Measuring the Production of Scientific Human Capital: New Data, Methods, and Evidence on the U.S. Scientific Training Ecosystem *

Dror Shvadron[†] University of Toronto
Hansen Zhang[‡] Duke University
Lee Fleming[§] UC Berkeley
Daniel P. Gross[¶‖] Duke University and NBER

March 2026

Abstract:

Using newly-collected data on the near-population of U.S. STEM PhD graduates since 1950, we develop a dissertation-based methodology for measuring PhD populations and use it to present new evidence on who funds PhD training, how many graduates are trained in areas of strategic national importance, and the effects of public investment in PhD training on the scientific workforce. The U.S. federal government is by far the largest source of financial and in-kind support for STEM PhD training in America. We identify universities and fields where PhD training has high rates of government, industry, or philanthropic support, and the organizations and universities that fund and train the most PhDs in critical technology areas such as AI, quantum information technology, and biotechnology. Leveraging variation in government support across agencies and over time, we provide evidence suggesting that increasing government-funded PhD trainees increases PhD production roughly one-for-one. To support further research, we provide public datasets at multiple levels of aggregation, reporting PhD graduates by (i) critical technology area and (ii) source of support. These data and methods complement existing data collection efforts by national statistics agencies, producing information which is otherwise costly to collect or not systematically observed, and can be extended forward in time and to other countries.

JEL Classification: I23, I28, O31, O32, O33, O38

Keywords: Science policy, PhD production, Scientific workforce, Science and technology indicators, Research funding

*We thank the editor and referees for comments that substantially improved this paper, Bhaven Sampat and Bruce Weinberg for helpful conversations, and audiences at the ICSSI annual conference, Summer School on Data and Algorithms for Science, Technology & Innovation Studies conference, NBER Investments in Early Career Scientists meeting, and Rotman School of Management for comments. We also thank James Dunham and colleagues at the Center for Security and Emerging Technology for insights related to technology classification; Michelle Qiu and Max Murakami-Moses for research assistance; and the Duke University Fuqua School of Business, UC Berkeley Technology Competitiveness and Industrial Policy Center, Alfred P. Sloan Foundation (Grant No. G-2023-21064), and National Science Foundation (Grant No. 2420824) for financial support. All errors are our own.

[†]University of Toronto, Rotman School of Management, Toronto, Canada; email: dror.shvadron@utoronto.edu.
[‡]Duke University, Fuqua School of Business, Durham, NC, USA; email: hansen.zhang@duke.edu.
[§]UC Berkeley, Haas School of Business, Berkeley, CA, USA; email: lfleming@berkeley.edu.
[¶]Duke University, Fuqua School of Business, Durham, NC, USA; email: daniel.gross@duke.edu.
[‖]Corresponding author

1 Introduction

One of the primary goals of international research policy, from its post-World War II origins to the present, is the deepening of countries' research capacity through investments in scientific human capital (e.g., Bush 1945, Steelman 1947, OECD 1963b). The size and shape of the scientific workforce influences not only the pace of scientific advance but also its reduction to practice—particularly in high-tech industries organized around the discovery and commercial application of new science. Reflecting these goals, science and technology (S&T) statistics and cross-national comparisons have long emphasized scientific manpower as a performance indicator (Godin 2002b, 2003). Accordingly, a substantial share of science funding in the U.S. and elsewhere is spent on training new researchers (e.g. National Center for Science and Engineering Statistics 2025b).

The doctoral trainee population has specifically been a subject of analysis since the 1940s, and modern innovation scholarship is increasingly focused on who enters science, to what effect, and how early careers evolve. Much of this research uses either public data, including government surveys and bibliometric datasets, or the UMETRICS sample (Lane et al. 2015). However, what is known is constrained by what is measured. Despite continuous progress—including recent improvements in access to administrative data—many important features of the scientific training ecosystem remain hard to measure. Questions like how many PhDs are trained in frontier or emerging subjects like artificial intelligence (AI) and quantum computing, where they are trained, who funds their training, and who subsequently harnesses this talent are difficult to evaluate using available data and measurement systems. National comparisons over long horizons are also challenged by data limitations. As a result, though the U.S. graduated >37,000 STEM PhDs in 2024 (National Center for Science and Engineering Statistics 2025a), it is difficult to assess whether the U.S. (or other countries) are producing the "right" or right number of PhDs, what determines the answer to that question, and how policy choices might cause that number to rise or fall.

In this paper, we introduce a new approach to measuring national doctorate production, using the content of PhD dissertations, and apply it to study several features of the U.S. scientific training ecosystem. Dissertation-based measurement presents three advantages: dissertation science provides a marker of the specific expertise a graduate embodies, and a predictor of future scientific research capacity[1]; dissertations are required of all graduates; and recent advances in large language models (LLMs) make their content considerably more accessible. We focus our analysis on U.S. PhD graduates between 1950 and 2022 in the natural sciences and engineering—over one million

[1]Even Nobel Laureates' future prize-winning science can sometimes be traced to their graduate research. For example, Jennifer Doudna (2020 laureate in chemistry, for the discovery of CRISPR gene editing technology) wrote a 1989 PhD dissertation on programmable RNA enzymes for RNA cutting and splicing.

graduates in total. Using data on the near-universe of these graduates, we use our methods to establish new evidence on (i) the set of organizations which finance doctoral research, (ii) the set of graduates producing science related to "critical technologies" considered important to U.S. national defense and economic security, including where they train, and who funds their work, and (iii) the relationship of public investments in PhD training to total U.S. PhD production.

The foundation of our analysis is the ProQuest Dissertations & Theses Global (PQDT) dissertation library, from which we obtain dissertation text and metadata for STEM graduates (i.e., those in the life sciences, physical sciences, mathematical sciences, and engineering).[2] The PQDT sample tracks administrative benchmarks closely in both total graduates and graduates by university and field, suggesting that it approximates the target population. For each of the 1.2 million dissertations in this sample, we observe metadata (e.g., year, subject, school, title) and for a majority we also have access to the full text—which is available for much of the sample pre-2000, and nearly 100% of the sample post-2000. We then develop two LLM-based procedures to measure new features. One assesses each dissertation's association to critical technology areas identified by the White House Office of Science and Technology Policy (OSTP 2024a) as U.S. national priorities. The other parses dissertations to identify acknowledgments, extracts the names of organizations which supported the graduate's work (financially or otherwise), and links these entities to the Research Organization Registry (ROR), an independent, consolidated registry of global research organizations. We validate the resulting measures against manual and external benchmarks.

Using these data, we characterize the evolution of these features of the U.S. scientific training ecosystem over the past 75 years. About 50% of STEM graduates currently acknowledge some source of support in their dissertations, the majority (42%) from the U.S. government—a share which is at its highest level since 1980, but significantly below its >50% peak in the 1960s. We view these shares as a lower bound on true (underlying) rates of support, which are potentially underreported in dissertations—a possibility we return to later in this paper. Disaggregating this evidence, we document the largest sources of support in the government, industry, and non-profit sectors. The four largest sponsors of U.S. STEM PhD trainees in our data are the National Science Foundation (NSF, with >90,000 graduates supported since 2000), National Institutes of Health (NIH, 80,000), Department of Defense (DoD, 34,000), and Department of Energy (DOE, 30,000). In all, 12 different government agencies each support more graduates than the largest firm (Intel) and the largest non-profit (HHMI)—reinforcing that U.S. scientific training, and in turn the future

[2]ProQuest until recently had agreements with nearly all U.S. universities to acquire rights to republish their graduates' dissertations, which are now available in ProQuest's single, consolidated repository. In recent years, U.S. universities have begun transitioning to self-publishing, an industry shift which will require a more diffuse effort to collect dissertations from institutional repositories (e.g., to extend our analysis) in the future.

scientific workforce, is heavily underwritten by the U.S. government.

Continuing this evidence, we document universities and subjects with higher and lower rates of government and industry support. The relative balance further indicates which universities and fields are more exposed to changes in funding from each sector—with subjects such as astronomy and astrophysics being particularly government-reliant, pharmaceutical sciences and automotive engineering being relatively more industry-supported, and subjects such as geology and materials science being supported by both. In nearly all subjects—and at all U.S. universities—government support outpaces private sector support, but the rate of government and private sector support correlate with each other, likely reflecting either subjects' scientific potential or resource requirements, but potentially also indicating correlation in demand across sectors.

Using our classification of graduates to critical and emerging technology areas, we identify the top training grounds and sources of support in AI, quantum science, and other areas since 2000. We measure the size of the U.S. PhD pipeline in these areas and highlight institutions and organizations which are ostensibly important to sustaining U.S. leadership in critical technology-related science. In nearly every critical technology area, NSF is the largest recognized source of support, and the top five sources are nearly always government agencies. The institutions which rank highest in their production of critical technology-related PhD graduates are MIT, Stanford, and UC Berkeley, though many other institutions rank highly in specific specialties.

Leveraging variation in individual government agencies' funding priorities (e.g., NIH primarily funds biomedical and health science, and DoD engineering) and each agency's total number of graduates supported each year (a reflection of budgets), we estimate the effect of government support for PhD training on total PhD production, using a shift-share design. The elasticities we estimate imply that across the U.S. higher education system, PhD production varies roughly one-for-one with the number of government-supported graduates in each cohort—similar to patterns which prior research has found for specific programs and settings (e.g., Blume-Kohout and Adhikari 2016).[3] This evidence adds an additional datapoint to the wider literature estimating the impacts of public R&D, in this case reinforcing that public investment in scientific training is one of the main factors shaping the size and composition of the future scientific workforce.

In introducing new measures of U.S. PhD production, this paper adds to a long history of

[3]This result is broadly consistent with the results of Blume-Kohout and Adhikari (2016), who find a roughly 1-for-1 effect of increasing NIH fellowships and traineeships on total PhD enrollment, and with other work relating public funding for PhD training to enrollment (e.g., Freeman et al. 2009) and career outcomes (e.g., Graddy-Reed et al. 2021, Kim et al. 2022). We expand on this work with population-level measures of funding sources and aggregate PhD production across all fields and over longer horizons—providing a view of the overall system—but with the limitation that we cannot reliably measure specific funding mechanisms (e.g., fellowships vs. assistantships), whose impacts can vary (Blume-Kohout and Adhikari 2016, Kim et al. 2022).

advances in S&T statistics, where the measurement of scientific manpower has been a long-running priority (e.g., see National Science Foundation 1959, Freeman 1962, Freeman and Young 1959 for early discussions, and Godin 2002b, 2003 for a more extended history). As Christopher Freeman (founder of this journal, one of the major early influences on the development of national S&T statistics, and author of the original Frascati Manual; see OECD 1963a and Godin 2008) has written, "The first elementary step towards improving the rationality of this process [science policy], towards making these choices more conscious and more carefully considered, is the systematic collection of statistics on the deployment of scientific manpower, and on the expenditures on different branches of scientific activity," further noting that "These statistics must be collected in a great variety of breakdowns" (Freeman 1968, as quoted in Godin 2002a, pp. 12-13).

Although the NSF administers large-scale surveys towards this end like the Survey of Earned Doctorates (SED, an annual approximate-census of graduating PhDs) and the Survey of Doctorate Recipients (SDR, a periodic/biennial survey which samples past graduates), surveys are expensive to administer, have gaps in what they measure, and are difficult to link to other data sources. UMETRICS has more detailed measures of how research is financed, but is limited to universities which have opted into reporting, reports financial transactions only (not in-kind and other support), and suppresses personally-identifying information (like names) for privacy reasons, making linkages to external datasets challenging.[4] Against these alternatives, dissertation-based data have multiple strengths (and some weaknesses, which we will discuss momentarily). Outside of the U.S., large surveys and UMETRICS-like projects may also be too expensive or practically infeasible in many countries. Dissertation samples provide an alternative, complementary lens into doctoral training and the early career scientific workforce which can be applied to any national education system—as other scholars have recently begun to do (e.g., Martínez et al. 2025).

With respect to measurement, this paper explores the capabilities that recent advances in LLM-enabled natural language processing offer for extracting information from large, unstructured texts (like dissertations) to identify target features. This can be done for many more features than we considered in this paper, as well as a wider range of scientific texts.[5] With this paper, we are also publishing new data with information on PhD production by university, field, and university-field over time, including measures of (i) graduates associated with each of the critical technology areas specified by the OSTP (2024a) and (ii) graduates supported by specific federal agencies and/or with industry or philanthropic support. Beyond this paper, our methods and data are also extensible to

[4] Although UMETRICS has been linked to PQDT, it omits PQDT identifiers to preserve anonymity, such that we were unable to link our measures to the UMETRICS data for comparison.

[5] Research is already moving in this direction, given new opportunities to evaluate publications (Dagdelen et al. 2024), lab notes (Jalali et al. 2024), and even LaTeX markup (Yang et al. 2025).

other universities and national education systems, to other degree levels, and to future graduates: the main input required is a sample of theses or dissertations.

There are nevertheless limitations to this approach. One is that it requires complete or representative samples of written work from the population studied. Though PQDT appears to contain the near-universe of U.S. STEM PhD graduates through the mid-2010s, a growing number of universities since then have stopped automatically submitting dissertations to ProQuest and instead publish them in open access university repositories, in response to federal guidance (see OSTP 2013, 2022) and the broader open science movement. This attrition is modest in our sample period, and not systematically related to research funding, making it unlikely to affect our findings beyond a (modest) underaccounting. However, filling these gaps and extending analysis of the U.S. doctoral training system forward will likely require expanding data sources to more of these repositories—a possibility we are exploring in ongoing work.[6] In our particular application of identifying research funders, a second limitation of our approach is that similar to survey data, it measures only the sources of support that graduates self-reported—and our evidence suggests there is some under-reporting. As in similarly-impacted survey settings, however, this limitation can be accommodated with some reweighting on reporting rates, an approach we explore in our analysis.

Notwithstanding these limitations, we believe this paper makes a useful advance in both methods and findings. Despite long-running interest in critical technology assessment across many corners of government (Fuchs 2022), systematic measurement techniques are still developing. The methods we develop can be adapted and applied in other settings. The evidence may also have applied value: evidence on the state and organization of critical technology scientific training in the U.S. can inform academic and public understanding of U.S. science, and potentially influence research funding from all sectors by identifying pockets of strength and gaps in capability. Evidence on sources of funding offer a reference point for understanding the potential impacts of large changes in public funding for doctoral training—and at the time of writing (June 2025), large cuts to U.S. graduate student funding are already in progress (Mervis 2025, Bhatia et al. 2025).

We proceed as follows. Section 2 connects our work to previous research on doctoral training and early career scientists. Section 3 describes our data sources, describes measurement methodologies, summarizes the performance of our methods in validation tests, and discusses limitations. Section 4 uses these measures to introduce new facts about U.S. STEM PhD production, examining the level

[6]The contents of this paper point to the value of national dissertation repositories, which some concurrent research is beginning to use—especially research under the European Doc-Track program (see Martínez et al. 2025 for a project report, and Corsini et al. 2025 for an application). National dissertation repositories are more common in countries with national university systems, though the effect could be replicated in the U.S. with a national depository program similar to the Library of Congress' mandatory deposit requirements.

and composition of new trainees and the impacts of public investment in scientific training on PhD production. Section 5 describes the data we are releasing with this paper. In Section 6, we briefly review the paper and then suggest directions for future research.

2 Prior Research on the Scientific Training Ecosystem

Scientific manpower has been a central concern of innovation policy since World War II, which has particularly emphasized the importance of public investment in scientific human capital to advancing national defense, economic, and public health objectives (e.g., Bush 1945, Atkinson 1990). Doctoral training has accordingly been a focus of policy studies throughout this period, including a large and growing academic literature. Economically, knowledge workers in STEM fields are broadly recognized as an engine of innovation (e.g., Bianchi and Giorcelli 2020, Lubczyk and Moser 2025), and their finite population a limiting factor on economic growth (Romer 1990, Jones 2009, Akcigit et al. 2025). Consistent with this view, prior research has empirically linked the size of the STEM workforce to broad-based productivity gains, profits, wages, entrepreneurship, and other economic improvements (Winters 2014, Peri et al. 2015, Balsmeier et al. 2025).

Identifying the forces that shape the STEM scientific workforce is thus a first-order interest of both research and policy. A large share of work on this question has gravitated to immigration policy, on the observation that immigrants comprise a large, disproportionate, and disproportionately impactful share of U.S. scientists (e.g., Stephan and Levin 2001), inventors (e.g., Hunt and Gauthier-Loiselle 2010, Kerr and Lincoln 2010, Bernstein et al. 2022), and entrepreneurs (e.g., Azoulay et al. 2022). In the realm of science, prior research has examined not only immigrant scientist productivity (e.g., Kahn and MacGarvie 2016, Agarwal et al. 2023) and spillovers (e.g., Ganguli 2015), but also the arguably more elementary question of which foreign students come to the U.S. for graduate studies, who stays, and why (e.g., Gaulé 2014, Ganguli and Gaulé 2019, Bostwick et al. 2024). A common theme in this literature is that immigration policy is a key determinant of the foreign-born STEM PhD population (Kahn and MacGarvie 2020). Alongside the aforementioned evidence on the impacts of foreign-born scientists, this recognition has driven recurring calls for visa reform as one way to grow the PhD population (Glennon 2024, Nice 2025).

Despite the emphasis on immigrants, a majority of U.S.-trained STEM PhDs are U.S. born. An alternative path to growing the STEM workforce is to encourage homegrown entrants, and a separate body of research has examined the effectiveness of policies supporting entry into STEM fields, from childhood education (Dynarski et al. 2013) to curricular reforms (Goodman 2019). One of the most direct levers for shaping the scientific workforce is public funding (Bush 1945, Steelman 1947). As Myers (2020) observes, funding can shift the direction of scientists' work. Though existing

researchers' switching costs are large (Goolsbee 1998, Myers 2020), trainees and potential trainees may be more responsive—especially if younger workers are flexible and adaptive to changes in demand, as is often the case in other sectors. Consistent with this idea, Dugoua et al. (2025) have found that windfalls to the Department of Energy's R&D budget have historically increased PhD production in DOE-related fields. More generous funding environments in graduate school may also affect students' career progression (Graddy-Reed et al. 2021).

Reflecting long-running recognition that science is important to national outcomes, the NSF has been tasked with measuring it since the agency's inception. The NSF-administered Survey of Earned Doctorates specifically focuses on measuring U.S. PhD production. Despite its long history, many features of the U.S. scientific training system are not systematically known or measured—in part due to the intrinsic limitations of surveys, including their length and complexity. Key questions—such as (i) how many graduates does the country train in frontier, nationally-important areas; (ii) where does this training occur; (iii) who funds this training; and (iv) how many trainees leave the U.S., and with what consequences—remain difficult to systematically answer with existing data. Where some of these gaps have been partially filled (e.g., Chang et al. 2019 present statistics on the funding sources of PhD graduates at seven midwestern U.S. universities in 2011-2012, using UMETRICS data, as a demonstration case), data limitations continue to make it difficult to produce comprehensive national accountings, and the limited opportunities to systematically measure graduates' science have similarly made it difficult to evaluate its nature or impacts.

These are among the gaps we aim to address in this study and a concurrent study of postgraduate migration by U.S.-trained PhDs (Shvadron et al. 2025a). We also introduce new methods and data which others can use to continue exploring these and other themes. Related research loci which we do not address here include scientist career choice (e.g., Roach and Sauermann 2010, 2017, Sauermann and Roach 2012, 2016), how training environments affect career outcomes (e.g., Gaule and Piacentini 2018, Fry and Glennon 2025), and the degree of and consequences of diversity in STEM scientific career pipelines (e.g., Ceci et al. 2014, Buffington et al. 2016, Kahn and Ginther 2017, Bostwick and Weinberg 2022, Stansbury and Rodriguez 2024). The data we are releasing with this paper may be useful for exploring these questions with respect to emerging science-based technologies and the research funding ecosystem.

3 Measuring the Scientific Training Ecosystem

Our first goal in this paper is to build a near-population sample of U.S. STEM (Science, Technology, Engineering, and Mathematics) PhD dissertations since the mid-20th century. Using this sample, we then aim to measure characteristics of PhD graduates which are not elsewhere

observed or are hard to observe at the same scale, and for which comprehensive statistics therefore cannot otherwise be produced. We accordingly aim to do so in systematic ways which can span universities, fields, funders, and time.[7] Our emphasis is mostly on two features and phenomena: (i) measures of graduates' financial sponsors, and (ii) measures associating graduates' dissertation science to technology areas which according to the White House Office of Science and Technology Policy (OSTP) are important to defense and/or economic security.[8]

To produce this sample and these measures, we combine data from three sources: ProQuest (a commercial provider of global dissertation data), OpenAlex (an open-access bibliometric database; Priem et al. 2022), and universities' own institutional repositories. From these sources we compile a nearly-complete sample of U.S. STEM PhD graduates over the past 75 years, including dissertation text for many of these graduates pre-2000 and nearly all post-2000. To evaluate sample completeness and some dissertation-derived measures we additionally collect more aggregated data from the Survey of Earned Doctorates (SED) and Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS) as well as individual-level data from one large, nationwide fellowship program, the NSF's Graduate Research Fellowship Program (GRFP).

3.1 Base sample: PQDT

Our main data source is ProQuest Dissertations & Theses Global (PQDT). For nearly a century, ProQuest has been acquiring and re-publishing dissertations of U.S. PhD graduates via licensing agreements with universities. The ProQuest dissertation database is the largest collection of digitally catalogued PhD dissertations in the world, and it includes extensive metadata on each dissertation (e.g., author, institution, degree earned, title, abstract, subject). We identify 1.2 million STEM PhD dissertations in PQDT between 1950 and 2022 using degree letters, which we manually filter to PhD and PhD-equivalents, and graduates' self-reported subjects, which we manually crosswalk to SED

[7] In addition to complementing existing data sources on U.S. PhD graduates—including administrative data, such as NCSES surveys—these data sources and methods can likewise substitute for some administrative measures in other countries or contexts where data collection infrastructure is costly and less developed.

[8] The term "critical technology" is indicative of technology that a policymaking organization considers important to national or economic security. The terminology traces to the U.S. Export Administration Act of 1979, which required the Department of Defense (DoD) to provide a list of "militarily critical technologies" to the Department of Commerce to ensure they are export-restricted. This requirement was eliminated by later legislation, and the terminology has morphed to other applications. The OSTP list we use was developed in 2024 by the U.S. National Science and Technology Council, a committee which coordinates science and technology policy across the executive branch. In parallel, DoD's Office of Strategic Capital has also published a list of critical technology areas it considers important to national security, many of which overlap with the OSTP list. The CHIPS and Science Act of 2022 enacted by Congress similarly identifies key technology focus areas, which also intersect. One advantage of the OSTP guidance is that it represents cross-governmental policy. A second is that it includes more detailed descriptions which are useful for linking PhD dissertations (or other textual inputs) to technology areas.

major fields and filter to natural sciences and engineering.[9] Using IPEDS data, we further restrict to PhDs issued by institutions which ever held an R1 or R2 Carnegie classification post-2000, as well as PhDs from doctorate-granting medical and engineering schools.[10]

By comparison, the SED reports roughly 1.3 million STEM PhD graduates over this period.[11] Relative to the SED (and its underlying, de-identified microdata), a dissertation sample like PQDT presents two distinctive opportunities for analyzing doctoral training systems. First, we observe the dissertation science itself. Second, it provides enough identifying information (most importantly, names) to link graduates to their postgraduate careers. This enables us to do so for more graduates than other NSF instruments such as the SDR, and in more ways.

The SED nevertheless provides a useful benchmark for evaluating the completeness of the PQDT sample. Figure A.1 compares annual PQDT STEM dissertation counts to SED totals (defining STEM as above; see Appendix A for a more detailed accounting of the subsumed fields). The PQDT sample tracks the SED population closely until roughly 2010, after which a modest gap emerges—though PQDT still accounts for 90% of the SED population through 2022.[12] Appendix Figure A.2 disaggregates this comparison to individual fields and finds similar patterns at the field level. Together, these comparisons suggest our PQDT-based sample approximates the underlying population, notwithstanding modest undermeasurement since 2010.

[Figure 1 about here]

3.2 Dissertation text

Existing understanding of scientific training in the U.S. innovation system is largely shaped by the available data. Surveys like the SED are nearly complete in their sampling and collect a rich

[9] To link subjects to fields, we use the SED's own crosswalk between research subjects and major fields to the extent possible—which covers most of the PQDT sample—and judgment otherwise.

[10] These institutions together account for over 99.5% of the unrestricted sample.

[11] The SED is an annual survey of graduating PhD students at U.S. universities, and has been administered to the population of graduates in all fields (nearly 60,000 in 2022) since 1958. In practice it is a near-census of U.S. PhD graduates, with a >95% response rate through the mid-1990s and 90% today.

[12] Closer examination of the PQDT sample suggests this difference is mostly attributable to a small but growing number of universities ending their relationship with PQDT, such that their graduates' data are no longer automatically included. This change in turn reflects a movement of universities towards making research freely available, typically on university repositories in addition to or in place of commercial publishers. Much of this movement is in response to either state law or to federal guidance from OSTP in 2013 requiring they do so for federally-funded research within one year of publication (OSTP 2013). Updated guidance in 2022 eliminated this embargo (OSTP 2022). PQDT sample completeness may thus decline going forward. In ongoing work, we are developing a data collection and processing pipeline to collect dissertations from university repositories, one at a time, initially to backfill gaps in our PQDT sample since ca. 2010—especially with respect to a handful of important universities which exit the PQDT sample early, such as Georgia Tech (which does so in 2013)—and eventually with the goal of developing a sustainable data collection program that can substitute for PQDT in scholarly research.

set of information on graduates' background and immediate future plans, but offer limited insight into their science. PQDT's tabular data provides basic identifying information about doctoral graduates (e.g., name, institution, subject, graduation year) which can be used in more flexible ways (e.g., linking to other data sources) and is increasingly being used in research, but these data are likewise limited in what they measure. To our knowledge, no prior work has made systematic use of dissertation text, which—as we will show—presents opportunities to measure features of doctoral research, training, and trainees that are not elsewhere observable.

An important feature of PQDT for our purposes is thus that in addition to metadata, it also provides dissertations' full text. An added benefit of ProQuest having been the main provider of dissertation publishing services over the last century is that dissertations from hundreds of different universities generally share a common document structure, which enforces some consistency in reported information beyond science as well. For example, most dissertations have an "Acknowledgments" section (among others), where authors thank funders and other supporters of their research and training. These acknowledgments are a crucial resource for this paper, as they present an opportunity to systematically measure PhD graduates' sources of support.[13]

Dissertation text (beyond metadata) is available for most, but not all, of our PQDT sample: of the 1.17 million dissertations in our 1950-2022 sample, we have the text of 870,000 (75%) in total, and roughly 96% since 2000, which will be a focal period for much of our analysis. Where we use pre-2000 data, gaps are unlikely to be a problem for our analysis provided there are no systematic differences in the set for which text is available: as long as sampling is as good as random, the data can be used to produce representative estimates. Even with systematic differences in data availability, sample weights (e.g., inverse propensity score weights) can be constructed and used to rebalance the full text sample to resemble the population (on observables).

An important question is then what explains why some dissertations have full text, and others do not. Guidance from the data provider and evidence from the data vary a bit with respect to this question. According to ProQuest, this variation is a result of (i) differences in indexing methods—with some dissertations historically having been stored by ProQuest with full text, some with abstracts, and some with identifying data only—and (ii) idiosyncratic variation in what dissertations have been retrieved from microfilm. Figure 2(A) shows the share of PQDT dissertations over time in each indexing category, as indicated by ProQuest's dissertation-level identifiers (see Appendix A for details), which suggests that dissertation text may be available for 90%+ of its sample since

[13]Though nearly all dissertations include acknowledgments, there are nevertheless exceptions. When an acknowledgments section is missing, funders are often mentioned in the dissertations' biography section in footnotes to the text—cases which we can also identify by processing dissertations' complete text.

1965. In practice, however, we do not reach this level of completeness until ca. 2000. Figure 2(B) shows full text availability in the data, with the share of graduates each year for whom we have dissertation text, by field. Full text is available for around 40% of our sample until 1990, when it discretely jumps to 60-80%, and 1996, when it jumps to ≈100%.

[Figure 2 about here]

We explore one additional effort to grow the full text sample: for seven universities with the most PQDT graduates without dissertation text (MIT, UC Berkeley, UCLA, Cornell, U. of Minnesota, Michigan State U., and U. of Maryland), we visited their online repositories to try to retrieve it. Though Figure 2 indicates that most of these gaps are pre-2000, one exception was MIT, an important institution whose graduates were missing dissertation text throughout our sample. Using this approach we backfilled a large number of dissertation texts for MIT and Michigan State graduates but saw more limited success with other universities (Appendix Table A.1). We incorporate these results into our sample, but given the high cost of one-by-one, university-specific full text retrieval and mostly meager results—and that this mainly fills gaps in the pre-2000 era—we have not expanded this effort to the full sample at this time.

3.3 Linking PhD graduates to technologies

Using the data inputs described above, we aim to produce two new measures. One is to associate PhD graduates to critical technology areas, based on the relationship of their dissertation science to the above-noted 18 technology areas identified by OSTP (2024a) as militarily and/or economically important. Similar to Aiken et al. (2024), we develop an unsupervised LLM-based classifier which (i) summarizes dissertations, (ii) uses this summary to evaluate whether each dissertation is related to each of the OSTP technology areas, and (iii) filters these candidate associations based on detailed subfields within these technology areas which the OSTP guidance describes.

Roughly 43% of dissertations in our sample post-2000 are linked to OSTP critical technology areas by this method. The largest is *Biotechnology*, the fastest-growing is *Artificial Intelligence* (AI), and other large categories include *Advanced Materials* and *Semiconductors and Microelectronics* (see Appendix Figure A.6 for a full list). Though critical technology classification lacks a ground truth to test against—"critical technologies" do not have an obvious or naturally occurring taxonomy, nor a clear rubric for evaluation—the method nevertheless validates well against both intuition and independent external measures. The technology linkages our procedure produces are intuitive when read side-by-side with the dissertation titles, abstracts, and text. Using OpenAlex data, we

also show in Appendix A that PhD graduates we associate to particular areas are systematically concentrated in universities with the most publications in those areas.[14]

This approach is one among several methods of linking science or scientists to specific categories of technology. A common alternative is patent citations (Marx and Fuegi 2020, 2022), which can be flexibly used with patent classifications as a measure of technology space, though it is limited to science that is cited by patents. More similar to our method, Fry and Glennon (2025) study a sample of PhD graduates in AI based on keyword searches of PQDT dissertations from the 20 highest-ranked U.S. universities in AI-related fields. Our LLM-based classifier essentially systematizes this approach, and may be particularly useful for multidimensional classifications of science (e.g., linking science to many technologies), for classifying large corpora, and for classifying scientific outputs that do not get cited by patents—such as dissertations.

3.4 Measuring dissertation sponsors

The second feature we aim to measure is PhD graduates' research sponsors, including both financial and in-kind support (examples of which might include research inputs, equipment usage, API credits, etc.). Our motivation for doing so is that existing data sources are limited in completeness or resolution. Though the SED includes questions on how respondents financed their doctoral education, the survey has not provided response options to indicate specific funders since 1997. Even then, response options were limited to only a few federal agencies—and as we will show, doctoral training is supported by a much wider range of organizations. Another resource for measuring graduates' sources of support is UMETRICS (Lane et al. 2015), which provides administrative data on 45 universities' financial transactions, including funding from specific sources and transfers to specific employees. UMETRICS provides higher resolution information than SED but for a smaller number of universities and years, and with a changing sample as universities join or leave the program (as of April 2024, 27 universities were actively contributing data). Data privacy considerations also limit the degree to which the data can be linked to external sources. As a result, despite its specificity, UMETRICS is also limited in its suitability for making broad assessments of the U.S. higher education system and evaluations over long horizons.

Dissertation acknowledgments present a potential opportunity to do so over a larger sample of graduates and for a wider range of organizations than existing data sources permit. More complete data on how doctoral training is funded is of both research and policy interest (e.g., Blume-Kohout and Adhikari 2016). For this reason, we develop a methodology to extract this information from

[14] We demonstrate this for quantum science in Appendix Figure A.8, plotting universities' number of quantum science PhD graduates against quantum science publications—whose correlation is 0.9.

dissertations. This approach may also have drawbacks relative to administrative data such as UMETRICS, including incomplete reporting rates, which we will explore below. On net, however, the evidence suggests that despite some imperfections, the data and methodology we develop towards this end yields a broader, more detailed, and more expansive record of how doctoral education is financed in the U.S. than is currently otherwise available.

Many readers of this paper may recognize dissertation acknowledgments from their own thesis. To distill the basic idea behind our measurement, Figure 3 provides an example acknowledgment from the dissertation of Ian Buck, a 2005 Stanford University computer science graduate (now a senior executive at NVIDIA) who in his PhD created the precursor to NVIDIA's CUDA platform, a parallel computing architecture that allows GPUs to be used for general computing and is essential to modern AI. Figure 3 identifies several organizations which funded Buck's work, including DARPA, the Department of Energy (DOE), the Advanced Simulation and Computing program (ASC, a supercomputing facility that is also under DOE), NVIDIA, and ATI Technologies (a former semiconductor company specializing in GPUs, later acquired by AMD). Our task is to identify these funders, consolidate them to their ultimate parents (e.g., grouping ASC up into DOE), and classify them into sectors (government, industry, and non-profits/foundations).

[Figure 3 about here]

To flexibly identify research sponsors across our dissertation corpus, we develop a pipeline that leverages natural language processing tools and large language models (LLMs).[15] We process dissertations one at a time, working initially with the entire text. Our procedure for each dissertation takes the following steps. First, we partition the dissertation into sentences and identify "acknowledgment sentences". To isolate potential acknowledgment sentences, we applied a rule-based string matching approach by splitting the text into sentences (using NLTK's Punkt Sentence Tokenizer; Bird et al. 2009) and identifying sentences with tokens such as "fund", "grant", "thank", and so on. This procedure tags roughly 103 million potential acknowledgment sentences.

We next identify named entities within these sentences, classify entities into sectors, and identify the type of support and grant identifiers (where provided), using the recent open-source LLMs Solar 10.7B and Smaug 34B (Kim et al. 2023, Pal et al. 2024). We established a procedure that uses the LLM to verify that a sentence acknowledges support from an external entity, identify the supporting entities, categorize them by organization and support type, and record grant and contract identifiers

[15] A more detailed description of our methodology, including the LLM prompts we use, and our approach to consolidation and classification, is provided in Appendix 3. Here we present an abridged summary.

where applicable. We used the VLLM Python package (Kwon et al. 2023) and the LM format enforcer (Gat 2023) for efficient batch inference and to ensure the model produces a well-structured JSON schema. The LLM output indicates that 11 million of our potential acknowledgment sentences (12%) indeed acknowledged support or were part of a biography section. Among all sentences, the LLM extracted 9.3 million mentions of organizations (4 million unique), with 4.8 million of these mentions originating from sentences classified as support or biography.

We then consolidate named entities (accounting for name and spelling variants by grouping them together) and link them to external firm registries. We consolidate extracted entities using the Smaug 34B model and match them with the Research Organization Registry (ROR) and Wikidata to disambiguate institution names and obtain external identifiers.

3.4.1 Example application

Table 1 demonstrates the performance of our procedure on the acknowledgment text in Figure 3. Each row is an extracted funder, and the columns report the name, sector, type of support, and the associated ROR organization (where matched to ROR; otherwise blank). Comparing the text to the first three columns of this table, we find that our procedure correctly extracts nearly all entities by name, and those it identifies are classified into the correct sectors and type of support. The one exception is ASC, which our procedure has difficulty classifying due to ambiguity of the acronym and its relative infrequency as a research funding organization—one which we ourselves were only able to define with further research and contextual understanding.

[Table 1 about here]

3.4.2 Validation

We validate this approach in multiple ways. The first is by comparing the extracted entities to a manually-processed random sample of 500 dissertations, for which we meticulously identified funders by hand. Figure 4 compares the share of this sample with government, non-profit, and industry support according to our manual and automated measurement (in red and teal, respectively). The results show a substantial degree of similarity, with 31-32% of this sample classified as reporting government support under both methods, 11-12% non-profit support, and 8-9% industry support—suggesting the LLM classifier approximates manual extraction.

[Figure 4 about here]

We undertake two additional validation exercises. In Appendix A.5.1, we collect data on all NSF Graduate Research Fellowships (GRFs) awarded since the beginning of the program, link awardees to PQDT, and evaluate how many of these graduates acknowledge NSF support in their dissertations. Based on a simple textual search, we find that among the matched NSF graduate fellows, 60.8% mention the NSF in their dissertation. Comparing these mentions to our LLM-based extraction, we find that our procedure identified and linked 98.6% of these mentions.[16] These results provide further validation for our approach while also highlighting some of the limitations we face, particularly with respect to underreporting—we return to this issue momentarily.

Finally, in Appendix A.5.2, we compare our sample to the NSF's Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). Under the GSS, university administrations annually report to NCSES the number of enrolled graduate (master's and doctoral) students in each major field with different sources of funding, including broad categories like institutional support, student loans, or self-support and particular federal agencies like HHS and NSF. Though the GSS sample (which consists of all enrolled students rather than only graduates, and includes master's students) is broader than the sample we collect in this paper, we can nevertheless test whether the data are broadly consistent across the two sources. We correlate across the two sources the share of federally-supported students with support from specific agencies (e.g., DoD, DOE, HHS, NASA, NSF, USDA), at the university-field-year level, between 2000 and 2022. Appendix Table A.11 shows that the two sources correlate closely, with the two sources correlating one-for-one for HHS- and NSF-funded shares, and near one-for-one for other federal agencies.

3.4.3 Limitations and opportunities for improvement

Collectively, this evidence suggests to us that our procedure can extract acknowledged research sponsors with a high degree of precision and completeness.[17] The resulting data nevertheless have limitations. One of these limitations is a result of our reliance on PQDT: as previously noted, despite being the most comprehensive source of information on U.S. PhD graduates over the last century, the set of universities submitting dissertations to PQDT appears to recently have begun shrinking. In ongoing work we are exploring ways to address this gap, including by collecting dissertations and metadata from individual universities' institutional repositories.

A second limitation more endemic to our approach is the risk of underreporting: graduates may forget or otherwise neglect to acknowledge all sources of support in their dissertations (much as graduates may do the same in their SED responses, or journal publication authors might in research

[16] The remaining 1.4% of mentions that were not picked up by our LLM-based procedure were due to unconventional wording in acknowledgments or mentions of the NSF without an indication of funding.

[17] Note that this process can also be applied to other scientific corpora as well.

articles, etc.). The comparison of our measures to official NSF GRFP award lists (Appendix A.5.1) suggests such underreporting can potentially even be substantial.

There are two strategies we consider to address underreporting. One is to simply report the data as they are—an approach which will generally yield conservative measures (i.e., underestimates) of the share of graduates supported by any given agency or organization. The second is to use our validation data to estimate the likelihood of accurate reporting as a function of observables (e.g., by university, field, and/or year), and to weight our analysis with inverse propensity weights (IPW), overweighting low propensity populations to obtain representative results—similar to how survey weights are used to attempt to make non-representative survey results reflect the underlying population, or how research using nonrepresentative linked samples reweights on link propensities (Bailey et al. 2020). Across the paper, we opt for reporting the raw data to prioritize transparency and conservatism, and present IPW-weighted results in Appendix B.[18] Our results should thus be treated as a lower bound on underlying (true) rates of support.

Despite these limitations, one final point of comparison is available from Chang et al. (2019), who link SED graduates from seven Midwestern universities between 2011-2012 to UMETRICS data in order to evaluate the share of SED graduates at these universities who were paid with federal funding in the final two years of their PhD—including through research assistantships, teaching assistantships, and fellowships. Much of Chang et al. (2019)'s analysis presents data for all graduates (rather than STEM graduates), but in passing reports that "over three in five students in physical sciences, agriculture/biology, and computer/mathematical sciences are on federal grants in the two years prior to getting their degree" (Chang et al. 2019, p. 1490). By comparison, our IPW-weighted measures indicate 60% of STEM graduates from these university-years report federal support in their dissertations (65% of graduates in physical sciences, 62% in life sciences, and 41% in math and computer science)—consistent with the UMETRICS data.

4 New Facts about U.S. STEM PhD production

4.1 Who funds PhD training in the United States?

Figure 5, Panel (A) shows the share of graduates supported by different organization types from 1950 to 2022. In 2022, approximately 50% of graduates in our sample acknowledged some source of support. U.S. federal agencies are the most common funders, supporting over 40% of graduates.

[18] In exploring the data descriptively, we have found that within our ground truth sample, variation in reporting cannot be explained by cross-institution differences (which might have been generated by, e.g., institutional dissertation templates or guidance), as most institutions have similarly middling reporting rates. Major field turns out to have the most explanatory power for reporting, and we base our estimated propensity off of these.

Roughly 15% of graduates are supported by non-profit organizations, and 10% by private firms. These patterns have changed substantially over time: government support increased rapidly in the 1950s and 1960s, peaked around 1968, retracted in the 1970s, and stabilized (at ≈40%) in 1980. It has since modestly grown (2000 onwards). Non-profit and firm support has also fluctuated: the share of graduates supported by firms peaked at 20% around 1960, whereas nonprofit support grew from 10% to 15% between 1980 and 2000 and has since stabilized.

Panels (B1) to (B4) disaggregate these patterns by broad field. Disaggregation reveals that government support was historically especially high in the physical sciences and engineering, but has since largely equalized across most fields. Non-profit support is somewhat higher in the life sciences, and industry support in engineering, though these patterns have also fluctuated: firms, for example, supported 30% of graduates in the physical sciences in 1960—which follow-up analysis indicates was led by Du Pont in chemistry—but under 10% today; conversely, non-profits supported 10% of life sciences graduates in 1980—but more than 20% today.

[Figure 5 about here]

Figure 5 is almost certainly a lower bound on true rates of support, which are undermeasured due to underreporting. In Appendix Figure B.5 we show IPW-weighted shares (i.e., attempting to adjust for this underreporting), which we treat as an upper bound, and which indicate that in 2022, as many as 67% of graduates may have had U.S. government support, 22% non-profit support, and 16% industry support. Using the raw and IPW-adjusted shares, we plot confidence bands, along with the midpoint. Though IPW-weighted values are larger than raw values, the variation across domains and over time remains similar. The appendix also evaluates other dimensions of heterogeneity, including by country of origin, which we observe for a subset of our sample (inferred from pre-doctoral institutions, which are reported in many dissertations, especially post-2000); in Appendix Tables B.2 and B.3, we show that a substantial share of both U.S. and foreign-origin students have U.S. government support, though substantially more US-origin students do—reflecting their eligibility for a wider range of federal support mechanisms.

Table 2 lists the top 15 sources of external support for U.S. PhD graduates, by sector (government, industry, non-profit). The National Science Foundation (NSF), National Institutes of Health (NIH), Department of Defense (DoD), and Department of Energy (DOE) rank high on this list, as do the National Aeronautics and Space Administration (NASA) and Department of Agriculture (USDA). Top non-profit funders include academic societies and philanthropic foundations. The top firms tend to be large, research-intensive multinational corporations. Intel and IBM top this list, but the most

represented industries are the chemical, oil & gas, and pharmaceutical industries—indicating that many major corporations in high-tech manufacturing industries invest in developing highly-trained researchers who might augment the stock of relevant knowledge or might contribute directly to their future R&D efforts as future employee researchers.

[Table 2 about here]

4.1.1 Heterogeneity across subjects and universities

Figure 5 points to heterogeneity in the level of support broad subjects receive from different categories of funders. Figure 6 explores these differences in finer detail, plotting the share of PhD graduates between 2000 and 2022 in specific subjects with government and industry support. We associate graduates to subjects based on their own self-reported subjects in the PQDT data (see Appendix A for further discussion). Figure 6 splits the sample into very large subjects (with >1200 graduates in this period; Panel A) and large subjects (400-1200 graduates; Panel B) to improve readability, omitting smaller subjects. The figure provides direct insight into which areas of science are predominantly funded by either firms or the federal government, and which receive balanced support. Scientific training in automotive engineering and pharmaceutical science, for example, is more heavily industry-supported than government supported. Astronomy and astrophysics, on the other hand, are heavily government-supported and receive essentially no private support. Geology and materials science receive a mix. In nearly all subjects, however, the federal government provides much more support (5 to 10x) of doctoral training than industry.

[Figure 6 about here]

Figure 7 provides similar analysis across universities, plotting the share of STEM PhD graduates between 2000 and 2022 in each university with government and industry support. We similarly split the sample into very large universities (with >2000 graduates in this period; Panel A) and large universities (1000-2000 graduates; Panel B). The figure shows a clear positive correlation in rates of government and industry support—suggesting that the doctoral funding system concentrates among the same institutions and students. Technical universities like MIT, Carnegie Mellon, Georgia Tech, and Virginia Tech tend to have higher rates of industry support, but as before, the federal government remains (by far) the larger sponsor of PhD training.

[Figure 7 about here]

In Table 3 we more systematically examine this visual correlation in government, industry, and non-profit rates of support. We produce a battery of regressions relating the government-supported share of graduates and the industry- and nonprofit-supported share at different levels of aggregation and with different assortments of fixed effects. Concretely, we estimate:

$$PctGovernment_{it} = \beta_1 \cdot PctIndustry_{it} + \beta_2 \cdot PctNonprofit_{it} + \alpha_i + \delta_t + \varepsilon_{it} \quad (1)$$

and variants thereof, with i indexing universities or fields (or ij indexing university-fields) and t indexing years, and the remaining parameters representing fixed effects. We find large and relatively precise correlations in support rates across sectors, with magnitudes that reflect the relative size of each sector as seen in Figure 5 above. The evidence suggests that the PhD education funding system in the U.S. is more concentrated than diversified, with sponsors from all sectors focusing on the same universities, fields, and ultimately students.

[Table 3 about here]

4.1.2 Specialization and variation over time at federal agencies

We next turn attention to specific sources of support—especially federal government agencies, focusing on the four largest government supporters of PhD training: NSF, HHS, DoD, and DOE. Figure 8 characterizes the share of graduates in each of 17 major fields funded by these agencies. NSF provides relatively even support across subjects (except for health sciences, reflecting NSF's prioritization of funding basic science) and is an especially large supporter of students in geological (i.e., earth, atmospheric, and ocean) sciences. Other agencies have a clear mission focus. HHS, for example, supports a large fraction of trainees in biological and health science and biomedical engineering, but few in other fields. DoD is the country's largest funder of doctoral training in aerospace engineering and a large supporter of electrical engineering and computer science. DOE emphasizes physics, chemical engineering, and materials science. Though not shown, similar patterns are present for other agencies—for example, USDA is a major sponsor of agricultural sciences; DOT, of civil engineering; and NASA, like DoD, of aerospace engineering.

[Figure 8 about here]

Figure 9 shows how this support has varied over time. The common pattern across most agencies is that their support for PhD training peaked in the late 1960s and has only recently recovered, if at all. Roughly 20% of STEM PhD graduates currently acknowledge support from NSF, 15%

from HHS, and 12% from DoD and DOE collectively (roughly 5-7% individually)—shares which we consider conservative (lower bounds, given the potential for underreporting, as discussed above), but which capture the relative rate of support across agencies.

[Figure 9 about here]

4.2 PhD production in critical technology areas

An additional important dimension of the U.S. STEM scientific training ecosystem is the cultivation of scientific expertise in critical technology areas. Using the measures introduced in Section 3 we tabulate the number of graduates in our sample associated with each of the 18 OSTP (2024a) critical and emerging technology areas by university and source of support between 2000 and 2022. Table 4 lists the top five universities and sources of support for graduates in each of the 12 largest technology areas (excluding AI, which we will examine in more detail momentarily).[19] Despite the ostensible importance of critical technology science to U.S. science policy and economic policy, to our knowledge, for most of these technologies the U.S. doctoral education system has not been systematically and nationally assessed to understand how many such scientists the U.S. trains, where they are trained, and who pays for their training.

[Table 4 about here]

The table offers many insights. First, we find that several universities rank highly in multiple technologies areas, including MIT (in the top 5 for 10 of 12 areas in the table), Stanford (9), and UC Berkeley (7)—highlighting that these universities play an outsize role in the U.S. innovation ecosystem. Maryland, Michigan, Purdue, and UCLA also rank highly for 4+ technology areas. Second, we observe that some universities have strength in specific subjects, some of which are a product of decades of public investment. MIT, Northwestern, NC State, Penn State, and UIUC are the top universities for training PhDs in advanced materials—and all were among the 12 institutions which were selected by DARPA to establish materials science laboratories in the 1960s, when the field was effectively born. Third, with little exception, U.S. federal agencies are the top sources of support for trainees in all of these areas. For all of these facts, it is important to keep in mind that a few universities which are important training grounds for scientists in these areas (e.g., Georgia Tech) may be omitted from the list due to their exiting the PQDT sample early (as seen in Appendix Table A.12). Appendix Table B.5 reproduces Table 4 with data through 2012 only to explore what

[19] Technology areas omitted from this table are: human-machine interfaces, positioning, navigation, and timing (PNT) technologies, advanced gas turbine engines, directed energy, and hypersonics.

some of these may be, but the comparison also elucidates how the relative importance of each university to these technology areas may have changed over time.

Table 5 provides a similar analysis for artificial intelligence specifically, expanding the list of reported universities and research sponsors. At the top of the list of universities training researchers in AI are many expected institutions (e.g., Stanford, MIT, and UC Berkeley), which further reinforces the face validity of the method. Also in the top 15 schools, however—and especially below the top 5—are many public universities, which collectively train more PhDs in AI than the top 5 schools. As before, a small number of schools which might have otherwise made this list (e.g., Georgia Tech) are likely absent because they exited the PQDT sample early—an intuition supported by Appendix Table B.6, which reproduces this analysis through 2012 (despite being prior to the recent AI renaissance) and finds Georgia Tech ranked in the top three.

[Table 5 about here]

The right half of Table 5 shows that the top six sources of support for AI PhD trainees are government agencies, with NSF and DoD atop the list. The next-largest source of support is large U.S. technology companies such as Google, Microsoft, IBM, Intel, and Nvidia. The differences in magnitude between government and private support, however, are very large: the NSF supports 10x as many graduates as Google, and 20x as many as Intel.

In Figure 10, we aggregate this information up to the technology level, documenting the share of graduates in each technology area reporting government support against the total number of graduates in that area between 2000 and 2022. We color code each technology by the principal agency providing support. Figure 10 shows that there are essentially three clusters of PhD graduates with respect to critical technologies: the defense cluster, relatively low in count but heavily DoD-supported; the civilian mission agency cluster (space, energy, and biotech); and the NSF cluster, which spans a range of subjects but is concentrated around advanced electronics, communications, and computing. NSF is the leading patron of two-thirds of the areas in the OSTP list, and two-thirds of the graduates we associate to any critical technology area.

[Table 10 about here]

4.3 Effect of government funding on PhD production

The very large role that federal funding has in supporting scientific training in STEM, including in critical technology areas—together with historical fluctuations in the level of public support,

including recent cuts (at the time of writing; Bhatia et al. 2025)—naturally raises the question of what happens to PhD production when federal research funding programs expand or contract. Does total PhD production change with it? If so how much?

To evaluate these questions, we use our data to count PhD graduates at the field level, and relate annual PhD production to the level of federal support. We estimate the following relationship, where i indexes fields and t indexes years; $PhDGraduates_{it}$ measures total graduates in field-year it; $SupportedGraduates_{it}$ measures observed graduates in field-year it with federal support; X_{it} represents a vector of controls; and α_i and δ_t are fixed effects, with an estimation sample spanning all 17 major fields in our data, and 1970 to 2022 (53 years):

$$\ln PhdGraduates_{it} = \beta \cdot \ln SupportedGraduates_{it} + X_{it}\phi + \alpha_i + \delta_t + \varepsilon_{it} \tag{2}$$

Our goal in estimating Equation (2) is to evaluate the relationship between public investment in PhD training and the size of the STEM PhD workforce over the past 50 years. Answering this question using this specification presents two challenges. The first is the possibility of a mechanical relationship in a specification where the outcome is total graduates and the explanatory variable is a category of graduates. The second is that this relationship may be confounded by unobserved factors. For example, PhD funding may target fields with high scientific potential, which attracts general enrollment—which would lead to overstating the effect of federal funding on PhD production. Conversely, federal agencies funding might seek to pull more students into fields that are important to their missions when they are undersubscribed, in which case the estimated relationship between federal funding and PhD production might be understated.

To confront both challenges, we use a shift-share design to construct an instrument for supported graduates (e.g., Goldsmith-Pinkham et al. 2020, Borusyak et al. 2022), exploiting the agencies' annual number of supported graduates (the shift) and each field's share of agency graduates in a base period (the share). This design exploits the variation in Figures 8 and 9—the combination of which creates variation in federal support in each field over time. For example, changes in total DoD and NIH research funding will (ostensibly) have different effects on aerospace engineering versus biology. The instrument will both break the definitional relationship between the LHS and RHS of Equation (2) and introduce arguably independent variation.

Concretely, we (i) calculate field i's share of supported graduates from each funding agency j in a base period t_0 (which we will define in two different ways); (ii) interact it with the log number of funder j's supported graduates in t, and (iii) take the sum across funders $j \in J$ to produce field i's

predicted log number of supported graduates in each year t:

$$\ln \widehat{SupportedGraduates}_{it} = \sum_{j \in J} \omega_{ij} \cdot \ln Graduates_{jt} \tag{3}$$

where ω_{ij} is the share of graduates in field i funded by agency j in the base period and $\ln Graduates_{jt}$ is the log number of graduates funded by agency j in year t, excluding those in field i (i.e., a leave-one-out shifter). We take two approaches to defining the base period for computing ω_{ij}: we calculate shares for agency j in year t by averaging over (i) the previous ten years, and (ii) six to ten years prior, preceding enrollment of the given cohort. For this exercise we also condense our set of funding agencies J to the six largest funding agencies and an "other agencies" category, and to ensure we do not double-count graduates with multiple federal sponsors we reallocate graduates with multiple agencies to a 'multiple agencies' category.[20]

Using this instrument, we estimate Equation (2) by two-stage least squares (2SLS). Our main results are presented in Table 6, where the instrument is computed from prior decade average shares (results for 6 to 10 year-ago shares are both quantitatively and statistically similar; see Appendix B). In Column (1) we relate total PhD production to the number of government supported graduates, conditioning on the number of other graduates. Holding the latter fixed, the results suggest that a 10% increase in government-supported graduates increases total PhD production by approximately 4%. Given that in our data, around 40% of graduates reported government support throughout our sample period (Figure 5), this result implies that PhD production varies one-for-one with public investment in it: funding one more student yields one more graduate.

[Table 6 about here]

Though this result is not definitional, it may be nearly mechanical, as conditional on unsupported graduates, an additional supported graduate must yield an additional graduate. Columns (2) and (3) relax this conditioning to consider crowd-in (or crowd-out), examining whether the marginal government-supported graduate increases or decreases the number of other graduates. Across both columns, we find evidence of potential crowd-in: a 10% increase in government-supported graduates increases total PhD production by 7.5% (Column 2), and other PhDs by roughly 6% (Column 3). Given the roughly 40-60 split in the base rates in Figure 5, both results imply that on a base of 100 PhDs, funding another 4 graduates (an increase of 10% on the initial 40) will also yield another $\approx$3.5 non-funded graduates (an increase of 6% on the initial 60; reflected in Column 3), for a total

[20] The final list of federal agencies for this exercises is DoD, DOE, HHS, NASA, NSF, and USDA, to which we add categories for all other agencies and for graduates supported by multiple agencies.

increase of 7.5 (7.5% on the initial 100; reflected in Column 2). However, this interpretation (of crowd-in) should be tempered by the caveat that it may be a consequence of underreporting: for example, if the true rate of government support in the population were 65% (a possibility suggested by our IPW-weighted results in Appendix Figure B.5), the results in Table 6 would imply that a 10% increase in government-supported graduates (6.5 on an implicit base of 65) total would increase total graduates by a statistically similar number (7.5 on a base of 100).[21]

Column (4) estimates a variant of our 2SLS specification at a finer unit of analysis—the university x field x year level—using similar variation and a similar strategy, but now with university-field fixed effects and standard errors clustered at the university level. This estimation is arguably more conservative, essentially evaluating the relationship between supported graduates and total graduates at the individual program level (assuming university-fields approximate academic departments), but finds quantitatively similar point estimates to those in Column (1).

In Columns (5) and (6) we adapt this exercise to examine the relationship of PhD production to scientific output. To do so, we count total publications in a given major field and year in OpenAlex, measuring research articles only (omitting reviews, letters, etc.) with at least one U.S.-based author, using CrossRef to correct occasional errors in publication dates, and associating publications to major fields based on each publication's OpenAlex-reported subfield, which we crosswalk to SED major fields. We then regress the log number of publications in field-year it on the log number of graduates trained in that field over the prior 20 years (Column 4) and the log number of government-supported graduates trained over the prior 20 years (Column 5). In both cases, we apply a variant of our instrument adapted to predict 20-year graduate counts (as the sum of annual instrumental counts), and given mechanical serial correlation in the independent variable, we cluster standard errors by major field, applying a wild bootstrap due to the small number of clusters (MacKinnon and Webb 2018)—though bootstrapped standard errors are similar to analytical ones, despite the low number of clusters. Columns (4) and (5) indicate that every 10% increase in the stock of recent (past 20 year) PhD graduates increases the flow of scientific research articles by 2.5%. That this increase is less than 1-for-1 could reflect either decreasing marginal returns to growing the PhD stock (e.g., if marginal graduates are less productive) or that the marginal PhD graduate is more likely to enter technical careers that do not specifically generate publishable science.[22]

[21] A second potential caveat is the residual risk that our instrument is not fully independent but rather correlates with unobserved factors that increase enrollment—especially the possibility that particular agencies, with particular priorities, may experience budget fluctuations based on the latent potential of the fields they support—though our leave-one-out instrument should largely accommodate this concern, as it bases instrument values off of shifts in agencies' level of support in *other* fields, omitting the given field.

[22] Such graduates may instead apply their science training in industrial innovation, business, or policy—activities which may still require or otherwise benefit from PhD-level training.

Comparisons to existing literature

The evidence in Table 6 is broadly consistent with prior research, which has found that fellowship applications and PhD enrollment are responsive to the availability of funding (e.g., Freeman et al. 2009, Blume-Kohout and Adhikari 2016). The closest prior analysis is Blume-Kohout and Adhikari, who relate variation in NIH-supported PhD students to total PhD enrollments in biomedical sciences between 1998 and 2010—a period which covers the early 2000s doubling of the NIH—using NSF GSS data. Blume-Kohout and Adhikari find that "NIH-funded traineeships and fellowships increase full-time graduate enrollment by essentially 1:1" (Blume-Kohout and Adhikari 2016, p. 1297), while finding still-large but more attenuated effects of NIH-funded research assistantships (at a ratio of 0.5 enrollees for every additional funded research assistantship).

Amidst this evidence, we view the results in Table 6 as adding an additional datapoint on the relationship of public funding to national PhD production (and the impacts of public R&D more broadly), particularly with a more systemic perspective based on population-level data, multiple funders, and long panels—notwithstanding the measurement limitations previously discussed. The results are broadly consistent, even as differences in samples, measures, variation, and specifications could potentially lead to some differences in point estimates.

It is also useful to recognize that the evidence of Blume-Kohout and Adhikari and others (e.g., Kim et al. 2022) suggests that these elasticities can vary for specific funding mechanisms. Compared to this work, our measures are indicative of overall levels, but because our methodology cannot reliably distinguish the mechanisms(s) through which students were funded or otherwise supported, we are unable to speak to this heterogeneity. Recent research has also found that PhD students' sources of funding can affect not only enrollments but also completion rates, academic placements, future publications, and academic networks (Graddy-Reed et al. 2021). For reasons of both the goals of this paper and data limitations, our analysis abstracts from this heterogeneity and these multiple dimensions of impact, instead emphasizing the total size and breadth of the graduate population as a broad indicator of evolving scientific capacity (OECD 2015).

5 Distribution Dataset

Using the above-described data, we compile three public-use datasets measuring PhD production by university, field, and university-field over time. In all three cases we limit the sample to universities which have ever held an R1 (very high research activity) or R2 (high research activity) classification post-2000, or which are dedicated medical schools (e.g., SUNY Downstate or UT Southwestern). For each dataset, we report the total number of PhD graduates in the natural sciences and engineering;

the number for which we have full text and for which we have detected acknowledgments (for normalization); the number we identify as supported by the U.S. government, by industry, or by non-profit organizations; and the number supported by specific federal agencies (DoD, DOE, HHS, etc.). For the university-level dataset we further report the number of graduates we associate to OSTP critical technology areas. Data and accompanying documentation have been posted to the Harvard Dataverse and are available for public use (Shvadron et al. 2025b).

The distribution data include information for 326 unique institutions between 1950 and 2019—though the sample is unbalanced, with roughly 100 institutions in the 1950s and nearly 300 by the end of the sample. The field and university-field datasets report graduates for the 17 focal SED major fields described earlier, spanning the life sciences, physical sciences, and engineering. Users of these data who wish to make use of the counts of students with particular sources of funding may want to normalize these counts by either (i) the number of graduates in the given university-year, field-year, or university-field-year for which we have dissertation text (variable `has_ft`), or (ii) the number for which we have identified acknowledgments (`has_ack`).

6 Conclusion

In this paper, we use information from the dissertations of nearly 1.2 million U.S. STEM PhD graduates between 1950 and 2022 to examine the U.S. scientific training ecosystem. Leveraging recent advances in LLMs, we (i) link each dissertation to 18 critical and emerging technology areas recently identified as national priorities by the White House OSTP and (ii) process dissertation acknowledgments to identify research sponsors. With these data we show that about half of recent graduates acknowledge external support, with the U.S. federal government supporting about 42% of graduates—far eclipsing industry (≈10%) and non-profits (≈15%). NSF and NIH alone individually account for more supported PhDs than the entire commercial sector, and their support is highly concentrated in universities that also dominate critical-technology training (e.g., MIT, Stanford, UC Berkeley, among others). Across the 18 OSTP domains, federal agencies are typically the top funders, though a range of universities train students in these fields.

Exploiting agency-specific variation (in both the primary fields supported and total number of graduates supported over time) in a shift-share design, we estimate an elasticity between government support and PhD production which implies that supporting more graduates increases total PhDs roughly one-for-one, underscoring public investment as a driver of expansions and contractions in the scientific workforce. These data, methods, and findings together provide a new, scientifically- and policy-relevant map of the U.S. scientific training ecosystem and a foundation for tracking how

funding choices influence national science and innovation capacity. The approach we develop can also be extended forward and to other national innovation systems.

Whether more or less public investment (and accordingly, more or less PhD trainees) is desirable is not a question this paper specifically speaks to: our goal is to document features and effects. In particular, as in other markets, public subsidies may distort market outcomes. Nelson (1959), Arrow (1962), and others have argued that public investment in R&D can correct inefficiencies more than it introduces them, and policy advisors as far back as Bush (1945) have made the case for funding scientific training. Yet a perennial question and point of concern is whether the U.S. innovation system trains too many PhD scientists. Critics point to the limited number of tenure-track faculty jobs relative to PhD graduates, large number of postdocs, length of postdocs, and the growth of contingent faculty as evidence of oversaturation (e.g., Cyranoski et al. 2011, Powell 2015), especially in light of the prevailing wages these workers accept (Stephan 2013).

On the other hand, the size of the scientific workforce is typically seen as the rate-limiting factor for scientific progress, and in turn for improvements in economic growth and material well-being (Steelman 1947, Atkinson 1990, Romer 1990, Jones 2009). Those who advocate growing the scientific workforce point out that many PhDs enter industry, where they can be important contributors to private sector innovation (e.g., OSTP 2024b). Firms are generally considered unlikely to invest in transferable skills like scientific training, since those skills are broadly marketable (Becker 1964). This intuition may explain why industry funds relatively few PhDs (Table 2)—and where it does, it often only supports science with specific private value (as our example in Figure 3 demonstrates). Given this logic, and evidence that the returns to public R&D (much of which funds training) are often found to be very high (Jones and Summers 2022, Fieldhouse and Mertens 2024), it is easy to argue for the merits of the case for growing the scientific workforce.

We do not stake a claim on the issue, but rather observe that the answer may hinge on which scientists we have in mind—and the data we produce can be helpful to assessing seemingly important categories of trainees. Moreover, our evidence reinforces that public funding plays a central role in determining their supply (Table 6; also cf. Dugoua et al. 2025), and that the data we introduce can be used to further adjudicate this debate—including the possibility that whether public subsidy corrects or distorts markets may vary over time and by field, with critical and emerging scientific fields being riper targets for public investment than more mature ones.

The methods and data of this paper point to other opportunities for future research. The methods we develop can be used more widely to evaluate the content of dissertation science or other scientific corpora directly from the text. Our entity recognition procedure can also be applied more

broadly, and in ongoing research (Gross et al. 2025), members of this author team are adapting this approach to extract research funders from historical publications.

Beyond methods, the data accompanying this paper can be used in many ways, including for evaluating where scientific training takes place, how it has evolved over time, what drives changes in its scale and composition, and what its effects are on institutions, regions, scientific fields, and more. For example, how does PhD production interact with regional economies? To what degree do public investments in scientific training spill over into (i) closely related local industries, vs. (ii) less-related industries in the same region, vs. (iii) other regions? How important are PhD programs to universities' overall research productivity—and how much does research suffer when PhD enrollment declines or programs end? How important are PhD programs to nearby firms—and conversely, what is the role of industry support in doctoral training? Do universities with more industry-funded PhDs produce more commercially oriented research, industry placements, or university startups? And backing out from these more localized effects: what is the importance of PhD production to the wider U.S. innovation system, and what happens to U.S. science and innovation as the PhD population ebbs and flows? In releasing data which maps the U.S. scientific training ecosystem over a long horizon, we hope to enable further research on these questions.

References

Agarwal, Ruchir, Ina Ganguli, Patrick Gaulé, and Geoff Smith. 2023. "Why US immigration matters for the global advancement of science." *Research Policy* **52**(1):104659.

Aiken, Catherine, James Dunham, Jennifer Melot, and Zachary Arnold. 2024. *Identifying Emerging Technologies in Research*. Center for Security and Emerging Technology working paper, available at https://cset.georgetown.edu/publication/identifying-emerging-technologies-in-research/.

Akcigit, Ufuk, Jeremy Pearce, and Marta Prato. 2025. "Tapping into talent: Coupling education and innovation policies for economic growth." *Review of Economic Studies* **92**(2):696-736.

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pp. 609-626. Princeton University Press.

Atkinson, Richard C. 1990. "Supply and demand for scientists and engineers: A national crisis in the making." *Science* **248**(4954):425-432.

Azoulay, Pierre, Benjamin F Jones, J Daniel Kim, and Javier Miranda. 2022. "Immigration and entrepreneurship in the United States." *American Economic Review: Insights* **4**(1):71-88.

Bailey, Martha J, Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How well do automated linking methods perform? Lessons from US historical data." *Journal of Economic Literature* **58**(4):997–1044.

Balsmeier, Benjamin, Lee Fleming, Matt Marx, and Seungryul Ryan Shin. 2025. "Startups, unicorns, and the local influx of inventors." *Review of Economics and Statistics* pp. 1–44.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

Bernstein, Shai, Rebecca Diamond, Abhisit Jiranaphawiboon, Timothy McQuade, and Beatriz Pousada. 2022.

The contribution of high-skilled immigrants to innovation in the United States. NBER Working Paper No. 30797.

Bhatia, Aatish, Irineo Cabreros, Asmaa Elkeurti, and Ethan Singer. 2025. *Trump Has Cut Science Funding to Its Lowest Level in Decades*. New York Times, available at https://www.nytimes.com/interactive/2025/05/22/upshot/nsf-grants-trump-cuts.html.

Bianchi, Nicola and Michela Giorcelli. 2020. "Scientific education and innovation: from technical diplomas to university STEM degrees." *Journal of the European Economic Association* **18**(5):2608-2646.

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Blume-Kohout, Margaret E and Dadhi Adhikari. 2016. "Training the scientific workforce: Does funding mechanism matter?" *Research Policy* **45**(6):1291-1303.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-experimental shift-share research designs." *The Review of Economic Studies* **89**(1):181-213.

Bostwick, Valerie, Joseph Staudt, and Bruce A. Weinberg. 2024. *Best and Brightest? The Selectivity of Foreign-Born Ph.D. Recipients in the US*. Working paper.

Bostwick, Valerie K and Bruce A Weinberg. 2022. "Nevertheless she persisted? Gender peer effects in doctoral STEM programs." *Journal of Labor Economics* **40**(2):397-436.

Buffington, Catherine, Benjamin Cerf, Christina Jones, and Bruce A Weinberg. 2016. "STEM training and early career outcomes of female and male graduate students: Evidence from UMETRICS data linked to the 2010 census." *American Economic Review* **106**(5):333-338.

Bush, Vannevar. 1945. *Science, the Endless Frontier: A report to the President*.

Ceci, Stephen J., Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams. 2014. "Women in academic science: A changing landscape." *Psychological Science in the Public Interest* **15**(3):75-141.

Chang, Wan-Ying, Wei Cheng, Julia Lane, and Bruce Weinberg. 2019. "Federal funding of doctoral recipients: What can be learned from linked data." *Research Policy* **48**(6):1487–1492.

Corsini, Alberto, Johannes Koenig, Burcu Ozgun, Andriy Romanyuk, Guido Buenstorf, Francesco Lissoni, Ernest Miguelez, Michele Pezzoni, and Catalina Martinez. 2025. *Research careers in Europe: New evidence from the Doc-Track database*. Working paper.

Cyranoski, David, Natasha Gilbert, Heidi Ledford, Anjali Nayar, and Mohammed Yahia. 2011. "Education: The PhD factory." *Nature* **472**:276-279.

Dagdelen, John, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. "Structured information extraction from scientific text with large language models." *Nature Communications* **15**(1):1418.

Dugoua, Eugenie, Todd Gerarden, Kyle R. Myers, and Jacquelyn Pless. 2025. *How DOEs Government Funding Fuel Scientists?* Working paper.

Dynarski, Susan, Joshua Hyman, and Diane Whitmore Schanzenbach. 2013. "Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion." *Journal of Policy Analysis and Management* **32**(4):692-717.

Fieldhouse, Andrew J. and Karel Mertens. 2024. *The returns to government R&D: Evidence from US appropriations shocks*. Federal Reserve Bank of Dallas Working Paper 2305.

Freeman, Christopher. 1962. "Research and development: A comparison between British and American industry." *National Institute Economic Review* **20**:21-32.

Freeman, Christopher. 1968. "Science and Economy at the National Level." In *Problems of Science Policy*. OECD.

Freeman, Christopher and Alison Young. 1959. "The Research and Development Effort in Western Europe, North America and the Soviet Union: An Experimental International Comparison of Research Expenditures and Manpower in 1962." Technical report, OECD.

Freeman, Richard B, Tanwin Chang, and Hanley Chiang. 2009. "Supporting "The best and brightest" in science and engineering: NSF Graduate Research Fellowships." In *Science and Engineering Careers in*

the United States: An Analysis of Markets and Employment, edited by Richard B Freeman and Daniel L Goroff, pp. 19-57. National Bureau of Economic Research.

Fry, Caroline and Britta Glennon. 2025. *In Good Company: Coethnic Advisors and Career Paths of Immigrant Ph.D. Students*. NBER Working Paper No. 33782.

Fuchs, Erica R. 2022. "Building the analytic capacity to support critical technology strategy." Technical report, Brookings Institution.

Ganguli, Ina. 2015. "Immigration and ideas: what did Russian scientists "bring" to the United States?" *Journal of Labor Economics* **33**(S1):S257-S288.

Ganguli, Ina and Patrick Gaulé. 2019. "Will the US keep the best and the brightest (as postdocs)? career and location preferences of foreign STEM PhDs." In *The Roles of Immigrants and Foreign Students in US Science, Innovation, and Entrepreneurship*, edited by Ina Ganguli, Shulamit Kahn, and Megan MacGarvie, pp. 49-69. University of Chicago Press.

Gat, Noam. 2023. "lm-format-enforcer." https://github.com/noamgat/lm-format-enforcer.

Gaulé, Patrick. 2014. "Who comes back and when? Return migration decisions of academic scientists." *Economics Letters* **124**(3):461-464.

Gaule, Patrick and Mario Piacentini. 2018. "An advisor like me? Advisor gender and post-graduate careers in science." *Research Policy* **47**(4):805-813.

Glennon, Britta. 2024. "Skilled immigrants, firms, and the global geography of innovation." *Journal of Economic Perspectives* **38**(1):3-26.

Godin, Benoit. 2002a. *Are Statistics Really Useful? Myths and Politics of Science and Technology Indicators*. Working Paper.

Godin, Benoit. 2002b. "The numbers makers: Fifty years of science and technology official statistics." *Minerva* **40**(4):375-397.

Godin, Benoit. 2003. "The emergence of S&T indicators: why did governments supplement statistics with indicators?" *Research Policy* **32**(4):679-691.

Godin, Benoit. 2008. *The Making of statistical standards: The OECD and the Frascati manual, 1962-2002*. Working Paper.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik instruments: What, when, why, and how." *American Economic Review* **110**(8):2586-2624.

Goodman, Joshua. 2019. "The labor of division: Returns to compulsory high school math coursework." *Journal of Labor Economics* **37**(4):1141-1182.

Goolsbee, Austan. 1998. "Does government R&D policy mainly benefit scientists and engineers?" NBER Working Paper No. 6532.

Graddy-Reed, Alexandra, Lauren Lanahan, and Jesse D'Agostino. 2021. "Training across the academy: The impact of R&D funding on graduate students." *Research Policy* **50**(5):104224.

Gross, Daniel P., Bhaven N. Sampat, and Hansen Zhang. 2025. *Retreating from Science: The Long-Run Effects of the 1970s U.S. Military Disinvestment in Research*. Working paper.

Hunt, Jennifer and Marjolaine Gauthier-Loiselle. 2010. "How much does immigration boost innovation?" *American Economic Journal: Macroeconomics* **2**(2):31-56.

Jalali, Mehrdad, Yi Luo, Lachlan Caulfield, Eric Sauter, Alexei Nefedov, and Christof Wöll. 2024. "Large language models in electronic laboratory notebooks: Transforming materials science research workflows." *Materials Today Communications* **40**:109801.

Jones, Benjamin F. 2009. "The Burden of Knowledge and the "Death of the Renaissance Man": Is innovation getting harder?" *The Review of Economic Studies* **76**(1):283-317.

Jones, Benjamin F. and Lawrence H. Summers. 2022. "A Calculation of the Social Returns to Innovation." In *Innovation and Public Policy*, edited by Austan Goolsbee and Benjamin F. Jones, pp. 13-59. University of Chicago Press.

Kahn, Shulamit and Donna Ginther. 2017. "Women and Science, Technology, Engineering, and Mathematics

(STEM): Are Differences in Education and Careers Due to Stereotypes, Interests, or Family?" In *The Oxford Handbook of Women and the Economy*, edited by Susan L. Averett, Laura M. Argys, and Saul D. Hoffman. Oxford University Press. Chapter 31.

Kahn, Shulamit and Megan J. MacGarvie. 2016. "How important is US location for research in science?" *Review of Economics and Statistics* **98**(2):397-414.

Kahn, Shulamit and Megan J. MacGarvie. 2020. "The impact of permanent residency delays for STEM PhDs: Who leaves and why." *Research Policy* **49**(9):103879.

Kerr, William R and William F Lincoln. 2010. "The supply side of innovation: H-1B visa reforms and US ethnic invention." *Journal of Labor Economics* **28**(3):473-508.

Kim, Dongbin, Sehee Kim, Amanda Flores, and Angela M Palek. 2022. "Are Primary Funding Sources and Debt Level Associated with Career Outcomes Among Recent STEM Doctoral Graduates?" *The Journal of Higher Education* **93**(5):792-817.

Kim, Dahyun, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. "SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling."

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. "Efficient Memory Management for Large Language Model Serving with PagedAttention." In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lane, Julia I., Jason Owen-Smith, Rebecca F. Rosen, and Bruce A. Weinberg. 2015. "New linked data on research investments: Scientific workforce, productivity, and public value." *Research Policy* **44**(9):1659-1671.

Lubczyk, Moritz and Petra Moser. 2025. *The Ms. Allocation of Talent*. Working paper.

MacKinnon, James G and Matthew D Webb. 2018. "The wild bootstrap for few (treated) clusters." *The Econometrics Journal* **21**(2):114-135.

Martínez, Catalina, Alberto Corsini, Luis Sanz-Menéndez, Laura Cruz-Castro, Ernest Miguélez, Francesco Lissoni, Andriy Romanyuk, Michele Pezzoni, Guido Buenstorf, Johannes Koenig, Burcu Ozgun, Andrea Morrison, Fabiana Visentin, Stefano Breschi, Cornelia Lawson, Xin Deng, An Yu Chen, and Liangping Ding. 2025. "STEM Doctoral Graduates and Inventive Activities in European Countries (DOC-TRACK): Final Technical Report." Technical report, European Patent Office Academic Research Program.

Marx, Matt and Aaron Fuegi. 2020. "Reliance on Science: Worldwide Front-page Patent Citations to Scientific Articles." *Strategic Management Journal* **41**(9):1572-1594.

Marx, Matt and Aaron Fuegi. 2022. "Reliance on Science by Inventors: Hybrid Extraction of In-text Patent-to-Article Citations." *Journal of Economics & Management Strategy* **31**(2):369-392.

Mervis, Jeffrey. 2025. "NSF slashes graduate fellowship program." *Science Advisor* .

Myers, Kyle. 2020. "The elasticity of science." *American Economic Journal: Applied Economics* **12**(4):103-34.

National Center for Science and Engineering Statistics. 2025a. *Survey of Earned Doctorates, 2024*. NSF 25-349, available at https://ncses.nsf.gov/pubs/nsf25349.

National Center for Science and Engineering Statistics. 2025b. *Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions, 2023*. NSF 25-339, available at https://ncses.nsf.gov/pubs/nsf25339.

National Science Foundation. 1959. *Methodological aspects of statistics on research and development manpower*. Report No. 59-36.

Nelson, Richard R. 1959. "The simple economics of basic scientific research." *Journal of Political Economy* **67**(3):297-306.

Nice, Amy. 2025. "Meeting US Defense Science and Engineering Workforce Needs: A Progress Report." In *Entrepreneurship and Innovation Policy and the Economy*, edited by Benjamin F. Jones and Josh Lerner, volume 4, pp. 179-215. University of Chicago Press.

OECD. 1963a. *Proposed Standard Practice for Surveys of Research and Development*.

OECD. 1963b. *Science and the Policies of Government*. Also known as the Piganiol Report.

OECD. 2015. *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*.

OSTP. 2013. *Memorandum for the Heads of Executive Departments and Agencies*. White House Office of Science and Technology Policy (OSTP), available at `https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf`.

OSTP. 2022. *Memorandum for the Heads of Executive Departments and Agencies*. White House Office of Science and Technology Policy (OSTP), available at `https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf`.

OSTP. 2024a. *Critical and Emerging Technologies List Update*. White House Office of Science and Technology Policy (OSTP), available at `https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/02/Critical-and-Emerging-Technologies-List-2024-Update.pdf`.

OSTP. 2024b. *Interview with Arati Prabhakar on Achieving America's Aspirations to Improve Health Outcomes*. White House Office of Science and Technology Policy (OSTP), available at `https://bidenwhitehouse.archives.gov/ostp/news-updates/2024/12/04/arati-prabhakar-on-achieving-americas-aspirations-to-improve-health-outcomes/`.

Pal, Arka, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. "Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive." *arXiv preprint arXiv:2402.13228* .

Peri, Giovanni, Kevin Shih, and Chad Sparber. 2015. "STEM workers, H-1B visas, and productivity in US cities." *Journal of Labor Economics* **33**(S1):S225-S255.

Powell, Kendall. 2015. "The future of the postdoc." *Nature* **520**(7546):144-148.

Priem, Jason, Heather Piwowar, and Richard Orr. 2022. "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts." *arXiv preprint arXiv:2205.01833* .

Roach, Michael and Henry Sauermann. 2010. "A taste for science? PhD scientists' academic orientation and self-selection into research careers in industry." *Research Policy* **39**(3):422-434.

Roach, Michael and Henry Sauermann. 2017. "The declining interest in an academic career." *PloS One* **12**(9):e0184130.

Romer, Paul M. 1990. "Endogenous technological change." *Journal of Political Economy* **98**(5, Part 2):S71-S102.

Sauermann, Henry and Michael Roach. 2012. "Science PhD career preferences: Levels, changes, and advisor encouragement." *PloS One* **7**(5):e36307.

Sauermann, Henry and Michael Roach. 2016. "Why pursue the postdoc path?" *Science* **352**(6286):663-664.

Shvadron, Dror, Hansen Zhang, Lee Fleming, and Daniel P. Gross. 2025a. *Foreign Migration Patterns among U.S.-trained PhD Scientists*. Working paper.

Shvadron, Dror, Hansen Zhang, Lee Fleming, and Daniel P. Gross. 2025b. *Panel Data on U.S. STEM PhD graduates, 1950-2022*. Harvard Dataverse.

Stansbury, Anna and Kyra Rodriguez. 2024. *The class gap in career progression: Evidence from US academia*. Working paper.

Steelman, John R. 1947. *Science and Public Policy*. Washington: U.S. Government Printing Office. Report to the President by John R. Steelman, Chairman, President's Scientific Research Board.

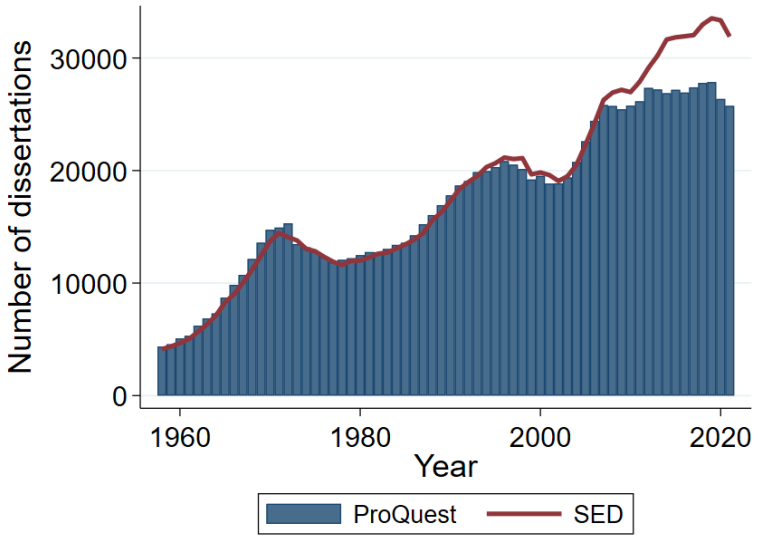
Stephan, Paula. 2013. "How to exploit postdocs." *BioScience* **63**(4):245-246.

Stephan, Paula E and Sharon G Levin. 2001. "Exceptional contributions to US science by the foreign-born and foreign-educated." *Population Research and Policy Review* **20**:59-79.

Winters, John V. 2014. "STEM graduates, human capital externalities, and wages in the US." *Regional Science and Urban Economics* **48**:190-198.

Yang, Yulin, Donna K. Ginther, and Lingfei Wu. 2025. *Large Teams Overshadow Individual Recognition*. Working paper.

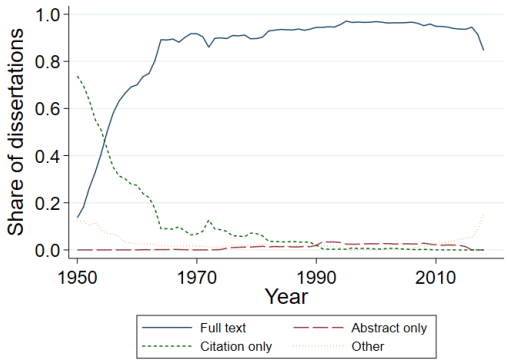
Figure 1: ProQuest vs. SED graduates in STEM fields, by year



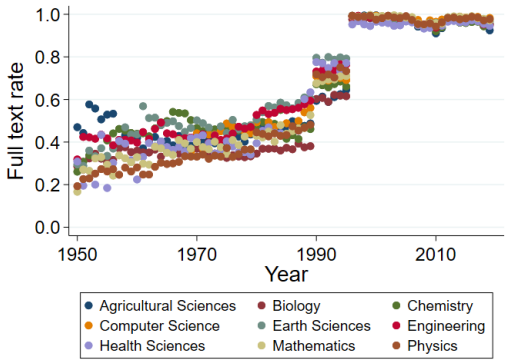
Notes: Figure shows annual PQDT graduate counts in the physical and life sciences and engineering (blue bars) and SED counts (red line) for comparison, from 1958 (the first year of the SED) to 2022.

Figure 2: Availability of dissertation text

Panel (A): Annual share of graduates in each dissertation indexing category



Panel (B): Full text availability in practice, over time and across major fields



Notes: Panel (A) shows the annual share of dissertations in our sample that ProQuest identifies as having been indexed full text (i.e., ProQuest received metadata (author, title, subject, etc.) and a full copy of the dissertation), abstract-only (metadata plus abstract), or citation only (metadata only). Panel (B) shows the annual share for which we actually have full text.

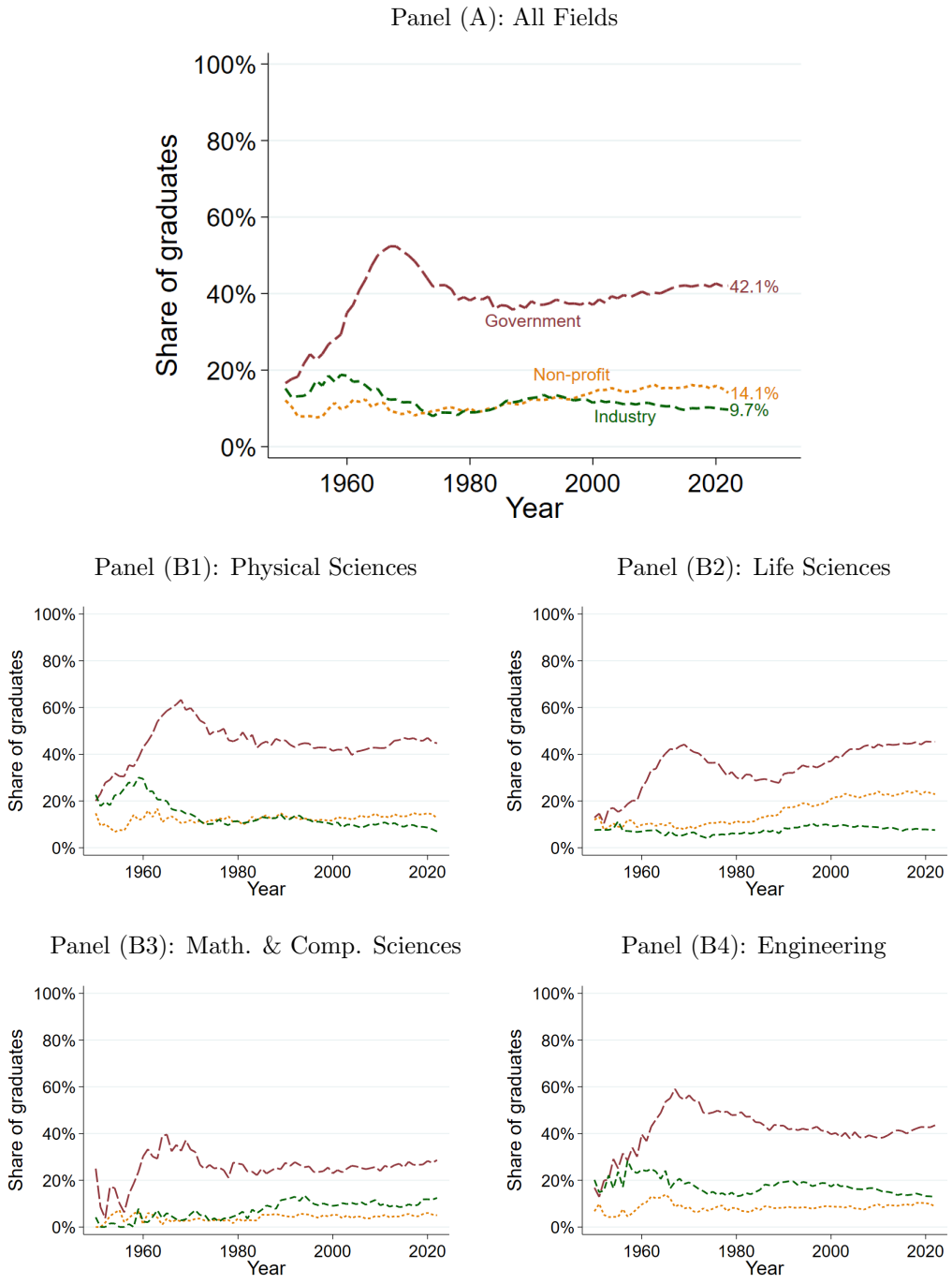
Figure 3: Example dissertation acknowledgment (Ian Buck, Stanford University, 2005)

> I have had a variety of funding sources throughout my graduate career including DARPA, DOE, ASC, NVIDIA, ATI. Specifically, I would like to thank Randy Frank, Bob Graybill, and Mike Macedonia for putting money where my mouth was. I also have been fortunate in receiving fellowships from the Stanford School of Engineering as well as NVIDIA.

Figure 4: Comparison of Manual and LLM Entity Recognition

Notes: Figure compares manual and automatic entity recognition from dissertation texts. The figure shows the percentages of dissertations acknowledging support by organization type for a sample of 500 dissertations, by classification source. Note that a dissertation can acknowledge support from multiple organization types.

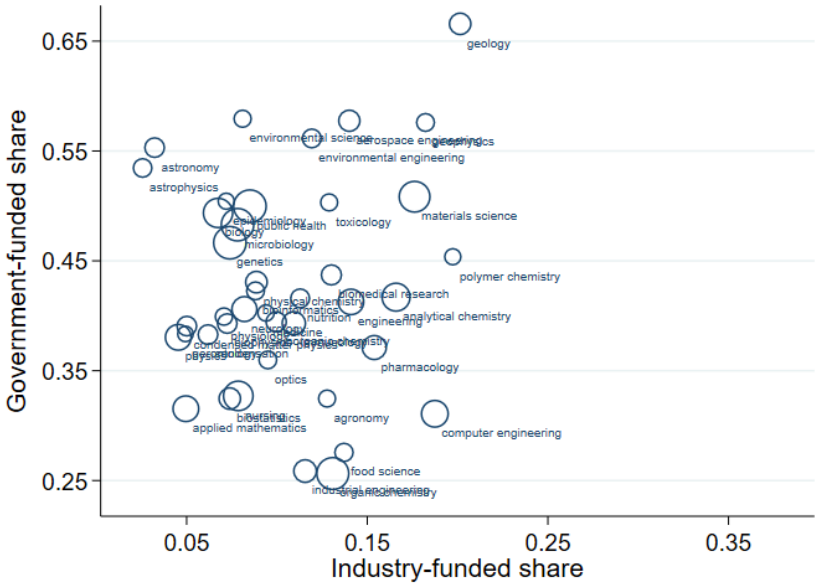
Figure 5: Share of PhD graduates supported over time, 1950-2022, by organization type, based on acknowledgments reported in dissertations (a lower bound)



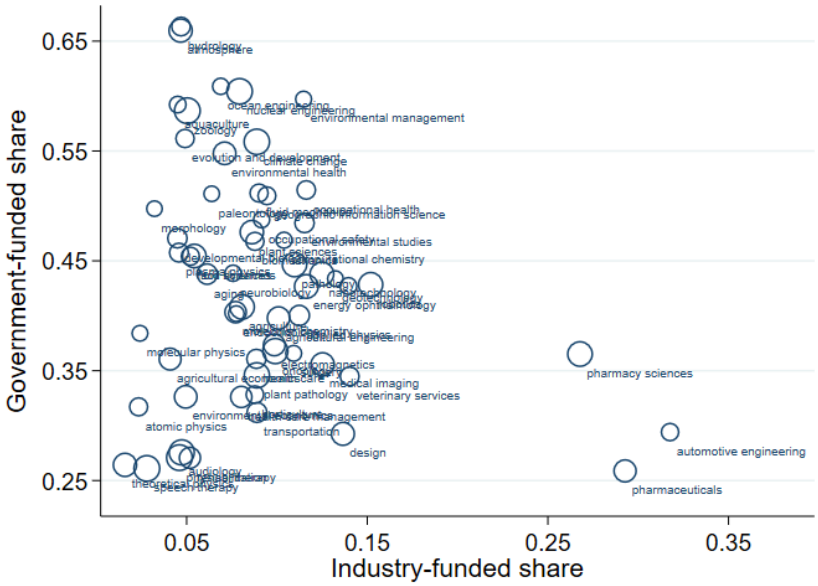
Notes: Figure shows the share of PhD graduates reporting their sources of support who are supported by (i) U.S. federal government agencies, (ii) firms, and (iii) non-profit organizations, from 1950 to 2022. Panel (A) presents overall results. Panels (B1) to (B4) present results by broad field (physical sciences, life sciences, mathematical and computer sciences, and engineering). Sample restricted to dissertations with acknowledgments (>98% of all dissertations). The shares shown here should be interpreted as a lower bound due to underreporting. See text for discussion, and see Appendix Figure B.5 for upper bounds, which use inverse propensity weighting to account for underreporting.

Figure 6: Share of graduates reporting government vs. industry support, by subject, 2000-2022

Panel (A): Very large subjects
(>1200 graduates total)



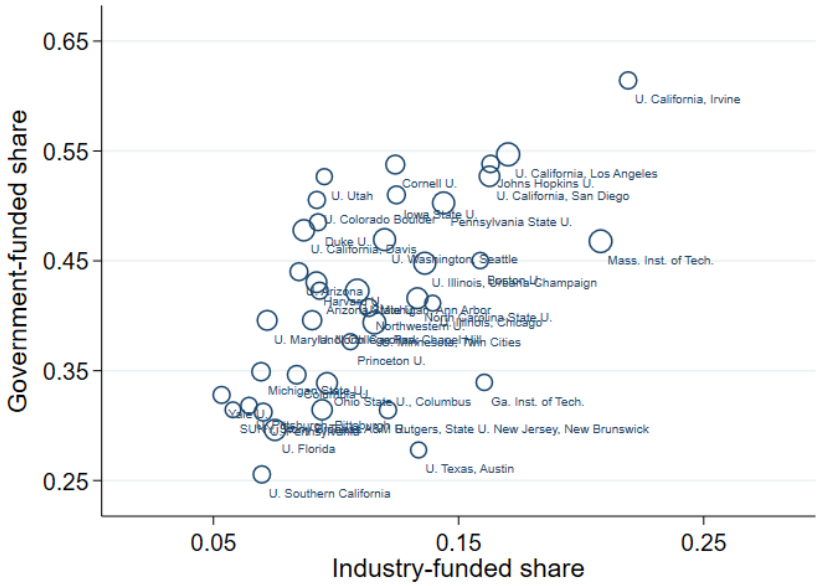
Panel (B): Large subjects
(400-1200 graduates total)



Notes: Figure plots the share of graduates in a given subject reporting government support against the share reporting industry support. Panel (A) does so for subjects with >1200 graduates between 2000 and 2022, and Panel (B) for subjects with 400 to 1200 graduates. Marker size proportional to subjects' number of graduates.

Figure 7: Share of graduates reporting government vs. industry support, by university, 2000-2022

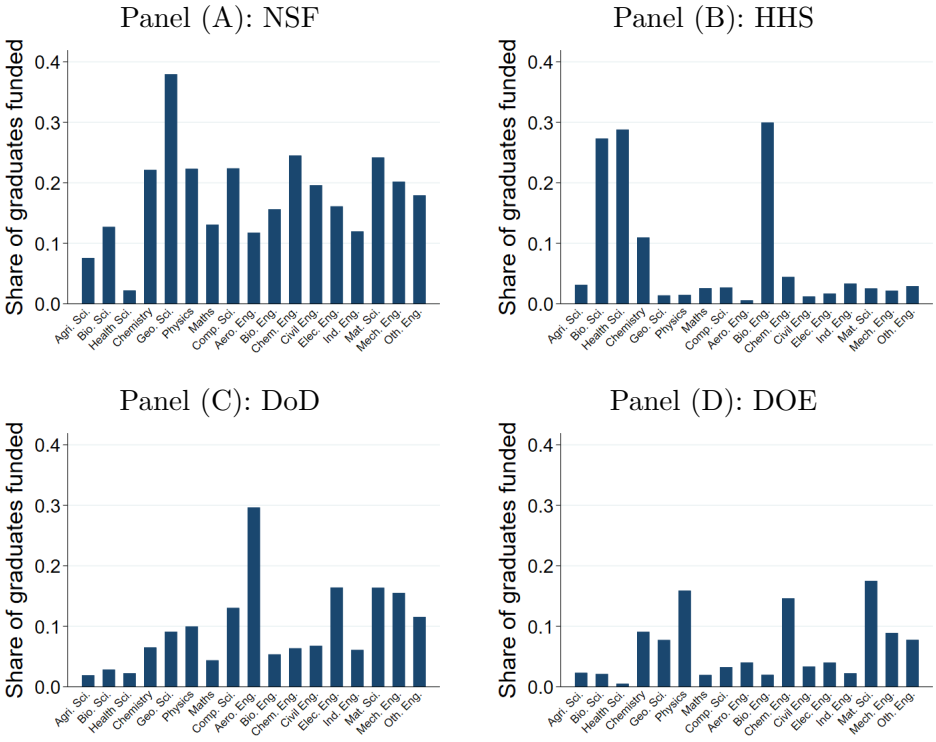
Panel (A): Very large universities
(>2000 graduates total)



Panel (B): Large universities
(1000-2000 graduates total)

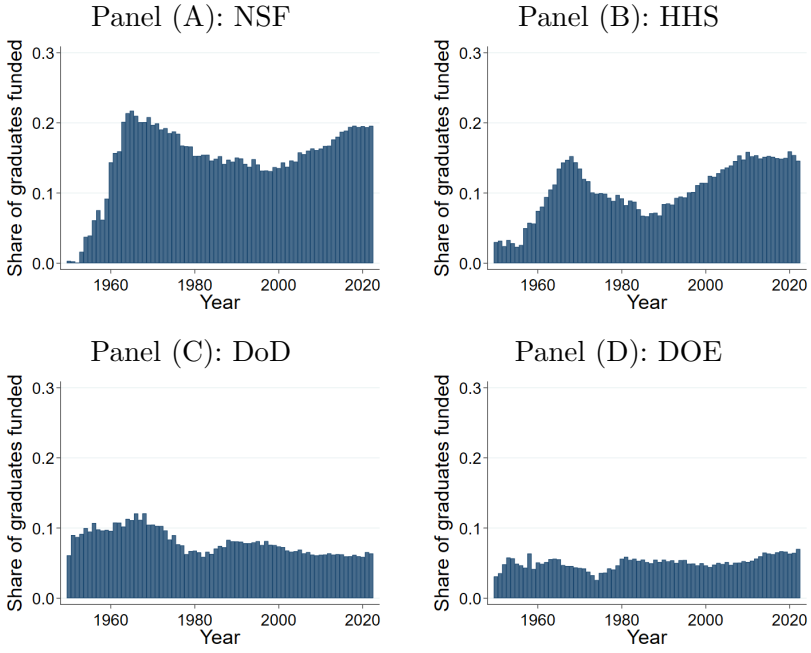
Notes: Figure plots the share of graduates from a given university reporting government support against the share reporting industry support. Panel (A) does so for universities with >2000 graduates between 2000 and 2022, and Panel (B) for universities with 1000 to 2000 graduates. Marker size proportional to universities' number of graduates.

Figure 8: Major fields' share of graduates supported by select federal agencies, 1950-2022



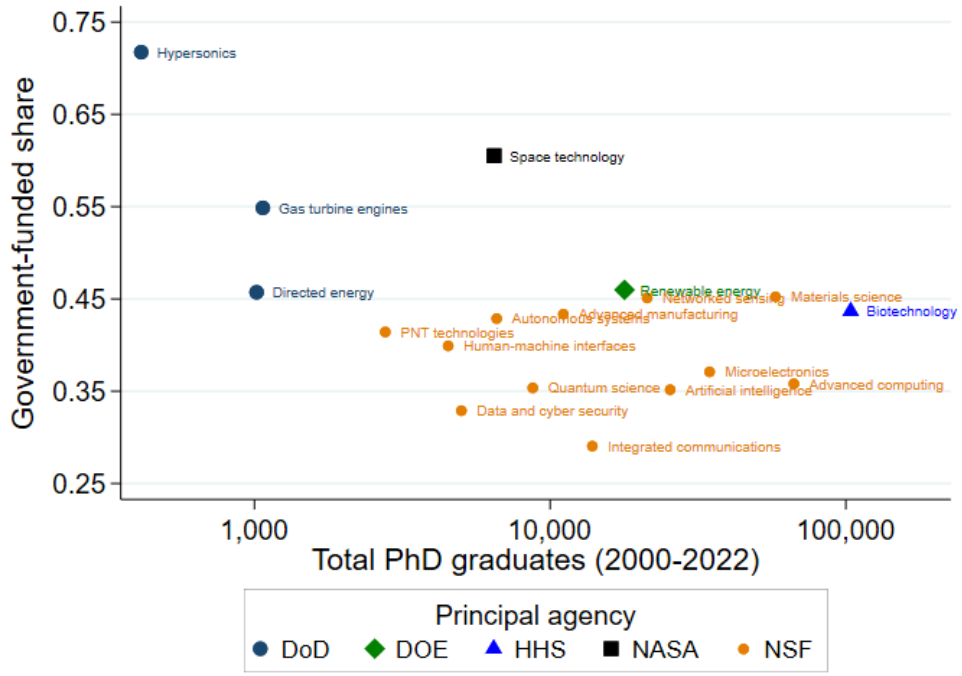
Notes: Figure shows the share of PhD graduates in individual major fields supported by each agency listed (NSF, HHS, DoD, DOE), illustrating differences in subject matter priorities across agencies.

Figure 9: Share of all graduates supported by select federal agencies over time, 1950-2022



Notes: Figure shows the share of PhD graduates supported over time by each agency listed (NSF, HHS, DoD, DOE), illustrating long-run changes in these agencies' investments in PhD training.

Figure 10: Share of graduates reporting government support, by critical technology area, 2000-2022



Notes: Figure plots the share of graduates in each critical technology area reporting government support against the number of graduates in that technology area, color-coding each technology area's principal source of government support (DoD, DOE, HHS, NASA, or NSF). See OSTP (2024a) for a complete list.

Table 1: Example extraction of acknowledged organizations for Ian Buck (Figure 3)

Entity Name	Entity Type	Support Type	ROR Organization
DARPA	government	funding	Defense Advanced Research Projects Agency
DOE	federal agency	funding	United States Department of Energy
ASC	unknown	funding	–
NVIDIA	private company	funding	Nvidia
ATI	private company	funding	Advanced Micro Devices
Stanford School of Engineering	academia	fellowship	–
NVIDIA	private company	fellowship	Nvidia

Notes: Table lists organizations identified in the text in Figure 3 as dissertation sponsors using our LLM-based procedure and provides their links to ROR. The LLM associates ATI with Advanced Micro Devices (AMD) because ATI Technologies was acquired by AMD in 2006.

Table 2: Top 15 acknowledged organizations, by sector, 2000-2022

Government agencies		Firms		Non-profit organizations	
Name		Name	Count	Name	Count
National Science Foundation	91,813	Intel	2,276	Howard Hughes Medical Institute	3,731
Department of Health and Human Services	77,989	IBM	1,942	American Heart Association	3,676
Department of Defense	34,081	Merck	1,417	Sigma Xi	2,787
Department of Energy	30,528	Google	1,300	American Cancer Society	1,568
National Aeronautics and Space Administration	15,038	Microsoft	1,221	American Chemical Society	1,453
Department of Agriculture	13,443	Pfizer	1,177	Geological Society of America	1,386
Department of Commerce	8,925	General Electric	977	Robert Wood Johnson Foundation	1,178
Department of the Interior	6,887	DuPont	873	W. M. Keck Foundation	1,104
Environmental Protection Agency	5,630	Dow Chemical	847	Fulbright Program	953
Department of Transportation	5,283	Eli Lilly	822	David and Lucile Packard Foundation	943
Department of Education	4,615	Chevron	820	Welch Foundation	910
Department of State	4,244	ExxonMobil	774	Burroughs Wellcome Fund	897
Department of Veterans Affairs	2,218	GlaxoSmithKline	764	Gordon and Betty Moore Foundation	856
Agency for International Development	1,615	Novartis	753	Ford Foundation	815
Department of Homeland Security	1,368	Boeing	718	National Geographic Society	759

Notes: Table lists the top 15 government, industry, and non-profit sponsors of graduates between 2000 and 2022, as measured by our LLM-based procedure. Government and other sponsors acknowledged in dissertations are consolidated into ultimate parent agencies (e.g., defense agencies are grouped up to the U.S. Department of Defense (DoD); the National Cancer Institute is grouped into the National Institutes of Health, which is in turn grouped up to the U.S. Department of Health and Human Services, etc.).

Table 3: Correlation of share of graduates who have govt. support with share who have industry or non-profit support, 2000-2022; estimated across various units of analysis

Panel (A): Estimated at the university-field and university-year level

	University-field				University-year			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Industry-funded share	0.373***	0.218***	0.227***	0.131**	0.461***	0.229***	0.486***	0.244***
	(0.046)	(0.048)	(0.056)	(0.053)	(0.028)	(0.026)	(0.028)	(0.026)
Nonprof-funded share	0.729***	0.637***	0.890***	0.763***	0.705***	0.471***	0.696***	0.454***
	(0.035)	(0.030)	(0.043)	(0.039)	(0.019)	(0.022)	(0.019)	(0.021)
N	4055	4043	4055	4043	6251	6241	6251	6241
R^2	0.33	0.60	0.49	0.72	0.33	0.73	0.35	0.75
Univ. FEs		Y		Y		Y		Y
Field FEs			Y	Y				
Year FEs							Y	Y

Panel (B): Estimated at the university-field-year level

	University-field-year							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Industry-funded share	0.220***	0.141***	0.204***	0.222***	0.140***	0.141***	0.209***	0.143***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Nonprof-funded share	0.541***	0.452***	0.545***	0.540***	0.422***	0.450***	0.540***	0.416***
	(0.008)	(0.007)	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)	(0.007)
N	56006	55997	56006	56006	55997	55997	56006	55997
R^2	0.16	0.30	0.23	0.16	0.35	0.30	0.23	0.36
Univ. FEs		Y			Y	Y		Y
Field FEs			Y		Y		Y	Y
Year FEs				Y		Y	Y	Y

Notes: Table estimates the correlation of government support rates against industry and non-profit support rates at the university-field and -year levels (Panel A) and the university-field-year level (Panel B). Each column controls for alternative sets of fixed effects, and the estimation period in all columns is 2000-2022. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table 4: Top 5 universities and sponsors of PhD graduates in U.S. critical technology areas, 2000-2022

Advanced Computing

Top universities

University	Count
Stanford U.	1,971
Mass. Inst. of Tech.	1,818
Purdue U., West Lafayette	1,617
U. California, Berkeley	1,572
U. California, Los Angeles	1,500

Top funders

Name	Sector	Count
NSF	Govt	13,571
DoD	Govt	7,038
HHS	Govt	4,400
DOE	Govt	3,411
NASA	Govt	1,975

Advanced Manufacturing

Top universities

University	Count
Mass. Inst. of Tech.	381
U. California, Berkeley	362
Purdue U., West Lafayette	326
U. Michigan, Ann Arbor	311
U. Illinois, Urbana-Champaign	296

Top funders

Name	Sector	Count
NSF	Govt	2,842
DoD	Govt	1,567
DOE	Govt	1,064
HHS	Govt	442
NASA	Govt	341

Advanced Materials

Top universities

University	Count
Mass. Inst. of Tech.	1,555
Northwestern U.	1,425
Pennsylvania State U.	1,345
North Carolina State U.	1,305
U. Illinois, Urbana-Champaign	1,290

Top funders

Name	Sector	Count
NSF	Govt	14,770
DOE	Govt	8,777
DoD	Govt	7,714
HHS	Govt	2,324
NASA	Govt	1,666

Autonomous Systems

Top universities

University	Count
Mass. Inst. of Tech.	327
Stanford U.	276
Carnegie Mellon U.	237
U. California, Berkeley	226
U. Michigan, Ann Arbor	203

Top funders

Name	Sector	Count
NSF	Govt	1,403
DoD	Govt	1,294
NASA	Govt	526
DOT	Govt	184
HHS	Govt	166

Biotechnology

Top universities

University	Count
U. Wisconsin-Madison	2,427
Harvard U.	2,308
Stanford U.	2,260
U. Washington, Seattle	2,082
U. California, Berkeley	2,006

Top funders

Name	Sector	Count
HHS	Govt	27,586
NSF	Govt	18,226
DoD	Govt	4,394
DOE	Govt	4,091
USDA	Govt	4,007

Clean Energy Generation and Storage

Top universities

University	Count
Stanford U.	528
U. California, Berkeley	519
Mass. Inst. of Tech.	506
Purdue U., West Lafayette	421
Pennsylvania State U.	415

Top funders

Name	Sector	Count
DOE	Govt	4,541
NSF	Govt	4,031
DoD	Govt	1,343
USDA	Govt	421
NASA	Govt	400

Communications and Networking

Top universities

University	Count
U. California, Los Angeles	500
Stanford U.	418
U. California, San Diego	384
Ga. Inst. of Tech.	361
Purdue U., West Lafayette	351

Top funders

Name	Sector	Count
NSF	Govt	2,512
DoD	Govt	1,692
Intel	Firm	322
DOE	Govt	314
NASA	Govt	284

Data Privacy and Cybersecurity

Top universities

University	Count
Purdue U., West Lafayette	164
U. Maryland, College Park	124
U. California, Los Angeles	101
Arizona State U.	95
Mass. Inst. of Tech.	95

Top funders

Name	Sector	Count
NSF	Govt	1,189
DoD	Govt	621
DOE	Govt	143
Google	Firm	122
Intel	Firm	109

Microelectronics and Semiconductors

Top universities

University	Count
Stanford U.	1,313
U. California, Berkeley	1,157
Mass. Inst. of Tech.	1,014
U. Illinois, Urbana-Champaign	857
U. California, Los Angeles	848

Top funders

Name	Sector	Count
NSF	Govt	7,680
DoD	Govt	4,661
DOE	Govt	3,698
Intel	Firm	904
NASA	Govt	754

Networked Sensing

Top universities

University	Count
Stanford U.	603
Purdue U., West Lafayette	572
Mass. Inst. of Tech.	527
U. Maryland, College Park	474
U. Michigan, Ann Arbor	472

Top funders

Name	Sector	Count
NSF	Govt	4,494
DoD	Govt	3,173
NASA	Govt	2,147
DOE	Govt	1,228
HHS	Govt	725

Quantum Science

Top universities

University	Count
Mass. Inst. of Tech.	382
Stanford U.	329
U. California, Berkeley	288
Harvard U.	277
U. Maryland, College Park	258

Top funders

Name	Sector	Count
NSF	Govt	1,985
DoD	Govt	1,044
DOE	Govt	1,017
HHS	Govt	164
NASA	Govt	155

Space Technology

Top universities

University	Count
U. Colorado Boulder	319
Stanford U.	221
Mass. Inst. of Tech.	220
U. Michigan, Ann Arbor	208
U. Maryland, College Park	201

Top funders

Name	Sector	Count
NASA	Govt	2,745
NSF	Govt	1,265
DoD	Govt	836
DOT	Govt	567
DOE	Govt	374

Notes: Table lists the top 5 universities and sponsors of graduates between 2000 and 2022 in the 12 largest OSTP critical technology areas. Graduates allocated to technology areas as described in text. For a small number of schools we have limited data from the mid-2010s onwards (see Appendix Table A.12). These schools are thus undercounted, and with more complete data a few of these schools might have appeared more frequently in this table. In Appendix B we regenerate this table using data through 2012 only, which truncates the sample early but avoids systematically undercounting these schools. The results are broadly similar, though they highlight Georgia Tech as a university which ranks highly in many of these technology areas but is missing in this table from most of them, likely due to sample truncation (Georgia Tech stops reporting to PQDT after 2012).

Table 5: Top 15 universities and sponsors of PhD graduates in AI, 2000-2022

Top universities		Top funders		
University	Count	Name	Sector	Count
Stanford U.	760	NSF	Govt	4,925
Mass. Inst. of Tech.	708	DoD	Govt	2,758
U. California, Berkeley	611	HHS	Govt	1,920
U. Maryland, College Park	578	DOE	Govt	963
Purdue U., West Lafayette	564	NASA	Govt	651
Carnegie Mellon U.	562	Google	Firm	484
U. Washington, Seattle	535	Microsoft	Firm	327
U. California, Los Angeles	533	DOC	Govt	287
U. Michigan, Ann Arbor	476	DOT	Govt	283
U. Illinois, Urbana-Champaign	443	IBM	Firm	265
U. California, San Diego	435	Intel	Firm	252
Arizona State U.	417	Nvidia	Firm	228
U. Southern California	404	USDA	Govt	216
North Carolina State U.	370	Amazon	Firm	209
U. Minnesota, Twin Cities	368	Meta	Firm	162

Notes: Table lists the top 15 universities and sponsors of graduates between 2000 and 2022 in artificial intelligence (AI). Graduates identified as AI-related based on critical technology assessment (see Section 3).

Table 6: Relationship of federal support to PhD production at the field-year level, 1970-2022

	Ln(PhD graduates)				Ln(Publications)	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	All	Non-USG	All		
Ln(USG-supported PhDs)	0.438***	0.773***	0.628***	0.471***		
	(0.015)	(0.021)	(0.037)	(0.007)		
Ln(Non-USG PhDs)	0.533***			0.538***		
	(0.021)			(0.003)		
Ln(Past 20 years' PhDs)					0.254***	
					(0.098)	
Ln(Past 20 years' USG PhDs)						0.247**
						(0.098)
N	901	901	901	57103	900	900
F-stat	130.44	147.35	147.35	184.79	611.32	337.36
Field FEs	Y	Y	Y		Y	Y
Univ-Field FEs				Y		
Year FEs	Y	Y	Y	Y	Y	Y

Notes: Table estimates the relationship of annual PhD production in field-years to government-supported PhD graduates in those fields. Columns (1) and (2) relate government-supported graduates to total PhD graduates, and Column (3) to other PhD graduates. In Column (4), we reproduce Column (1) at the university-field-year level, in an analogous specification with university-field and year fixed effects. Columns (5) and (6) relate the stock of recent (past 20 year) PhD graduates in a given field to the annual flow of scientific output in that field. All columns estimate relationships by two-stage least squares, using a shift-share instrument as described in the text. Estimation sample covers the post-1970 period. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by major field (via wild bootstrap) in parentheses in Columns (1) to (3) and (5) to (6). SEs clustered by university in Column (4).

Online Appendix

Funding the U.S. Scientific Training Ecosystem: New Data, Methods, and Evidence

Dror Shvadron, Hansen Zhang, Lee Fleming, and Daniel P. Gross

Table of Contents

A Data Appendix

A.1 Core sample: ProQuest dissertations

Our goal in this paper is to build as complete a sample of U.S. STEM (Science, Technology, Engineering, and Mathematics) PhD graduates since the mid-20th century as possible. Using this sample, we develop new methods of measuring characteristics of PhD trainees that complement existing data sources on U.S. PhD graduates and can substitute for administrative data in other countries or contexts where data collection infrastructure is less developed.

Our main data source for doing so is ProQuest Dissertations & Theses (PQDT).[1] For nearly a century, ProQuest has been acquiring and re-publishing dissertations of U.S. (and other countries') PhD graduates. Its dissertation collection has since been digitally catalogued and includes extensive metadata on each dissertation (e.g., author, institution, degree name, title, abstract, subject), and the full text of many of these dissertations. PQDT data have been used in a range of recent research in the science of science, economics of innovation, and other fields studying scientific trainees (e.g., Toole and Czarnitzki 2010, Bikard et al. 2015, Buffington et al. 2016, Jiang et al. 2023, Antman et al. 2023, Arora et al. 2023, among others). As we show below, its contents approximate the universe of U.S. PhD graduates in our focal fields, which is a result of extensive, long-running licensing agreements it has held with many universities across the country.

At the time of acquisition (in 2023), the PQDT database contained nearly six million global master's theses and doctoral dissertations, including in the natural sciences (i.e., physical and life sciences), engineering, humanities, social sciences, and more. For this paper, we sought to build a sample of doctoral dissertations at U.S. institutions in the natural sciences and engineering—which we term STEM fields as shorthand—for graduates from 1950 to 2022.

We do so as follows. Beginning with the full PQDT corpus, we first removed (i) non-U.S. dissertations and (ii) non-doctoral theses and dissertations. We also manually reviewed degree names to identify those corresponding to research degrees in the natural sciences and engineering.[2] We then crosswalk these graduates' self-reported subjects (as provided in the PQDT data) to "major fields" defined by the U.S. National Center for Science and Engineering Statistics (NCSES) and filter our sample to dissertations in 17 major fields which are typically considered STEM subjects.[3]

[1] Note that portions of this appendix—particularly those documenting our sample and our classification of PhD graduates to technology areas—share content in common with the data appendix of Shvadron et al. (2025a), where we use the same data source to study postgraduate migration of U.S. scientific trainees.

[2] Though most doctoral degrees are PhDs, PQDT also includes other doctorates such as DScs or field-specific doctorates (e.g., D.CS. (Doctor of Computer Science) or Chem. D. (Doctor of Chemistry)), as well as certain doctoral professional degrees (e.g., PsyD (Doctor of Psychology) or D.N.P. (Doctor of Nursing Practice)).

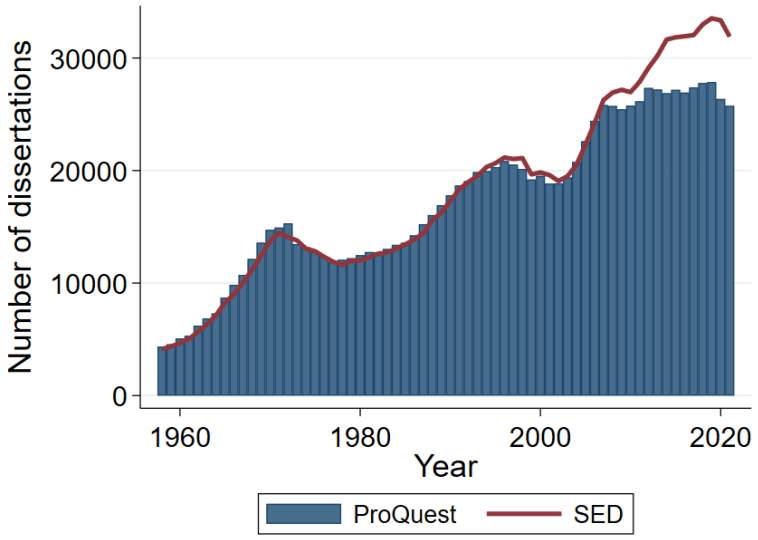
[3] Our focal major fields are: Agricultural sciences and natural resources, Biological and biomedical sciences, and Health sciences (belonging to the broad field "Life Sciences"); Physics and astronomy, Geosciences, atmospheric sciences, and ocean sciences, and Chemistry (belonging to "Physical Sciences"); Computer and information sciences and Mathematics and statistics (belonging to "Mathematical and Computer Sciences"); and Aerospace, aeronautical, and astronautical engineering, Bioengineering and biomedical engineering, Chemical engineering, Civil engineering, Electrical and electronics engineering, Industrial and manufacturing engineering, Materials science engineering, Mechanical engineering, and Other engineering (belonging to "Engineering"). In early years, PQDT reports a single

Lastly, we restrict the sample to graduates of institutions which ever held an R1 or R2 Carnegie classification post-2000, as well as of doctorate-granting medical and engineering schools. Our final sample contains 1.17 million dissertations from 1950 to 2022.

To evaluate the completeness and representativeness of our data, we compare the annual number of PQDT graduates in our final sample to annual counts of U.S. doctoral graduates from the Survey of Earned Doctorates (SED), an annual census of research doctorate recipients from U.S. universities which has been administered by NSF since 1958. Figure A.1 shows that PQDT (the blue bars) tracks the SED (red line) closely to the late 2000s. Even when they diverge slightly post-2005, PQDT still sums to >90% of the SED totals. Figure A.2 disaggregates this comparison to individual fields and shows a similar degree of consistency within them.

Figure A.3, Panel (A) extends these comparisons, presenting a binned scatterplot of the log number of graduates in our PQDT sample at the university-field-year level against the log number of graduates according to the SED data. Panel (B) repeats this comparison at the university-year level. The sample in these charts is by construction limited to observations with a nonzero number of graduates in both PQDT and SED—though for the most part, when one source reports zero graduates, and the other nonzero, the latter reports only one or two.

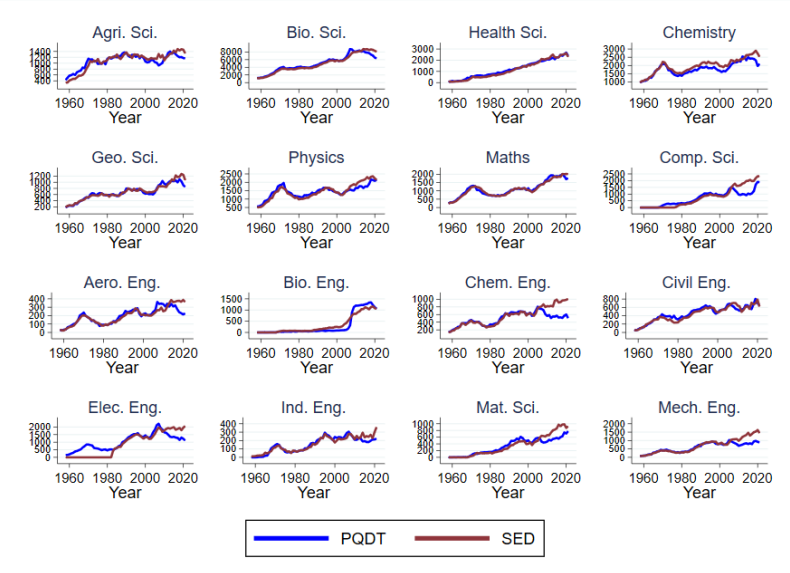
Figure A.1: ProQuest vs. SED graduates in STEM fields, by year



Notes: Figure shows annual PQDT graduate counts in the physical and life sciences and engineering (blue bars) and SED counts (red line) for comparison, from 1958 (the first year of the SED) to 2022.

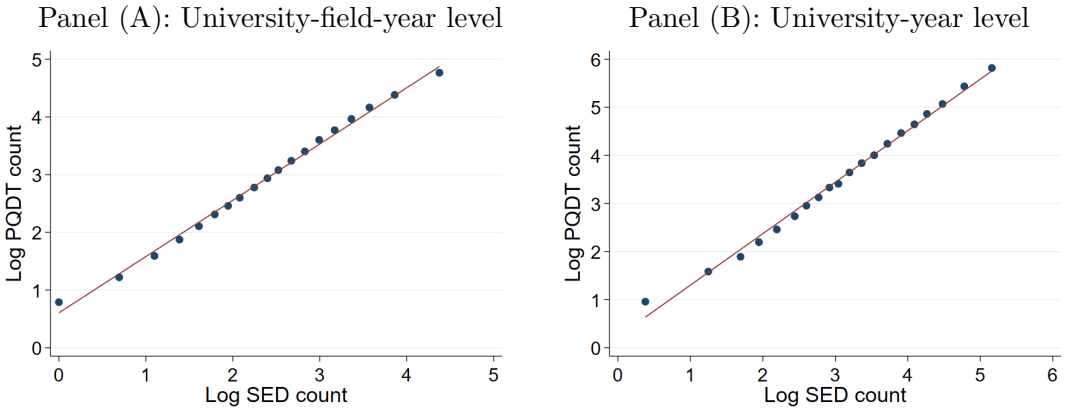
subject for each graduate, but in later years, multiple reported subjects are common. When multiple subjects are provided, we treat the first-listed subject as the graduate's primary subject. We map PQDT subjects to major fields primarily using crosswalks provided by NCSES and used to process its own surveys.

Figure A.2: ProQuest vs. SED graduates in STEM fields, by field and year



Notes: Figure shows annual PQDT graduate counts in the natural sciences and engineering (blue line) and SED counts (red line) for comparison, from 1958 (the first year of the SED) to 2022, disaggregated across 16 SED major fields (all major fields in the sample except for "Other Engineering"). We crosswalk PQDT graduates to these major fields on their author-provided, first-listed subjects, using a manually-curated crosswalk developed from the concordance used by SED to map its PhD graduate survey responses to these fields.

Figure A.3: Correlation of ProQuest vs. SED graduate counts at the university-field-year and university-year level



Notes: Figure shows a binned scatterplot of log PQDT vs. log SED graduate counts at the university-year level (left panel) and university-field-year level (right panel).

Obtaining dissertation text

An important feature of PQDT is that in addition to metadata, it also has dissertations' full text. Access to the contents of a dissertation present an opportunity to obtain direct insight into a graduate's doctoral training and output. Our focus in this paper is extracting information on graduates' sources of support—essentially, measuring who pays for scientific training—which to our

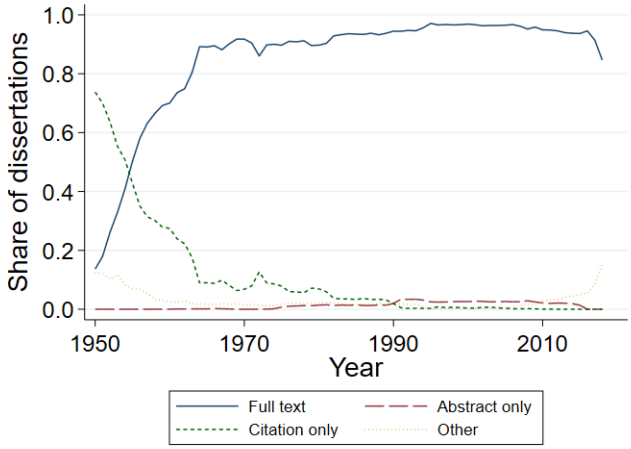
knowledge is not otherwise systematically measured or reported.[4]

Dissertation structures have for many decades followed a standard publishing template specified by ProQuest, and most dissertations have an "Acknowledgments" section, where authors thank funders and other supporters. These acknowledgments are a crucial resource for this paper: they provide an opportunity to systematically measure sources of PhD graduate support. Even when an acknowledgments section is missing, funders tend to be mentioned in the biography section of the dissertation or in footnotes in the text. Access to the full text of dissertations allows us to identify these cases regardless of where they are mentioned in the text.

Though PQDT provides the full text for most modern dissertations, this is not always the case historically, for two reasons. First, some early dissertations were indexed citation-only. Second, until recently, dissertations were stored on microfilm—not all of which has yet been digitized. Using ProQuest document numbering, we can determine how a dissertation was indexed, and whether full text *may* be available from PQDT. In Figure A.4, we plot these frequencies over time. Though early dissertations were primarily indexed citation-only, by the mid-1960s, 90%+ dissertations have full text indexing. In principle, this set should be an upper bound to an obtainable full text sample. In practice, not all such dissertations have yet been digitized. Figure A.5 shows this, plotting the annual share of dissertations for which full text is available from PQDT, by field. Between 1950 and 1990, the observable full text share of dissertations is around 30-50% in all fields, gradually increasing over time. This rate discretely jumps in 1990 and 1996—to 60-80% and then 100%—and holds near 100% thereafter. Given the quality of recent coverage and intrinsic interest in more recent data, much of our analysis will emphasize the post-2000 period.

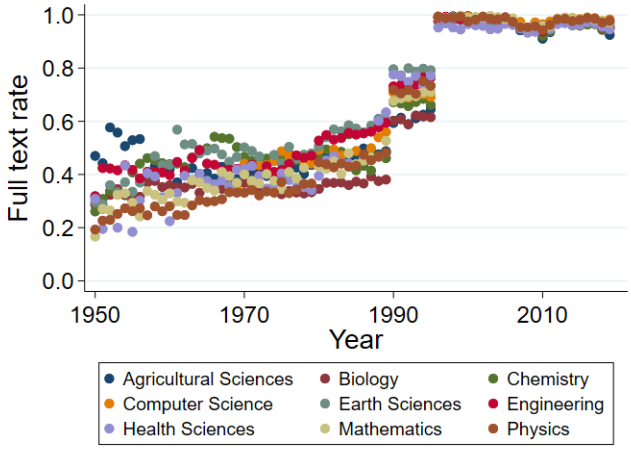
[4]The closest substitute that we are aware of is a question that was historically asked by the NSF's Survey of Earned Doctorates (SED) questionnaire on graduates' principal sources of support, where the response options changed over time but generally included a few specific government agencies or funding programs, a few specific philanthropies, and options for business, foreign government, state government, and own resources. However, this question was removed from the SED in 1997, creating an extended gap in measurement. Moreover, the response options were fairly limited in scope relative to what is reported in full text. To better understand these measures, we also applied for and were granted access to restricted-use SED microdata which include responses to this question. Though the SED itself is a near-census of U.S. PhD graduates, we found response rates to this particular question highly variable, and rates of support from specific organizations and categories of support fluctuate substantially year-to-year, whereas our dissertation text-based measures change more gradually. The combination of the question not having been asked in nearly 30 years, and the questionable quality of the responses it received, make us hesitant to use these data—and create a need to find other ways to measure who pays for scientific training.

Figure A.4: ProQuest dissertations by indexing category (according to ProQuest identifiers)



Notes: Figure shows the annual share of PQDT dissertations in our sample that ProQuest identifies has having been indexed full text (i.e., ProQuest received metadata (author, title, subject, etc.) and a full copy of the dissertation), abstract-only (metadata plus abstract), or citation only (metadata only).

Figure A.5: ProQuest dissertations with full text available, by field



Notes: Figure shows the annual share of PQDT dissertations for which we have access to full text: the subset of full-text indexed dissertations (Figure A.4) which were stored electronically or have been digitized from their original physical storage medium (microfilm).

We undertake one additional effort to fill gaps in dissertation availability: retrieving full-text dissertation files for individuals in our PQDT sample from online university repositories. We initially did so for (i) the six universities with the most graduates lacking full text dissertations, and (ii) the University of California system, whose constituent universities' repositories share a common platform. The most important of the back-filled universities is MIT, which had no full text dissertations in PQDT. Other universities for which PQDT was missing at least 10,000 dissertation texts during our sample period include UC Berkeley, UCLA, Cornell, University of Maryland, and University of

Minnesota—though as Figure A.5 indicates, most of these gaps in the available full text dissertations are pre-2000. For each of these institutions we crawl the associated repository, obtain dissertation metadata and PDFs, extract their contents, and merge these records into PQDT on name, field, and year, disambiguating manually as needed. The only universities for which we were able to find a significant number of dissertation texts which were not already available from PQDT were MIT and Michigan State University (see Table A.1). Due to the substantial effort required to crawl university repositories for historical dissertations and the meager returns, we subsequently ended this effort, but incorporated the content we already retrieved into our analysis.

Table A.1: Full text collected from university repositories for PQDT graduates

	PQDT PhDs,	Missing full text		Found in repository	
University	1950-2019	Number	Percent	Number	Pct. of missing
Cornell U.	25,061	12,112	48%	847	7%
Mass Inst. Tech.	27,546	27,546	100%	16,052	58%
Michigan State U.	15,907	6,902	43%	6,384	92%
U. of Maryland	18,682	6,183	33%	288	5%
U. of Minnesota	23,085	10,729	46%	197	2%
U. of Texas	15,232	6,085	40%	840	14%
UC Berkeley	32,037	14,456	45%	20	0%
UC Davis	14,936	4,944	33%	69	1%
UC Irvine	7,596	366	5%	93	25%
UCLA	19,893	6,570	33%	61	1%
UC Merced	439	27	6%	10	37%
UC Riverside	5,605	1,856	33%	75	4%
UC San Diego	239	28	12%	1	4%
UC San Francisco	4,460	1,166	26%	882	76%
UC Santa Barbara	6,529	1,180	18%	50	4%
UC Santa Cruz	3,159	641	20%	19	3%

Notes: Table lists the set of universities for which we searched for full text dissertations at online university thesis repositories, in attempt to fill gaps in full text availability in the PQDT sample. Columns report the number of PQDT STEM doctoral graduates associated with each university, the number missing full text in PQDT, and the results of our effort to fill these gaps with university data.

A.2 Characteristics of PhD graduates

We measure several characteristics of the PhD graduates in our sample. The first is their PhD institution, which we crosswalk to IPEDS UnitIDs (the official institution identifier of the U.S. Department of Education's Integrated Postsecondary Education Data System). Using these UnitIDs, we can merge in university characteristics from IPEDS—the principal one being each institution's Carnegie classification (measured as of 2021), which we will use at times to filter our sample to R1 universities (which are the roughly 150 most research intensive U.S. institutions). We also measure graduates' principal subjects and major fields (as described above).

We supplement the PQDT-reported measures with measures of each graduate's association with 18 critical technology areas as specified by the White House Office of Science and Technology Policy (OSTP) in 2024. The White House identified these 18 technologies as national security priorities in order to inform science and technology policy across the executive branch. This linkage serves two

purposes: (i) it produces a measure of specific policy interest, enabling us to track the development of human capital across specific technology domains that are currently considered important for national security and technological competitiveness, and (ii) it also demonstrates the possibilities presented by our PQDT-based sample and textual analysis methods.

A.3 Mapping PhD graduates to critical technology areas

A.3.1 Step 1: Reference list of critical technology areas

Our implementation draws from OSTP (2024a)'s enumeration of 18 critical and emerging technology areas and 107 constituent subfields. To improve the precision and interpretability of our classification framework, we apply a filtering process to remove subfields that are ambiguous, duplicative, or overly generic.[5] These refinements allow us to focus on a subset of subfields with clearer technical boundaries and greater specificity, ensuring that the resulting classifications are more meaningful for the purposes of policy analysis and measurement of scientific labor. Table A.2 provides a list of the 18 technology areas, with example subfields.

Table A.2: List of OSTP critical technologies

Technology	Subtechnology	Included?
Advanced Computing		
	Advanced supercomputing, including for AI...	Yes
	Edge computing and devices	Yes
	Advanced cloud services	Yes
	High-performance data storage and data centers	Yes
	Advanced computing architectures	Yes
	Advanced modeling and simulation	No
	Data processing and analysis techniques	No
	Spatial computing	Yes
Advanced Engineering Materials		
	Materials by design and material genomics	Yes
	Materials with novel properties...	Yes
	Techniques for material property characterization...	Yes
Advanced Gas Turbine Engine Technologies		
	Aerospace, maritime, and industrial development...	Yes
	Full-authority digital engine control...	Yes
Advanced and Networked Sensing and Signature Management		
	Payloads, sensors, and instruments	Yes
	Sensor processing and data fusion	Yes
	Adaptive optics	Yes
	Remote sensing of the Earth	Yes
	Geophysical sensing	Yes
	Signature management	Yes
	Detection and characterization of pathogens ...	Yes
	Transportation-sector sensing	No
	Security-sector sensing	No
	Health-sector sensing	No
	Energy-sector sensing	No

[5]For example, under *Advanced Computing*, we exclude "Advanced modeling and simulation" and "Data processing and analysis techniques". Although these terms reflect important computational practices, they are used broadly across many disciplines, making it difficult to reliably associate such labels with work situated specifically within advanced computing. Similarly, within *Advanced and Networked Sensing and Signature Management*, we remove high-level application categories such as "Health-sector sensing", "Energy-sector sensing", "Manufacturing-sector sensing", etc.—all of which lack specificity to make useful distinctions.

Manufacturing-sector sensing	No
Building-sector sensing	No
Environmental-sector sensing	No
Advanced Manufacturing	
Advanced additive manufacturing	Yes
Advanced manufacturing technologies and techniques...	Yes
Artificial Intelligence (AI)	
Machine learning	Yes
Deep learning	Yes
Reinforcement learning	Yes
Sensory perception and recognition	Yes
AI assurance and assessment techniques	Yes
Foundation models	Yes
Generative AI systems, multimodal and large language models	Yes
Synthetic data approaches for training, tuning, and testing	Yes
Planning, reasoning, and decision making	No
Technologies for improving AI safety, trust, security...	Yes
Biotechnologies	
Novel synthetic biology including...	Yes
Multi-omics and other biometrology, bioinformatics, computational biology...	Yes
Engineering of sub-cellular, multicellular, and multi-scale systems	Yes
Cell-free systems and technologies	Yes
Engineering of viral and viral delivery systems	Yes
Biotic/abiotic interfaces	Yes
Biomanufacturing and bioprocessing technologies	Yes
Clean Energy Generation and Storage	
Renewable generation	Yes
Renewable and sustainable chemistries, fuels, and feedstocks	Yes
Nuclear energy systems	Yes
Fusion energy	Yes
Energy storage	Yes
Electric and hybrid engines	Yes
Batteries	Yes
Grid integration technologies	Yes
Energy-efficiency technologies	Yes
Carbon management technologies	Yes
Data Privacy, Data Security, and Cybersecurity Technologies	
Distributed ledger technologies	Yes
Digital assets	Yes
Digital payment technologies	Yes
Digital identity technologies...	Yes
Communications and network security	No
Privacy-enhancing technologies	Yes
Technologies for data fusion and interoperability...	Yes
Distributed confidential computing	Yes
Computing supply chain security	No
Security and privacy technologies in AR/VR	Yes
Directed Energy	
Lasers	Yes
High-power microwaves	Yes
Particle beams	Yes
Highly Automated, Autonomous, and Uncrewed Systems (UxS), and Robotics	
Surface	Yes
Air	Yes
Maritime	Yes
Space	Yes
Supporting digital infrastructure...	Yes
Autonomous command and control	Yes
Human-Machine Interfaces	
Augmented reality	Yes

	Virtual reality	Yes
	Human-machine teaming	Yes
	Neurotechnologies	Yes
Hypersonics		
	Propulsion	Yes
	Aerodynamics and control	Yes
	Materials, structures, and manufacturing	Yes
	Detection, tracking, characterization, and defense	Yes
	Testing	Yes
Integrated Communication and Networking Technologies		
	Radio-frequency (RF) and... components	Yes
	Spectrum management and sensing technologies	Yes
	Future generation wireless networks	Yes
	Optical links and fiber technologies	Yes
	Terrestrial/undersea cables	Yes
	Satellite-based and stratospheric communications	Yes
	Delay-tolerant networking	Yes
	Mesh networks/infrastructure independent communications...	Yes
	Software-defined networking and radios	Yes
	Modern data exchange techniques	No
	Adaptive network controls	No
	Resilient and adaptive waveforms	Yes
Positioning, Navigation, and Timing (PNT) Technologies		
	Diversified PNT-enabling technologies...	Yes
	Interference, jamming, and spoofing detection technologies...	Yes
	Disruption/denial-resisting and hardening technologies	Yes
Quantum Information and Enabling Technologies		
	Quantum computing	Yes
	Materials, isotopes, and fabrication techniques for quantum devices	Yes
	Quantum sensing	Yes
	Quantum communications and networking	Yes
	Supporting systems	No
Semiconductors and Microelectronics		
	Design and electronic design automation tools	Yes
	Manufacturing process technologies and manufacturing equipment	Yes
	Beyond complementary metal-oxide-semiconductor (CMOS) technology	Yes
	Heterogeneous integration and advanced packaging	Yes
	Specialized/tailored hardware components for artificial intelligence...	Yes
	Novel materials for advanced microelectronics	Yes
	Microelectromechanical systems (MEMS), Nanoelectromechanical systems (NEMS)	Yes
	Novel architectures for non-Von Neumann computing	Yes
Space Technologies and Systems		
	In-space servicing, assembly, and manufacturing...	Yes
	Technology enablers for ... reusable space launch systems	Yes
	Technologies that enable ... cislunar space and/or novel orbits	Yes
	Sensors and data analysis tools for space-based observations	Yes
	Space propulsion	Yes
	Advanced space vehicle power generation	Yes
	Novel space vehicle thermal management	Yes
	Crewed spaceflight enablers	Yes
	Resilient and path-diverse space communication systems...	Yes
	Space launch, range, and safety technologies	Yes

Notes: Table lists critical technology areas in OSTP (2024a) and specific implementations, occasionally abridged to improve readability. Complete list available at `https://www.govinfo.gov/content/pkg/CMR-PREX23-00185928/pdf/CMR-PREX23-00185928.pdf`. Right column indicates whether the given subfield is used in our classification procedure (details below).

A.3.2 Step 2: Zero-shot classification to technology areas

We begin our classification with the full population of U.S. STEM PhD graduates between 2000 and 2022. Rather than directly classifying PhD graduates based on their dissertations' raw titles, keywords, or abstract text, we first generate a concise, one-sentence summary for each dissertation using their titles and abstracts. This approach parallels recent work by (Aiken et al. 2024), who employed large language models to annotate scientific publications for topic relevance in emerging technologies. Specifically, we use the `gpt-4o-mini` model with a tailored prompt designed to extract the motivation, task, and methodology of each dissertation, while excluding evaluative or impact-based language.[6] The prompt we use is as follows:

Listing 1: LLM prompt for summarizing a dissertation

```
You are an expert in science and engineering.

**Task**: Write a one-sentence summary of a dissertation based on the
    title and abstract. Include the research motivation if evident, and
    clearly describe the task(s) and method(s) used. Do not describe the
    importance or benefits.

**Input**:
{title} and {abstract}

### Return the summary as plain text only.
```

This generative summary step allows us to standardize input across a large, heterogeneous corpus of dissertations and provides a foundation for downstream classification into critical technology areas with several methodological advantages. First, it helps normalize heterogeneous textual formats: dissertation abstracts vary widely in length, structure, and verbosity, often containing extraneous information unrelated to core research contributions. Second, summarization encourages the model to foreground research motivation, task, and method, thereby aligning the input more closely with the latent conceptual criteria used in topic classification. Third, as emphasized in Aiken et al. (2024)'s application of large language models to emerging technology classification, generating structured summaries via prompt-based language models enhances both consistency and zero-shot classification performance by filtering out irrelevant content and emphasizing comparable semantic structure across documents. This preprocessing step can thus reduce noise, mitigate bias from field-specific jargon, and improve classification accuracy.

Building on these summaries, we then evaluate whether each PhD graduate's dissertation science is related to each of the 18 critical and emerging technology areas, allowing a PhD graduate to associate to multiple areas. For this task, we prompt `gpt-4o-mini` to act as an expert in defense

[6] We use the July 18, 2024 version of the `gpt-4o-mini` model. gpt-4o-mini is optimized for high-throughput, low-latency tasks and supports both text and image inputs, though only text output was used in our application. Its affordability and responsiveness made it particularly well-suited for summarization of hundreds of thousands of dissertations. See https://platform.openai.com/docs/models/gpt-4o-mini for technical details.

and emerging technologies and make an assessment. For each PhD graduate, we supply the model with their dissertation title, summary, subject terms, and paper keywords, and instruct it to return a binary "Yes/No" judgment for each technology area, with instructions which prioritize precision over recall. The user prompt lists all technology categories and presents the metadata for one dissertation at a time. An example prompt is as follows:

Listing 2: LLM prompt for classifying a dissertation

```
System: You are an expert in defense and emerging technologies. Your task
    is to determine whether a dissertation is relevant to each of the
    listed critical technology categories. For each category, answer "Yes"
    if it clearly applies; otherwise, answer "No". If uncertain, answer "No
    ". Return only a JSON object mapping each category to "Yes" or "No".

User: You will determine which of the following critical technologies this
     dissertation is related to.

Critical Technology Categories:
1. Advanced Computing
2. Advanced Engineering Materials
3. Biotechnologies
...
18. Space Technologies and Systems

Dissertation Information:
Dissertation Title: [Title here]
Summary: [Generated summary]
Subject Terms: [Controlled vocabulary terms]
Paper Keywords: [Author-provided keywords]

Please return a JSON object exactly in the following format:
{
  "Advanced Computing": "Yes",
  "Advanced Engineering Materials": "No",
  "Biotechnologies": "No",
  ...
}
Only return this JSON (no extra text).
```

This approach enables us to assess dissertations in a scalable and consistent way. The end result is a structured dataset where each dissertation can be tagged with a binary "Yes" or "No" measure of its relation to each of the OSTP (2024a) technology areas.

A.3.3 Step 3: Technology subfield matching and filtering

The top-level technology area classification is a first pass at our intended classification. It is also an intentionally broad filter. Several of the OSTP-defined domains—for example, *Advanced Computing*, *Advanced Engineered Materials*, *Advanced Manufacturing*, and *Biotechnologies*—create a broad catchment. The details of the OSTP guidance, however, make clear that the technologies

or applications its authors have in mind are often more specific.

To refine our measures, we implement a second-stage classification that further classifies each PhD graduate within a main technology area to each of its subfields—essentially re-applying our method for more specific target classes. This step helps us not only filter the initial classification to the specific subdomains that are considered "critical", but also to distinguish them: for example, this step will distinguish AI science related to "Generative AI systems, multimodal and large language models" versus "Technologies for improving AI safety, trust, security, and responsible use", or biotechnology science related to "Novel synthetic biology including nucleic acid, genome, epigenome, and protein synthesis and engineering" versus "Biotic/abiotic interfaces".

This second-stage classification is implemented by prompting `gpt-4o-mini` with the previously generated one-sentence summary, along with the dissertation title, subject terms, and keywords. The model is asked to assess whether the work is relevant to each subfield within a given technology area. For example, a dissertation identified under *Semiconductors and Microelectronics* is individually evaluated for its relevance to the following subfields: "Design and electronic design automation tools", "Manufacturing process technologies and manufacturing equipment", "Beyond complementary metal-oxide-semiconductor (CMOS) technology", "Heterogeneous integration and advanced packaging", "Specialized/tailored hardware components for artificial intelligence", natural and hostile radiation environments, RF and optical components, high-power devices, and other critical applications", "Novel materials for advanced microelectronics", "Microelectromechanical systems (MEMS) and Nanoelectromechanical systems (NEMS)", "Novel architectures for non-Von Neumann computing". The full prompt for this step can be seen in Listing 3.

Listing 3: LLM prompt for second-stage classification

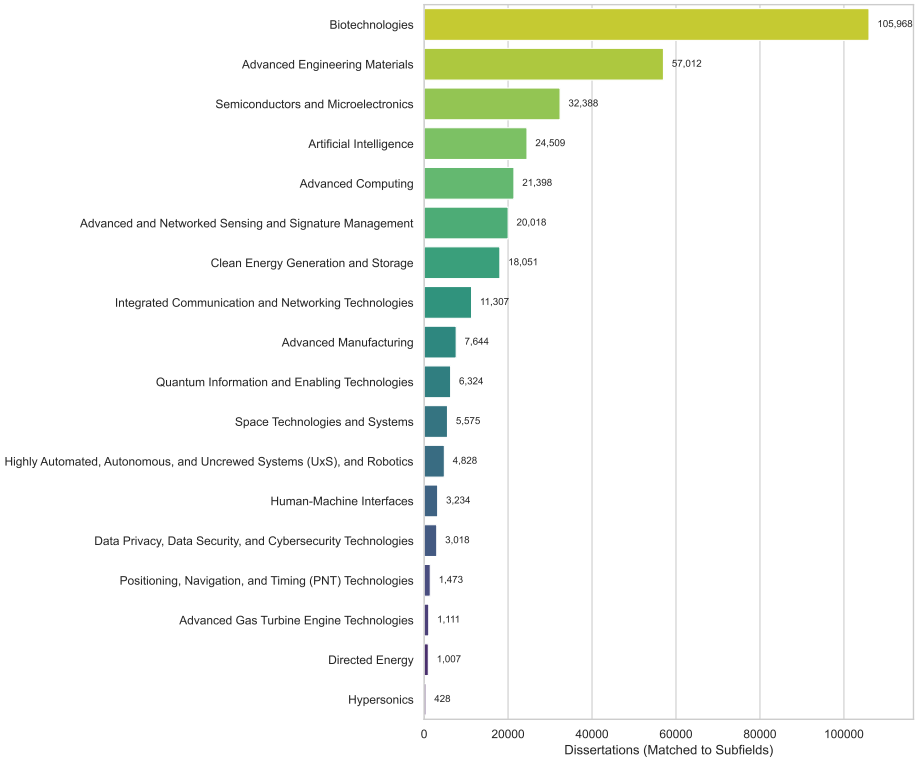
```
System: You are an expert in defense and emerging technologies. Determine
    whether a dissertation is relevant to a specific subtechnology within a
     main technology category.
Return a JSON object mapping "{tech} - {subfield}" to "Yes" or "No".

User: You will determine whether the following dissertation contributes
to the subfield "{subfield}" under the main technology category "{tech}":

Title: {title}
Summary: {summary}
Terms: {subjterms}
Keywords: {disspaperkwd}

Return exactly one JSON object (no extra text):
{ "{tech} - {subfield}": "Yes" }
or
{ "{tech} - {subfield}": "No" }
```

A.3.4 Results

Across the full set of 585,226 PhD graduates in our sample, 454,030 (77.6%) are classified to at least one of the 18 OSTP critical technology areas in the first stage, and 249,674 PhD graduates (42.7%) also classify to at least one subfield in the second stage.

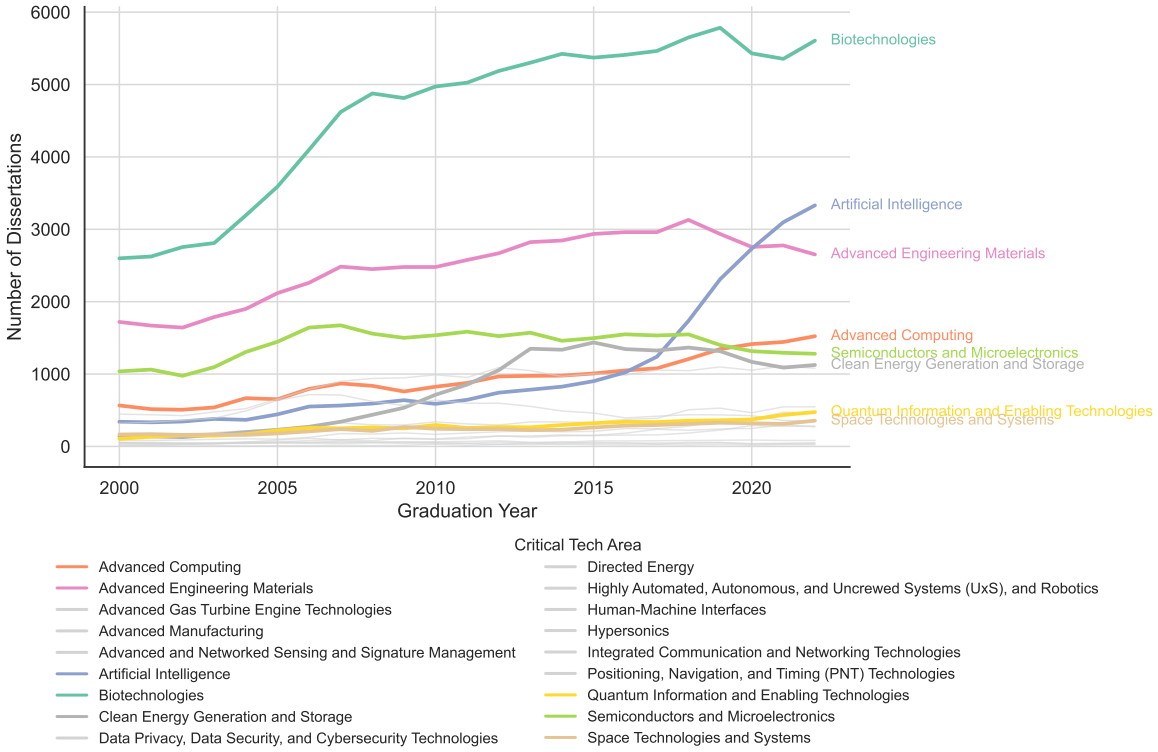
Figure A.6 shows the total number of dissertations that pass the second-stage classification by technology area, and Figure A.7 shows the number over time. *Biotechnologies* is the largest technology area by a significant margin, with roughly 106,000 matched PhD graduates, reflecting the size of the life sciences in the U.S. more generally (Figure A.2, for example, shows that Biological Sciences is the largest SED field of U.S. doctoral graduates, by a similarly wide margin). The second largest (and fastest growing) technology area is AI, which was ranked sixth as recently as 2015 but experienced an inflection point ca. 2017. Other large areas include *Advanced Engineering Materials* and *Semiconductors and Microelectronics*, both of which have been associated with sustained federal and private investment. At the other end of the distribution, technology areas like *Directed Energy*, *Hypersonics*, and *Advanced Gas Turbine Engines* have far fewer associated graduates. This difference could reflect limited university research capacity, greater concentration in non-academic institutions (e.g., national labs, defense contractors), or simply their specificity.

Figure A.6: Total number of graduates by critical technology area, 2000-2022



Notes: Figure shows the total number of U.S. natural science and engineering PhD dissertations between 2000 and 2022 classified to each of 18 OSTP-defined critical technology areas, after applying both the first- and second-stage filters. See text for explanation of methodology.

Figure A.7: Annual number of graduates by critical technology area, 2000-2022



Notes: Figure shows the annual number of U.S. natural science and engineering PhD dissertations between 2000 and 2022 classified to each of 18 OSTP-defined critical technology areas, after applying both the first- and second-stage filters. See text for explanation of methodology.

Unlike SED fields, critical technologies are not mutually exclusive: the OSTP technology list lists categories with coarse and overlapping boundaries. For example, modern advanced computing technology such as GPUs can be closely related to artificial intelligence. Consistent with this logic, under our procedure, dissertations can be classified to multiple areas, and can be classified as such. Of the dissertations we link to critical technology categories, roughly half link to two or more categories. At the extreme, one dissertation (out of hundreds of thousands) is classified to 8 technologies ("Bio-inspired VLSI systems: From synapse to behavior", by Peng Xu, of the University of Maryland; from the title alone, it is clear this dissertation would link to biotechnology, advanced computing, human-machine interfaces, and advanced sensing; it also gets classified to AI, autonomy, microelectronics, and positioning technologies).

In Table A.3, we show how these categories correlate (coincide). In general, pairwise correlations are low. We highlight in boldface the three pairs with the highest correlations: AI and advanced computing (0.58), AI and microelectronics (0.39), and advanced computing and communications and networking (0.33). These correlations are, in our view, intuitive.

Table A.3: Correlation across critical technology classifications

		(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
(1)	Advanced Computing	1.00								
(2)	Advanced Manufacturing	0.00	1.00							
(3)	Artificial Intelligence	**0.58**	-0.01	1.00						
(4)	Autonomous Systems	0.21	0.02	0.24	1.00					
(5)	Biotechnology	0.00	-0.02	-0.01	-0.04	1.00				
(6)	Communications and Networking	**0.33**	-0.01	0.04	0.06	-0.07	1.00			
(7)	Data Privacy and Cybersecurity	0.25	-0.01	0.10	0.01	-0.04	0.18	1.00		
(8)	Directed Energy	0.00	0.03	-0.01	0.00	-0.02	0.00	0.00	1.00	
(9)	Clean Energy	0.03	0.03	0.00	0.00	-0.04	0.00	0.00	0.00	1.00
(10)	Hypersonics	0.03	0.00	0.00	0.00	-0.01	0.00	0.00	0.01	0.00
(11)	Human-Machine Interfaces	0.17	0.01	0.19	0.19	0.01	0.01	0.03	0.00	-0.01
(12)	Advanced Materials	-0.08	0.28	-0.06	-0.02	-0.05	-0.04	-0.03	0.02	0.17
(13)	Microelectronics	0.08	0.22	-0.03	-0.02	-0.07	0.07	0.00	0.01	0.07
(14)	PNT Technologies	0.12	0.00	0.05	0.22	-0.03	0.08	0.03	0.00	-0.01
(15)	Quantum Science	0.06	0.01	-0.02	-0.01	-0.04	0.00	0.01	0.01	0.01
(16)	Networked Sensing	0.20	0.03	0.10	0.16	-0.03	0.19	0.06	0.00	0.00
(17)	Space Technology	0.06	0.00	0.01	0.08	-0.04	0.02	-0.01	0.01	-0.01
(18)	Advanced Turbine Engines	0.01	0.01	0.00	0.00	-0.02	-0.01	0.00	0.00	0.01

		(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
(10)	Hypersonics	1.00								
(11)	Human-Machine Interfaces	0.00	1.00							
(12)	Advanced Materials	0.00	-0.02	1.00						
(13)	Microelectronics	-0.01	0.00	**0.39**	1.00					
(14)	PNT Technologies	0.00	0.03	-0.02	0.00	1.00				
(15)	Quantum Science	0.00	-0.01	0.08	0.29	0.00	1.00			
(16)	Networked Sensing	0.00	0.06	0.01	0.05	0.21	0.00	1.00		
(17)	Space Technology	0.07	0.00	-0.01	-0.01	0.12	-0.01	0.12	1.00	
(18)	Advanced Turbine Engines	0.01	0.00	0.02	-0.01	0.00	-0.01	0.00	0.02	1.00

Notes: Table reports pairwise correlations in dissertations' technology classifications. The three largest correlations are highlighted in boldface.

A.3.5 Validation: Evidence and limitations

Validating the results of this classification is made difficult by the absence of any ground truth sample. Whether a given dissertation's science is related to a particular technology is a simultaneously broad and nuanced question, for which there is no clear rubric, and thus a question that even experts may disagree on.[7] The procedure we outlined above developed iteratively as we assessed the results of different rule-based screens, LLM-based screens, and prompt language. As we did so, we (informally) evaluated the results, comparing title and abstracts to the LLM

[7] "Critical technology areas" itself is not an objectively category or taxonomy—as reflected by the fact that different government offices have produced different definitions of critical technologies over time, and have different (but overlapping) assessments of what technologies should be considered critical even today, including with varying degrees of specificity. The concept has traditionally instead been used as a general framework for policy choices and guidance around R&D, immigration, export control, and other implied domains.

classification. Though this experimentation and evaluation was not systematic or scientific, it helped build understanding of the data and some *prima facie* confidence in the results.

To more systematically assess the validity of our LLM-based classification, we conduct a case study focused on dissertations matched to the *Quantum Information and Enabling Technologies* area (henceforth, "quantum science"), examining where these PhD graduates are trained. Quantum science offers a useful testbed due to its topical specificity and its relatively well-documented institutional footprint, research centers, and publication patterns.

We begin by identifying the universities with the most quantum science dissertations between 2000 and 2022. We then examine whether the top PhD-producing universities in quantum science are also home to known quantum computing research centers, institutes, or degree programs. Table A.4, for example, shows that every one of the top 20 universities for quantum science PhDs over the past 20 years currently has a quantum science center.

Table A.4: OpenAlex Top 20 U.S. universities for quantum science, 2000-2022

Institution	Quantum Research Center	Dissertations	Publications
Massachusetts Institute of Technology	MIT Center for Quantum Engineering (CQE)	271	8,833
University of California, Berkeley	Berkeley Quantum Info. & Comput. Center	225	7,728
Stanford University	Stanford Q-FARM	221	6,643
Harvard University	Harvard Quantum Initiative & QSE	220	5,202
University of Maryland, College Park	JQI & QuICS (UMD/NIST)	217	6,145
University of California, Santa Barbara	UCSB Quantum Foundry	172	6,476
Princeton University	Princeton Quantum Initiative	164	6,381
University of Illinois at Urbana–Champaign	Illinois Quantum Info. Sci. & Tech Center	160	5,847
University of Michigan	U-M Quantum Engineering Sci. & Tech	144	5,318
California Institute of Technology	Institute for Quantum Info. and Matter	129	4,586
University of Wisconsin, Madison	Wisconsin Quantum Institute	123	3,213
Northwestern University	Inst for Quantum Info. Res. & Engineering	120	3,603
University of Washington	UW QuantumX	121	3,009
Purdue University	Purdue Quantum Sci. & Eng. Inst.	106	3,302
University of Rochester	Center for Coherence and Quantum Science	96	2,571
University of New Mexico	UNM Center for Quantum Info. and Control	95	2,214
University of Chicago	Chicago Quantum Exchange	91	3,226
Yale University	Yale Quantum Institute	115	2,983
Cornell University	Cornell Quantum Science & Engineering	86	3,716
University of California, Los Angeles	UCLA Quantum Sci. & Engineering	86	3,255

Notes: Table lists the top 20 universities by quantum science PhD graduates between 2000 and 2022, as measured by our LLM-based procedure. For each university we also list what we assess to be its (current) principal quantum science research center, based on publicly available information.

Though affirming, this evidence is also limited by its lack of variation: given rapidly growing interest in quantum computing and its underlying science, it may just be that many universities now have quantum science centers. To distinguish universities with stronger and weaker quantum science programs or research environments, we shift our attention from this extensive margin to the intensive margin. To do so, we measure universities' volume of scientific publications in quantum science. To do so we use OpenAlex data, filtering by publications whose OpenAlex-defined topics include the keyword "quantum" (see Table A.5 for a full list).

Using these data, Figure A.8 plots the total number of quantum science publications at individual universities between 2000 and 2022 against the total number of quantum science PhD graduates

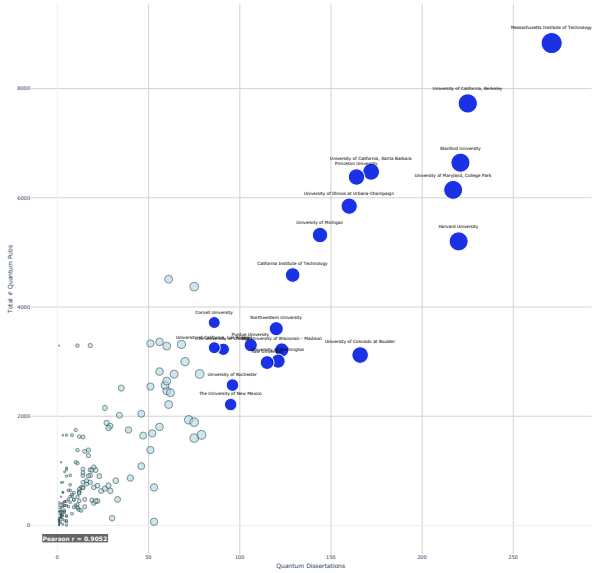
over the same period. The top 20 universities from Table A.4 are colored in blue. The figure shows a very strong correlation between the two (ρ=0.9), indicating that the LLM-based procedure captures meaningful variation across universities and PhD graduates in their relation to quantum science. Leading schools such as MIT, UC Berkeley, Stanford, Harvard, and the University of Maryland rank highly on both axes, reinforcing to us the face validity of the approach.

Table A.5: OpenAlex "quantum" research topics + parent fields and subfields

Field	Subfield	Topic
Chemistry	Physical and Theoretical Chemistry	Chemical Reactions Involving Quantum Tunneling
Computer Science	Artificial Intelligence	Quantum Computing and Simulation
Computer Science	Artificial Intelligence	Quantum Information and Computation
Computer Science	Computational Theory and Mathematics	Design and Simulation of Quantum-dot Cellular Automata
Materials Science	Materials Chemistry	Applications of Quantum Dots in Nanotechnology
Materials Science	Materials Chemistry	Synthesis and Applications of Carbon Quantum Dots
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Foundations of Electromagnetic Theory and Quantum Field Theory
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Foundations of Quantum Mechanics and Interpretations
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Parity-Time Symmetry in Optics and Quantum Mechanics
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Quantum Coherence in Photosynthesis and Aqueous Systems
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Quantum Dot Devices and Semiconductors
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Quantum Effects in Helium Nanodroplets and Solids
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Quantum Many-Body Systems and Entanglement Dynamics
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Quantum Size Effects in Metallic Nanostructures
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Semiconductor Spintronics and Quantum Computing
Physics and Astronomy	Atomic and Molecular Physics, and Optics	Slow Light Propagation and Quantum Memory
Physics and Astronomy	Condensed Matter Physics	Quantum Spin Liquids in Frustrated Magnets
Physics and Astronomy	Statistical and Nonlinear Physics	Cantorian-Fractal Theory of Quantum Physics
Physics and Astronomy	Statistical and Nonlinear Physics	Characterization of Chaotic Quantum Dynamics and Structures
Physics and Astronomy	Statistical and Nonlinear Physics	Characterization of Chaotic Quantum Dynamics and Structures
Physics and Astronomy	Statistical and Nonlinear Physics	Quantum Gravity and Noncommutative Field Theories

Notes: Table lists OpenAlex publication topics which include the word "quantum". According to OpenAlex, publications in its database are associated with "Topics" using an automated procedure that takes into account information about the work, including title, abstract, source (journal) name, and citations.

Figure A.8: University quantum science publications vs. PhD graduates, 2000-2022



Notes: Figure plots universities' number of quantum science publications between 2000 and 2022 (according to OpenAlex) against their number of quantum science PhD graduates (according to our LLM-based classification). Marker size proportional to an university's number of PhD graduates. The 20 universities with the most graduates colored in blue and labeled.

Additional validation tests We complement this evidence with validation tests across three more technology areas: AI, Biotechnology, and Space Technology. We select these areas to expand the exercise to fields with a wider range of maturity, scale, and rate of change, which is useful in assessing whether quantum science is idiosyncratic in its classifiability.

We follow the same approach as before: for each of these technology areas, we compare institution-level dissertation counts between 2000 and 2022 and the institution-level number of publications in the same area over the same period. We identify associated publications in a similar manner, using OpenAlex topics (see Tables A.6 to A.8 for a list), using the same version of the OpenAlex database (July 2024) for consistency. Similar to Figure A.8, Figures A.9 to A.11 plot the total number of publications in each of these three additional areas (AI, biotechnology, space technology) against the total number of associated PhD graduates over the same period, coloring the top 20 universities in blue. These figures continue to show a very strong correlation between the two (ρ=0.89 for AI, ρ=0.83 for biotechnology, ρ=0.85 for space technology), reinforcing that the LLM-based procedure appears to capture the target features and variation effectively.

Table A.6: OpenAlex "AI" research topics + parent fields and subfields

Field	Subfield	Topic
Computer Science	Artificial Intelligence	Neural Network Fundamentals and Applications
Computer Science	Artificial Intelligence	Natural Language Processing
Computer Science	Artificial Intelligence	Statistical Machine Translation and Natural Language Processing
Computer Science	Artificial Intelligence	Semantic Web and Ontology Development
Computer Science	Artificial Intelligence	Machine Learning for Mineral Prospectivity Mapping
Computer Science	Artificial Intelligence	Quantum Information and Computation
Computer Science	Artificial Intelligence	Text Compression and Indexing Algorithms
Computer Science	Artificial Intelligence	Program Analysis and Verification Techniques
Computer Science	Artificial Intelligence	Model-Based Clustering with Mixture Models
Computer Science	Artificial Intelligence	Advanced Cryptographic Schemes and Protocols
Computer Science	Artificial Intelligence	Logic Programming and Knowledge Representation
Computer Science	Artificial Intelligence	Anomaly Detection in High-Dimensional Data
Computer Science	Artificial Intelligence	Active Learning in Machine Learning Research
Computer Science	Artificial Intelligence	Privacy-Preserving Techniques for Data Analysis and Machine Learning
Computer Science	Artificial Intelligence	Deep Learning in Medical Image Analysis
Computer Science	Artificial Intelligence	Learning and Inference in Bayesian Networks
Computer Science	Artificial Intelligence	Cryptography and Error-Correcting Codes
Computer Science	Artificial Intelligence	Quantum Computing and Simulation
Computer Science	Artificial Intelligence	Particle Filtering and Nonlinear Estimation Methods
Computer Science	Artificial Intelligence	Machine Learning for Internet Traffic Classification
Computer Science	Artificial Intelligence	Speech Recognition Technology
Computer Science	Artificial Intelligence	Dialogue Act Modeling for Spoken Language Systems
Computer Science	Artificial Intelligence	Machine Learning for Earthquake Early Warning Systems
Computer Science	Artificial Intelligence	Automatic Keyword Extraction from Textual Data
Computer Science	Artificial Intelligence	Scientific Computing and Data Analysis with Python
Computer Science	Artificial Intelligence	Artificial Intelligence Planning and Reasoning
Computer Science	Artificial Intelligence	Language-based Information Flow Security
Computer Science	Artificial Intelligence	Autonomic Computing and Self-Adaptive Systems
Computer Science	Artificial Intelligence	Reinforcement Learning Algorithms
Computer Science	Artificial Intelligence	Advances in Transfer Learning and Domain Adaptation
Computer Science	Artificial Intelligence	Adversarial Robustness in Deep Learning Models
Computer Science	Artificial Intelligence	Application of Genetic Programming in Machine Learning
Computer Science	Artificial Intelligence	Deep Learning Applications in Healthcare
Computer Science	Artificial Intelligence	Automated Detection of Hate Speech and Offensive Language
Computer Science	Artificial Intelligence	Effectiveness of Intelligent Tutoring Systems
Computer Science	Artificial Intelligence	Learning with Noisy Labels in Machine Learning
Computer Science	Artificial Intelligence	Application of Fuzzy Cognitive Maps in Modeling
Computer Science	Artificial Intelligence	Photonic Reservoir Computing for Neural Computation
Computer Science	Artificial Intelligence	Methods and Techniques for Agent-Based Modeling

Computer Science	Artificial Intelligence	Graph Neural Network Models and Applications
Computer Science	Artificial Intelligence	Swarm Intelligence Optimization Algorithms
Computer Science	Artificial Intelligence	Optimization Methods in Machine Learning
Computer Science	Artificial Intelligence	Gaussian Processes in Machine Learning
Computer Science	Artificial Intelligence	Cryptanalysis of Block Ciphers and Hash Functions
Computer Science	Artificial Intelligence	Game Artificial Intelligence Research
Computer Science	Artificial Intelligence	Machine Learning Methods for Solar Radiation Forecasting
Computer Science	Artificial Intelligence	Type-2 Fuzzy Logic Systems and Applications
Computer Science	Artificial Intelligence	Explainable Artificial Intelligence
Computer Science	Artificial Intelligence	Sentiment Analysis and Opinion Mining
Computer Science	Artificial Intelligence	Adaptation to Concept Drift in Data Streams
Computer Science	Artificial Intelligence	Automatic Text Simplification and Readability Assessment
Computer Science	Artificial Intelligence	Artificial Intelligence and Expert Systems
Computer Science	Artificial Intelligence	Multi-label Text Classification in Machine Learning
Computer Science	Artificial Intelligence	Handling Imbalanced Data in Classification Problems
Computer Science	Artificial Intelligence	Statistical Computing and Data Analysis in R
Computer Science	Artificial Intelligence	Data Clustering Techniques and Algorithms
Computer Science	Artificial Intelligence	Artificial Intelligence in Service Industry
Computer Science	Artificial Intelligence	Cyberbiosecurity and Legal Implications of AI Technology
Computer Science	Artificial Intelligence	Theory and Applications of Extreme Learning Machines
Computer Science	Artificial Intelligence	Compilation and Analysis of Spoken Language Corpora
Computer Science	Artificial Intelligence	Authorship Attribution and User Profiling in Text
Computer Science	Artificial Intelligence	Deep Learning for Wireless Signal Classification
Computer Science	Artificial Intelligence	Inductive Modeling in Scientific Research
Computer Science	Artificial Intelligence	Theoretical Framework of Cognitive Informatics and Computational Intelligence
Computer Science	Artificial Intelligence	Machine Learning in Smart Healthcare
Computer Science	Artificial Intelligence	Fuzzy Computing and Intelligent Systems
Computer Science	Artificial Intelligence	Robotics Programming Education for Students
Computer Science	Artificial Intelligence	Artificial Intelligence in Education and Technology
Computer Science	Artificial Intelligence	Knowledge Base Graph Embedding for Visual Question Answering
Computer Science	Artificial Intelligence	Artificial Intelligence and Technology Innovation
Computer Science	Artificial Intelligence	Neuro-Symbolic Networks in Automation and Robotics
Computer Science	Artificial Intelligence	Development and Application of Expert Systems
Computer Science	Artificial Intelligence	Experience-Based Knowledge Representation and Management
Computer Science	Artificial Intelligence	Cybernetics and Information Theory
Computer Science	Artificial Intelligence	Digital Image Processing and Artificial Neural Networks
Computer Science	Artificial Intelligence	Optimization of Big Data Processing and Analysis
Computer Science	Artificial Intelligence	Enhancing E-Learning with Intelligent Agents and Analytics

Notes: Table lists topics within the OpenAlex publication subfield for AI.

Figure A.9: University AI publications vs. PhD graduates, 2000-2022

Notes: Figure plots universities' number of AI publications between 2000 and 2022 (according to OpenAlex) against their number of AI PhD graduates (according to our LLM-based classification). Marker size proportional to an university's number of PhD graduates. The 20 universities with the most graduates colored in blue and labeled.

Table A.7: OpenAlex "biotechnology" research topics + parent fields and subfields

Biochemistry	Biotechnology	Engineering Bacteria for Cancer Treatment, Genetics and Molecular Biology
Biochemistry	Biotechnology	Production of Recombinant Pharmaceuticals in Plants, Genetics and Molecular Biology
Biochemistry	Biotechnology	Microbial Enzymes and Biotechnological Applications, Genetics and Molecular Biology
Biochemistry	Biotechnology	Applications of Electroporation in Biotechnology and Food Processing, Genetics and Molecular Biology
Biochemistry	Biotechnology	Listeria Monocytogenes Pathogenesis and Food Safety, Genetics and Molecular Biology
Biochemistry	Biotechnology	Sponge-Associated Microorganisms and Biotechnological Potential, Genetics and Molecular Biology
Biochemistry	Biotechnology	Microbial Pigments and Their Applications, Genetics and Molecular Biology
Biochemistry	Biotechnology	Biotechnological Production of Vanillin, Genetics and Molecular Biology
Biochemistry	Molecular Biology	Role of Carbonic Anhydrases in Medicine and Biotechnology, Genetics and Molecular Biology
Immunology and Microbiol.	Applied Microbiol. and Biotech.	Global Burden of Antimicrobial Resistance
Immunology and Microbiol.	Applied Microbiol. and Biotech.	Microbial Tannase Production and Applications
Agr. and Biol. Sciences	Plant Science	Agricultural Biotechnology and Nutrition
Agr. and Biol. Sciences	Plant Science	Plant-Microbe Interactions in Agriculture and Biotechnology

Notes: Table lists OpenAlex topics we associate to biotechnology.

Figure A.10: University biotechnology publications vs. PhD graduates, 2000-2022

Notes: Figure plots universities' number of biotechnology publications between 2000 and 2022 (according to OpenAlex) against their number of biotechnology PhD graduates (according to our LLM-based classification). Marker size proportional to an university's number of PhD graduates. The 20 universities with the most graduates colored in blue and labeled.

Table A.8: OpenAlex "space technology" research topics + parent fields and subfields

Field	Subfield	Topic
Engineering	Aerospace Engineering	Accelerator Technology and Superconducting Cavities
Engineering	Aerospace Engineering	Simultaneous Localization and Mapping
Engineering	Aerospace Engineering	Antenna Design and Applications
Engineering	Aerospace Engineering	Nuclear Reactor Technology and Development
Engineering	Aerospace Engineering	Metasurfaces for Antenna and Radar Applications
Engineering	Aerospace Engineering	Aeroacoustic Analysis of Jet Noise
Engineering	Aerospace Engineering	Cryogenic Fluid Storage and Management
Engineering	Aerospace Engineering	Space Suit Design and Ergonomics for EVA
Engineering	Aerospace Engineering	Radiometric Calibration and Performance Monitoring
Engineering	Aerospace Engineering	Aluminium Alloys for Aerospace and Automotive Appl.
Engineering	Aerospace Engineering	Challenges and Applications of Detonation Propulsion Tech.
Engineering	Aerospace Engineering	Hybrid Rocket Propulsion and Stability Analysis
Engineering	Aerospace Engineering	Optimization Techniques for Antenna Arrays
Engineering	Aerospace Engineering	Evolution and Applications of CubeSat Missions
Engineering	Aerospace Engineering	Infrared Small Target Detection and Tracking
Engineering	Aerospace Engineering	Space Debris Removal and On-Orbit Servicing Technologies
Engineering	Aerospace Engineering	Biological and Biomimetic Flight Dynamics
Engineering	Aerospace Engineering	Air Traffic Management and Conflict Resolution
Engineering	Aerospace Engineering	Aerodynamics of High-Speed Trains and Vehicles
Engineering	Aerospace Engineering	Thermal Barrier Coatings for Gas Turbines
Engineering	Aerospace Engineering	Global Navigation Satellite Systems (GNSS)
Engineering	Aerospace Engineering	Optimization of Spacecraft Trajectories and Formations
Engineering	Aerospace Engineering	Aerodynamics and Heat Transfer in Turbomachinery
Engineering	Aerospace Engineering	Plasma Actuators for Aerodynamic Flow Control
Engineering	Aerospace Engineering	Unmanned Aerial Vehicles for Wind Estimation and Soaring
Engineering	Aerospace Engineering	Multiple-Input Multiple-Output Radar Systems
Engineering	Aerospace Engineering	Inertial Navigation Systems and Sensor Fusion Techniques
Engineering	Aerospace Engineering	Missile Guidance and Control Strategies
Engineering	Aerospace Engineering	Morphing Aircraft Technology
Engineering	Aerospace Engineering	Wind Energy Technology and Aerodynamics
Engineering	Aerospace Engineering	Synthetic Aperture Radar (SAR) Technology and Appl.
Engineering	Aerospace Engineering	Airborne Wind Energy Systems and High-Altitude Platforms
Engineering	Aerospace Engineering	Synthetic Aperture Radar Interferometry
Engineering	Aerospace Engineering	Autonomous Aerial Refueling Systems for UAVs
Engineering	Aerospace Engineering	Nuclear Thermal Hydraulics in Passive Systems
Engineering	Aerospace Engineering	Satellite Communication Networks and Systems
Engineering	Aerospace Engineering	Unmanned Aerial Vehicle Communications
Engineering	Aerospace Engineering	Radar Wave Propagation and Refractivity Estimation
Engineering	Aerospace Engineering	Electromagnetic Launch Science and Technology
Engineering	Aerospace Engineering	Icing Mitigation Techniques for Wind Turbines and Aircraft
Engineering	Aerospace Engineering	Optimization of Weapon-Target Assignment Problems
Engineering	Aerospace Engineering	Inertially Stabilized Platform Technology Concepts
Engineering	Aerospace Engineering	Hypersonic Perception and Physiological Response
Engineering	Aerospace Engineering	Modernization of Air Force in Warfare
Physics and Astronomy	Astronomy and Astrophysics	Stellar Astrophysics and Exoplanet Studies
Physics and Astronomy	Astronomy and Astrophysics	Space Weather and Magnetospheric Physics
Physics and Astronomy	Astronomy and Astrophysics	Solar Physics and Space Weather
Physics and Astronomy	Astronomy and Astrophysics	Radio Astronomy Techniques and Instruments
Physics and Astronomy	Astronomy and Astrophysics	Superconducting Detectors for Astrophysical Observations
Physics and Astronomy	Astronomy and Astrophysics	Global Lightning Distribution and Physics
Physics and Astronomy	Astronomy and Astrophysics	Space Exploration and Governance
Physics and Astronomy	Nuclear and High Energy Physics	High-Energy Astrophysics and Particle Acceleration Studies
Physics and Astronomy	Instrumentation	Astronomical Instrumentation and Spectroscopy
Physics and Astronomy	Atomic Physics and Optics	Adaptive Optics in Astronomy and Space Telescopes
Physics and Astronomy	Atomic Physics and Optics	Acousto-Optic Interaction in Crystalline Materials

Notes: Table lists OpenAlex topics we associate to space technology.

Figure A.11: University space technology publications vs. PhD graduates, 2000-2022

Notes: Figure plots universities' number of space technology publications between 2000 and 2022 (according to OpenAlex) against their number of space technology PhD graduates (according to our LLM-based classification). Marker size proportional to an university's number of PhD graduates. The 20 universities with the most graduates colored in blue and labeled.

A.3.6 Robustness and reproducibility of LLM-based classification

A central concern when using LLMs for scientific classification is the stability of their outputs over time. Even when provided with identical prompts and inputs, model-internal randomness or backend updates may result in inconsistent classifications.

To validate the reliability of our GPT-based pipeline, we conduct a two-part robustness check that compares classification outcomes across temporally separated runs using the same prompts, preprocessing, and model (`gpt-4o-mini`). We randomly sampled 2% of all dissertations (9,592 in total) that were already classified in the original full run and reprocessed them at a later time. Table A.9 reports counts for each technology area in each run. Recall that a PhD graduate is counted as a match to a technology area if their dissertation was classified to at least one of its associated subfields (see Sections A.3.2 and A.3.3 for classification details).

The results demonstrate near-identical technology area counts across runs on the same sample, suggesting that GPT-based classification is stable at this level. To assess stability at finer granularity, Table A.10 compares subfield-level classifications for a representative subset of technologies: *Artificial Intelligence*, *Biotechnologies*, and *Quantum Information and Enabling Technologies*. The results further confirm the reproducibility of classification outcomes, even at a detailed level, suggesting that reproducibility is unlikely to be a major concern.

Table A.9: Consistency of LLM-based classification in main run and 2% sample: technology areas

Technology	Match (Full)	Match (2%)
Biotechnologies	2243	2249
Advanced Engineering Materials	1262	1252
Semiconductors and Microelectronics	613	614
Artificial Intelligence	606	600
Advanced and Networked Sensing and Signature Management	472	471
Advanced Computing	452	443
Clean Energy Generation and Storage	410	414
Integrated Communication and Networking Technologies	238	239
Advanced Manufacturing	172	171
Quantum Information and Enabling Technologies	148	148
Space Technologies and Systems	121	120
Highly Automated, Autonomous, and Uncrewed Systems (UxS), and Robotics	102	101
Human-Machine Interfaces	70	66
Data Privacy, Data Security, and Cybersecurity Technologies	66	63
Positioning, Navigation, and Timing (PNT) Technologies	29	29
Directed Energy	22	23
Advanced Gas Turbine Engine Technologies	20	20
Hypersonics	7	7

Notes: Table compares technology area classification outcomes across two independent runs for a randomly sampled subset (2%) of 9,592 dissertations. The comparison assesses whether LLM-generated classification results remain stable when rerun on identical inputs at different times.

Table A.10: Consistency of LLM-based classification in main run and 2% sample: subfields

Technology	Subfield	Match (Full)	Match (2%)
Artificial Intelligence	AI assurance and assessment techniques	16	15
Artificial Intelligence	Deep learning	199	199
Artificial Intelligence	Foundation models	14	16
Artificial Intelligence	Generative AI systems, multimodal and large language models	17	19
Artificial Intelligence	Machine learning	556	548
Artificial Intelligence	Reinforcement learning	48	47
Artificial Intelligence	Sensory perception and recognition	132	129
Artificial Intelligence	Synthetic data approaches for training, tuning, and testing	22	23
Artificial Intelligence	Technologies for improving AI safety, trust, security, and responsibility	60	59
Biotechnologies	Biomanufacturing and bioprocessing technologies	108	110
Biotechnologies	Biotic/abiotic interfaces	881	873
Biotechnologies	Cell-free systems and technologies	23	21
Biotechnologies	Engineering of sub-cellular, multicellular, and organ-level systems	993	1004
Biotechnologies	Engineering of viral and viral delivery systems	157	158
Biotechnologies	Multi-omics and other biometeorology, bioinformatics, and -omics platforms	520	519
Biotechnologies	Novel synthetic biology including nucleic acid technologies	394	399
Quantum Info. and Enabling Tech.	Materials, isotopes, and fabrication techniques for quantum devices	86	85
Quantum Info. and Enabling Tech.	Quantum communications and networking	14	16
Quantum Info. and Enabling Tech.	Quantum computing	54	55
Quantum Info. and Enabling Tech.	Quantum sensing	35	36

Notes: Table compares subfield classification outcomes across two independent runs for a randomly sampled subset (2%) of 9,592 dissertations. We do so for three representative technology areas: *Artificial Intelligence*, *Biotechnologies*, and *Quantum Information and Enabling Technologies*. The comparison assesses whether LLM-generated classification results remain stable when rerun on identical inputs at different times.

A.4 Identifying dissertation sponsors

A second category of features we aim to measure is PhD graduates' sources of financial (and other) support during their doctoral training—information which can be used to evaluate who pays for scientific training. Using our STEM PhD graduate sample and dissertation text, we seek to identify research sponsors in this text. This is in essence a named entity recognition problem. However, because acknowledgments are highly variable and unstructured, a fully rule-based computational approach is ineffective. We instead use large language models to flexibly identify research supporters acknowledged in dissertation text.

To retrieve this information, we develop a six-step data processing pipeline:

Step 1. *Isolate potential acknowledgment sentences*

Step 2. *Identify named entities in these sentences*

Step 3. *Classify entities into sectors (government, industry, non-profit)*

Step 4. *Identify support type and extract grant identifiers, where available*

Step 5. *Consolidate named entities into parent organizations*

Step 6. *Match extracted organizations to external registries (ROR)*

The procedure we develop uses LLMs in several steps, particularly Upstage.ai's Solar 10.7B and Abacus.ai's Smaug 34B models (Kim et al. 2023, Pal et al. 2024).

A.4.1 Isolating acknowledgment sentences in dissertation full text

To isolate sentences that acknowledge funding in the dissertation texts, we employ a rule-based string matching approach. First, we split the raw dissertation text into sentences using NLTK's Punkt Sentence Tokenizer (Bird et al. 2009). We modify the tokenizer to ignore common academic abbreviations such as "Dr.," "Prof.," and "univ." Next, we iterate through each sentence, searching for words that indicate a potential acknowledgement or support mention. After careful refinement, we arrived at the list of keywords as appears in Listing 4 below.

Listing 4: Keyword list for acknowledgement sentence extraction

```
keywords = ["contract", "fund", "grant", "financ", "fellow", "scholarship
    ", "internship", "award", "generous", "generos", "support", "thank", "
    grateful", "appreciate", "assist", "help", "mentor", "journey", "
    guidance", "effort", "encourag", "opportunity", "provide", "gratitude",
     "committee", "advisor", "sponsor", "appreciat", "paying", "investing",
     "invested", "indebt", "obtain", "money", "stipend", "facilitat", "
    blessed", "acknowledgements", "biography", "vita", "preface"]
```

A.4.2 Identifying and classifying named entities in acknowledgment text

In steps 2 through 4, we utilize a Named Entity Recognition (NER) process employing a generative large language model (LLM). Specifically, we use Upstage.ai's Solar 10.7B Instruct v1.0 (Kim et al. 2023), a fine-tuned model based on the Mistral 7B model (which, in turn, is based on the

Llama-2 architecture). We chose these models in early 2024 based on their cost and performance characteristics, which at this time compared well to others. Open source (rather than proprietary) LLMs were selected for both cost and reproducibility reasons.

After considerable experimentation and refinement, we established a procedure that takes each sentence as input and initially evaluates if the sentence acknowledges support from an external entity. If so, we then prompted the model to identify the supporting entities mentioned in this text, categorize them by sector and type of support (funding vs. other/in-kind support), and retrieve grant and contract identifiers when available. For batch inference, we utilize the VLLM Python package (Kwon et al. 2023) and the LM format enforcer (Gat 2023). These tools facilitate efficient inference and require the model to produce a well-structured JSON schema that we predefined. Our prompt to the model is structured as appears in Listing 5 below.

Listing 5: LLM prompt for extracting supporting entities

```
### System:
You are a helpful assistant. Always answer as helpfully as possible.

### User:
Below is a sentence taken from a doctoral dissertation.

Instructions:
1. Determine whether in the provided text, the author expresses gratitude
   or mentions support from some organization.
    - set ‘gratitude_support‘ to true if yes, and false otherwise.
2. Determine whether in the provided text is part of a CV or biography.
   Set ‘biography_vita‘ to true if yes and false otherwise.
    - Note that biographies typically list awards and recognitions.
3. Extract names of supporting organizations mentioned in the text. Notes:
    - Skip people’s names. I don’t care about advisors, mentors, committee
       members etc.
    - Ensure complete extraction of grant, contract, and award identifiers
      , if present.
    - Where necessary, infer the organization’s name and type based on the
       grant name.
    - Represent each financial support and funding source as a distinct
      entity.
    - Ignore doctoral committees, academic departments, colleagues, lab
      members, unaffiliated individuals, family and friends.
    - Ensure that each organization is listed separately.
    - Return an empty list if no supporting organizations are mentioned.

Text:
{sentence}

You MUST answer using the following json schema:
{extractionFormat.schema_json()}

### Assistant:
```

To validate the structure of the LLM output, we use two Pydantic models as described in Listing 6 below. The `orgFormat` model includes fields for the organization's name and type, as well as the type of support provided. The organization and support types are restricted to a set of predefined values. The `extractionFormat` model incorporates the `orgFormat` model as a list, allowing for multiple organizations to be associated with each analyzed sentence. Additionally, it includes boolean fields for indicating whether the sentence expresses gratitude or support, or is a part of a biography. These models serve the purpose of standardizing the representation of extracted support entities and facilitate subsequent data processing.

Listing 6: Pydantic model for LLM output

```
class extractionFormat(BaseModel):
    gratitude_support: bool
    biography_vita: bool
    organizations: List[orgFormat]

class orgFormat(BaseModel):
    organization_name: str
    organization_type: Literal['government', 'federal agency', 'non-profit
        ', 'private company', 'foundation', 'academia', 'research lab', '
        consortium', 'center', 'association', 'other', 'unknown']
    support_type: Literal['funding', 'grant', 'financial support', '
        internship', 'scholarship', 'fellowship', 'training', 'materials',
        'facilities', 'data', 'in-kind support', 'technical support', '
        administrative support', 'employment', 'unspecified','other', '
        unknown']
    grant_contract_numbers: str
```

After running the first step of this procedure on our full text dissertation sample, we recovered over 92 million sentences that potentially included acknowledgments of supporting entities. We then performed batch inference on these sentences, one at a time, using language models as described above. The language model output indicated that 11 million sentences (about 12%) were indeed sentences that acknowledged support or were part of a graduate's biography. Among all sentences, the language model extracted 9.3 million mentions of organizations, with 4.8 million of these mentions originating from sentences classified as support or biography. Collectively, the retrieved entities consist of 4 million unique organization names.

A.4.3 Consolidate named entities and match to external registries

The resulting set of organizations includes many inter-related entities (e.g., parents and subsidiaries of firms and government agencies) and practically innumerable spelling variants. The next step in the data processing pipeline is therefore to consolidate these names (grouping strings representing the same organization) for subsequent analysis.

To do so, we match these entities to two external registries. The first is the Research Organization Registry (ROR), which maintains a registry of over 100,000 global research organizations, with

persistent identifiers. ROR was created to disambiguate institution names and to be used in bibliometric datasets linking researchers, research organizations, and research outputs. It also identifies parent/child relationships between ROR entities. Together, these features make it a near-ideal reference for consolidating the organizations we extract from doctoral dissertations. For named entities which we could not match to ROR, we attempt to link them to Wikidata, a free and open knowledge base hosted by Wikimedia that acts as a central database of topics, concepts, and objects referenced in Wikimedia projects (like Wikipedia).

Government Entities

We consolidate organizations that are part of the U.S. government and match them to ROR entities. First, we filter our sample for names that the LLM classified as "government" or "federal agency". We apply some standardization using regex (e.g., dropping the "U.S." at the beginning of strings) and cluster together similar names. Among 739,494 clustered government organization names, we end up with 242,382 names are associated with at least one sentence that the LLM classified as related to support or part of the author's biography.

To further consolidate this set, we run it through a large language model again. For this task, we use Abacus.ai's Smaug 34B v0.1 model (Pal et al. 2024). This model is larger than the one we used for the first round. It requires more resources to run, but performs better in correctly identifying organizations and linking them to their parent organizations. Our prompt for doing so is shown in Listing 7. The required schema includes fields for the organization name, the parent organization, indicators for federal government agencies, and location.

Listing 7: LLM prompt for consolidating government organizations

```
### System:
You are a helpful assistant. Always answer as helpfully as possible.

### User:
Below is a name of a funding organization, most likely in the US.
Return a json schema that includes:
1. The official, clean, unabbreviated name of the organization.
2. The ultimate federal government department or agency (if relevant).
3. Indicators for whether the organization is a federal organization and/
   or an independent agency.
4. Country and state. If irrelevant return "-".

Don't make stuff up. If you are unsure, return "unknown".

You MUST answer using the following json schema:
{entityFormat.schema_json()}

Organization name: {organization}

### Assistant:
 """
```

Among the initial consolidated list of 242,382 entities, the model identified 90,655 as agencies and organizations in the U.S. federal government. The model produced 3,550 unique parent organization names. We matched these names to a list of government organizations from the ROR dataset. Weighted by the number of mentions, the match resolved 81% of the mentions of U.S. government organizations and matched them with a ROR identifier.

Private Companies

The list of extracted entities includes 448,745 unique names that the LLM classified as private companies. Within them, 170,009 names are associated with at least one sentence that the LLM classified as related to support or part of the author's biography. We start the matching process by using the Smaug 34B LLM to consolidate these names. The results were further validated using an additional run through the LLM and manual comparisons.

Listing 8: LLM prompt for consolidating private companies

```
### System:
You are a helpful assistant. Always answer as helpfully as possible.

### User:
Below is a name of a grant or award.

Return a json schema that includes:
1. The ticker name for the firm. Return past ticker where relevant.
2. The name of the firm, associated with the ticker.
3. The relevant stock exchange.

Don't make stuff up. If you are unsure, return "unknown".

You MUST answer using the following json schema:
{entityFormat.schema_json()}

name: {organization}

### Assistant:
 """
```

The consolidation process resulted in 10,802 names. Working with these results, we then we used the ROR REST API to search for ROR identifiers associated with these names. Second, we used the SemTab API to find additional matches with the Wikidata identifiers (Nguyen et al. 2021). Third, we ran additional API searches using the original, unconsolidated list of names, filtered to names that were not previously matched. This step allowed for matches of firms that are not publicly traded and startups. Finally, we validated the list of resulting matches using similar techniques to those described above. The resulting match includes a ROR or Wikidata ID for 217,013 mentions of firms, corresponding to 123,366 dissertations. The matched records are associated with 3,275 unique ROR identifiers and 7,638 unique Wikidata identifiers.

Non-profit Organizations

We matched mentions of non-profit organizations in our data with external identifiers. The original list included 549,329 unique strings that were classified as foundations, philanthropic organizations, non-profits, associations or centers. Among them, 322,901 names were associated with at least one sentence that the LLM classified as related to support or part of the author's biography. For matching these mentions, we used a similar process to the one used for firms. Overall, we were able to match 333,603 mentions, corresponding to 157,884 dissertations. These matches are linked to 2,139 unique ROR identifiers and 8,126 Wikidata identifiers.

A.5 Validation

A.5.1 Validating measures against NSF graduate fellowships

To validate these measures, we need a systematic ground truth sample to compare our extracted sponsors against. Though ground truth data are hard to find—the lack of systematic, administrative data is the very reason we are developing these methods—large, organized fellowship programs with published lists of awardees offer a potential benchmark. One such program is the NSF graduate research fellowship (GRF) program, which spans many fields and has been administered continuously from 1952 to the present. Specifically, we wish to know how many of NSF GRF awardees are in our PQDT sample, how many do we measure as having NSF support, and for those we don't, why not: is it due to incomplete measurement or incomplete reporting?

To answer these questions, we retrieve all NSF GRF awardees from the NSF's website[8], link them to PQDT, and evaluate our measures against this link. To prepare for linking, we first clean the PQDT and NSF data under common parameters, standardizing names and separating names into first names, surnames, and middle names/initials. We then crosswalk the subjects reported by each awardee to the SED major fields and institutions to which we have harmonized our PQDT data. Importantly, the usefulness of this information, and in turn the quality of the links, is limited by the fact that awardees' reported fields often represent intended (rather than actual) fields of study. Nevertheless, these fields have information, which we will selectively use.[9] The NSF data present other challenges, including that individuals are sometimes awarded multiple times (implying that they chose not to use the award in one year, and later applied again); to deal with these challenges we attempt to de-duplify the data by identifying awards to same-named individuals within ten years and restricting the sample to the latest award for every individual.

We then link NSF GRF awardees to PQDT in an iterative procedure. We first attempt to link on last name (LN), first name (FN), and middle initial (MI). Among the unlinked set, we then link on LN and FN; then, on LN, first initial (FI), and MI. At every step we keep only links where the PQDT graduation year was within 10 years of the NSF GRF award year, and we discard links where

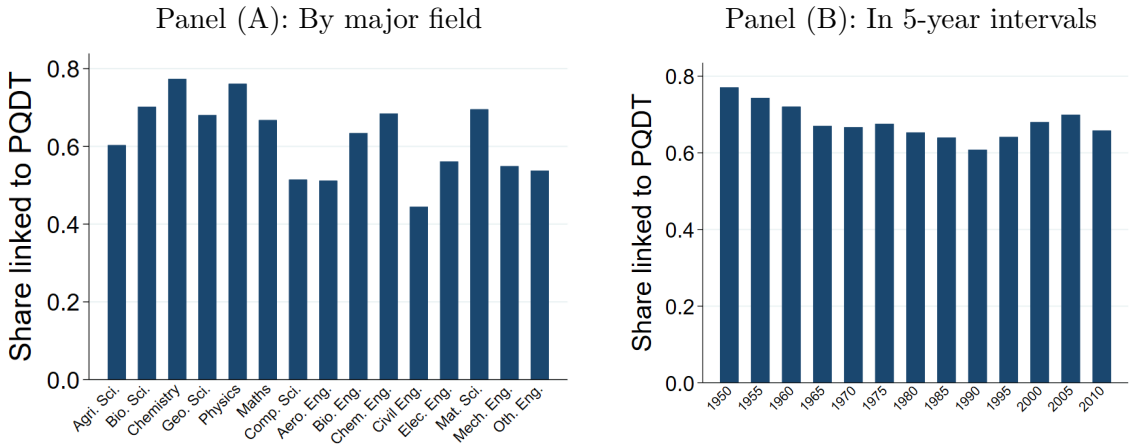
[8]See https://www.research.gov/grfp/AwardeeList.do?method=loadAwardeeList.

[9]The NSF data also include reported institutions; however, because many GRF applicants are pre-PhD, awardees' current institutions are often their baccalaureate or master's program rather than their PhD program. We found the institution data to be too ambiguous to use in this exercise.

available, unused information would rule them out (e.g., when linking on LN and FN, if MI does not match we discard the link). When multiple links are made from the NSF GRF list to PQDT or vice versa, we use subjects to disambiguate, and we then attempt to manually disambiguate any remaining ambiguous cases. Our goal is to produce a highly precise link for validation, including at the expense of recall—and given these conservative choices, the share of GRF awardees which we link to PQDT will necessarily be lower than it could otherwise be.

Through this process we successfully (i.e., without ambiguity) link 65% of GRF awardees to our PQDT sample. Figure A.12 shows how this link rate varies across NSF GRF awardees' reported subjects and over time, where it varies between 40% and 75% across subjects and steadily declines over time, from ≈75% in the 1950s to ≈55% in the 2010s. There are several reasons why these link rates may be below 100%. The most salient among them are that the NSF sample can include awardees who are not accepted to PhD programs, decline to enroll, or never graduate. Given this, a 100% link rate is an unrealistic (and incorrect) benchmark. A second reason links are less than 100% is residual ambiguity in name-based links. However, because the NSF and PQDT samples are specific slices of the general population, sampled on a common feature (potential or actual PhD trainees in the hard sciences), and because full names and specializations are provided in both datasets, we believe this link is relatively precise and mostly complete.

Figure A.12: Link rates of NSF GRF awardees to our PQDT sample (awards through 2014)

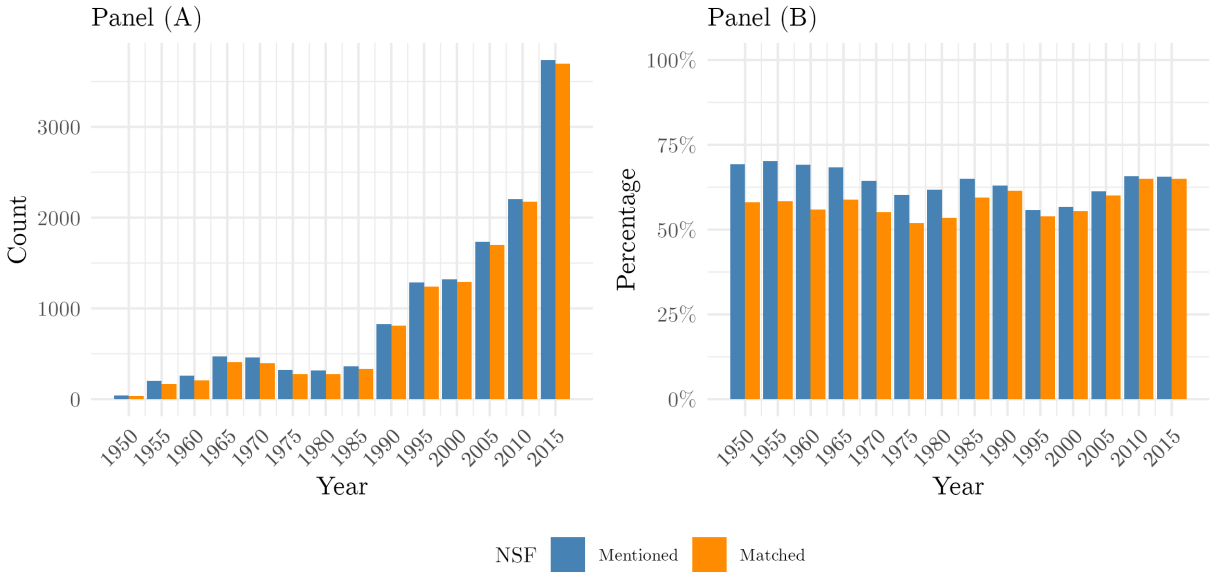


Notes: Figure shows the share of NSF graduate research fellows that we are able to link the PQDT sample, by major field (Panel A) and in five-year intervals (Panel B), where every year shown represents the upcoming five years (i.e., 2010 = 2010-2014). The sample in these charts is restricted to awards through 2014 to allow at least five years, such that all awardees can graduate and be present in our PQDT sample. Link rates below unity may reflect awardees who are not accepted to PhD programs, decline to enroll, or never graduate.

We then undertake our main validation exercise: evaluating how our text-based measurement of dissertation sponsors performs within this linked sample. Within our sample of roughly 900,000 dissertations with full text, we identify 25,039 dissertations whose authors are linked to the NSF GRF awardee list. Figure A.13 presents an analysis of NSF dissertations mentions and the ROR link rates within this sample. Panel (A) shows five-year counts of NSF-acknowledging dissertations and

of the subset of these cases where we can link to the ROR (in blue and orange, respectively). Panel (B) shows the share of our linked NSF-PQDT sample for which we find NSF acknowledgments and can link to the ROR. Among these 25,039 graduates, 15,229 (60.8%) mention the NSF somewhere in the text.[10] Our subsequent matching process links 15,026 (98.6%) of these mentions to the NSF identifier in ROR. Panel (B) shows that these rates are relatively stable over time, with acknowledgment rates between 55-65% over the past 70 years.

Figure A.13: NSF graduate fellows matched to NSF funding



Notes: Figure shows the counts (Panel A) and shares (Panel B) of NSF Graduate Fellows mentioning NSF funding support (in blue) and matched to the NSF by our matching process (in orange). Sample restricted to NSF fellows which we can link to PQDT.

Importantly, although these reporting rates are less than 100%, they are stable over time, even as the number of awardees (and PhD graduates) has grown significantly. In our analysis, there are two ways to approach potential underreporting: one is to restrict analysis to the reporting population—i.e., to PhD graduates with full text dissertations that include acknowledgments. The other is to estimate observable predictors of reporting, project the odds of reporting from observables, and use inverse propensity weighting to reweight the measured sample to represent the population. We report weighted results in the body of the paper and unweighted results in Appendix B, though our findings are qualitatively similar under both approaches.

A.5.2 Validating measures against GSS

A second opportunity for validation is possible using the NSF's Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). The GSS is an annual census of U.S. institutions

[10] For this test, we perform a simple search for "National Science Foundation" or "NSF" in acknowledgment sentences we extract from the dissertation text to identify NSF-acknowledging dissertations.

granting research-based master's degrees or doctorates in science and engineering (National Science Foundation 2022). Usefully for our purposes, the data report primary sources of support for graduate students at doctorate-granting institutions, providing annual counts of graduates in each major field with specific sources of funding: DoD, DOE, NIH and other HHS funders, NASA, NSF, USDA, other government agencies, institutional support, other U.S. or foreign sources, and self-support. Our goal here is to compare the field-year share of graduates in GSS funded by each of these agencies to the share we measure in our PQDT-based dissertation sample.

Though these comparisons make use of systematic administrative data, this exercise faces several structural limitations. First is the obvious caveat that the GSS measures enrolled students, whereas our PQDT sample measures graduates. An implication is that the GSS sample in a given year is much larger than PQDT. To a first order approximation, however, we expect that doctoral students' sources of support in a given year will correlate strongly with graduates' sources of support. A second, potentially more important caveat is that although we restrict the GSS measures to doctorate-granting institutions, this sample will include master's and doctoral students at these institutions. Sources of financial support can differ significantly across degree levels, and federal funding is particularly focused on supporting doctoral training (whereas master's training is often self-funded). In light of this, we will focus on comparing agency shares of federally-funded students (rather than all students) in GSS and our PQDT sample. Finally, we restrict our attention to 2000 to 2022, when GSS reported supports for all six of our focal agencies for this exercise (DoD, DOE, HHS, NASA, NSF, USDA; DOE wasn't added until 1999).

Specifically, we regress (i) the share of federally-funded PhD graduates in a given field and year which we measure as being supported by each of these agencies, on (ii) the share of federally-funded GSS-reported students at doctoral institutions primarily supported by these agencies. If the samples perfectly aligned, and our measurement were accurate, we would expect these to correlate roughly one-for-one (i.e., $\beta = 1$). Given sampling differences, as well as the fact that GSS measures primary sources of funding and our methodology will measure all sources of support reported in dissertations, these shares may not vary one-for-one, but they should correlate.

Table A.11 provides the results. We see strong and highly significant correlations across the table. For every 10 percentage point (p.p.) increase in federally-funded students in GSS with DoD support, we see an 9.1 p.p. higher share of DoD funding in our PQDT data. The analogous results for a 10 p.p. increase in other agency support are: 14.1 p.p. for DOE, 10.1 p.p. for HHS, 14.0 p.p. for NASA, 10.2 p.p. for NSF, and 8.1 p.p. for USDA. For HHS and NSF (the two largest funders of doctoral training today), we cannot reject that the GSS and PQDT-based measures correlate one-for-one. More broadly, we view the correlation between the GSS and PQDT-based measures as validation of our approach to measuring research sponsors.

Table A.11: Comparing PQDT-based measures to GSS

	(1) DoD	(2) DOE	(3) HHS	(4) NASA	(5) NSF	(6) USDA
Share of GSS graduates	0.879***	1.392***	1.014***	1.437***	1.021***	0.834***
	(0.043)	(0.051)	(0.018)	(0.050)	(0.026)	(0.028)
N	374	374	374	374	374	374
R^2	0.70	0.78	0.95	0.89	0.82	0.94
P-val: $\beta = 1$	0.01	0.00	0.42	0.00	0.43	0.00

Notes: Table correlates (i) the share of PhD graduates in each major field and year which we measure as having been supported by a given agency to (ii) the GSS-reported share of enrolled PhD students in that field and year reporting support from that agency. Sample includes 17 major fields measured between 2000 and 2022. Variables are not perfectly comparable because one measures PhD graduates and the other measures enrolled students, but if our dissertation-based measurement of PhD support is precise, we anticipate they will strongly correlate, and if complete, regression coefficients ($\hat{\beta}$) should be near 1. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

A.6 Limitations

As we note in the paper, one limitation of PQDT is a recent decline in the number of universities which submit dissertations to ProQuest. This attrition is not necessarily problematic for most of our analysis: the PQDT sample still appears to cover 90% of the population in recent years, and the sample attrition is not related to funding sources and is therefore unlikely to influence our findings. On the margin, however, it does preclude a complete accounting. In Table A.12 we list the principal universities in our sample and provide a window into which ones appear to have ended automatic or mandatory ProQuest licensing. For parsimony we limit the list to institutions with at least 20 STEM graduates every year between 2000 and 2022. For each such institution we report the number of PQDT graduates in our post-2000 sample, the number of SED graduates, and the difference and ratio, as well as the last year we substantially observe the university in the PQDT sample (defined as the last year where the number of PQDT graduates is >10% of the SED total[11]). Note that minor differences in PQDT and SED counts may arise from discrepancies in STEM field definitions and accounting periods (calendar years vs. academic years, respectively).

Consistent with Figure A.1, the evidence in Table A.12 suggests that the PQDT sample is relatively complete. Some exceptions nevertheless stand out, including the following schools (among others): Texas Tech (which exits the sample after 2004), UT Southwestern (2004), U. of Tennessee (2007), Brown U. (2010), U. of Central Florida (2010), Georgia Tech (2012), U. of Connecticut (2012), U. of Texas Medical Branch (2012), U. of Virginia (2013), Oregon State U. (2013), U. of Oklahoma (2013), UMass Amherst (2014), Georgia State U. (2014), Virginia Tech (2016), and Baylor College of Medicine (which is never reported in PQDT).

[11] We set this threshold to reflect observed evidence that even after a university ends its agreement with ProQuest, some graduates occasionally submit their own dissertations to PQDT anyway

Table A.12: PQDT vs. SED STEM PhD counts for principal universities in this paper's sample

University	Carnegie Classification	Total graduates, 2000-2022 PQDT Count	SED Count	PQDT:SED	Years Missing	Final year in PQDT
Arizona State U.	R1	4,778	4,184	114%	0	
Auburn U.	R1	2,763	2,778	99%	0	
Baylor C. of Medicine	Med.	1	1,560	0%	23	n.a.
Boston U.	R1	4,281	4,081	105%	0	
Brandeis U.	R1	936	862	109%	0	
Brown U.	R1	1,013	2,394	42%	12	2010
C. of William and Mary	R2	627	634	99%	0	
CUNY Graduate Center	R1	2,546	2,399	106%	0	
Cal. Inst. of Tech.	R1	3,156	3,728	85%	0	
Carnegie Mellon U.	R1	3,339	4,559	73%	0	
Case Western Reserve U.	R1	3,015	3,266	92%	0	
Clemson U.	R1	2,738	2,715	101%	0	
Colorado School of Mines	R1	1,549	1,497	103%	0	
Colorado State U.	R1	3,507	3,384	104%	0	
Columbia U.	R1	5,757	5,744	100%	0	
Cornell U.	R1	6,704	7,262	92%	0	
Weill Cornell Med. Coll.	Med.	1,126	1,104	102%	0	
Dartmouth C.	R1	1,529	1,556	98%	0	
Drexel U.	R1	2,036	2,017	101%	0	
Duke U.	R1	4,691	4,492	104%	0	
Emory U.	R1	2,292	2,306	99%	0	
Florida State U.	R1	2,804	2,689	104%	0	
George Mason U.	R1	1,740	1,764	99%	0	
George Washington U.	R1	2,122	1,920	111%	0	
Georgetown U.	R1	926	875	106%	0	
Ga. Inst. of Tech.	R1	4,549	8,743	52%	10	2012
Georgia State U.	R1	566	1,166	49%	8	2014
Harvard U.	R1	7,548	7,318	103%	0	
Howard U.	R1	639	697	92%	0	
Icahn School of Medicine	Med.	487	707	69%	11	
Ill. Inst. of Tech.	R2	1,303	1,191	109%	0	
Indiana U.	R1	3,184	3,100	103%	0	
Iowa State U.	R1	5,223	5,037	104%	0	
Johns Hopkins U.	R1	5,221	7,405	71%	2	2020
Kansas State U.	R1	2,313	2,268	102%	0	
Kent State U.	R1	937	1,103	85%	0	
Lehigh U.	R1	1,435	1,459	98%	0	
Louisiana State U.	R1	3,509	3,351	105%	0	
Loyola U. of Chicago	R1	617	588	105%	0	
Mass. Inst. of Tech.	R1	10,036	10,551	95%	0	
Michigan State U.	R1	5,577	5,474	102%	0	
Mississippi State U.	R1	1,712	1,680	102%	0	
Missouri U. of Sci. and Tech.	R2	1,576	1,623	97%	0	
NJ Inst. of Tech.	R1	907	1,310	69%	1	
New Mexico State U.	R1	1,142	1,146	100%	0	
New York U.	R1	3,272	3,080	106%	0	
North Carolina State U.	R1	7,149	6,995	102%	0	
North Dakota State U.	R1	1,218	1,183	103%	0	
Northeastern U.	R1	1,719	2,041	84%	6	
Northwestern U.	R1	5,278	5,199	102%	0	
Nova Southeastern U.	R2	448	864	52%	1	
Ohio State U.	R1	8,701	8,700	100%	0	
Ohio U.	R1	1,032	982	105%	0	
Oklahoma State U.	R1	2,148	2,134	101%	0	
Old Dominion U.	R1	1,125	1,069	105%	0	
Oregon State U.	R1	1,880	3,493	54%	9	2013
Pennsylvania State U.	R1	8,258	8,646	96%	0	
Princeton U.	R1	4,077	3,958	103%	0	
Purdue U.	R1	10,022	10,253	98%	0	
Rensselaer Polytech. Inst.	R1	2,603	2,586	101%	0	
Rice U.	R1	2,779	2,672	104%	0	
Rutgers U., New Brunswick	R1	4,762	5,413	88%	0	
Binghamton U.	R1	1,290	1,201	107%	0	
Stony Brook U.	R1	4,106	3,876	106%	0	

U. Albany	R1	1,200	1,159	104%	0	
U. Buffalo	R1	3,681	3,636	101%	0	
Southern Illinois U.	R1	1,086	1,039	105%	0	
Stanford U.	R1	10,391	10,594	98%	0	
Syracuse U.	R1	1,357	1,328	102%	0	
Temple U.	R1	1,662	1,594	104%	0	
Texas A&M U.	R1	8,388	9,385	89%	0	
Texas Tech U.	R1	336	2,262	15%	18	2004
Texas Woman's U.	R2	630	800	79%	0	
Tufts U.	R1	1,582	1,794	88%	0	
Tulane U.	R1	1,483	1,505	99%	0	
U. Akron	R2	1,294	1,585	82%	0	
U. Alabama, Birmingham	R1	2,717	2,640	103%	0	
U. Alabama, Huntsville	R1	596	712	84%	0	
U. Alabama, Tuscaloosa	R1	1,482	1,362	109%	0	
U. Arizona	R1	5,305	5,214	102%	0	
U. Arkansas	R1	1,784	1,779	100%	0	
U. California, Berkeley	R1	10,831	11,243	96%	0	
U. California, Davis	R1	7,830	7,736	101%	0	
U. California, Irvine	R1	4,907	4,707	104%	0	
U. California, Los Angeles	R1	8,897	8,499	105%	0	
U. California, Riverside	R1	3,086	2,944	105%	0	
U. California, San Diego	R1	6,852	6,695	102%	0	
U. California, San Francisco	R2	2,426	2,463	98%	0	
U. California, Santa Barbara	R1	3,795	3,616	105%	0	
U. California, Santa Cruz	R1	1,972	1,930	102%	0	
U. Central Florida	R1	774	2,572	30%	12	2010
U. Chicago	R1	3,427	3,235	106%	0	
U. Cincinnati	R1	3,184	3,111	102%	0	
U. Colorado, Boulder	R1	4,805	4,672	103%	0	
U. Colorado, Denver	R1	1,685	1,618	104%	0	
U. Connecticut	R1	2,086	3,883	54%	10	2012
U. Delaware	R1	2,949	2,842	104%	0	
U. Florida	R1	8,826	9,678	91%	0	
U. Georgia	R1	883	4,055	22%	16	
U. Hawaii, Manoa	R1	1,892	1,709	111%	0	
U. Houston	R1	1,680	2,961	57%	7	2020
U. Idaho	R1	1,052	1,054	100%	0	
U. Illinois, Chicago	R1	4,118	4,066	101%	0	
U. Illinois, Urbana-Champaign	R1	8,769	10,340	85%	0	
U. Iowa	R1	3,870	3,834	101%	0	
U. Kansas	R1	2,568	2,493	103%	0	
U. Kentucky	R1	2,277	3,385	67%	4	2018
U. Maine	R1	750	801	94%	0	
U. Maryland, Baltimore	R2	1,073	1,479	73%	0	
U. Maryland, Baltimore County	R1	996	1,153	86%	1	
U. Maryland, College Park	R1	6,902	6,798	102%	0	
U. Massachusetts, Amherst	R1	2,034	3,354	61%	8	2014
U. Miami	R1	1,909	2,018	95%	2	2020
U. Michigan, Ann Arbor	R1	9,343	11,154	84%	0	
U. Minnesota, Twin Cities	R1	8,770	8,700	101%	0	
U. Mississippi, Oxford	R1	1,171	1,206	97%	0	
U. Missouri, Columbia	R1	2,179	3,114	70%	0	
U. Nebraska, Lincoln	R1	2,878	2,727	106%	0	
U. Nevada, Reno	R1	1,298	1,231	105%	0	
U. New Hampshire	R1	897	861	104%	0	
U. New Mexico, Albuquerque	R1	2,148	2,110	102%	0	
U. North Carolina, Chapel Hill	R1	6,295	6,050	104%	0	
U. North Texas	R1	866	1,197	72%	0	
U. Notre Dame	R1	2,477	2,441	101%	0	
U. Oklahoma	R1	990	2,290	43%	9	2013
U. Oregon	R1	1,254	1,152	109%	0	
U. Pennsylvania	R1	5,388	5,104	106%	0	
U. Pittsburgh	R1	5,145	5,196	99%	0	
U. Rhode Island	R1	1,182	1,162	102%	0	
U. Rochester	R1	2,962	2,990	99%	0	
U. South Carolina	R1	2,860	2,876	99%	0	
U. South Florida	R1	2,973	2,731	109%	0	

U. Southern California	R1	5,203	5,772	90%	0	
U. Tennessee, Knoxville	R1	795	3,553	22%	15	2007
U. Texas HSC, Houston	Med.	1,228	2,283	54%	4	2018
U. Texas HSC, San Antonio	Med.	587	815	72%	0	
U. Texas Medical Branch	Med.	395	821	48%	10	2012
U. Texas Southwestern Med. Center	Med.	197	1,379	14%	18	2004
U. Texas, Arlington	R1	1,909	2,031	94%	0	
U. Texas, Austin	R1	4,636	8,773	53%	6	
U. Texas, Dallas	R1	2,062	1,981	104%	0	
U. Toledo	R1	1,250	1,346	93%	0	
U. Utah	R1	4,250	4,203	101%	0	
U. Virginia, Charlottesville	R1	2,123	3,759	56%	9	2013
U. Washington, Seattle	R1	8,791	8,751	100%	0	
U. Wisconsin, Madison	R1	10,070	9,926	101%	0	
U. Wisconsin, Milwaukee	R1	1,246	1,221	102%	0	
U. Wyoming	R1	1,059	1,044	101%	0	
Utah State U.	R1	1,167	1,090	107%	0	
Vanderbilt U.	R1	3,156	3,498	90%	0	
Virginia Commonwealth U.	R1	1,500	2,058	73%	4	2018
Virginia Polytech. Inst.	R1	3,146	5,974	53%	6	2016
Wake Forest U.	R2	950	954	100%	0	
Washington State U.	R1	2,875	2,781	103%	0	
Washington U., St. Louis	R1	3,706	3,568	104%	0	
Wayne State U.	R1	2,532	2,547	99%	0	
West Virginia U.	R1	1,994	2,013	99%	0	
Yale U.	R1	5,172	4,205	123%	0	

Notes: Table lists universities in our PQDT sample, filtering to those with least 20 STEM graduates every year from 2000 to 2022. Table provides number of associated PQDT- and SED-reported STEM grads over this period, as well as the number of years for which we observe graduates in PQDT and the final year we observe the university in the PQDT sample, if the university permanently exits the sample before 2022. We consider a university-year missing if the PQDT graduate count is less than 10% of the SED count. Carnegie designations are as follows: R1 = Very high research activity universities, R2 = High research activity universities, Med. = Medical schools and centers.

Though these gaps are modest, there is nevertheless a possibility of filling them from other sources—especially by retrieving (scraping) dissertations and information about their authors from university repositories. Doing so is a substantial challenge, due to structural differences across them, which necessitates a custom program for each university.

To better understand the feasibility of doing so, and how many additional graduates and dissertations it could deliver, we undertook this data collection. We specifically crawled online repositories for the universities listed in Table A.13 for years where believed we were missing a significant number of graduates (based on our SED benchmark), and gathered data on all graduates from those years, targeting all fields PQDT provides (e.g., name, title, abstract, subject(s), and dissertation text if available). In this case, we mapped subjects to SED major fields using an LLM. We then used these data to identify doctoral degrees issued in STEM fields. Reproducing Figure A.1 in Figure A.14, we see that this procedure appears to close roughly half the gap between our PQDT sample and SED graduate counts in recent years.

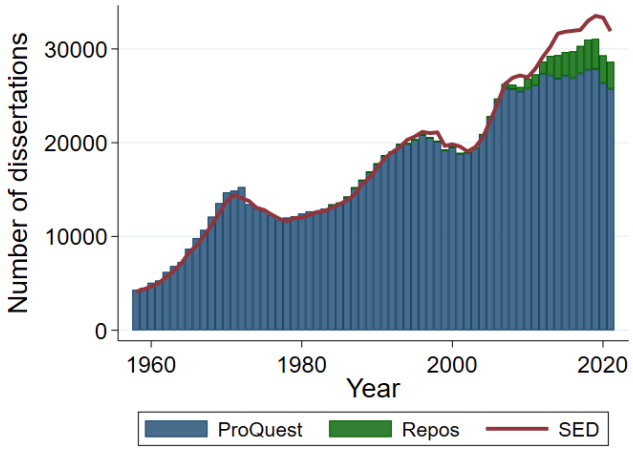
We do so for now as a proof of concept, especially for assessing whether this sampling can be extended forward. We opt to not append these graduates to the PQDT-derived sample in this paper primarily to maintain a consistent sampling procedure, especially with respect to the individuals accounted for, the subjects they are associated with (whose data generating process may vary between PQDT and university repositories), and how we allocate them to SED major fields, which will affect the calculated overall STEM graduate counts.

Table A.13: Backfilled graduates retrieved from university repositories

University	Retrieved graduates	First year	Last year
The U. of Texas at Austin	6,050	2010	2024
Ga. Inst. of Tech.	5,754	2013	2024
U. Tennessee, Knoxville	3,689	2007	2025
Virginia Polytech. Inst.	3,429	2016	2025
U. of Georgia	2,783	2002	2019
U. Central Florida	2,446	2010	2025
U. Houston	2,137	2012	2025
Texas Tech U.	2,128	2005	2025
U. Massachusetts, Amherst	1,868	2014	2025
U. Connecticut	1,532	2013	2020
U. Texas Southwestern Med. Center	1,357	2005	2024
U. Kentucky	1,163	2018	2025
U. Oklahoma	1,137	2014	2024
U. Texas HSC, Houston	813	2010	2025
U. Louisville	804	2014	2025
Virginia Commonwealth U.	638	2018	2025
Rockefeller U.	580	2004	2025
Wright State U.	388	2011	2024
Med. Coll. Ga.	191	2012	2022
South Dakota State U.	172	2011	2025
U. Texas Medical Branch	65	2013	2024

Notes: Table lists the set of universities and years which are absent from PQDT entirely (no information on their graduates; per Table A.12) and for which we can backfill data on their STEM PhD graduates from university repositories.

Figure A.14: ProQuest vs. SED graduates in STEM fields, with backfilled graduates



Notes: Figure shows annual PQDT graduate counts in the physical and life sciences and engineering (blue bars) plus graduates we can backfill from university repositories (stacked green bars, see Table A.13) versus SED counts (red line) for comparison. This backfill can close roughly half the gap between PQDT and SED in recent years.

B Supplementary Results

B.1 Foreign share of U.S. STEM PhD graduates

In this paper, we have restricted our attention to which PhDs the U.S. higher education system trains and who supports them. An important parallel set of questions is where they come from and where they go after their studies conclude. Immigrant scientists are a subject of perennial interest, and a substantial body of research indicates that immigrant scientists (including those trained in the U.S.) make outsize contributions to national science and innovation.

Though not our principal focus here—we instead examine patterns in PhD graduate out-migration in concurrent work (Shvadron et al. 2025b)—some context on the foreign-born PhD graduate population may be useful. As Figure B.1 shows, the foreign national share of U.S. STEM PhD graduates is currently around 45%, based on citizenship reported in the SED (an annual census of U.S. PhD graduates). These shares vary by field, with higher shares in engineering and lower shares in biomedicine, but across all fields are large and have increased since the beginning of our study period. Foreign nationals are not eligible for some categories of U.S. federal research funding (namely, fellowship funding, such as from the NSF or NIH), but they are eligible for other categories of federal support, including research grants, assistantships, and more.

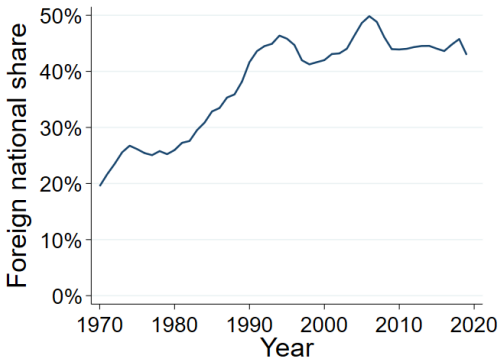
A structural limitation of the PQDT data is that they do not systematically measure or report country of origin (nor any other demographic features). However, this data gap does not necessarily rule out the possibility of identifying foreign-origin graduates; as MacGarvie (2007) has argued, information included in students' dissertations reporting lower degree institutions can suggestively indicate graduates' country of origin. Though some graduates may complete a bachelor's degree in a country other than their home country—including many in the U.S. itself—this information presents an opportunity to make progress in identifying foreign origin students through their dissertations (and more generally, in the absence of administrative/SED data).

We processed the dissertations in our sample to extract lower degree institutions, measuring the degree level (bachelor's and master's equivalents, henceforth denoted as BA and MA, though many degrees will be in science and engineering, including BS and MS degrees) and country where listed. We find this information in dissertations of many students (around 10% of the sample, and nearly 20% of post-2000 graduates) at many universities (70-90 per year after 2000)—substantially more than the 11 universities recognized by MacGarvie (2007) as being systematically reported. Because foreign trainees and immigration patterns are not a principal focus of this paper, we have used these data only in exploratory analysis. But within the sample of graduates for whom we have lower degree institution data, we can, for example, calculate the number of such graduates by field or with dissertation work related to specific critical technologies, and the share of such graduates with non-U.S. lower degree institutions, suggesting foreign origin.

In Table B.2 we do so by broad and major field, and in Table B.3 for critical technology areas. Note that these data comprise a roughly 20% sample, and may not be representative, but

they nevertheless provide a datapoint. The (apparent) foreign-origin share of U.S. PhD graduates varies substantially across fields, with roughly 20% in the biological sciences having a non-US BA (or equivalent), and roughly 60% in civil engineering. Across all fields and technology areas, a substantially larger share of graduates with US BAs have government support during their PhD than those with foreign BAs—partly due to eligibility restrictions. However, non-US BA graduates also report US government support for their research at high rates, reflecting that the U.S. government funds PhD research broadly. Moreover, the returns to these investments often accrue domestically: as we show in Shvadron et al. (2025b), a substantial share of these (apparent) foreign-origin graduates continue working in the U.S. for a decade or more post-graduation.

Figure B.1: Foreign share of U.S. STEM PhD graduates, 1970-2019 (SED)



Notes: Figure shows share of STEM PhD graduates over time who are foreign nationals, from 1970 to 2019. Foreign national share computed as the share of graduates who are temporary or permanent U.S. residents, using the NCSES Doctorate Record File (which is in turn compiled from the SED).

Table B.1: Foreign share of PhD graduates by field and decade

	Field	1970s	1980s	1990s	2000s	2010s
Broad fields	Life Sci.	19%	20%	34%	32%	31%
	Physical Sci.	21%	27%	40%	44%	41%
	Math/Comp. Sci.	22%	42%	51%	56%	56%
	Engineering Sci.	39%	55%	57%	63%	59%
Major fields	Agr. Sci.	36%	38%	51%	47%	46%
	Bio. Sci.	14%	15%	32%	32%	30%
	Health Sci.	17%	18%	25%	26%	25%
	Chemistry.	19%	25%	39%	43%	41%
	Geo. Sci.	19%	22%	35%	37%	34%
	Physics	23%	34%	45%	50%	45%
	Maths.	22%	43%	52%	54%	50%
	Comp. Sci.	22%	40%	50%	59%	61%
	Aero. Eng.	35%	56%	46%	55%	40%
	Bio. Eng.	18%	26%	37%	40%	36%
	Chem. Eng.	43%	48%	52%	54%	53%
	Civil Eng.	53%	64%	67%	69%	69%
	Elec. Eng.	35%	54%	57%	71%	72%
	Ind. Eng.	33%	64%	61%	71%	72%

Notes: Figure shows share of STEM PhD graduates across broad and major fields who are foreign nationals, by decade, from 1970 to 2019. Foreign national share is computed as the share of graduates who are temporary or permanent U.S. residents, using the NCSES Doctorate Record File (which is in turn compiled from the SED).

Table B.2: Foreign pre-PhD degree share of graduates, by field, 2000-2019

			Share of graduates with non-US ...		Share w/ USG support, by BA country:	
	Field	Sample size	BA	MA	US	Non-US
Broad fields	Life Sci.	26,916	23%	13%	46%	34%
	Physical Sci.	12,010	33%	16%	49%	37%
	Math/Comp. Sci.	7,464	44%	19%	29%	22%
	Engineering Sci.	18,092	48%	22%	47%	37%
Major fields	Agr. Sci.	3,031	35%	20%	42%	34%
	Bio. Sci.	18,361	21%	12%	48%	34%
	Health Sci.	5,524	19%	10%	41%	34%
	Chemistry	5,254	35%	17%	40%	33%
	Geo. Sci.	2,397	28%	17%	73%	57%
	Physics	4,359	34%	16%	45%	34%
	Maths.	4,140	39%	17%	25%	17%
	Comp. Sci.	3,324	51%	21%	36%	26%
	Aero. Eng.	891	29%	17%	62%	46%
	Bio. Eng.	2,261	37%	15%	46%	32%
	Chem. Eng.	1,603	44%	18%	51%	45%
	Civil Eng.	1,733	61%	37%	49%	38%
	Elec. Eng.	3,881	53%	22%	41%	31%
	Ind. Eng.	528	57%	26%	35%	21%

Notes: Table lists (i) the number of U.S. PhD graduates in each broad and major field over the 2000-2019 period for whom we can measure pre-PhD locations, (ii) the share with a foreign pre-PhD degree (BA or MA, and equivalents), and the share of those with a US vs. non-US BA who acknowledge some U.S. federal government support.

Table B.3: Foreign pre-PhD degree share of graduates, by technology, 2000-2019

		Share of graduates with non-US ...		Share w/ USG support, by BA country:	
Technology	Sample size	BA	MA	US	Non-US
Advanced computing	8,307	49%	21%	43%	31%
Advanced manufacturing	1,400	50%	23%	55%	48%
Artificial intelligence	3,031	50%	22%	42%	33%
Autonomous systems	854	41%	20%	51%	37%
Biotechnology	11,860	31%	15%	47%	34%
Integrated communications	1,770	57%	24%	37%	28%
Data and cyber security	653	58%	20%	39%	26%
Directed energy	156	35%	19%	38%	38%
Hypersonics	29	28%	14%	76%	50%
Human-machine interfaces	571	40%	17%	43%	36%
Materials science	6,349	47%	22%	53%	42%
Microelectronics	3,960	49%	21%	44%	36%
PNT technologies	409	37%	18%	51%	32%
Quantum science	1,119	38%	15%	38%	32%
Renewable energy	2,124	45%	21%	51%	42%
Networked sensing	2,870	42%	19%	53%	41%
Space technology	972	27%	15%	64%	54%
Gas turbine engines	134	47%	17%	56%	43%

Notes: Table lists (i) the number of U.S. PhD graduates in each critical technology area over the 2000-2019 period for whom we can measure pre-PhD locations, (ii) the share with a foreign pre-PhD degree (BA or MA, and equivalents), and the share of those with a US vs. non-US BA who acknowledge some U.S. federal government support.

B.2 Female share of U.S. STEM PhD graduates

In addition to our measures of foreign vs. U.S. origin, we also develop measures of PhD graduate gender, using the name-based classifier `genderize.io`, which assigns names a gender probability based on empirical frequencies in a training corpus. In Figures B.2 and B.3 we report the gender composition of graduates in our sample between 2000 and 2022 by major field and by technology area, respectively. In doing so, we restrict to the 80% of our sample that our classifier predicts is either male or female with >90% probability.

Figure B.2: Female share of graduates, by field, 2000-2022

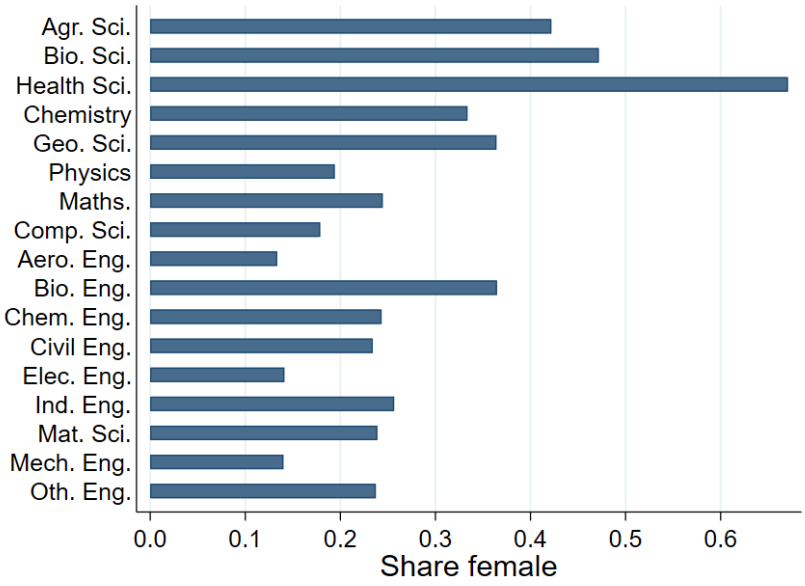
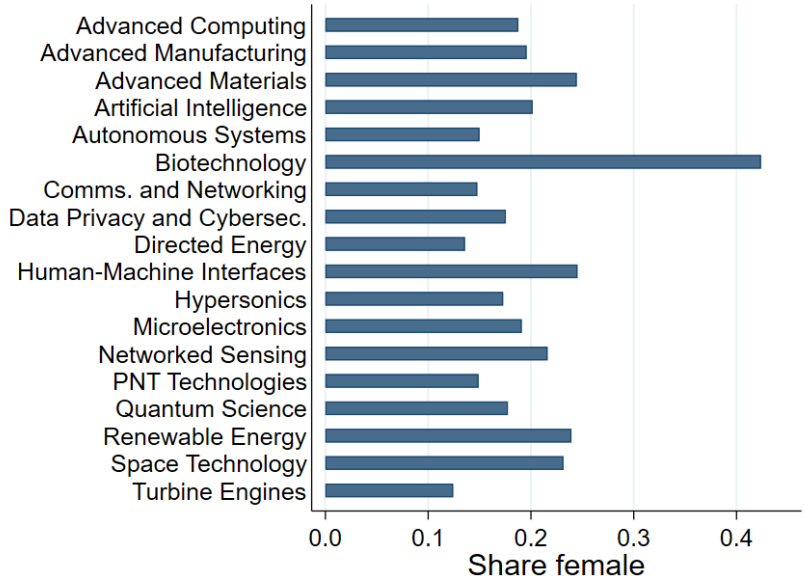


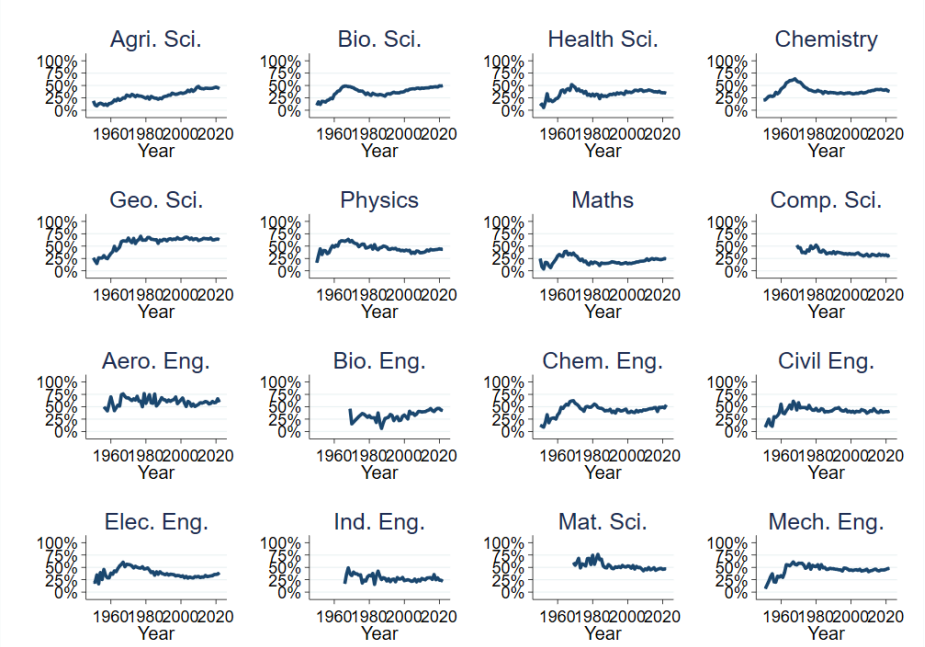
Figure B.3: Female share of graduates, by technology, 2000-2022



B.3 U.S. government support over time, by field

Figure B.4 reports the share of graduates in each major field between 1950 and 2022 acknowledging U.S. government support, conditional on having any dissertation acknowledgments.

Figure B.4: U.S. government-funded share of dissertations, by field, 1950-2022



Notes: Figure shows government-funded share of dissertations by major field, from 1950 to 2022.

B.4 Support rates adjusted for potential underreporting

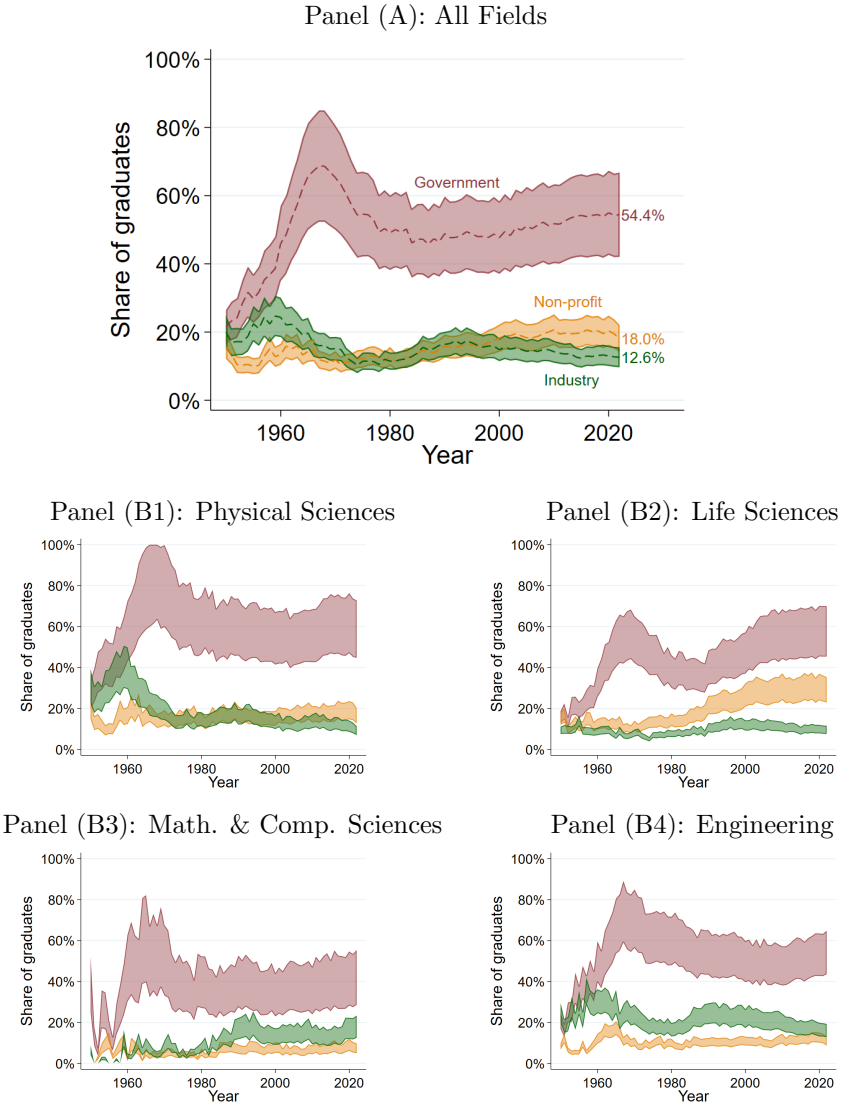
One challenge to using dissertations to measure PhD graduates' sources of support is the risk of underreporting, which we conjecture is more likely inadvertent than deliberate. Insofar as such omissions are as good as random, correlational results are unlikely to be affected, but the raw data will underestimate true rates of support from different organizations.

One approach to correct for underreporting is to reweight the sample to produce representative measures. We use our NSF GRFP validation data to estimate the likelihood of accurate reporting as a function of observables. More concretely, we run a probit regression of whether a GRFP awardee reports NSF support in their dissertation on degree field fixed effects, and use the results to estimate the likelihood of an accurate report.[12] We use these probabilities to calculate inverse propensity weights (IPW), which we can apply to our data to overweight low propensity-to-report groups (and underweight others)—similar to how survey weights are used to attempt to make non-representative survey results reflect the underlying population. Though re-weighting is imperfect and second-best to accurate reports, it helps us put an upper bound on latent values.

[12]This propensity specification was selected after evaluating a range of alternatives. Some predictors which we could in principle include, like university fixed effects, are too sparse to use in propensity estimation because they would create an incidental parameters problem (e.g., for small institutions). Others, like decade fixed effects, have large standard errors and are thus prone to misdirecting the reweighting.

Figures B.5 to B.7 and Table B.4 reproduce results from similar charts and tables in the paper using IPW-weighted shares of graduates with government, industry, and non-profit support. Of these, Figure B.5 is in our view the key figure, adapting Figure 5 of the paper to show both raw and IPW-weighted shares of graduates with each source of support over time as lower and upper bounds, and the midpoint between them (dashed line). Figures B.8 and B.9 provide IPW-weighted shares of graduates with support from specific federal agencies.

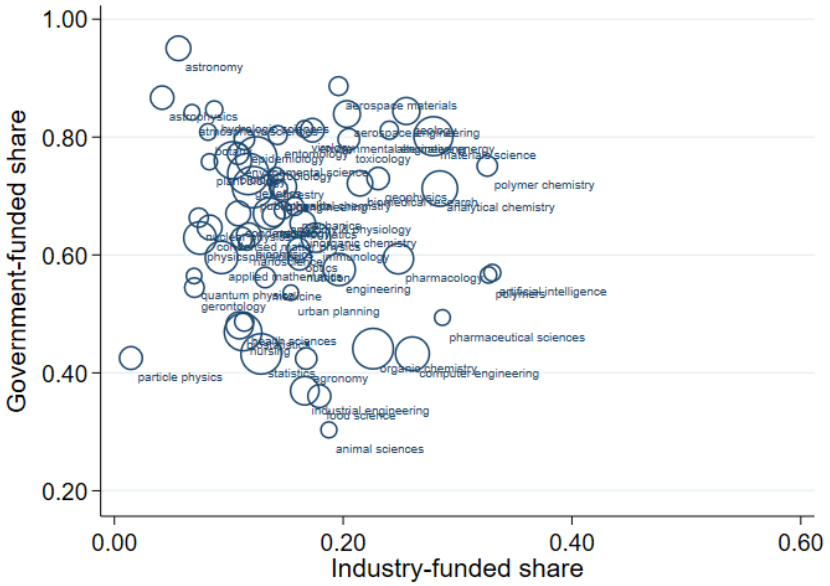
Figure B.5: Bounds on share of PhD graduates supported over time, 1950-2022



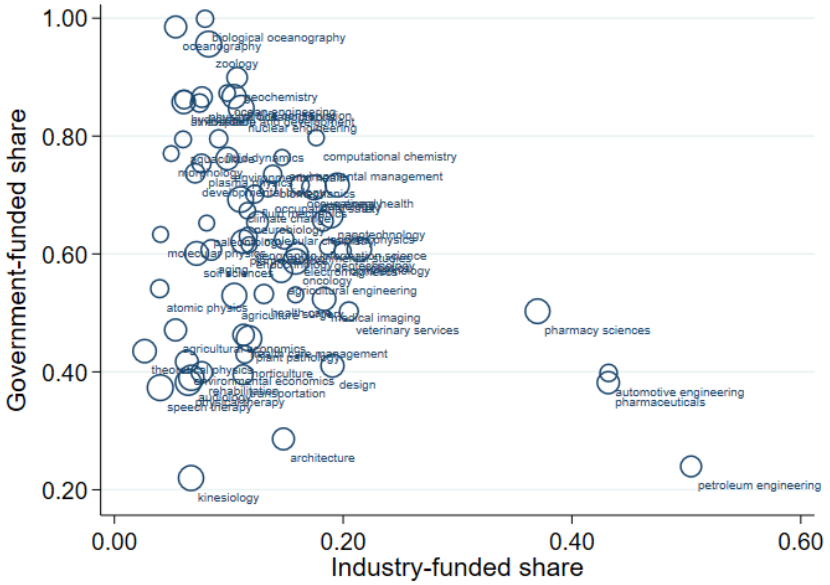
Notes: Figure shows the share of PhD graduates reporting their sources of support who are supported by (i) U.S. federal government agencies, (ii) firms, and (iii) non-profit organizations, from 1950 to 2022. Shaded area demarcates lower and upper bounds, where lower bounds are shares observed in the raw data and upper bounds use inverse propensity weighting to account for underreporting; dotted lines mark the midpoint between bounds. Panel (A) presents overall results. Panels (B1) to (B4) present results by broad field (physical sciences, life sciences, mathematical and computer sciences, and engineering). Sample restricted to dissertations with acknowledgments (>98% of all dissertations).

Figure B.6: Share of graduates with government vs. industry support, by subject, 2000-2022; estimated with inverse propensity weighting to account for underreporting

Panel (A): Very large subjects
(>1200 graduates total)



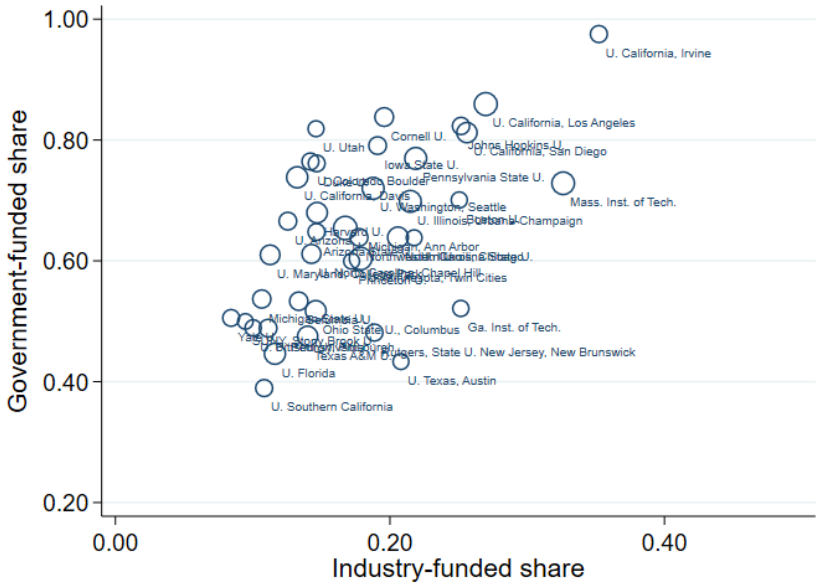
Panel (B): Large subjects
(400-1200 graduates total)



Notes: Figure plots the share of graduates in a given subject with government support against the share with industry support. Panel (A) does so for subjects with >1200 graduates between 2000 and 2022, and Panel (B) for subjects with 400 to 1200 graduates. Marker size proportional to subjects' number of graduates. Population shares calculated using inverse propensity weights to account for potential underreporting.

Figure B.7: Share of graduates with government vs. industry support, by university, 2000-2022; estimated with inverse propensity weighting to account for underreporting

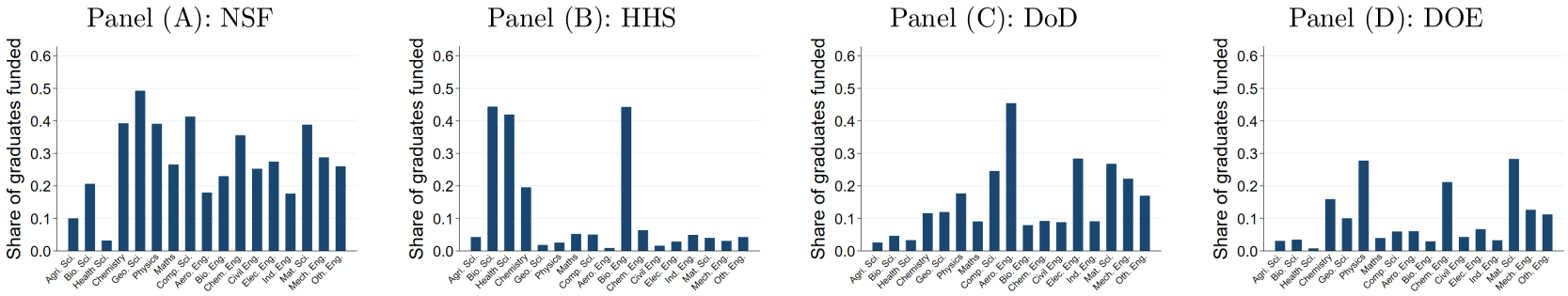
Panel (A): Very large universities
(>2000 graduates total)



Panel (B): Large universities
(1000-2000 graduates total)

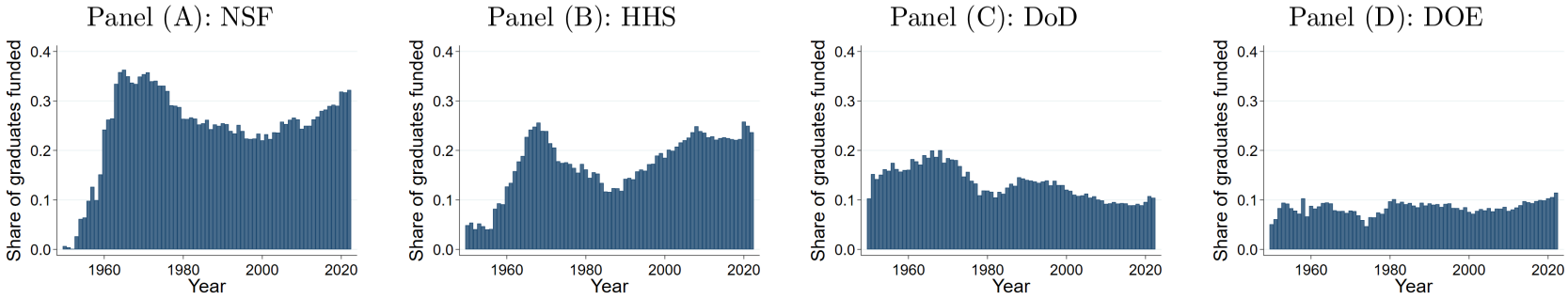
Notes: Figure plots the share of graduates from a given university with government support against the share with industry support. Panel (A) does so for universities with >2000 graduates between 2000 and 2022, and Panel (B) for universities with 1000 to 2000 graduates. Marker size proportional to universities' number of graduates. Population shares calculated using inverse propensity weights to account for potential underreporting.

Figure B.8: Major fields' share of graduates supported by select federal agencies, 1950-2022; estimated with inverse propensity weighting to account for underreporting



Notes: Figure shows the share of PhD graduates in individual major fields supported by each agency listed (NSF, HHS, DoD, DOE), illustrating differences in subject matter priorities across agencies. Population shares calculated using inverse propensity weights to account for potential underreporting.

Figure B.9: Share of all graduates supported by select federal agencies over time, 1950-2022; estimated with inverse propensity weighting to account for underreporting



Notes: Figure shows the share of PhD graduates supported over time by each agency listed (NSF, HHS, DoD, DOE), illustrating long-run changes in these agencies' investments in PhD training. Population shares calculated using inverse propensity weights to account for potential underreporting.

Table B.4: Correlation of share of graduates who have govt. support with share who have industry or non-profit support, 2000-2022; estimated across various units of analysis and estimated with inverse propensity weighting to account for underreporting

Panel (A): Estimated at the university-field and university-year level

	University-field				University-year			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Industry-funded share	0.373***	0.218***	0.227***	0.131**	0.461***	0.229***	0.486***	0.244***
	(0.046)	(0.048)	(0.056)	(0.053)	(0.028)	(0.026)	(0.028)	(0.026)
Nonprof-funded share	0.729***	0.637***	0.890***	0.763***	0.705***	0.471***	0.696***	0.454***
	(0.035)	(0.030)	(0.043)	(0.039)	(0.019)	(0.022)	(0.019)	(0.021)
N	4055	4043	4055	4043	6251	6241	6251	6241
R^2	0.33	0.60	0.49	0.72	0.33	0.73	0.35	0.75
Univ. FEs		Y		Y		Y		Y
Field FEs			Y	Y				
Year FEs							Y	Y

Panel (B): Estimated at the university-field-year level

	University-field-year							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Industry-funded share	0.220***	0.141***	0.204***	0.222***	0.140***	0.141***	0.209***	0.143***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Nonprof-funded share	0.541***	0.452***	0.545***	0.540***	0.422***	0.450***	0.540***	0.416***
	(0.008)	(0.007)	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)	(0.007)
N	56006	55997	56006	56006	55997	55997	56006	55997
R^2	0.16	0.30	0.23	0.16	0.35	0.30	0.23	0.36
Univ. FEs		Y			Y	Y		Y
Field FEs			Y		Y		Y	Y
Year FEs				Y		Y	Y	Y

Notes: Table estimates the correlation of government support rates against industry and non-profit support rates at the university-field and -year levels (Panel A) and the university-field-year level (Panel B). Each column controls for alternative sets of fixed effects, and the estimation period in all columns is 2000-2022. Population shares calculated using inverse propensity weights to account for potential underreporting. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

B.5 Top universities in critical technologies, using data through 2012 only

Tables 4 and 5 of the paper list the top universities in specific technology areas by number of graduates produced post-2000. One limitation of this accounting, however, is that for a small number of schools—including some potentially important schools—we have limited data from the mid-2010s onwards (e.g., Georgia Tech exits the PQDT sample after 2012, and Virginia Tech after 2016). These schools are thus undercounted in more recent years.

The systematic patterns in our paper (such as the overall distribution of support, or regression estimates of its relationship to PhD production) are unaffected by this attrition, given that it is relatively modest. However, tables that rank individual universities (e.g., in their training of PhDs related to specific technologies) may see their specific contents affected: schools that exit the sample early may be omitted from these lists merely due to this attrition.

In order to evaluate how this sample melt affects our university rankings, in Tables B.5 and B.6 we reproduce these tables for 2000-2012 only. Though this is an imperfect comparison—measuring PhD production through 2012 in technology areas which were considered critical or emerging over a decade later in 2024—it can mitigate the impacts of attrition on our findings. The results are similar to what was found for the full 2000-2022 sample.

Table B.5: Top 5 universities and sponsors of PhD graduates in U.S. critical technology areas, 2000-2012

Advanced Computing

Top universities

University	Count
Mass. Inst. of Tech.	915
Ga. Inst. of Tech.	897
Stanford U.	885
U. Illinois, Urbana-Champaign	809
U. California, Berkeley	738

Top funders

Name	Sector	Count
NSF	Govt	5,240
DoD	Govt	3,405
HHS	Govt	1,417
DOE	Govt	1,105
NASA	Govt	959

Advanced Manufacturing

Top universities

University	Count
Mass. Inst. of Tech.	247
U. California, Berkeley	241
Ga. Inst. of Tech.	240
U. Michigan, Ann Arbor	212
U. Illinois, Urbana-Champaign	197

Top funders

Name	Sector	Count
NSF	Govt	1,366
DoD	Govt	851
DOE	Govt	462
HHS	Govt	169
NASA	Govt	151

Advanced Materials

Top universities

University	Count
Ga. Inst. of Tech.	865
Mass. Inst. of Tech.	861
U. Illinois, Urbana-Champaign	824
Pennsylvania State U.	780
U. Florida	732

Top funders

Name	Sector	Count
NSF	Govt	6,908
DoD	Govt	4,051
DOE	Govt	3,867
HHS	Govt	989
NASA	Govt	859

Autonomous Systems

Top universities

University	Count
Carnegie Mellon U.	143
Mass. Inst. of Tech.	141
Stanford U.	108
Ga. Inst. of Tech.	106
U. California, Berkeley	75

Top funders

Name	Sector	Count
DoD	Govt	547
NSF	Govt	434
NASA	Govt	223
HHS	Govt	58
DOT	Govt	53

Biotechnology

Top universities

University	Count
U. Wisconsin-Madison	1,181
U. California, Berkeley	1,107
Mass. Inst. of Tech.	1,021
Johns Hopkins U.	1,016
Stanford U.	1,009

Top funders

Name	Sector	Count
HHS	Govt	12,545
NSF	Govt	7,650
DoD	Govt	2,113
DOE	Govt	1,869
USDA	Govt	1,662

Clean Energy Generation and Storage

Top universities

University	Count
U. California, Berkeley	200
Mass. Inst. of Tech.	182
Ga. Inst. of Tech.	175
Pennsylvania State U.	150
Texas A&M U.	138

Top funders

Name	Sector	Count
DOE	Govt	1,331
NSF	Govt	983
DoD	Govt	526
NASA	Govt	133
USDA	Govt	122

Communications and Networking

Top universities

University	Count
Ga. Inst. of Tech.	360
U. California, Los Angeles	331
Stanford U.	303
U. Southern California	236
U. California, Berkeley	224

Top funders

Name	Sector	Count
NSF	Govt	1,298
DoD	Govt	1,065
Intel	Firm	226
NASA	Govt	196
IBM	Firm	136

Data Privacy and Cybersecurity

Top universities

University	Count
Ga. Inst. of Tech.	55
Purdue U., West Lafayette	55
U. Maryland, College Park	47
U. California, Berkeley	45
Mass. Inst. of Tech.	44

Top funders

Name	Sector	Count
NSF	Govt	366
DoD	Govt	247
IBM	Firm	47
Intel	Firm	45
Microsoft	Firm	39

Microelectronics and Semiconductors

Top universities

University	Count
Stanford U.	754
U. California, Berkeley	731
Ga. Inst. of Tech.	682
Mass. Inst. of Tech.	596
U. Illinois, Urbana-Champaign	586

Top funders

Name	Sector	Count
NSF	Govt	3,790
DoD	Govt	2,671
DOE	Govt	1,633
Intel	Firm	612
NASA	Govt	427

Networked Sensing

Top universities

University	Count
Stanford U.	301
Mass. Inst. of Tech.	288
Ga. Inst. of Tech.	278
U. California, Berkeley	255
U. Michigan, Ann Arbor	248

Top funders

Name	Sector	Count
NSF	Govt	1,927
DoD	Govt	1,672
NASA	Govt	1,047
DOE	Govt	525
DOC	Govt	325

Quantum Science

Top universities

University	Count
Mass. Inst. of Tech.	183
Stanford U.	173
U. Illinois, Urbana-Champaign	157
U. California, Berkeley	140
Harvard U.	115

Top funders

Name	Sector	Count
NSF	Govt	830
DoD	Govt	515
DOE	Govt	378
HHS	Govt	78
NASA	Govt	78

Space Technology

Top universities

University	Count
U. Colorado Boulder	127
U. Michigan, Ann Arbor	120
Mass. Inst. of Tech.	116
Stanford U.	110
U. Maryland, College Park	92

Top funders

Name	Sector	Count
NASA	Govt	1,297
NSF	Govt	503
DoD	Govt	439
DOT	Govt	244
DOE	Govt	148

Notes: Table lists the top 5 universities and sponsors of graduates between 2000 and 2012 in the 12 largest OSTP critical technology areas. Graduates allocated to technology areas as described in text.

Table B.6: Top 15 universities and sponsors of PhD graduates in AI, 2000-2012

Top universities		Top funders		
University	Count	Name	Sector	Count
Carnegie Mellon U.	296	NSF	Govt	1,280
Mass. Inst. of Tech.	284	DoD	Govt	1,003
Ga. Inst. of Tech.	240	HHS	Govt	455
U. Illinois, Urbana-Champaign	204	NASA	Govt	227
Stanford U.	190	DOE	Govt	173
U. California, Berkeley	179	IBM	Firm	103
U. Southern California	176	Microsoft	Firm	90
U. Maryland, College Park	163	DOC	Govt	87
U. Washington, Seattle	136	Google	Firm	71
Purdue U., West Lafayette	123	Intel	Firm	66
U. Florida	123	DOT	Govt	63
U. California, Los Angeles	118	USDA	Govt	43
Ohio State U., Columbus	112	DOEd	Govt	42
Pennsylvania State U.	112	Boeing	Firm	39
Columbia U.	108	Department of the Interior	Govt	39

Notes: Table lists the top 15 universities and sponsors of graduates between 2000 and 2012 in artificial intelligence (AI). Graduates identified as AI-related based on critical technology assessment (see Section 3).

B.6 Regression results: robustness checks

B.6.1 First-stage regressions

Table B.7 presents first-stage regression results underlying Table 6 of the paper, relating the shift-share instrument for government-supported graduates to observed values. Columns in this table correspond one-for-one to columns in Table 6.

Table B.7: First-stage estimation underlying Table 6 of the paper

	Ln(USG-supported PhDs)				Ln(Past 20 years' ...)	
	(1)	(2)	(3)	(4)	(5)	(6)
					All PhDs	USG PhDs
Ln(USG-supported PhDs), shift-share	0.108***	0.226***	0.226***	4.763***		
	(0.009)	(0.019)	(0.019)	(0.350)		
Ln(Non-USG PhDs)	0.832***			0.264***		
	(0.057)			(0.008)		
Ln(Past 20 years' USG PhDs), shift-share					0.656***	0.676***
					(0.069)	(0.092)
N	901	901	901	57103	900	900
R^2	0.96	0.89	0.89	0.66	0.99	0.98
Field FEs	Y	Y	Y		Y	Y
Univ-Field FEs				Y		
Year FEs	Y	Y	Y	Y	Y	Y

Notes: Table presents first-stage estimations underlying Table 6, relating the shift-share instrument for government-supported graduates to observed values. Columns in this table correspond to columns in Table 6 and follow their specifications (see that table's notes for details). Column headers indicate the variables being predicted by the instrument, which appear in Table 6 as explanatory variables. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by major field (via wild bootstrap) in parentheses in Columns (1) to (3) and (5) to (6). SEs clustered by university in Column (4).

B.6.2 Robustness checks: Alternative instrument

Table B.8 provides a set of robustness checks on Table 6 of the paper, using a shift-share instrument based on agencies' field shares in the 6 to 10 years prior to a given graduation year (rather than in the 1 to 10 years prior, as in the paper). This instrument thus precedes enrollment of a given cohort but may nevertheless predict its level of support. The results are quantitatively and statistically similar, with point estimates which continue to imply a roughly one-for-one relationship between government-funded PhD graduates and total PhD graduates.

Table B.8: Relationship of federal support to PhD production at the field-year level, 1970-2022, with 6 to 10 year-ago instrument

	Ln(PhD graduates)				Ln(Publications)	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	All	Non-USG	All		
Ln(USG-supported PhDs)	0.445***	0.770***	0.615***	0.480***		
	(0.016)	(0.028)	(0.050)	(0.009)		
Ln(Non-USG PhDs)	0.527***			0.536***		
	(0.022)			(0.004)		
Ln(Past 20 years' PhDs)					0.332**	
					(0.146)	
Ln(Past 20 years' USG PhDs)						0.319**
						(0.142)
N	901	901	901	57103	890	890
F-stat	85.85	114.26	114.26	213.11	122.58	109.19
Field FEs	Y	Y	Y		Y	Y
Univ-Field FEs				Y		
Year FEs	Y	Y	Y	Y	Y	Y

Notes: Table estimates the relationship of annual PhD production in field-years to government-supported PhD graduates in those fields. Columns (1) and (2) relate government-supported graduates to total PhD graduates, and Column (3) to other PhD graduates. In Column (4), we reproduce Column (1) at the university-field-year level, in an analogous specification with university-field and year fixed effects. Columns (5) and (6) relate the stock of recent (past 20 year) PhD graduates in a given field to the annual flow of scientific output in that field. All columns estimate relationships by two-stage least squares, using a shift-share instrument as described in the text. Estimation sample covers the post-1970 period. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by major field (via wild bootstrap) in parentheses in Columns (1) to (3) and (5) to (6). SEs clustered by university in Column (4).

B.6.3 Robustness checks: Stable university sample

Table B.9 provides an additional set of robustness checks on Table 6 of the paper, constructing the sample solely from universities which are present in PQDT in all years of our study period. The results remain quantitatively and statistically similar to those in Table 6.

Table B.9: Relationship of federal support to PhD production at the field-year level, 1970-2022, estimated on a sample constructed from universities present in PQDT throughout this period

	Ln(PhD graduates)				Ln(Publications)	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	All	Non-USG	All		
Ln(USG-supported PhDs)	0.439***	0.771***	0.622***	0.470***		
	(0.013)	(0.021)	(0.038)	(0.009)		
Ln(Non-USG PhDs)	0.535***			0.538***		
	(0.019)			(0.006)		
Ln(Past 20 years' PhDs)					0.247**	
					(0.098)	
Ln(Past 20 years' USG PhDs)						0.243**
						(0.098)
N	900	900	900	52931	900	900
R^2	138.79	140.66	140.66	65.28	549.38	320.14
F-stat	Y	Y	Y		Y	Y
Field FEs				Y		
Univ-Field FEs	Y	Y	Y	Y	Y	Y

Notes: Table estimates the relationship of annual PhD production in field-years to government-supported PhD graduates in those fields. Columns (1) and (2) relate government-supported graduates to total PhD graduates, and Column (3) to other PhD graduates. In Column (4), we reproduce Column (1) at the university-field-year level, in an analogous specification with university-field and year fixed effects. Columns (5) and (6) relate the stock of recent (past 20 year) PhD graduates in a given field to the annual flow of scientific output in that field. All columns estimate relationships by two-stage least squares, using a shift-share instrument as described in the text. Estimation sample covers the post-1970 period; underlying data constructed from universities which are present in PQDT throughout this period. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by major field (via wild bootstrap) in parentheses in Columns (1) to (3) and (5) to (6). SEs clustered by university in Column (4).

Appendix References

Aiken, Catherine, James Dunham, Jennifer Melot, and Zachary Arnold. 2024. *Identifying Emerging Technologies in Research.* Center for Security and Emerging Technology working paper, available at https://cset.georgetown.edu/publication/identifying-emerging-technologies-in-research/.

Antman, Francisca M, Xuechao Qian, Kirk Doran, and Bruce A Weinberg. 2023. *Innovation Nation: Evidence from Broadening Access to Ph.D. Training in the U.S.*. Working Paper.

Arora, Ashish, Sharon Belenzon, Larisa C Cioaca, Lia Sheer, and Hansen Zhang. 2023. *The Effect of Public Science on Corporate R&D.* NBER Working Paper No. 31899.

Bikard, Michaël, Fiona Murray, and Joshua S Gans. 2015. "Exploring trade-offs in the organization of scientific work: Collaboration and scientific reward," *Management Science*, Vol. 61, No. 7, pp. 1473–1495.

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python.*: O'Reilly Media Inc.

Buffington, Catherine, Benjamin Cerf, Christina Jones, and Bruce A Weinberg. 2016. "STEM training and early career outcomes of female and male graduate students: Evidence from UMETRICS data linked to the 2010 census," *American Economic Review*, Vol. 106, No. 5, pp. 333–338.

Gat, Noam. 2023. "lm-format-enforcer," https://github.com/noamgat/lm-format-enforcer.

Jiang, Xuan, Joseph Staudt, and Bruce A Weinberg. 2023. *A tale of two fields? STEM career outcomes.* NBER Working Paper No. 31835.

Kim, Dahyun, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. "SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling."

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. "Efficient Memory Management for Large Language Model Serving with PagedAttention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles.*

MacGarvie, Megan. 2007. *Using Published Dissertations to Identify Graduates' Countries of Origin.* Working paper.

National Science Foundation. 2022. *Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS).* Available at https://ncses.nsf.gov/surveys/graduate-students-postdoctorates-s-e/2022.

Nguyen, Phuc, Ikuya Yamada, Natthawut Kertkeidkachorn, Ryutaro Ichise, and Hideaki Takeda. 2021. "SemTab 2021: Tabular Data Annotation with MTab Tool," in *SemTab@ISWC 2021*, Vol. 3103 of CEUR Workshop Proceedings, pp. 92–101: CEUR-WS.org, URL: http://ceur-ws.org/Vol-3103/paper8.pdf.

OSTP. 2024. *Critical and Emerging Technologies List Update.* White House Office of Science and Technology Policy (OSTP), available at `https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/02/Critical-and-Emerging-Technologies-List-2024-Update.pdf`.

Pal, Arka, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. "Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive," *arXiv preprint arXiv:2402.13228.*

Shvadron, Dror, Hansen Zhang, Lee Fleming, and Daniel P. Gross. 2025a. *Foreign Migration Patterns among U.S.-trained PhD Scientists.* Working paper.

Shvadron, Dror, Hansen Zhang, Lee Fleming, and Daniel P Gross. 2025b. "A Quarter of US-Trained Scientists Eventually Leave. Is the US Giving Away Its Edge?" *arXiv preprint arXiv:2512.11146.*

Toole, Andrew A and Dirk Czarnitzki. 2010. "Commercializing science: Is there a university "brain drain" from academic entrepreneurship?" *Management Science*, Vol. 56, No. 9, pp. 1599–1614.