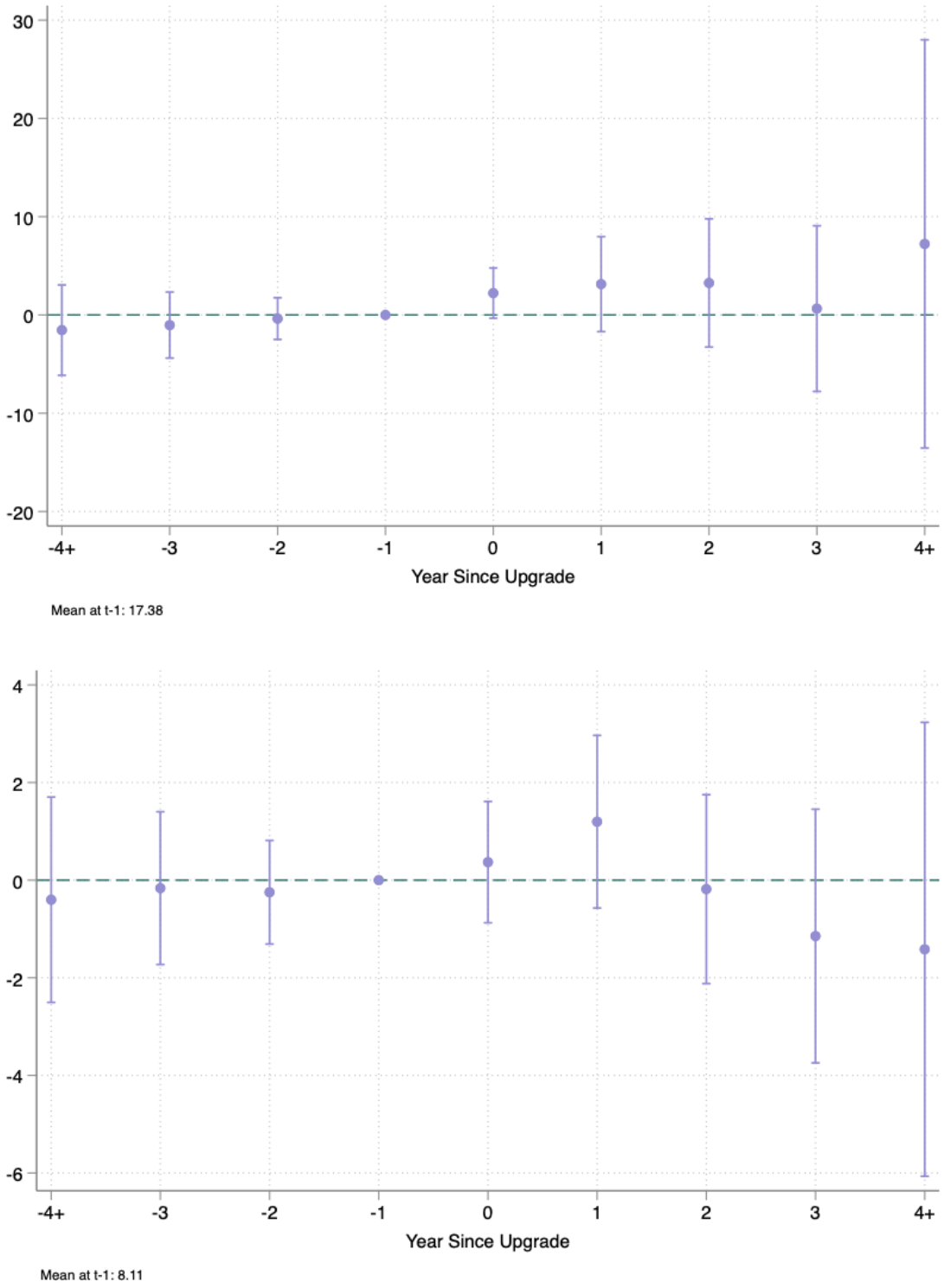


Figure 8: Event Study of AI Publications for All and Incumbents. Interaction-weighted event-study estimates (Sun and Abraham 2021) of equation (5.1) for the number of AI-related publications, estimated on the 53 manually verified universities (30 treated, with controls drawn from incremental-change and not-yet-treated institutions). An event is a within-year GPU-capacity increase exceeding five FP16 TFLOPS per research-personnel headcount. Coefficients are relative to the year before the upgrade ($t-1$, normalized to zero); the -4 and $+4$ endpoints are accumulated bins. The top panel uses all researchers; the bottom restricts to incumbents (quasi-fixed researchers affiliated at least one year before the event). Points are equal-author-weighted paper counts, and vertical bars are 95% confidence intervals from standard errors clustered by university. *Mean at* $t-1$ is the outcome level in the reference year, against which coefficients can be read as a share.

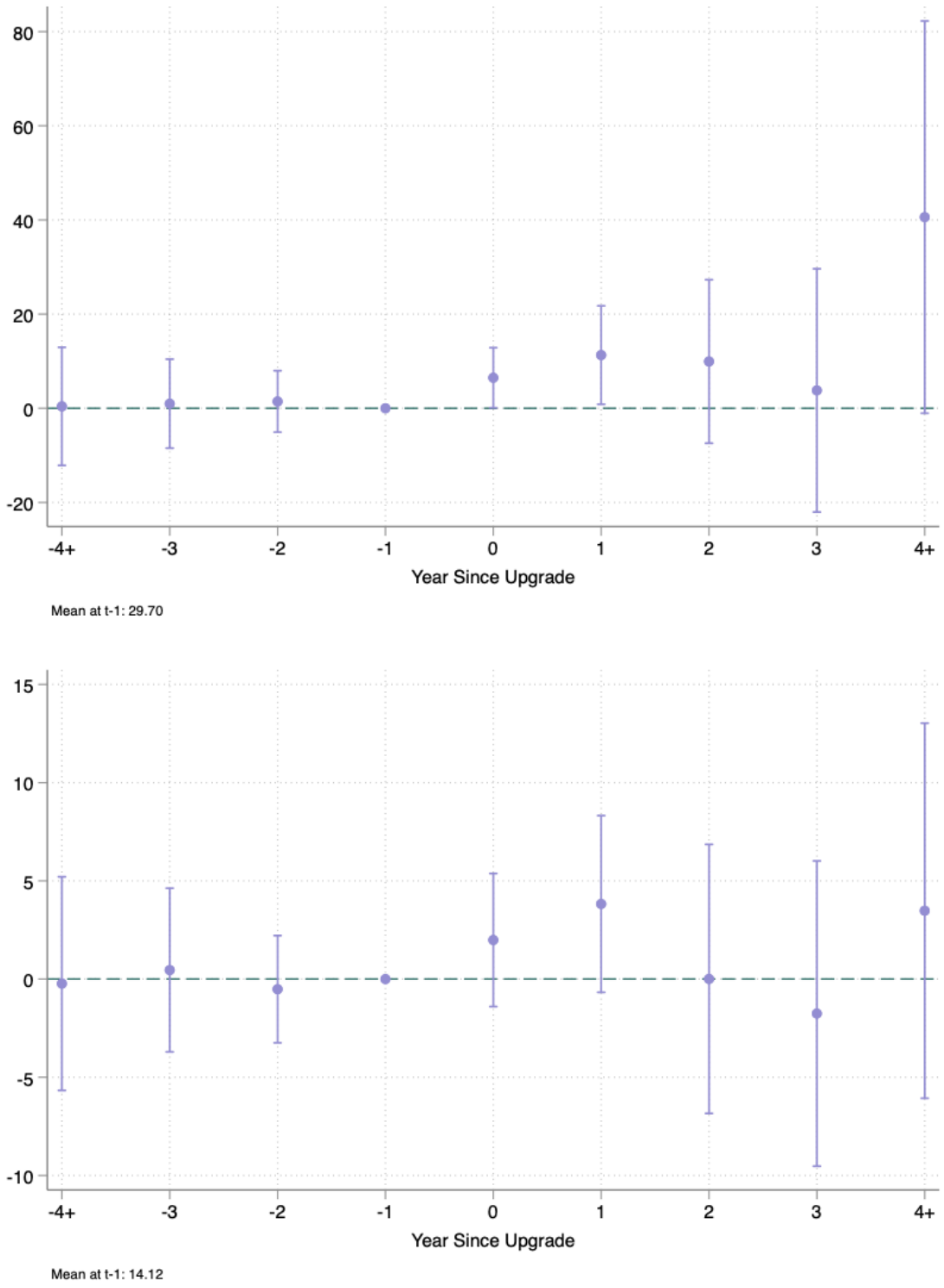


Figure 9: Event Study of High-Tier AI Publications for All and Incumbents. As in Figure 8, restricted to high-tier universities (top third of 2022 baseline endowment). Top panel: all researchers; bottom panel: incumbents. Estimator, sample construction, and confidence intervals follow Figure 8.

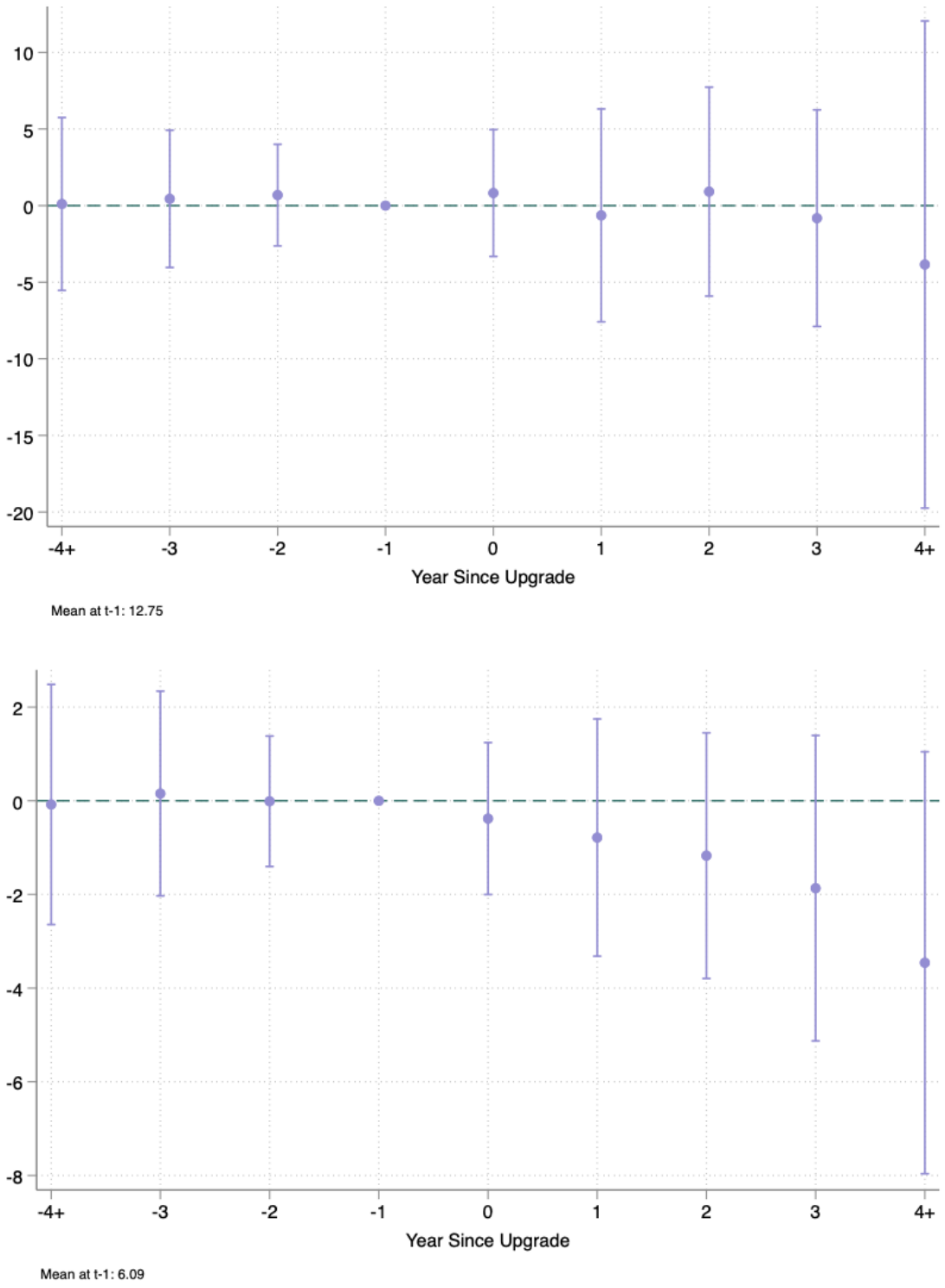


Figure 10: Event Study of Medium-Tier AI Publications for All and Incumbents. As in Figure 8, restricted to medium-tier universities (middle third of 2022 baseline endowment). See Figure 8 for the estimator, sample, and inference.

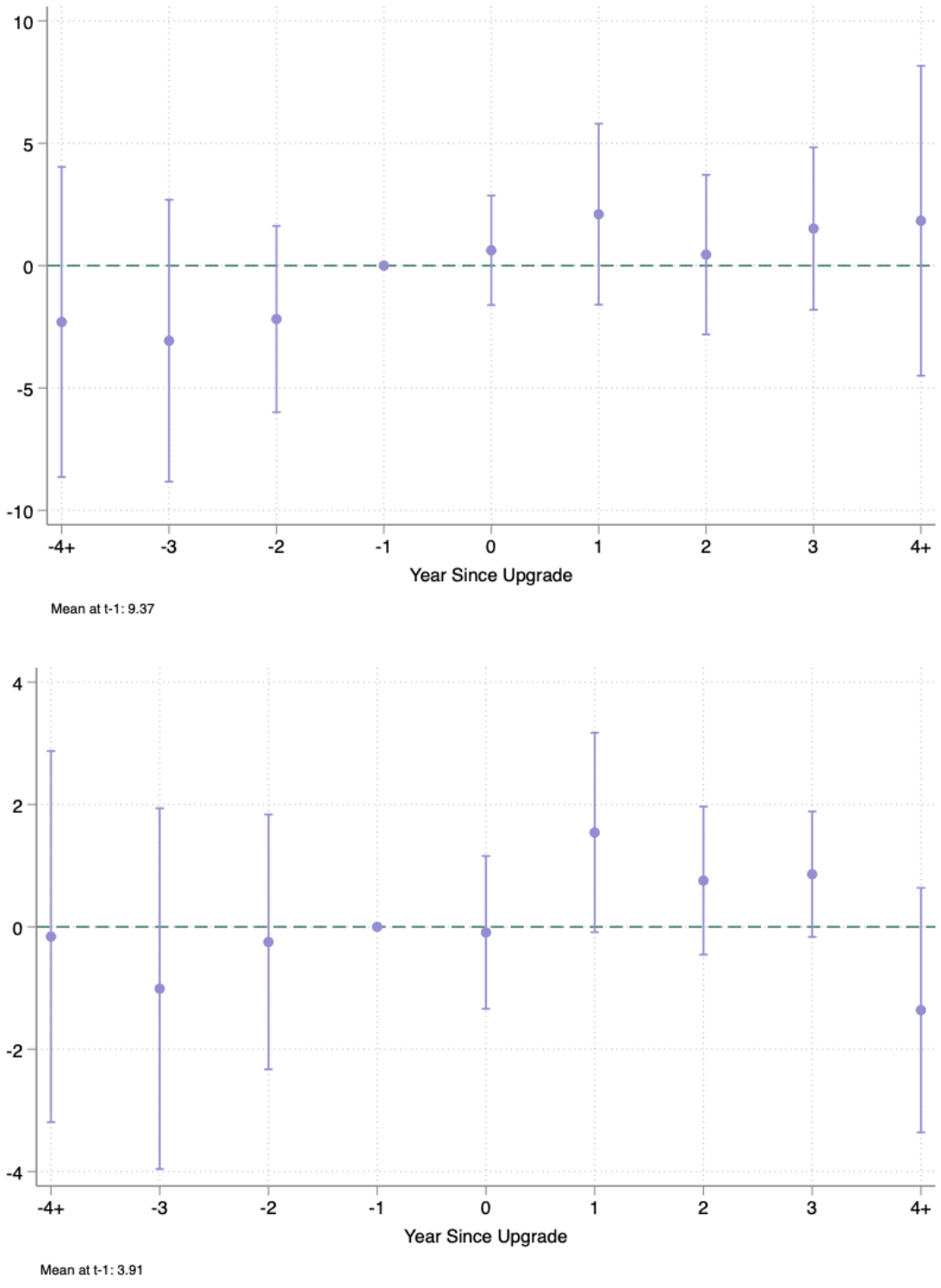


Figure 11: Event Study of Low-Tier AI Publications for All and Incumbents. As in Figure 8, restricted to low-tier universities (bottom third of 2022 baseline endowment). See Figure 8 for specification and inference.

The pattern is consistent with diminishing marginal returns to computing capital as well as dynamic crowd-in. 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 and encourage dynamic investment that persists over time.

Furthermore, half of publication increases are driven by incumbents across university tiers, suggestive evidence that GPU infrastructure serves 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).

5.4 NSF-Funded Capacity Increases

Since most NSF grants are distributed to lower-tier universities, we now restrict the event-study sample to NSF-funded upgrade events for low-tier universities. We remove all universities in the control group who received NSF infrastructure funding at any point in time. Figure 15 shows a significant increase for NSF-funded events in period $t+1$ for low-tier universities with flat pre-trends, indicating that there is a resource constraint that was relieved by external funding.

6 Model of University Production and Investment

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 that 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 it 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

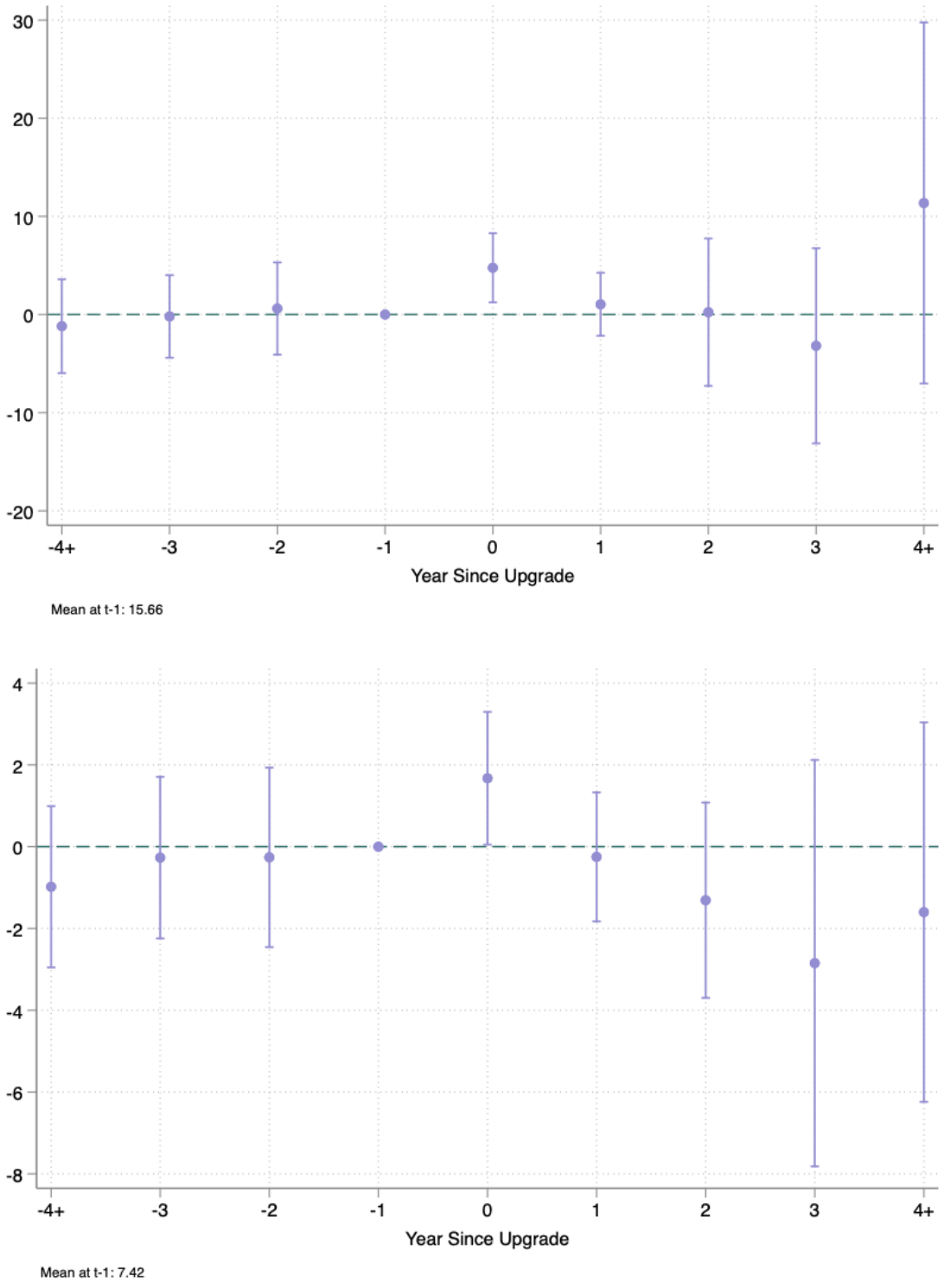


Figure 12: Event Study of High-Tier Computer Science Publications for All and Incumbents. As in Figure 8, but the outcome is computer science publications, and the sample is restricted to high-tier universities (top third of 2022 baseline endowment). Top panel: all researchers; bottom panel: incumbents. Estimator and inference follow Figure 8.

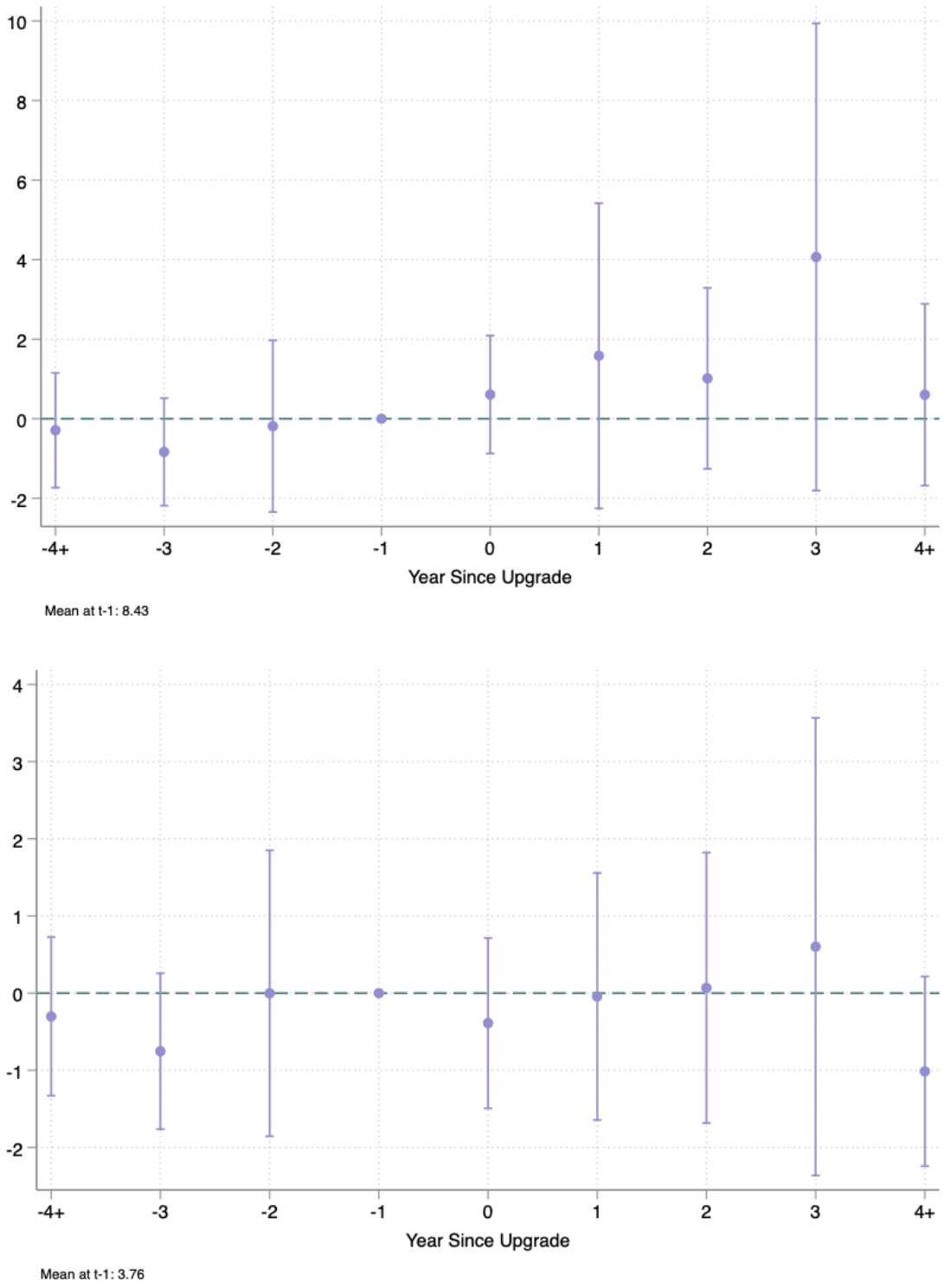


Figure 13: Event Study of Medium-Tier Computer Science Publications for All and Incumbents. As in Figure 12, restricted to medium-tier universities. The outcome is computer science publications; estimator and inference follow Figure 8.

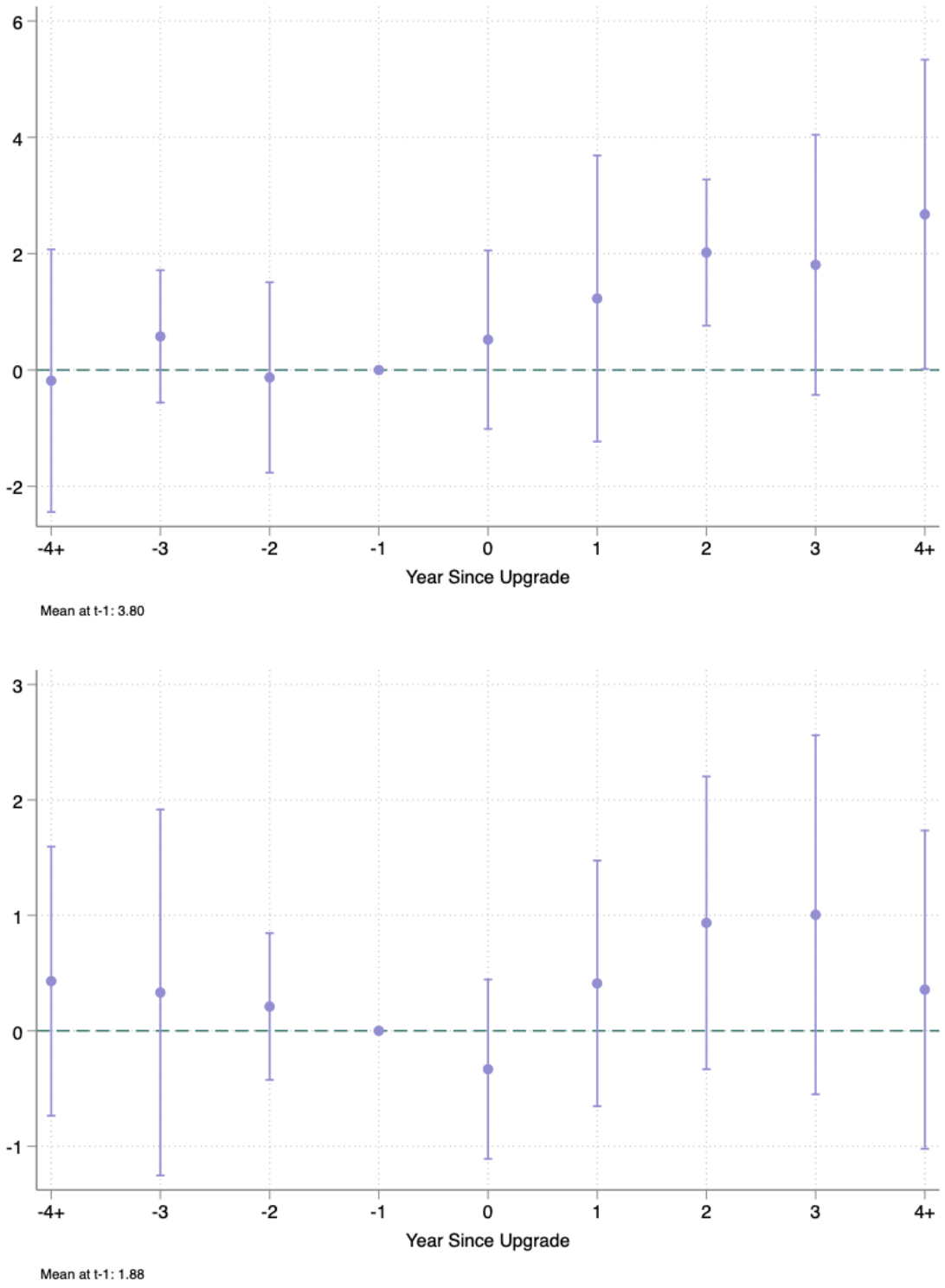


Figure 14: Event Study of Low-Tier Computer Science Publications for All and Incumbents. As in Figure 12, restricted to low-tier universities. The outcome is computer science publications; estimator and inference follow Figure 8.

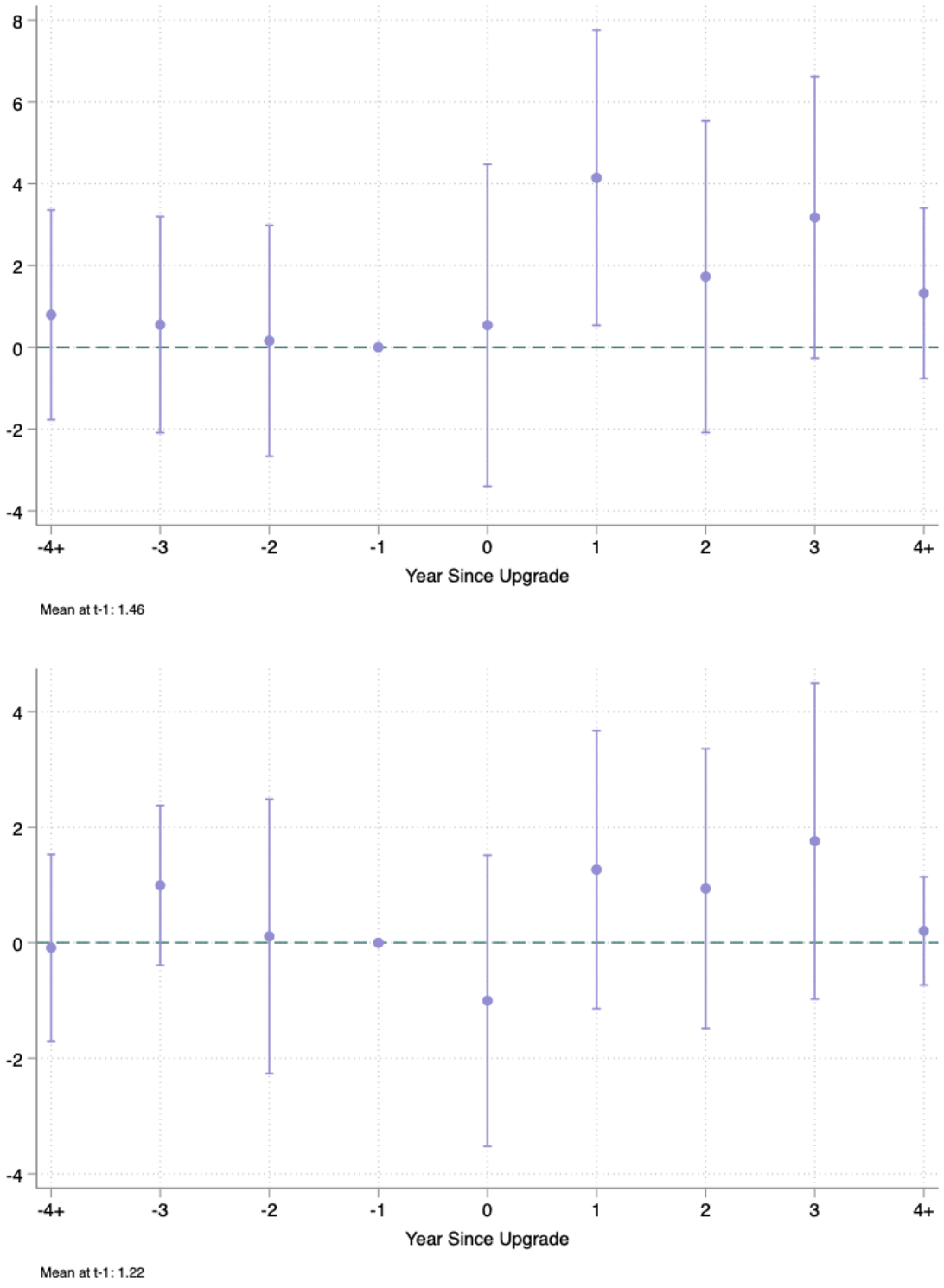


Figure 15: NSF-Funded Event Study of Low-Tier Computer Science Publications for All and Incumbents. As in Figure 14, but treatment is restricted to NSF-funded upgrade events at low-tier universities, and control institutions that ever received NSF infrastructure funding are dropped. The outcome is computer science publications; estimator and inference follow Figure 8.

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 static first-order conditions. Since we do not impose a normative stance on the university's objective function, we estimate a reduced-form investment policy function instead of the full dynamic Bellman.

6.1 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 & geosciences, and engineering. We exclude mathematics, social sciences, psychology, and humanities 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:

$$Y_{jft} = \Omega_{jft}\, \mathcal{K}_{jft}^{\alpha_g^K}\, \mathcal{L}_{jft}^{\alpha_g^L}, \qquad \nu_g \equiv \alpha_g^K + \alpha_g^L \tag{6.1}$$

$$\mathcal{K}_{jft} = \left[\gamma_g\, C_{jft}^{\rho_K} + (1-\gamma_g)\left(G_{jft}\right)^{\rho_K}\right]^{1/\rho_K}, \qquad \rho_K \equiv \frac{\sigma_K - 1}{\sigma_K} \tag{6.2}$$

$$\mathcal{L}_{jft} = \left[\phi_g\, (L^V_{jft})^{\rho_L} + (1-\phi_g)\, (L^F_{jft})^{\rho_L}\right]^{1/\rho_L}, \qquad \rho_L \equiv \frac{\sigma_L - 1}{\sigma_L} \tag{6.3}$$

$$\omega_{jft} = \varrho_g \omega_{jft-1} + \xi_{jft}, \qquad \xi_{jft} \text{ i.i.d} \tag{6.4}$$

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, and other fields. Within-nest capital and labor substitution elasticities (σ_K, σ_L) are not group-specific, imply-

ing that the rate of substitution between CPU and GPU capital is 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).

6.2 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 postdocs) 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:

$$V(C_{jt}, G_{jt}, \mathbf{L}^F_{jt}, \boldsymbol{\omega}_{jt}) = \max_{\substack{I^C_{jt}, I^G_{jt}, \{H_{jft}\}_f, \\ \{L^V_{jft}\}_f, \{C_{jft}\}_f, \{G_{jft}\}_f}} \left\{ \sum_f \lambda_{jf}\, Y_{jft} + \beta\, \mathbb{E}\big[V\big(C_{jt+1}, G_{jt+1}, \mathbf{L}^F_{jt+1}, \boldsymbol{\omega}_{jt+1}\big) \,\big|\, \boldsymbol{\omega}_{jt}\big] \right\} \tag{6.5}$$

subject to the laws of motion for the state variables:

$$C_{jt+1} = (1 - \delta_C)\, C_{jt} + I^C_{jt} \tag{6.6}$$

$$G_{jt+1} = (1 - \delta_G)\, G_{jt} + I^G_{jt} \tag{6.7}$$

$$L^F_{jft+1} = (1 - \delta_F)\, L^F_{jft} + H_{jft} \tag{6.8}$$

$$\omega_{jft+1} = \varrho_g \omega_{jft} + \xi_{jft+1} \tag{6.9}$$

where δ_C, δ_G, and δ_F are depreciation rates for CPU capital, GPU capital, and faculty, respectively. Productivity evolves according to (6.9), 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:

$$\sum_f C_{jft} \leq C_{jt} \tag{6.10}$$

$$\sum_f G_{jft} \leq G_{jt} \tag{6.11}$$

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

$$\sum_f W^V_{jft}\, L^V_{jft} + \sum_f W^F_{jft}\, L^F_{jft} + P^C_t\, I^C_{jt} + P^G_t\, I^G_{jt} \leq B_{jt} \tag{6.12}$$

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 does not carry a separate cost beyond the cost of faculty salaries.

6.3 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 (6.10), the GPU constraint (6.11), and the budget constraint (6.12), respectively. The period-t Lagrangian for university j is:

$$\begin{aligned}\mathcal{L} = &\sum_f \lambda_{jf}\, Y_{jft} + \beta\, \mathbb{E}[V(\cdot) \mid \boldsymbol{\omega}_{jt}] \\ &- \mu^C_{jt}\left(\sum_f C_{jft} - C_{jt}\right) - \mu^G_{jt}\left(\sum_f G_{jft} - G_{jt}\right) \\ &- \eta_{jt}\left(\sum_f W^V_{jft}\, L^V_{jft} + \sum_f W^F_{jft}\, L^F_{jft} + P^C_t\, I^C_{jt} + P^G_t\, I^G_{jt} - B_{jt}\right) \end{aligned} \tag{6.13}$$

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:

$$\mu^C_{jt}\, C_{jft} = \alpha^K_g\; s^{\mathcal{K}}_C(jft) \cdot \lambda_{jf} Y_{jft}, \tag{6.14}$$

$$\mu^G_{jt}\, G_{jft} = \alpha^K_g\; s^{\mathcal{K}}_G(jft) \cdot \lambda_{jf} Y_{jft}, \tag{6.15}$$

$$\eta_{jt}\, W^V_{jft}\, L^V_{jft} = \alpha^L_g\; s^{\mathcal{L}}_V(jft) \cdot \lambda_{jf} Y_{jft} \tag{6.16}$$

with within-nest value shares

$$s^{\mathcal{K}}_C(jft) = \gamma_g \Big(\frac{C_{jft}}{\mathcal{K}_{jft}}\Big)^{\rho_K}, \qquad s^{\mathcal{K}}_G(jft) = (1-\gamma_g)\Big(\frac{G_{jft}}{\mathcal{K}_{jft}}\Big)^{\rho_K} = 1 - s^{\mathcal{K}}_C(jft), \tag{6.17}$$

$$s^{\mathcal{L}}_V(jft) = \phi_g \Big(\frac{L^V_{jft}}{\mathcal{L}_{jft}}\Big)^{\rho_L}, \qquad s^{\mathcal{L}}_F(jft) = 1 - s^{\mathcal{L}}_V(jft) \tag{6.18}$$

From (6.16), $\lambda_{jf} Y_{jft} = \eta_{jt} W^V_{jft} L^V_{jft} / [\alpha^L_g\; s^{\mathcal{L}}_V(jft)]$, which we substitute into (6.14) and (6.15) 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:

$$C_{jft} = C_{jt} \cdot \frac{\psi^C_{jft}\, W^V_{jft}\, L^V_{jft}}{\sum_h \psi^C_{jht}\, W^V_{jht}\, L^V_{jht}}, \qquad \psi^C_{jft} \equiv \frac{\alpha^K_g}{\alpha^L_g} \cdot \frac{s^{\mathcal{K}}_C(jft)}{s^{\mathcal{L}}_V(jft)} \tag{6.19}$$

$$G_{jft} = G_{jt} \cdot \frac{\psi^G_{jft}\, W^V_{jft}\, L^V_{jft}}{\sum_h \psi^G_{jht}\, W^V_{jht}\, L^V_{jht}}, \qquad \psi^G_{jft} \equiv \frac{\alpha^K_g}{\alpha^L_g} \cdot \frac{s^{\mathcal{K}}_G(jft)}{s^{\mathcal{L}}_V(jft)} \tag{6.20}$$

Equations (6.19)–(6.20) 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 (6.19)–(6.20) 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-labor wages are constant across fields within a university-year:

$$W^V_{jft} = W^V_{jt} \quad \forall f \tag{6.21}$$

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

$$\boxed{C_{jft} = C_{jt} \cdot \frac{\psi^C_{jft} L^V_{jft}}{\sum_h \psi^C_{jht} L^V_{jht}}} \tag{6.22}$$

$$\boxed{G_{jft} = G_{jt} \cdot \frac{\psi^G_{jft} L^V_{jft}}{\sum_h \psi^G_{jht} L^V_{jht}}} \tag{6.23}$$

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}. In Appendix A1, we use self-reported data from `phdstipends.com` and verify that the majority of variance in PhD stipends is across university $\times$ year compared to across fields in physical sciences, geosciences, life sciences, engineering, and computer science. Social sciences, psychology, and math stipends are on average 15% lower.

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.2 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.

7 Estimation

We first estimate the nested CES production function parameters, then proceed to estimate a reduced-form investment policy function. Production function estimation is complicated by the fact that we do not have a closed-form allocation rule. There are three sets of production function parameters, which we call labor nest $\boldsymbol{\theta_L}$, capital nest $\boldsymbol{\theta_K}$, and outer Cobb–Douglas (CD) $\boldsymbol{\theta_{CD}}$:

$$\boldsymbol{\theta_L} = (\sigma_L, \{\phi_g\}_{g\in\mathcal{G}}) \tag{7.1}$$

$$\boldsymbol{\theta_K} = (\sigma_K, \{\gamma_g\}_{g\in\mathcal{G}}) \tag{7.2}$$

$$\boldsymbol{\theta_{CD}} = \{\alpha_g^K, \alpha_g^L, \varrho_g\}_{g\in\mathcal{G}} \tag{7.3}$$

We proceed in three stages. First, we estimate the labor nest parameters $\boldsymbol{\theta_L}$ from the flexible-labor first-order condition using nonlinear least squares. We also test the null hypothesis that $\sigma_L = 1$, under which the labor nest is Cobb–Douglas. Second, conditional on labor nest parameters, we take a guess of capital nest parameters $\boldsymbol{\theta_K}$ and estimate outer CD parameters $\boldsymbol{\theta_{CD}}$ using an iterated linear IV (until field-level capital allocations converge). Finally, we loop over guesses of $\boldsymbol{\theta_K}$ to minimize the same group-blocked GMM objective, evaluated at the concentrated $\hat{\boldsymbol{\theta}}_{CD}(\boldsymbol{\theta_K})$ with the weight matrix held fixed across guesses; because the linear stage uses the $\boldsymbol{\theta_{CD}}$ moment directions, $\boldsymbol{\theta_K}$ is identified off the overidentifying restrictions alone.

7.1 Production Function Estimation

Taking logs of the aggregate production function we have

$$y_{jft} = \alpha_g^K \log \mathcal{K}_{jft}(\boldsymbol{\theta}) + \alpha_g^L \log \mathcal{L}_{jft} + \omega_{jft} + \varepsilon_{jft} \tag{7.4}$$

where ε_{jft} is an iid measurement error, labor nest $\mathcal{L}_{jft}$ is computed from observed (L_{jft}^V, L_{jft}^F), 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 quasi-differencing:

$$y_{jft} - \varrho_g\, y_{jf,t-1} = \alpha_g^K\, [\log \mathcal{K}_{jft}(\boldsymbol{\theta}) - \varrho_g \log \mathcal{K}_{jf,t-1}(\boldsymbol{\theta})] + \alpha_g^L\, [\log \mathcal{L}_{jft} - \varrho_g \log \mathcal{L}_{jf,t-1}] + u_{jft} \tag{7.5}$$

$$u_{jft} \equiv\ \xi_{jft} + \varepsilon_{jft} - \varrho_g\, \varepsilon_{jf,t-1}. \tag{7.6}$$

Even when ε_{jft} is iid, u_{jft} is generally serially correlated because it contains lagged measurement error. We therefore form moments using predetermined instruments and cluster inference at the university level.

Stage 1: Labor First-Order Condition. Substitution elasticity parameters are difficult to identify off of productivity residuals alone, so we incorporate information from the first-order conditions. Ideally, we would have data on relative wages and directly estimate the relationship between $\log\left(\frac{L^V_{jft}}{L^F_{jft}}\right)$ and $\log\left(\frac{W^V_{jft}}{W^F_{jft}}\right)$. Absent wage data, we take logs of the first-order condition with respect to flexible labor L^V_{jft} in (6.16) and rearrange:

$$\log\left(\frac{Y_{jft}}{L^V_{jft}}\right) = \log(\eta_{jt}) + \log(W^V_{jft}) - \log(\alpha^L_g) - \log(s^{\mathcal{L}}_V(jft)) - \log(\lambda_{jf}), \tag{7.7}$$

where

$$\log(s^{\mathcal{L}}_V(jft)) = \log(\phi_g) + \rho_L \log(L^V_{jft}) - \log\left(\phi_g \left(L^V_{jft}\right)^{\rho_L} + (1-\phi_g)\left(L^F_{jft}\right)^{\rho_L}\right). \tag{7.8}$$

Under the common wage assumption, W^V_{jft} simplifies to W^V_{jt}. Because group g is nested within field f, the terms involving α^L_g, ϕ_g, and λ_{jf} that enter additively are absorbed by the $j \times f$ fixed effect. Recall that $\rho_L \equiv (\sigma_L - 1)/\sigma_L$, and the labor first-order condition therefore simplifies to

$$\boxed{\log\left(\frac{Y_{jft}}{L^V_{jft}}\right) = \log\left[\phi_g + (1-\phi_g)\left(\frac{L^F_{jft}}{L^V_{jft}}\right)^{\rho_L}\right] + a_{jf} + b_{jt} + e^L_{jft}} \tag{7.9}$$

where

$$a_{jf} = \log(\phi_g) - \log(\alpha^L_g) - \log(\lambda_{jf}) \tag{7.10}$$

$$b_{jt} = \log(\eta_{jt}) + \log(W^V_{jt}) \tag{7.11}$$

e^L_{jft} captures classical measurement error in Y_{jft}. We estimate $\boldsymbol{\theta}_L \equiv (\sigma_L, \{\phi_g\}_{g\in\mathcal{G}})$ jointly using nonlinear least squares. For each candidate $\boldsymbol{\theta}_L$, we absorb the $j \times f$ and $j \times t$ fixed effects from the residual. Concretely, write $y^L_{jft} = \log(Y_{jft}/L^V_{jft})$ and $l_{jft} = \log\left[\phi_g + (1-\phi_g)\left(\frac{L^F_{jft}}{L^V_{jft}}\right)^{\rho_L}\right]$. The estimator solves the full problem over the parameter of interest *and* every fixed effect,

$$\min_{\boldsymbol{\theta}_L,\, \{a_{jf}\},\, \{b_{jt}\}} \sum_{j,f,t} \Big[\, y^L_{jft} - a_{jf} - b_{jt} - m(l_{jft}; \boldsymbol{\theta}_L) \Big]^2, \tag{7.12}$$

but the FEs never enter the numerical search. Fix a candidate $\boldsymbol{\theta}_L$ and define the partial residual $r_{jft}(\boldsymbol{\theta}_L) \equiv y^L_{jft} - m(l_{jft}; \boldsymbol{\theta}_L)$. Given $\boldsymbol{\theta}_L$, (7.12) is ordinary least squares of r on the FE dummies alone, and its first-order conditions say each fixed effect equals a cell mean of r net of the other effect:

$$\hat{a}_{jf} = \frac{1}{T_{jf}} \sum_t \left(r_{jft} - \hat{b}_{jt}\right), \qquad \hat{b}_{jt} = \frac{1}{F_{jt}} \sum_f \left(r_{jft} - \hat{a}_{jf}\right). \tag{7.13}$$

We iterate on (7.13): demean r within each (j,f) cell, then within each (j,t) cell, and repeat until nothing changes. So the fixed point is the residual from projecting r on the span of both dummy families. Substituting $\hat{a}, \hat{b}$ back into (7.12) gives the criterion now as a function of $\boldsymbol{\theta}_L$ alone:

$$Q_N(\boldsymbol{\theta}_L) \;=\; \frac{1}{N} \sum_{j,f,t} \left(\tilde{y}^L_{jft} - \tilde{m}_{jft}(\boldsymbol{\theta}_L)\right)^2 \tag{7.14}$$

Stage 2: Iterated Linear IV. Conditional on $\hat{\boldsymbol{\theta}}_L$, collect the remaining parameters as $\boldsymbol{\theta}_P \equiv (\boldsymbol{\theta}_K, \boldsymbol{\theta}_{CD})$. $\boldsymbol{\theta}_K$ enters nonlinearly, through the capital aggregator (6.2) and the allocations, while conditional on the allocations, the residual is bilinear in $\boldsymbol{\theta}_{CD}$. Writing $v_{jft} \equiv y_{jft} - \alpha_g^K \log \mathcal{K}_{jft} - \alpha_g^L \log \mathcal{L}_{jft}$ for the pre-quasi-difference residual, $u_{jft} = v_{jft} - \varrho_g \, v_{jf,t-1}$ is linear in (α_g^K, α_g^L) given ϱ_g and linear in ϱ_g given (α_g^K, α_g^L). We thus exploit this linearity: holding a guess of $\boldsymbol{\theta}_K$ fixed, we estimate $\boldsymbol{\theta}_{CD}$ by alternating a closed-form linear IV–GMM solve with a recomputation of the implied capital allocations. Instead of estimating $\boldsymbol{\theta}_P$ jointly in a GMM that must search over a high-dimensional parameter space, we estimate $\boldsymbol{\theta}_{CD}$) as a linear IV and shrink the search space to only $\boldsymbol{\theta}_K$ in the next stage.

The capital allocations have no closed form except at $\sigma_K = 1$ (Cobb–Douglas), so even fixing all parameters the capital allocations require a nested fixed-point to solve. This inner loop is therefore part of *constructing* the regressor $\log \mathcal{K}_{jft}(\boldsymbol{\theta})$ at fixed parameters; it is not additional estimation. The Cobb–Douglas allocation in step 1 below is exactly the $\sigma_K = 1$ closed form and serves as the initializer. Throughout this stage, fix a guess of $\boldsymbol{\theta}_K$. Then for some value of $\boldsymbol{\theta}_{CD}$, we calculate field-level capital $\{C_{jft}, G_{jft}\}_{f \in \mathcal{F}}$ as follows:

1. Initialize $\hat{C}_{jft}^{(0)}, \hat{G}_{jft}^{(0)}$ at the Cobb–Douglas allocation under the common wage assumption (6.21):
$$\hat{C}_{jft}^{(0)} = \frac{C_{jt} \cdot \left(\alpha_g^K \gamma_g / (\alpha_g^L \hat{\phi}_g)\right) L_{jft}^V}{\sum_h \left(\alpha_{g(h)}^K \gamma_{g(h)} / (\alpha_{g(h)}^L \hat{\phi}_{g(h)})\right) L_{jht}^V},$$
and analogously for $\hat{G}_{jft}^{(0)}$, replacing γ_g with $1 - \gamma_g$.

2. Compute $\mathcal{K}_{jft}^{(k)}$ from (6.2) using $\hat{C}_{jft}^{(k)}, \hat{G}_{jft}^{(k)}$, and compute $\mathcal{L}_{jft}$ from (6.3) using observed labor and $\hat{\boldsymbol{\theta}}_L$.

3. Compute shares $s_C^{\mathcal{K},(k)}(jft), s_G^{\mathcal{K},(k)}(jft)$ and $s_V^{\mathcal{L}}(jft; \hat{\boldsymbol{\theta}}_L)$.

4. Update $\psi_{jft}^{C,(k)}$ and $\psi_{jft}^{G,(k)}$.

5. Update allocations from (6.22)–(6.23).

6. Stop when $\left\| \hat{\mathbf{C}}_{jt}^{(k+1)} - \hat{\mathbf{C}}_{jt}^{(k)} \right\| < 10^{-8}$ and similarly for $\hat{\mathbf{G}}_{jt}$. We use a tight tolerance of 10^{-8} because this allocation solver is nested within the linear IV iteration and the outer loop below.

Given fixed-point allocations, construct $\log \mathcal{K}_{jft}(\boldsymbol{\theta}_P; \hat{\boldsymbol{\theta}}_L)$, form residuals from (7.5), and let Z_{jft}^P be the second-stage instrument vector. We define a set of moments for each broad group $g \in \mathcal{G}$:
$$\hat{m}_g(\boldsymbol{\theta}_P) \;=\; \frac{1}{N_g} \sum_{jt} \sum_{f:\, g(f)=g} u_{jft}(\boldsymbol{\theta}_P; \hat{\boldsymbol{\theta}}_L) \otimes Z_{jft}^P, \qquad N_g = \#\{(j,f,t) : g(f) = g\}, \tag{7.15}$$

and stack the blocks into $\hat{m}_P(\boldsymbol{\theta}_P) = \left[\hat{m}'_{g_1}, \ldots, \hat{m}'_{g_{|\mathcal{G}|}}\right]'$. For any weight matrix W, the bilinearity of u_{jft} makes the minimizers of $\hat{m}'_P W \hat{m}_P$ over $\boldsymbol{\theta}_{CD}$ available in closed form: given $\{\varrho_g\}$, the $\{\alpha_g^K, \alpha_g^L\}$ solve linear IV–GMM equations in the quasi-differenced variables (separable across groups when W is block-diagonal). Given the α's, stacking $a \equiv$ the moments of v_{jft} and $B \equiv$ the group-blocked moments of $v_{jf,t-1}$ gives $\hat{\boldsymbol{\varrho}} = (B'WB)^{-1}B'Wa$. The iterated linear IV alternates these solves with the allocation solver. The iterated linear IV is then the parameter-space loop: the allocation solver constructs $\log \mathcal{K}$ at *fixed parameters*, while this loop updates the parameters $\boldsymbol{\theta}_{CD}$ themselves:

1. Given $\boldsymbol{\theta}_{CD}^{(m)}$, solve the allocation fixed point and construct $\log \mathcal{K}_{jft}^{(m)}$.
2. Given $\log \mathcal{K}_{jft}^{(m)}$, update $\boldsymbol{\theta}_{CD}^{(m+1)}$ by the closed-form solves above.
3. Stop when $\|\boldsymbol{\theta}_{CD}^{(m+1)} - \boldsymbol{\theta}_{CD}^{(m)}\| < 10^{-4}$.

At convergence the $\boldsymbol{\theta}_{CD}$ moment conditions hold exactly at allocations consistent with the estimated parameters themselves. In our data the fixed point is reached in a handful of iterations, and the result reproduces direct nonlinear minimization of the same objective, at the cost of a few linear solves rather than a numerical search over $3|\mathcal{G}|$ parameters.

Stage 3: GMM Outer Loop over $\boldsymbol{\theta}_K$. The map $\boldsymbol{\theta}_K \mapsto \hat{\boldsymbol{\theta}}_{CD}(\boldsymbol{\theta}_K; W)$ defined by Stage 2 concentrates the linear block out of the objective, the exact analogue of Stage 1: just as the fixed effects never enter the Stage-1 NLLS search, $\boldsymbol{\theta}_{CD}$ never enters the outer search.

$$Q_P(\boldsymbol{\theta}_K) \;=\; \hat{m}_P\big(\boldsymbol{\theta}_K, \hat{\boldsymbol{\theta}}_{CD}(\boldsymbol{\theta}_K; W)\big)' \, W \, \hat{m}_P\big(\boldsymbol{\theta}_K, \hat{\boldsymbol{\theta}}_{CD}(\boldsymbol{\theta}_K; W)\big). \tag{7.16}$$

We minimize (7.16) by derivative-free multistart search over the $|\mathcal{G}| + 1$ dimensions of $\boldsymbol{\theta}_K$. Within each pass the same W is used by the inner closed-form solves and the outer criterion, and is held fixed across candidate $\boldsymbol{\theta}_K$, so that criterion values are comparable across guesses. $\boldsymbol{\theta}_K$ is then identified only off the remaining overidentifying directions. This is demanding in practice: where Q_P is flat or monotone in $\boldsymbol{\theta}_K$ (as we find it is), we calibrate that component and report instead that $\hat{\boldsymbol{\theta}}_{CD}$ is invariant across its plausible range. Standard errors are computed by block-bootstrap on universities ($B = 200$ draws), re-estimating the labor NLLS, the iterated linear IV, and the outer $\boldsymbol{\theta}_K$ search on each draw so that first- and second-stage estimation error propagates. Each bootstrap replication costs a few linear solves per outer evaluation rather than a full nonlinear multistart over $\boldsymbol{\theta}_P$.

7.1.1 Instruments

The production function residual instrument vector must be orthogonal to u_{jft} in (7.6). Current capital and quasi-fixed labor are predetermined state variables under the timing assumptions in Section 6.2; their validity additionally requires that the measurement error ε_{jft} is not observed when inputs are chosen. Theoretically, we can then use current quasi-fixed inputs and lagged input measures. Due to the presence of publication lags, we use a conservative instrument set that takes $t-2$ lags of all inputs:

$$Z_{jft}^{\text{inputs}} = \left\{ \log L_{jft-2}^{V}, \log L_{jft-2}^{F}, \log C_{jt-2}, \log G_{jt-2} \right\} \tag{7.17}$$

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

$$Z_{jft}^{\text{HERD}} = \left\{ \log R_{jft-2} \right\} \tag{7.18}$$

Capital substitution σ_K. The capital substitution parameter is estimated jointly with the remaining production-function parameters rather than from a separate capital first-order condition since we do not have observed data on field-level capital allocations. To identify σ_K, we use NVIDIA architecture transitions interacted with predetermined university exposure to older 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:

$$\Delta p_t^G \equiv \log P_t^G - \log P_{t-1}^G, \tag{7.19}$$

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

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

$$\text{share}_{jt-1}^{\text{prev}} \equiv \frac{\sum_{v \in \mathcal{V}^{\text{prev}}(t)} \text{TFLOPS}_{jt-1}^{v}}{G_{jt-1}}. \tag{7.20}$$

The instrument is

$$Z_{jt}^{\text{NVDA}} \equiv \Delta p_t^G \cdot \text{share}_{jt-1}^{\text{prev}}. \tag{7.21}$$

The identifying assumption is that, conditional on fixed effects and baseline field composition and GPU intensity, the national price shock interacted with predetermined vintage exposure affects field-level output through its effect on university GPU investment and the CPU–GPU capital mix, rather

than through exposure-specific productivity shocks. We assess this assumption by reporting the first stage for GPU investment and the GPU–CPU capital ratio, controlling for baseline GPU intensity interacted with year effects, and testing for effects on non-GPU publication outcomes.

The Stage 2 instrument vector is

$$Z^P_{jft} = Z^{\text{inputs}}_{jft} \cup Z^{\text{HERD}}_{jft} \cup Z^{\text{NVDA}}_{jt}. \tag{7.22}$$

7.2 Investment Policy Function

Since investment is a dynamic choice variable, universities can also endogenously invest in response to external capital grants. We estimate a reduced-form behavioral response function, a proxy for the investment policy function if we were to estimate the full dynamic model. We define GPU investment as $I^G_{jt} = G_{jt+1} - (1-\delta_G)G_{jt}$. On average, most data-center hardware expires in 5 to 6 years, so we set $\delta_G = 0.2$. We estimate the following behavioral response function:

$$I^G_{jt} = \alpha^G_j + \delta^G_t + \beta^G N^G_{jt} + \sum_{\tau \geq 2} \beta^G_\tau \big(N^G_{jt} \times \mathbf{1}\{\text{tier}_j = \tau\}\big) + \pi G_{jt} + X'_{jt}\gamma^k + \varepsilon^G_{jt}, \tag{7.23}$$

where α^G_j and δ^G_t are university and year fixed effects and tier_j is the financial tier (i.e., mid, low, high) of the university. The most-constrained tier ($\tau = 1$) is the omitted category, so β^G is its funding response and $\beta^G + \beta^G_\tau$ is the response of tier τ. Because the tier is time-invariant, its level is absorbed by α^G_j and only the interaction with N^G_{jt} is identified. We control for baseline levels of GPU capital G_{jt} and covariates X_{jt} that vary at the institution–year level: log non-NSF federal research funding, log STEM faculty, log graduate enrollment, and log endowment payout.

Table 1 tests whether lagged investment contains additional predictive information once we condition on the observable state variables entering the investment policy. In particular, we control for current GPU stock G_{jt}, NSF computing awards N^G_{jt}, and university financial and research inputs.

$$I^G_{jt} = \alpha^G_j + \delta^G_t + \rho_1 I^G_{j,t-1} + \rho_2 I^G_{j,t-2} + \beta^G N^G_{jt} + \pi G_{jt} + X'_{jt}\gamma^k + u_{jt} \tag{7.24}$$

Among universities that have already adopted GPU capacity, neither $I^G_{j,t-1}$ nor $I^G_{j,t-2}$ predicts current investment under either depreciation assumption. Thus, conditional on the current state, lagged investment does not appear to constitute an additional state variable for the intensive-margin investment decision. This supports approximating the post-adoption investment policy with the contemporaneous reduced-form policy function in Equation (7.23), rather than estimating a separate dynamic law of motion for investment.

When pre-adoption years are included, however, $I^G_{j,t-2}$ is positive and statistically significant. The coefficient disappears entirely in the post-adoption sample, suggesting that this persistence is

associated with the lumpy extensive-margin transition into GPU ownership rather than persistence in investment conditional on an existing GPU stock. We therefore interpret Equation (7.23) as a reduced-form policy for the intensive-margin investment response of existing GPU adopters.

Table 1: Autoregressions of GPU Investment Flows I_{jt}^G

	(1) $\delta_G = 0$	(2) $\delta_G = 0.2$	(3) $\delta_G = 0$	(4) $\delta_G = 0.2$
$I_{j,t-1}$	−0.01	0.03	0.03	0.03
	(0.03)	(0.04)	(0.02)	(0.03)
$I_{j,t-2}$	0.13***	0.16***	−0.01	−0.01
	(0.02)	(0.03)	(0.02)	(0.03)
University FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Years Included	All	All	Post-Adoption	Post-Adoption
Observations	318	318	132	132

Notes. Within-university autoregressions of internal GPU investment flows, conditional on the policy-function state: NSF award dollars N_{jt}^G, the GPU stock level G_{jt}, log non-NSF federal R&D, log STEM faculty, log graduate enrollment, and log endowment. Standard errors clustered by university in parentheses. *, **, *** denote significance at the 10%, 5%, and 1% levels.

NSF computing capital is scarce in the panel: across the sample window the MRI program makes on the order of one hundred computing-related awards in total, so most institution–years have $N_{jt}^G = 0$ and the variation is thin. The award that does occur is also endogenous; institutions that win are those positioned and inclined to invest, so N_{jt}^G is correlated with the policy-function innovation ε_{jt}^G, biasing the OLS towards crowd-in. We therefore instrument N_{jt}^G (and its tier interactions) with a combination of instruments:

1. **The CHIPS cost-share waiver.** Section 10320 of the CHIPS and Science Act of 2022 eliminated the 30% MRI cost-share requirement for PhD-granting institutions for five years beginning with the FY2023 competition. The waiver lowered the effective price of an award differentially across institutions according to how binding the match had been, giving $Z_{jt}^{\text{chips}} = \text{FinConstraint}_{j,\text{pre}} \times \mathbf{1}\{t \geq 2023\}$.

2. **An MRI and CC* program-emphasis shift-share.** $Z_{jt}^{\text{cc}} = \sum_f s_{j,\text{pre}}^f \tilde{e}_{ft}$, where $s_{j,\text{pre}}^f$ is j's predetermined field composition and $\tilde{e}_{ft}$ is the MRI and CC* domain-emphasis in year t, residualized on contemporaneous national field research funding so that only the administratively idiosyncratic component of program targeting enters. This instrument varies across the full pre-2023 panel from year to year and is the primary source of identification prior to 2023.

Given a GPU grant of N_{jt}^{G}, the counterfactual field-level output can be calculated as:

$$\widehat{Y}_{jft}(C_{jft}, G_{jft} + (1 + \beta^{G} + \beta_{\tau}^{G})N_{jt}^{G}, L_{jft}^{V}, L_{jft}^{F}) \tag{7.25}$$

taking the baseline levels of CPU capital, GPU capital, quasi-fixed labor, and flexible labor as given. Since we do not estimate the full dynamic model, we cannot simulate out the path of endogenous investment in other dynamic choice variables.

8 Results

We present results from the production function estimation procedure. We calibrate and hold fixed the within-nest shares, the capital substitution elasticity, and the autoregressive parameters. We calibrate $\gamma_g = \gamma = 0.5$, $\phi_g = \phi = 0.25$, $\sigma_K = 0.8$, and $\varrho_g = \varrho = 1$. This assumes that CPU and GPU capital are complements and unobserved productivity is a fixed effect that we first-difference out. The focus will be on identifying the two types of labor as strong substitutes ($\sigma_L > 1$) and the outer Cobb–Douglas parameters α^K and α^L.

We measure output as *computational papers*; flexible labor is enrolled PhD students plus postdocs from our spliced PQDT and NSF GSS surveys; and quasi-fixed labor is computational faculty from OpenAlex authorship records. Capital is CPU cores and dense-FP16 GPU TFLOPS. Estimation is at the university × field × year level.

8.1 Labor Nest Parameters

Unfortunately, within-cell NLLS drives σ_L to its upper bound of ∞ in every specification due to the within-cell variation in output being unrelated to the measured flexible labor. We run a modified between-cell version of the NLLS by taking the long-run averages. Each university × broad-field cell is collapsed to its mean over the $T_{jf} \geq 5$ years with valid data, and we estimate the steady-state analogue of (7.9):

$$\bar{y}_{jf} = \kappa_f + \delta_j + \log\left[\phi + (1-\phi)\, e^{\rho_L \bar{l}_{jf}}\right] + \bar{e}_{jf}, \qquad \bar{y}_{jf} = \overline{\log(Y/L^V)}_{jf}, \;\; \bar{l}_{jf} = \overline{\log(L^F/L^V)}_{jf}, \tag{8.1}$$

Recall that $\rho_L \equiv (\sigma_L - 1)/\sigma_L$. The panel cell effect a_{jf} cannot be absorbed between cells; it is decomposed as $\kappa_f + \delta_j + \eta_{jf}$, with broad-field effects κ_f (technology, wage premia, field-level coverage) and university effects δ_j (wages, shadow values, university-wide coverage) concentrated out as before, and the interaction η_{jf} left in the error. Identification requires university-field productivity deviations to be orthogonal to long-run composition, so the estimate is best read as an upper bound within its

confidence set. Our final sample has 397 observations and median $T = 9$. Standard errors are clustered by university.

With Y the equal-credit count of compute-benefiting papers and L^F the computational faculty stock, the between-cell NLLS gives $\hat{\sigma}_L = 5.3$. Hypotheses on σ_L are tested with the criterion-difference statistic $D = N[\,Q_N(\sigma_0) - Q_N(\hat{\sigma}_L)\,]/Q_N(\hat{\sigma}_L)$, whose null distribution is simulated by a wild cluster bootstrap under various nulls:

H_0: $\sigma_L =$	1 (Cobb–Douglas)	1.4	2	100
D	275.5	88.7	24.4	4.3
bootstrap p	0.005	0.010	0.095	0.220

Cobb–Douglas is rejected, as is $\sigma_L = 1.4$. $\sigma_L = 2$ is not rejected at the 5% level ($p = 0.095$), and the criterion cannot distinguish large values. The 95% criterion-based confidence set is therefore approximately $\sigma_L \in [1.7, \infty)$ with the point estimate at roughly 5. The data speak against complementarity and unit elasticity, and favor strong substitutability.

Literature Benchmarks. Labor-market estimates generally suggest greater substitutability among more similar worker groups. Between broad education groups, Katz and Murphy (1992) estimate an elasticity of substitution of 1.41 between college and high-school labor equivalents, while Goldin and Katz (2008) obtain a preferred estimate of 1.64, with estimates around 1.6–1.8. Acemoglu and Autor (2011) summarize most estimates between college and non-college workers as lying between 1.4 and 2. Allowing workers of different ages to be imperfect substitutes within education groups, Card and Lemieux (2001) estimate a substantially larger elasticity of approximately 4–6; the same paper estimates an elasticity of roughly 2–2.5 between college and high-school labor. These age groups are not a direct analogue to faculty versus PhD students and postdocs, but provide a useful benchmark for substitution among workers with similar skills at different career stages.

8.2 Outer CD Parameters

We present estimates of α^K and α^L using both $\sigma_L = 2$ and $\sigma_L = 4$ in Table 2. The parameters do not vary much under either σ_L, so we examine the case when $\sigma_L = 2$. Computer science exhibits the highest α^K at 0.26, followed by life sciences at 0.098. We see slightly increasing returns to labor for engineering and life sciences. Since the outer CD parameters are estimated via iterated linear IV, parameter bounds are not artificially imposed.

Table 2: Output elasticities of HPC capital and research labor

	α^K	α^L	J (p)
Panel A: $\sigma_L = 2$			
Computer science	0.262*	0.277	10.4 (0.065)
	(0.078)	(0.410)	
Engineering + materials	0.087	1.019*	6.5 (0.261)
	(0.073)	(0.463)	
Geo/physics/chemistry	0.036	0.654*	7.3 (0.199)
	(0.064)	(0.312)	
Life sciences	0.098*	1.100*	3.1 (0.685)
	(0.028)	(0.284)	
Panel B: $\sigma_L = 4$			
Computer science	0.252*	0.323	10.3 (0.067)
	(0.087)	(0.446)	
Engineering + materials	0.030	0.544	11.5 (0.042)
	(0.058)	(0.489)	
Geo/physics/chemistry	0.014	0.634	7.6 (0.180)
	(0.059)	(0.356)	
Life sciences	0.080*	0.894*	8.0 (0.156)
	(0.027)	(0.333)	

Notes: Calibration: $\sigma_K = 0.8$, $\gamma = 0.5$, $\phi = 0.25$, $\varrho = 1$; each panel re-solves the capital-allocation fixed point at its σ_L. Instruments: $\{1, \log C_{t-2}, \log G_{t-2}, \log L^V_{t-2}, \log L^F_{t-2}, \log R_{t-2}\}$. Standard errors (in parentheses) clustered by university; * $p < 0.05$. J is the per-group two-step Hansen statistic (dof 5). Estimates are insensitive to the calibrated nests.

Instrument diagnostics. The NVIDIA instrument was rebuilt using launch prices and dense tensor throughput from Epoch AI (2026) and TechPowerUp (2026). For the first stage, we match to the moment timing (first differences, year effects, clustered by university) and find that $\Delta \log(G/C)$ on Z^{NVDA} gives $\beta = -1.00$ (F-stat= 9.2). The sign is the replacement-cycle prediction: universities holding old vintages upgrade when the new generation cuts the price per TFLOP. Unfortunately, a quantity response to the price shock moves $\log G$ but does not trace out within-nest curvature at this grain since σ_K is unidentified. Thus, we calibrate $\sigma_K = 0.8$ to be complements for now.

Benchmarking. We plan to compare the marginal product of GPU capital implied by the structural production function with the reduced-form effects from the event-study design.

8.3 Investment Policy Function

We present a very preliminary discussion on the grant-to-investment response since the instruments are not yet valid, especially pre-2023. An OLS of GPU investment on NSF compute-award dollars is uninformative at obligation-year timing since it's mechanically negative as hardware arrives one to

two years later. Due to endogeneity between winning a grant and investment (which both respond to high unobserved productivity), without instruments we instead do a matched event-study design. 38 capital-verified award events at the 53-school HPC panel show flat pretrends, then a build-out cumulating to roughly +33,000 TFLOPS by $t+5$, about +25,000 TFLOPS net of the hardware the award itself purchased. This translates to an institutional amplification of roughly 2.7×. A placebo using non-compute MRI awards shows no award-timed response, so the effect is compute-specific. Amplification rises steeply with financial tiers: low-tier schools show pass-through only (net ≈ 0), the middle tier roughly 5×, and high-tier schools 8× (six events). One caveat on units: the net-of-grant calculation converts award dollars to TFLOPS at a placeholder price of $83 per FP16 TFLOP and the crowd-in verdict flips only if capacity costs less than about $23 per TFLOP at award time (below that price the average award could have purchased the entire observed build-out). Historical GPU prices exceed this threshold for nearly all award years, but a year-specific per-TFLOP price series still remains a priority.

9 Counterfactuals

In order to understand the returns to external infrastructure grants, there are two quantities that must be measured. First, the extent of the university's endogenous private investment I_{jt}^G given an external grant of size N_{jt}^G. Second, the net returns to net change in computing capital: $N_{jt}^G + I_{jt}^G$. Our production function framework in Section 6 allows us to estimate the net returns under various levels of inputs, and we approximate the endogenous investment response with a reduced-form policy response function outlined in Section 7.2. We first formalize the planner's efficiency–equity tradeoff and then describe counterfactual allocation exercises. Since we do not have estimates for the university's behavioral response, we present very preliminary results assuming no endogenous private investment.

9.1 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 $N_j^G \geq 0$ for each university j, and the counterfactual output will be realized according to

(7.25). The planner's objective is an Atkinson-style social welfare function over scientific output:

$$W(\rho) = \sum_{j=1}^{J} \frac{\mathcal{Y}_j(N_j^G; \boldsymbol{\theta})^{1-\rho}}{1-\rho} \tag{9.1}$$

where

$$\mathcal{Y}_j(N_j^G; \boldsymbol{\theta}) \equiv \widehat{Y}_{jt}\Big(G_{jt} + (1 + \beta^G + \beta^G_{\tau(j)})N_j^G; \boldsymbol{\theta}\Big) \tag{9.2}$$

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:

$$\begin{aligned} \max_{\{N_j^G\}_{j=1}^J} \quad & \sum_{j=1}^{J} \frac{\mathcal{Y}_j(N_j^G; \boldsymbol{\theta})^{1-\rho}}{1-\rho} \\ \text{s.t.} \quad & \sum_{j=1}^{J} p_G\, N_j^G \leq \bar{B}_{\text{NSF}}, \qquad N_j^G \geq 0 \;\; \forall\, j \end{aligned} \tag{9.3}$$

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

$$\mathcal{Y}_j(N_j^G; \boldsymbol{\theta})^{-\rho} \cdot \frac{1}{p_G}\, MPG_j^{\text{post}}\Big(G_{jt} + (1 + \beta^G + \beta^G_{\tau(j)})N_j^G; \boldsymbol{\theta}\Big) = \lambda \tag{9.4}$$

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, which directs more resources toward universities with lower baseline output.

9.2 Counterfactual Reallocation of NSF Computing Grants

We use the estimated production function to ask how the observed NSF computing portfolio compares with an output-maximizing allocation of the same budget. The planner reallocates the FY2021–2025 MRI and CC* compute-award dollars observed at the estimation panel (B = \$30.5M across

22 universities) taking the baseline year t^* as given. Grant dollars convert one-for-one into GPU capacity at a benchmark price p^G chosen so that the total budget purchases one aggregate 2020 stock ($83 per dense-FP16 TFLOP). Recall that there is *no* behavioral investment response in this section; each grant dollar becomes exactly $1/p^G$ TFLOPS of installed capacity since the preliminary results in Section 8.3 suggest substantial institutional crowd-in.

Within-university allocations across fields re-solve through the model's capital-allocation fixed point at the counterfactual stock. Unobserved productivity Ω_{jf} (recall we imposed $\varrho = 1$ for now, so unobserved productivity is a fixed effect) is recovered residually from observed output at t^*. Ω_{jf} is held fixed and labor and CPU capital do not adjust. The efficient benchmark allocates the budget by greedy equalization of dollar marginal products across universities implied by Equation 9.4.

Table 3 reports the headline comparison at two baseline years. At $t^* = 2021$, the observed allocation raises predicted computational output by 242 papers relative to a no-grant baseline of 7,302, while the output-maximizing allocation of the same budget raises it by 544: **the NSF's allocation captures 45% of the attainable output gain** (7.9 versus 17.8 papers per $M). The 2020 baseline gives 49%. The difference reflects *who* is funded rather than how much: the rank correlation between observed and efficient allocations is −0.12, only 14 of the 22 funded universities receive anything from the planner, and $14.5M of observed dollars flow to universities to which the planner allocates nothing. The efficient allocation concentrates two-thirds of the budget in ten universities with high baseline computational output but small installed capacity (Pittsburgh, Georgia, Maryland, Caltech), reflecting the marginal-product dispersion.

Table 3: Counterfactual NSF allocation: observed vs. output-maximizing (no behavioral response, $m = 1$)

	$t^* = 2020$	$t^* = 2021$
Budget B (observed, $M)	29.9	30.5
Baseline $\hat{Y}$ (no grants)	7,052	7,302
Gain, observed allocation	+246	+242
Gain, efficient allocation	+503	+544
Papers per $M (observed / efficient)	8.2 / 16.8	7.9 / 17.8
Share of attainable gain captured	49%	45%
Universities funded (observed / efficient)	21 / 27	22 / 32

Notes: Point estimates at $\hat{\boldsymbol{\theta}}$ from Table 2, Panel A ($\sigma_L = 2$); no bootstrap inference yet. Output is predicted computational papers. Budget is calculated as observed FY2021–2025 MRI and CC* compute awards at panel universities. Awards at non-panel universities ($60M, roughly two-thirds of the national portfolio) are outside the estimation sample and excluded. p^G = $83 per FP16 TFLOP (calibrated). Efficient allocation is greedy equalization of dollar marginal products within a numerical tolerance.

Table 4 decomposes the $t^* = 2021$ gains by field. Computer science expands most in proportional terms under both allocations (+10% observed, +24% efficient), consistent with its high α^K and small baseline. Life sciences contributes the largest level gain (+353 of the +544 efficient-allocation papers) through the combination of a moderate, precisely estimated α^K and the largest baseline output. Geosciences, physics, and chemistry are nearly inert: their α^K of 0.036 implies little scope for computing capital to move output at this margin.

Table 4: Predicted computational papers by field: baseline vs. counterfactual allocations ($t^* = 2021$)

	Baseline	Observed allocation		Efficient allocation	
		Level	%	Level	%
Life sciences	4,704	4,856	+3.2	5,057	+7.5
Engineering + materials	1,114	1,148	+3.1	1,183	+6.2
Geo/physics/chemistry	1,078	1,091	+1.3	1,104	+2.5
Computer science	407	449	+10.4	502	+23.5
Total	7,302	7,544	+3.3	7,847	+7.5

Notes: Field-level output re-solved through the allocation fixed point at each university's counterfactual GPU stock; productivity fixed at its t^* backout. Percent changes relative to the no-grant baseline.

Finally, we estimate where the NSF currently sits on the efficiency–equity frontier. Taking logs of the interior FOC (9.4), the optimality condition for post-adoption universities that received NSF funding becomes:

$$\ln\left(MPG_{jt}^{\text{post}}\right) = \alpha_t + \rho \ln\left(\widehat{Y}_{jt}^{\text{base}}\right) + u_{jt}, \tag{9.5}$$

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.

Estimating (9.5) on the funded, post-adoption universities in the panel yields $\hat{\rho} = 0.40$ at the 2021 baseline ($n = 16$) and $\hat{\rho} = 0.98$ at the 2020 baseline ($n = 14$). Unsurprisingly, under either

baseline the NSF is revealed to accept lower marginal products at smaller universities, with an implied inequality aversion between moderate curvature and approximately logarithmic preferences over university output. The smallest funded universities are pushed toward saturation (low MPG^{post} at low $\widehat{Y}^{\text{base}}$), while the highest-output funded universities retain marginal products an order of magnitude higher, the same pattern that generates the roughly 45% efficiency capture in the reallocation exercise.

10 Conclusion

This paper studies the returns to centrally provisioned GPU capital and the allocation of federal cyberinfrastructure funding across universities. The central economic tradeoff is that well-resourced universities may have a high technical return to compute because their researchers and complementary inputs are more productive, but they also begin with larger input stocks and are better able to purchase GPUs on their own. Less-resourced universities may have lower baseline productivity, but smaller existing capital stocks and tighter financing constraints make an external grant more additional. The return to a grant therefore combines three objects: the technical marginal product of compute, the university's private investment response, and the adjustment of complementary research labor.

We construct a new panel of university CPU and GPU capacity and link it to field-level publications and researcher histories. Event studies around large capacity expansions show modest average effects on AI-related publications but substantial heterogeneity in computer science. Low-tier universities experience a persistent increase of roughly 50% in computer-science output, while the initial response at high-tier universities dissipates quickly. Roughly half of the low-tier gain comes from incumbent researchers, with the remainder reflecting the entry of new researchers. These results are consistent with an increased return to compute at constrained institutions and with infrastructure affecting the location of complementary talent.

However, we also find preliminary evidence of substantial crowd-in at middle- and high-tier universities using a matched event study with 38 capital-verified NSF awards. Installed capacity rises by substantially more than the benchmark amount purchased by the grant itself, implying average institutional amplification of roughly 2.7×. Because award receipt is endogenous and the net-of-grant calculation still relies on a provisional historical price series, these results should be treated as

descriptive rather than causal.

The structural estimates provide a complementary measure of the technical return to research inputs. The labor first-order condition favors strong substitution between flexible and quasi-fixed researchers, with a point estimate of $\widehat{\sigma}_L \simeq 5.3$ and a confidence set that rules out Cobb–Douglas. Conditional on the current nest calibrations, the output elasticity of HPC capital is largest in computer science, at approximately 0.26, followed by life sciences, at approximately 0.10. The current data do not separately identify the CPU–GPU substitution elasticity, and persistent productivity is treated as a fixed effect, so these estimates remain preliminary.

Using these estimates, the static counterfactual suggests that the observed NSF portfolio within the estimation panel is substantially more dispersed than an "efficient" output-maximizing allocation. At the 2021 baseline, the observed \$30.5 million allocation raises predicted output by 242 papers, while the output-maximizing allocation raises it by 544 papers. The observed portfolio captures about 45% of the attainable gain. The efficient allocation favors universities with high baseline computational output but relatively little installed GPU capacity. The NSF's observed choices imply a revealed inequality-aversion parameter between 0.40 and 0.98, although this estimate also absorbs objectives such as geographic dispersion, institutional capacity building, and researcher training that are not explicitly modeled.

Currently, the reduced-form and structural analyses point to different implications. The event study finds the largest persistent publication response at low-tier universities, while the structural counterfactual assigns most of the output-maximizing budget to relatively high-output institutions. The descriptive investment response also indicates more crowd-in at middle- and high-tier universities. Part of this difference is conceptual: the event study measures the realized response to large upgrades, including labor entry and other dynamic complementarities, whereas the production function isolates a technical marginal product, and the current counterfactual holds labor fixed and assumes no endogenous private investment. We therefore do not interpret the current counterfactual as overturning the reduced-form result per se, but the discrepancy does identify the main task for the next draft.

References

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4B, pp. 1043–1171. Amsterdam: Elsevier.

Advanced Micro Devices (2021). Amd instinct mi200 series accelerator datasheet. https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/instinct-mi200-datasheet.pdf.

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

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

Arellano, M. and S. Bond (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., J. S. Graff Zivin, D. Li, and B. N. Sampat (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. and F. Gaessler (2026, January). The returns to physical capital in knowledge production: Evidence from lab disasters. *American Economic Journal: Applied Economics 18*(1), 248–85.

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

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

Bresnahan, T. F., E. Brynjolfsson, and L. M. Hitt (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics 117*(1), 339–376.

Card, D. and T. Lemieux (2001). Can falling supply explain the rising return to college for younger men? A cohort-based analysis. *Quarterly Journal of Economics 116*(2), 705–746.

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

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

Emeras, J., S. Varrette, V. Plugaru, and P. Bouvry (2019). Amazon elastic compute cloud (EC2) versus in-house HPC platform: A cost analysis. *IEEE Transactions on Cloud Computing 7*(2), 456–468.

Epoch AI (2026). Data on machine learning hardware. https://epoch.ai/data/machine-learning-hardware. CC-BY 4.0. Accessed July 2026.

Furman, J. L. and S. Stern (2011). Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review 101*(5), 1933–1963.

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

Goldin, C. and L. F. Katz (2008). *The Race between Education and Technology*. Cambridge, MA: Belknap Press of Harvard University Press.

Hausman, N. (2022). University innovation and local economic growth. *Review of Economics and Statistics 104*(4), 718–735.

Hofstra, B., V. V. Kulkarni, S. Munoz-Najar Galvez, B. He, D. Jurafsky, and D. A. McFarland (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of*

Sciences 117(17), 9284–9291.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics 124*(4), 1403–1448.

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

Katz, L. F. and K. M. Murphy (1992). Changes in relative wages, 1963–1987: Supply and demand factors. *Quarterly Journal of Economics 107*(1), 35–78.

McKeon, J. and J. Boudou (2025, 06). Innovation under resource constraints: Supercomputing in scientific research. *SSRN Electronic Journal*.

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

Nagaraj, A., E. Shears, and M. de Vaan (2020). Improving data access democratizes and diversifies science. *Proceedings of the National Academy of Sciences 117*(38), 23490–23498.

NAIRR Task Force (2023). Strengthening and democratizing the U.S. artificial intelligence innovation ecosystem: An implementation plan for a national artificial intelligence research resource. Technical report, National Artificial Intelligence Research Resource Task Force.

National Science Foundation (2021). FY 2022 budget request to congress: Directorate for computer and information science and engineering. Technical report, National Science Foundation. https://www.nsf.gov/about/budget/fy2022/.

National Science Foundation (2022). NSF ACCESS awardees will advance innovations in cyberinfrastructure accessibility, user support and integration services. Press release. https://www.nsf.gov/news/nsf-access-awardees-will-advance-innovations.

National Science Foundation (2023). Major research instrumentation program: Instrument acquisition or development. Program Solicitation NSF 23-519, National Science Foundation. https://www.nsf.gov/funding/opportunities/mri-major-research-instrumentation-program/nsf23-519/solicitation.

National Science Foundation (2024a). Campus cyberinfrastructure (CC*). Program Solicitation NSF 24-530, National Science Foundation. https://www.nsf.gov/funding/opportunities/cc-campus-cyberinfrastructure.

National Science Foundation (2024b). Democratizing the future of AI R&D: NSF to launch national AI research resource pilot. Press release, January 24, 2024. https://www.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai.

National Science Foundation, Office of Advanced Cyberinfrastructure (2021). Transforming science through cyberinfrastructure: NSF's blueprint for a national cyberinfrastructure ecosystem for science and engineering in the 21st century. Technical report, National Science Foundation. https://www.nsf.gov/cise/oac/vision/blueprint-2019/.

NVIDIA Corporation (2017). Nvidia tesla v100 gpu architecture. Whitepaper WP-08608-001. https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

NVIDIA Corporation (2021). Nvidia ampere ga102 gpu architecture. Whitepaper, v2.1. https://www.nvidia.com/content/PDF/nvidia-ampere-ga-102-gpu-architecture-whitepaper-v2.1.pdf.

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

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

Shvadron, D., H. Zhang, L. Fleming, and D. P. Gross (2025, June). Measuring the production of scientific human capital: New data, methods, and evidence on the u.s. scientific training ecosystem. Working Paper 33944, National Bureau of Economic Research.

Suh, J. (2026, June). Instrumental burden: Rising capital intensity in science and consequences for innovation. SSRN working paper, New York University, Stern School of Business. Available at SSRN: https://ssrn.com/abstract=7017658.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199. Themed Issue: Treatment Effect 1.

TechPowerUp (2026). Gpu specs database. https://www.techpowerup.com/gpu-specs/. Accessed July 2026.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy 15*(6), 285–305.

Teodoridis, F. (2018). Understanding team knowledge production: The interrelated roles of technology and expertise. *Management Science 64*(8), 3625–3648.

U.S. Department of Energy, Office of Science (2011). The Magellan report on cloud computing for science. Technical report, U.S. Department of Energy, Office of Advanced Scientific Computing Research. https://science.osti.gov/-/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf.

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.

Yue, D. (2024, November). I, google: Estimating the impact of corporate involvement on ai research. Research Paper 5033334, Georgia Tech Scheller College of Business.

A1 Verifying the Within-University Equal-Wage Assumption

The model assumes that the wage w_{jt} of flexible research labor (graduate students, postdocs, and non-faculty researchers) varies by university × year but stays common across all fields f. We test this assumption using self-reported PhD stipends from `phdstipends.com`, a crowd-sourced site on which students report their university, department, academic year, and annual stipend. The raw file contains 17,933 reports spanning 2002 to 2026. We keep reports from 2011 onward (before which coverage is thin), drop exact duplicates, and truncate pay at the within-year 1st and 99th percentiles to remove data-entry errors. We string-match university names to the IPEDS database and end up with 15,350 self-reports across 334 U.S. universities from 2011 to 2025. Department names are matched to eight broad fields: computer science, engineering, geosciences, life sciences, mathematics and statistics, physical sciences, psychology, and social sciences. The final sample consists of 11,513 self-reports across 314 universities.

School identity explains stipend variation. Table A.1 reports the R^2 (and adjusted R^2) from regressing log stipend on combinations of year, university, and field fixed effects. University identity alone explains 24% of the variance in log stipends, and the broad field alone explains only 2%. Adding field fixed effects on top of university × year effects raises the R^2 by just 1.2 percentage points. In other words, once we know the university and the year, the broad field a student is in is nearly uninformative about their stipend.

A natural concern is that the high university R^2 is mechanical with 314 university dummies. Two features of Table A.1 rule this out. First, the adjusted R^2 (0.246) on the additive university specification is similar to the raw R^2 (0.267). Second, column (3) re-runs the decomposition on a random sample of 20 universities drawn from the 100 most-reported institutions (for a total of 2,068 observations). The adjusted and raw R^2 are nearly identical (0.172 vs. 0.179), and university still has two to three times the explanatory power that field does. Column (2) additionally drops the roughly 3% of surviving reports below $8,000, which are concentrated in monthly or partial-year figures. These reports inflate the residual variance and mechanically depress every R^2 level, but they leave the university-versus-field *ranking* intact. The university explains 48% of the variance, and field, added on top of university × year, still adds only 4 percentage points.

Within-university laboratory-science premia. Within university × year cells, we estimate the gap in stipend across fields:

$$\log w_r \;=\; \alpha_{s(r)\times t(r)} \;+\; \sum_g \beta_g \,\mathbf{1}\{f(r) \in g\} \;+\; \varepsilon_r, \tag{A.1}$$

where r is a self-report, $s(r)$ its university, $t(r)$ its year, and g its broad field, with life sciences as the omitted base. β_g measures the mean log stipend gap of field g relative to life sciences in the *same* university and year. Standard errors are clustered by university. Table A.2 reports all seven premia. Laboratory sciences (engineering, physical sciences, geosciences, and computer science) are paid within about three log points of life sciences, and none of these gaps is economically meaningful once the sub-$8,000 reports are removed (column 2). The systematic departures are in the non-laboratory fields: mathematics and statistics, psychology, and social sciences are paid 14 to 16 log

points less than life sciences within the same university × year across both samples.

Appendix Tables

Table A.1: Variance decomposition of log PhD stipends

Fixed effects	(1) All	(2) $\geq$ \$8k	(3) 20 of top 100
Year	0.026 (0.025)	0.084 (0.083)	0.044 (0.038)
University	0.238 (0.217)	0.479 (0.464)	0.179 (0.172)
Broad field	0.021 (0.021)	0.068 (0.068)	0.058 (0.055)
Year + University	0.267 (0.246)	0.575 (0.563)	0.220 (0.208)
+ Broad field	0.284 (0.262)	0.630 (0.619)	0.267 (0.253)
University × Year	0.403 (0.253)	0.679 (0.597)	0.314 (0.222)
+ Broad field	0.415 (0.268)	0.720 (0.649)	0.352 (0.262)
Observations	11,513	11,164	2,068
Universities	314	309	20

Notes. Each cell reports the R^2 (and, in parentheses, the adjusted R^2) from a regression of log stipend on the indicated fixed effects. "Broad field" is the eight-way field classification. Column (1) is the full sample; column (2) additionally drops the $\approx 3\%$ of reports below \$8,000 (largely monthly or partial-year amounts); column (3) is a random draw of 20 universities from the 100 most-reported institutions. The sample is is from `phdstipends.com`, 2011–2025.

Table A.2: Within university-year field premia in log PhD stipends

Field (vs. life sciences)	(1) All S&E	(2) S&E, $\geq$ \$8k
Engineering	-0.052^{***}	0.010
	(0.020)	(0.011)
Physical sciences	-0.020	0.004
	(0.018)	(0.009)
Geosciences	-0.050	-0.027
	(0.032)	(0.018)
Computer science	-0.109^{***}	-0.036^{**}
	(0.035)	(0.015)
Mathematics & statistics	-0.141^{***}	-0.142^{***}
	(0.027)	(0.016)
Psychology	-0.190^{***}	-0.162^{***}
	(0.022)	(0.014)
Social sciences	-0.155^{***}	-0.164^{***}
	(0.018)	(0.014)
Observations	10,324	9,998
University × year cells	1,408	1,385

Notes. Coefficients from equation (A.1): log stipend on broad-field indicators, absorbing university × year fixed effects, with life sciences the omitted category. Estimation uses only university-year cells containing at least two distinct broad fields. Standard errors clustered by university in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Appendix Figures

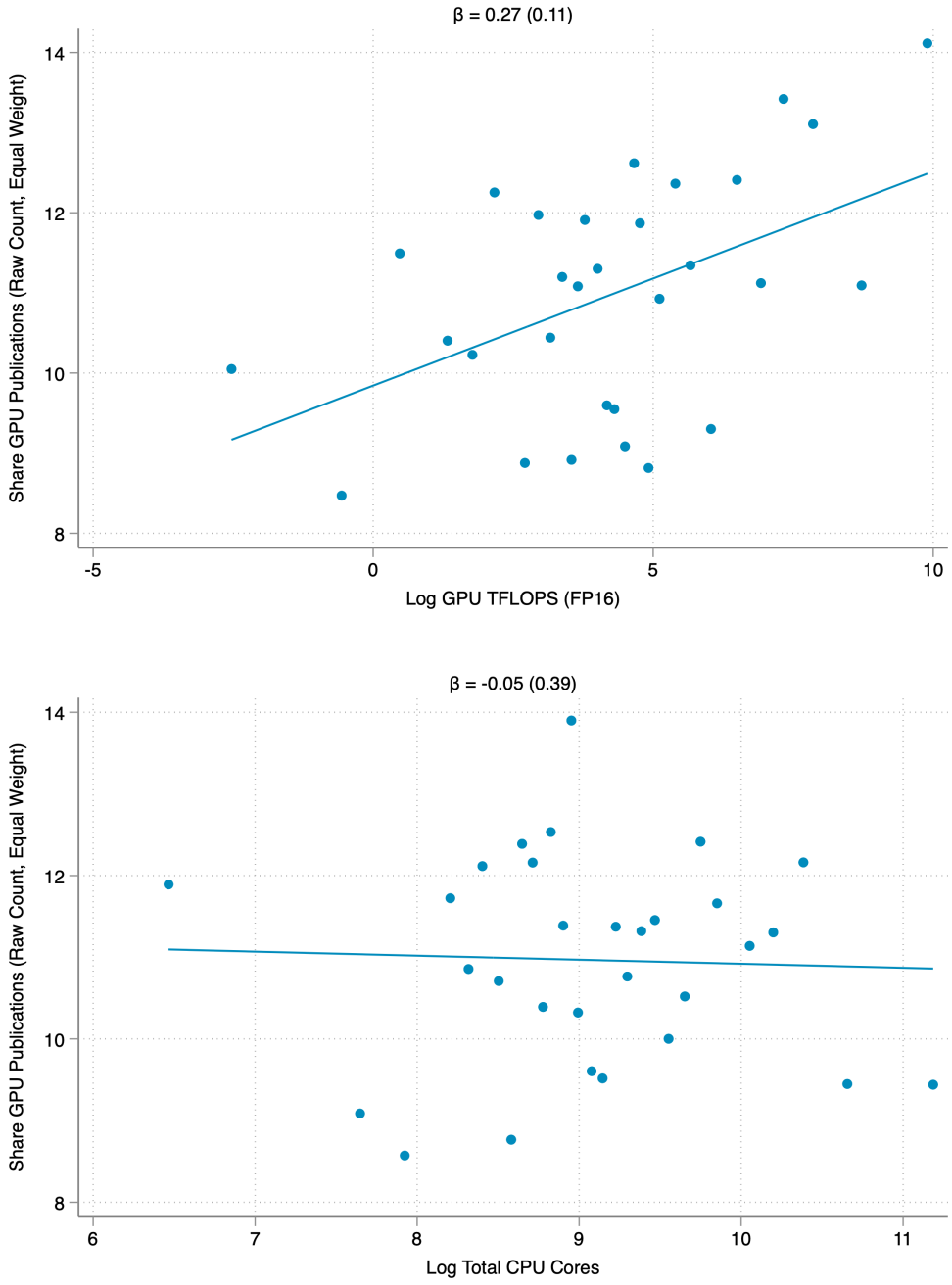


Figure A.1: Binscatter of Share GPU-Relevant Publications Against Log GPU and CPU Capacity (Across Institutions). Binned scatter of an institution's share of GPU-relevant publications against log GPU capacity (FP16 TFLOPS; top panel) and log CPU capacity (total cores; bottom panel), pooled across institutions and absorbing year fixed effects. Slopes and standard errors are from the corresponding regressions. The GPU-relevant share is positive and significant in GPU capacity but statistically indistinguishable from zero in CPU cores, supporting the text-based classification as capturing GPU-specific rather than general-purpose computing output.

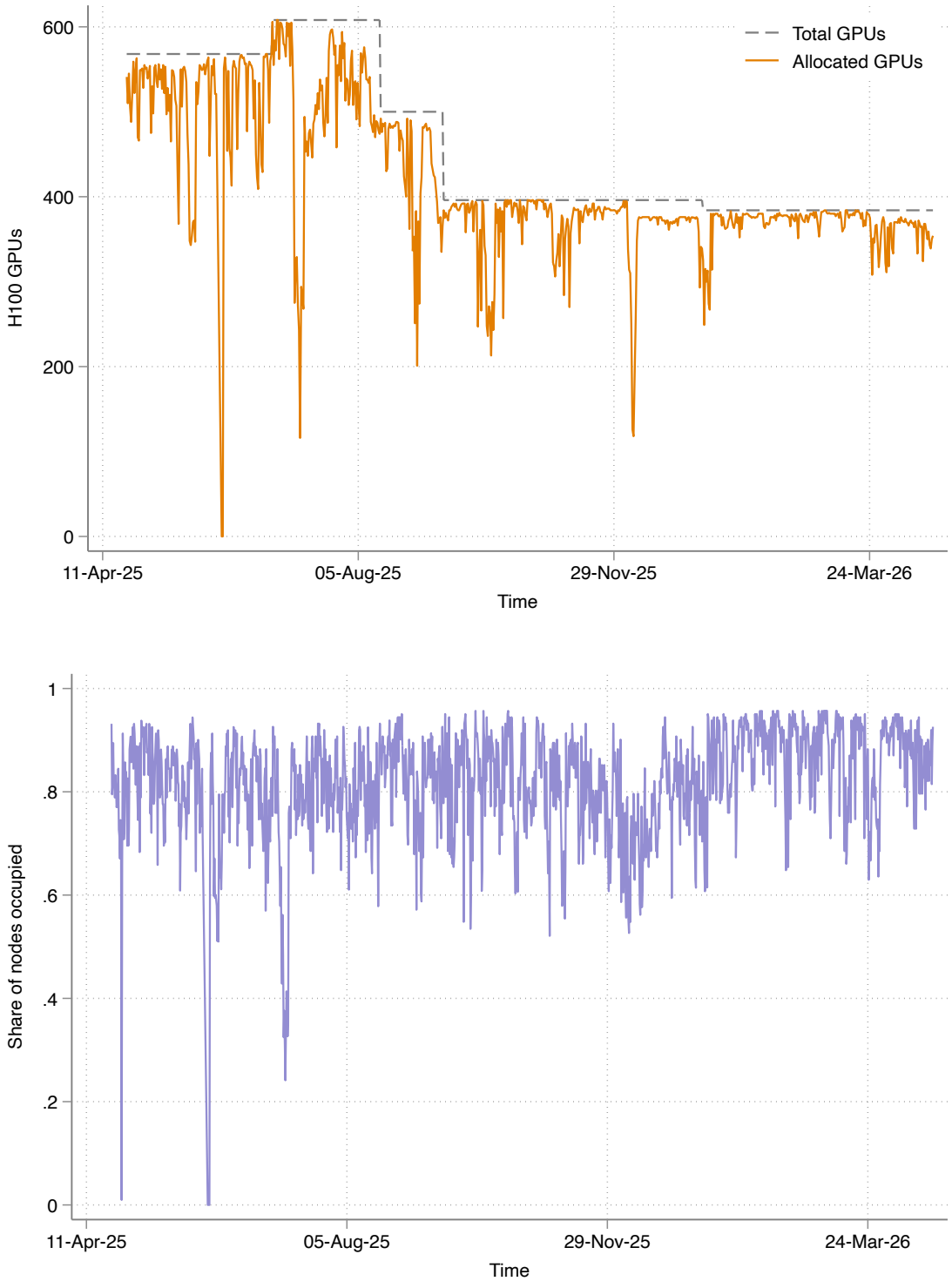


Figure A.2: Harvard Research Computing Total Utilization for NVIDIA H100 GPUs and All Nodes. High-frequency utilization on Harvard's FASRC cluster, April 2025–March 2026. Top panel: total installed NVIDIA H100 GPUs (dashed) versus GPUs allocated to jobs (solid); allocation tracks the available stock closely, and the total steps down as the pool is resized. Bottom panel: share of all compute nodes occupied, fluctuating around 0.8. The series indicate that GPUs run near their binding capacity while overall utilization is stable, motivating the binding-capacity and stable-utilization assumptions in Section 6.