PNAS RESEARCH ARTICLE | APPLIED PHYSICAL SCIENCES SOCIAL SCIENCES OPEN ACCESS

The rise of large language models and the direction and impact of US federal research funding

Yifan Qian[a,b,c,d], Zhe Wen[a,b,c,d], Alexander C. Furnas[a,b,c,d], Yue Bai[a,b,c,d], Erzhuo Shao[a,b,c,e], and Dashun Wang[a,b,c,d,e,1]

Affiliations are included on p. 8.

Edited by Mark Granovetter, Stanford University, Stanford, CA; received January 20, 2026; accepted June 26, 2026

Federal research funding shapes the direction, diversity, and impact of the US scientific enterprise. Large language models (LLMs) are rapidly diffusing into scientific practice, holding substantial promise while raising widespread concerns. Despite growing attention to AI use in scientific writing and evaluation, little is known about how the rise of LLMs is reshaping the public funding landscape. Here, we examine LLM involvement at key stages of the federal funding pipeline by combining two complementary data sources: confidential NSF and NIH proposal submissions from two large US R1 universities, including funded, unfunded, and pending proposals, and the full population of publicly released NSF and NIH awards. We find that LLM use rises sharply beginning in 2023 and exhibits a bimodal distribution, indicating a clear split between minimal and substantive use. Across both private submissions and public awards, higher LLM involvement is consistently associated with lower semantic distinctiveness, positioning projects closer to recently funded work within the same agency. The consequences of this shift are agency-dependent. LLM use is positively associated with proposal success and higher early-stage publication output at NIH, whereas no comparable associations are observed at NSF. Notably, the productivity gains at NIH are concentrated in nonhit papers rather than the most highly cited work. Together, these findings provide large-scale evidence that the rise of LLMs is reshaping how scientific ideas are positioned, selected, and translated into publicly funded research, with implications for portfolio governance, research diversity, and the long-run impact of science.

federal funding | generative AI | large language model | science of science | computational social science

Significance

How is generative AI shaping the public funding landscape? Using confidential NSF and NIH proposal submissions from two large US research universities, as well as the full population of publicly released NSF and NIH grant awards, we demonstrate that large language model (LLM) use has increased rapidly and is associated with proposals that are less semantically distinctive from recently funded work. At NIH, such proposals are more likely to be funded and to yield more publications, whereas at NSF, no such relationships are observed. These patterns indicate that LLM use is already reshaping how scientific ideas are articulated and evaluated in the federal funding system, with implications for research diversity, transparency, and public trust in the stewardship of taxpayer-supported science.

Federal research funding is the primary mechanism through which the United States converts public resources into scientific knowledge (1, 2). Through competitive grant programs at agencies such as the NSF and the NIH, decisions about the allocation of taxpayer dollars influence which scientific ideas receive support, which fields advance, and which investigators are able to initiate and sustain research programs (3–10). Thus, the federal funding process plays a central role in shaping the direction, diversity, and long-run impact of the US scientific enterprise and—given America's scientific leadership—the global research landscape as well (3–10). Understanding forces that influence this process is therefore essential not only for science policy and the rate and direction of scientific progress, but also for the stewardship and accountability of public investment in research (11–14).

Despite growing attention to the use of AI in science (15–27), little is known about how the rise of large language models (LLMs) is reshaping the public funding landscape. This gap is consequential, as funding decisions operate upstream of publication, hiring, and recognition, shaping which scientific ideas are pursued long before other forms of evaluation occur. Yet proposal texts, especially unfunded submissions, are typically confidential, making this stage of the funding pipeline difficult to study systematically at scale. As a result, we lack systematic evidence on how LLM adoption enters the funding pipeline, how it relates to proposal evaluation, and how it shapes the research that ultimately receives public support.

Two broader questions follow from recent work on AI in science (16, 26). One is whether the productivity gains associated with generative AI primarily reflect lower communication costs or broader improvements in scientific execution. Another is whether generative AI matters only for scientific outputs, such as manuscripts and citations, or is already reshaping the upstream processes through which ideas are proposed, evaluated, and selected for public support. Federal research funding provides a useful

Author contributions: Y.Q. and D.W. designed research; Y.Q., Z.W., A.C.F., Y.B., E.S., and D.W. performed research; Y.Q. contributed new reagents/analytic tools; Y.Q. and Z.W. analyzed data; and Y.Q., A.C.F., and D.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2026 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: dashun.wang@kellogg.northwestern.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2601439123/-/DCSupplemental.

Published August 11, 2026.

Downloaded from https://www.pnas.org by NORTHWESTERN UNIVERSITY; SERIALS DEPARTMENT on November 18, 2026 from IP address 165.124.85.138.

setting for examining both questions, because proposal review sits at the interface between idea articulation and public investment, yet has remained difficult to observe at scale.

LLMs represent a potentially transformative development for scientific work (15–21). Recent large-scale evidence shows that the diffusion of LLMs has already been accompanied by sharp increases in scientific production across major preprint platforms, underscoring their potential to substantially raise output at scale (16). By dramatically reducing the cost of drafting, synthesizing, and revising complex text, LLMs may help scientists articulate ideas more clearly, navigate the growing burden of knowledge, and explore unfamiliar domains more efficiently (28, 29). In principle, such tools could accelerate the pace of discovery, lower barriers to entry, facilitate interdisciplinary recombination, and support more effective pivoting across research directions (22, 29–33). From this perspective, LLMs hold the promise of expanding scientific exploration and improving the efficiency with which ideas are generated, refined, and communicated (34, 35).

Yet these same features also raise critical concerns (34). Because LLMs are trained on large corpora of existing scientific text, including published papers and public records of funded projects, they may tend to produce high-probability, discipline-typical language. When used in proposal preparation, such tools may enhance fluency and alignment with prevailing norms while reducing rhetorical and topical variance, making proposed projects more closely resemble those that have recently succeeded (23, 26, 36). In this way, LLM use could shift the balance between exploration and exploitation in scientific funding, pulling proposals toward the center of existing funding patterns and potentially narrowing the diversity of ideas that are submitted for or receive public support, even as it may lower the barriers to producing competitive applications (30, 37). Recent work highlights this tension more broadly, showing that AI can expand individual scientists' productivity and impact while simultaneously contracting the collective focus of science, underscoring the possibility that gains at the individual level may coexist with reduced diversity at the system level (26).

Beyond these concerns, a bottlenecks perspective further suggests caution in extrapolating writing-related productivity gains to downstream scientific outputs (38). Discovery is constrained by multiple frictions, from data collection and experimental throughput to coordination, implementation, and more, many of which LLM involvement does not directly resolve. Because the outputs of funded projects depend on both communication and execution, there are reasons to believe that productivity gains observed in manuscript writing may not translate to the impact of funded projects. This distinction makes federal research funding a particularly important context to evaluate how LLM adoption reshapes scientific activity.

Reflecting the policy relevance of this uncertainty, federal agencies have begun to issue guidance on generative AI in proposal preparation. In July 2025, NIH clarified that applications "substantially developed by AI" will not be considered applicants' original work (see NOT-OD-25-132), while NSF has emphasized investigator responsibility for accuracy and originality and encouraged disclosure of generative AI use (see Notice to Research Community). These actions underscore both the growing salience of LLMs in public funding and the need for empirical evidence on how they are being used and with what consequences.

Here, we examine LLM use at key stages in the federal funding pipeline by employing two complementary levels of analysis. First, we collect datasets containing the complete set of NSF (D1) and NIH (D2) proposal submissions (hereafter private NSF and NIH data) from two large US R1 universities with request start dates from 2021 to 2025, including funded, unfunded, and pending proposals. These data offer us a unique opportunity to observe proposal language at the point of submission. Second, we analyze the full population of publicly released NSF (D3) and NIH (D4) award abstracts (hereafter public NSF and NIH data) over the same period, together with publications supported by these awards. These data allow us to systematically analyze the funded awards (see *Materials and Methods* for details on datasets). Overall, these data enable us to characterize the evolution of LLM use in federal funding and to examine how it relates to the positioning of scientific ideas, their selection for funding, and their translation into publicly supported research output.

Rapid Rise and Bimodal Distribution of LLM Use in US Federal Research Funding

To estimate LLM use in federal grant proposals and awards, we analyze textual traces in their abstracts. Specifically, we leverage an established detection method proposed by Liang et al. (15). Using public grant abstracts with start dates in 2021, which preceded the widespread use of LLMs, we estimate the word distribution of human-written text. We also estimate a corresponding word distribution of LLM-generated text by prompting OpenAI's GPT-3.5-turbo-0125 model to rewrite the same abstracts, simulating LLM involvement in proposal language. Comparing these two distributions allows us to identify systematic differences between human and LLM-modified writing (see *SI Appendix*, section S1.1 for details on model training).

We then apply the detection method at both the corpus and individual grant levels. At the corpus level, we pool grant abstracts for each focal month and its two adjacent months, and we estimate the fraction of LLM-modified sentences (α) within this combined set in each dataset. At the individual grant level, we estimate α separately for each grant abstract in each dataset. These two levels represent complementary estimates, as the corpus-level estimates leverage large volumes of text to capture aggregate LLM use over time, whereas grant-level estimates enable grant-level analyses that capture heterogeneity across grants, allowing us to examine how LLM use relates to semantic distinctiveness and other grant outputs.

Fig. 1 *A–D* trace the evolution of LLM involvement (α) in NSF and NIH grants across two stages of the funding pipeline, proposal submission and award funding, drawing on confidential proposal submission abstracts (private data) and publicly released award abstracts (public data), respectively. Across all four datasets, α is relatively stable prior to late 2022, and then rises sharply beginning in 2023, coinciding with the widespread availability of ChatGPT. This increase is visible both at the point of submission, when ideas first enter the system, and among funded awards that have passed peer review, indicating that LLM use is not confined to a single stage of the pipeline. Over this period, NIH exhibits the same temporal pattern as NSF but with overall lower average levels of LLM involvement.

At the level of individual grants, Fig. 1 *E–H* show that α is not smoothly distributed. Among grants starting in 2023–2025, we observe a bimodal distribution, with one mode near zero and a second centered around roughly 10 to 15%. This pattern reveals a clear split: one group of grant abstracts that largely avoids LLM use and another that incorporates LLMs more substantively into proposal preparation. These results indicate

Downloaded from https://www.pnas.org by NORTHWESTERN UNIVERSITY; SERIALS DEPARTMENT on November 18, 2026 from IP address 165.124.85.138.

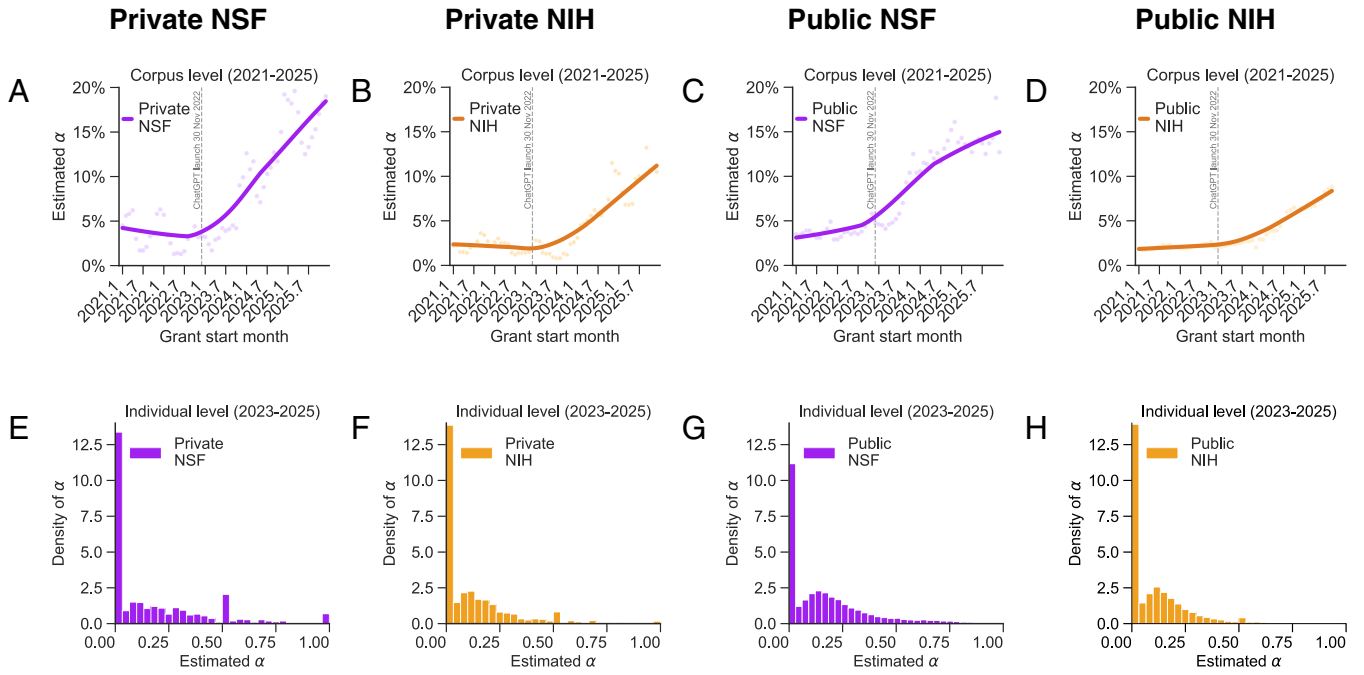

Fig. 1. Rapid rise and bimodal distribution of LLM use in US federal research funding. (*A–D*) Corpus-level estimates of LLM use (α) for private and public NSF and NIH grants from 2021 to 2025, computed using rolling three-month windows (points). Solid lines show locally weighted regressions. The vertical dashed line marks November 30, 2022, corresponding to the public release of ChatGPT. (*E–H*) Distributions of individual-grant α for private and public NSF and NIH proposals and awards with start dates between 2023 and 2025, showing a bimodal pattern consistent with a split between minimal and substantive LLM use across grants.

that LLM adoption has rapidly become a detectable feature of federal research proposals and awards, but it remains highly uneven across them.

LLM Use and Semantic Distinctiveness in US Federal Research Funding

The rapid and uneven adoption of LLMs prompts another critical line of inquiry: Is greater LLM involvement associated with a shift in the kinds of ideas that enter, and ultimately shape, the federal research portfolio? To quantify the positioning of ideas, we measure the semantic distinctiveness of each proposal or award abstract (D1–D4) relative to the agency's recent funding frontier, building on recent work using transformer-based embeddings (26, 39). Specifically, for grants starting between 2023 and 2025, we compute each abstract's average cosine distance [using SPECTER2 (40) embeddings, an established embedding method for scientific texts] to all abstracts funded in the previous year within the same agency. We then convert these distances into within-year percentiles, where higher values indicate more distinctive ideas relative to other grants in that agency-year. We adopt this distance transformation for interpretability, so that the coefficients can be read as shifts in relative position within the empirical distribution of proposals in the same agency-year.

We relate LLM involvement (α) to distinctiveness, estimating regressions that include grant start year, field, and investigator (PI and co-PI) fixed effects, along with controls for requested or awarded funding amount (*SI Appendix*, section S1.2). These fixed effects absorb time-invariant heterogeneity at the field and investigator levels while accounting for common shocks across grant start years. Across all four datasets, including private submissions and public awards for both NSF and NIH, higher α is consistently associated with lower distinctiveness percentiles (Fig. 2 and *SI Appendix*, section S1.2 and Tables S1–S4), indicating that proposals and awards with greater LLM involvement are positioned closer to the grants that the agency funded most recently. For example, for NSF awards, moving from low LLM involvement (25th percentile) to high LLM involvement (75th percentile) corresponds to a ∼5-point decrease in the distinctiveness percentile; for NIH awards, the decrease is ∼4 points. Because distinctiveness is measured as a within-year percentile rank relative to the full distribution of proposals and awards in the same agency and year, this shift indicates that proposals and awards with higher LLM involvement are repositioned meaningfully among their immediate peers. Importantly, the investigator fixed effects compare each investigator only to themselves across proposals and awards rather than comparing different investigators to one another. Thus, these estimates capture within-investigator changes in positioning: For the same investigator's proposals and awards, when the LLM involvement is higher, their proposals and awards become systematically less distinctive. These findings suggest that as LLM use rises, the proposal and award portfolio shifts toward the center of recently funded work, providing empirical evidence of a consistent reduction in semantic distinctiveness in federal research funding that is already occurring across both agencies. Note that this analysis exploits within-investigator variation in LLM involvement, reporting conditional correlations within investigators, not causal estimates. We therefore interpret LLM involvement as associated with the positioning, selection, and downstream output of federally funded research, not as a cause of these outcomes.

Does this association merely reflect LLM-induced stylistic smoothing, whereby LLMs homogenize surface-level language across proposals, or a substantive convergence in the ideas themselves? To distinguish between these channels, we apply our distinctiveness pipeline to LLM-rewritten versions of baseline abstracts from 2021, holding the underlying scientific content

Downloaded from https://www.pnas.org by NORTHWESTERN UNIVERSITY; SERIALS DEPARTMENT on November 18, 2026 from IP address 165.124.85.138.

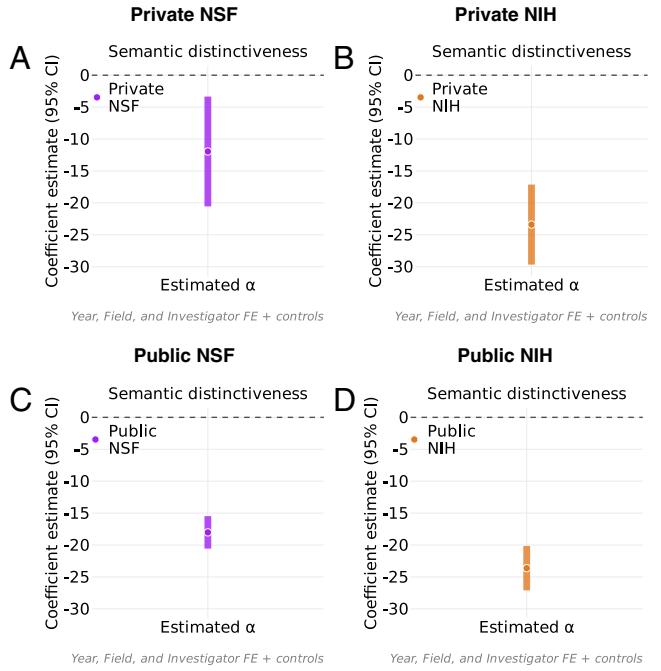

Fig. 2. LLM use and semantic distinctiveness in US federal research funding. (*A*–*D*) Regression estimates relating grant-level LLM use (α) to semantic distance from abstracts funded in the prior year within the same agency, expressed as within-year percentiles. Panels show results separately for private NSF (*A*), private NIH (*B*), public NSF (*C*), and public NIH (*D*) grants. All regressions include grant start year, field, and investigator fixed effects, as well as controls for funding amount. Points indicate coefficient estimates, and bars denote 95% CIs. Negative coefficients correspond to proposals and awards that are positioned closer, in semantic space, to recently funded work within the same agency.

constant while substantially altering the surface language. The average semantic distinctiveness of original and LLM-rewritten abstracts is nearly indistinguishable for both NSF and NIH (*SI Appendix*, section S2.1 and Fig. S1). This finding suggests that our distinctiveness measure is not simply driven by LLM-mediated linguistic changes, and that the evidence is consistent with convergence in substantive positioning rather than surface language alone.

LLM Use and US Federal Research Proposal Success

We next ask whether LLM adoption is associated with a first-order consequence in the funding pipeline: which proposals actually receive funding. Using confidential submissions that include both funded and unfunded proposals (D1–D2), we regress proposal success on proposal-level LLM involvement (α) for request start years 2023–2025, estimating models with request start year, field, and investigator fixed effects and controlling for requested funding (Fig. 3 and *SI Appendix*, section S1.2 and Tables S5 and S6). These fixed effects absorb time-invariant heterogeneity across research areas and investigators, while accounting for common shocks across years. For example, the investigator fixed effects hold constant time-invariant differences in investigators' ability, style, and topic choice, asking whether an investigator's proposals are more likely to be funded when their LLM involvement is higher.

We find no statistically significant association between α and proposal success for NSF proposals. In contrast, for NIH submissions, α is positively and significantly associated with proposal success. Specifically, moving from low LLM involvement (25th percentile) to high LLM involvement (75th percentile) corresponds to a ~4-percentage-point jump in NIH funding probability. Overall, these results indicate that the consequences of LLM adoption for proposal selection are agency-dependent: While LLM use does not systematically advantage proposals at NSF, it is associated with a higher likelihood of being funded at NIH, even for the same investigators. This agency-dependent selection pattern raises the question of translation: Conditional on being funded, is LLM adoption associated with changes in downstream outputs from these funded projects?

LLM Use and US Federal Research Funding Outputs

Using the public NSF and NIH award data (D3–D4), we next relate grant-level LLM involvement (α) to subsequent publication output for grants starting in 2023–2024, estimating regression models that control for funding amount and include grant start year, field, and investigator fixed effects (Fig. 4 and *SI Appendix*, section S1.2 and Tables S7–S12).

We again find no statistically significant relationship between α and publication output for NSF awards (Fig. 4*A*). However, for NIH awards, higher α is strongly associated with more resulting publications (Fig. 4*B*). Moving from low LLM involvement (25th percentile) to high LLM involvement (75th percentile) predicts ~5% more publications for NIH grants. Moreover, this NIH productivity premium appears mostly concentrated in publication volume rather than high-impact outputs: When we restrict our analyses to highly cited papers (e.g., top 5% papers in Fig. 4 *C* and *D*; top 1% in *SI Appendix*, Fig. S2), the relationship attenuates and is no longer statistically significant at conventional levels. Taken together, these results suggest that the downstream consequences of LLM adoption are also agency-dependent. At NIH, higher α is associated with more follow-on publications, but the additional output is concentrated in nonhit papers rather than the most highly cited work. We emphasize that these estimates are based on grants starting in 2023–2024 and therefore capture only the first one to two years of postaward publication output, reflecting early-stage rather than long-run productivity. Descriptive analyses of older cohorts suggest that such early windows appear to capture a meaningful share of a grant's eventual output, and that early output is overall predictive

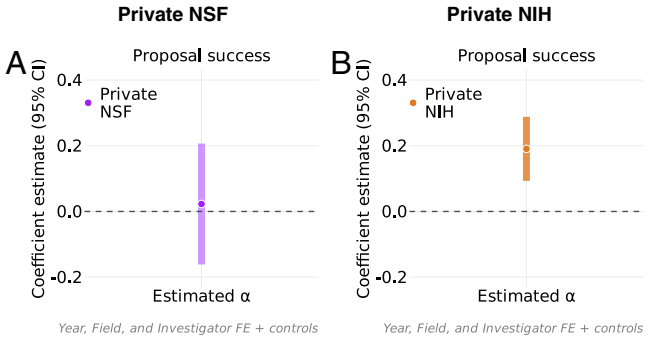

Fig. 3. LLM use and US federal research proposal success. Based on private NSF and NIH proposal submissions from two large US R1 universities, this figure examines the relationship between LLM use at submission (α) and proposal success. (*A*) Regression estimates for NSF submissions. (*B*) Corresponding estimates for NIH submissions. All regressions include proposal request start year, field, and investigator fixed effects, as well as controls for requested funding amount. Points indicate coefficient estimates, and bars denote 95% CIs.

Downloaded from https://www.pnas.org by NORTHWESTERN UNIVERSITY; SERIALS DEPARTMENT on November 18, 2026 from IP address 165.124.85.138.

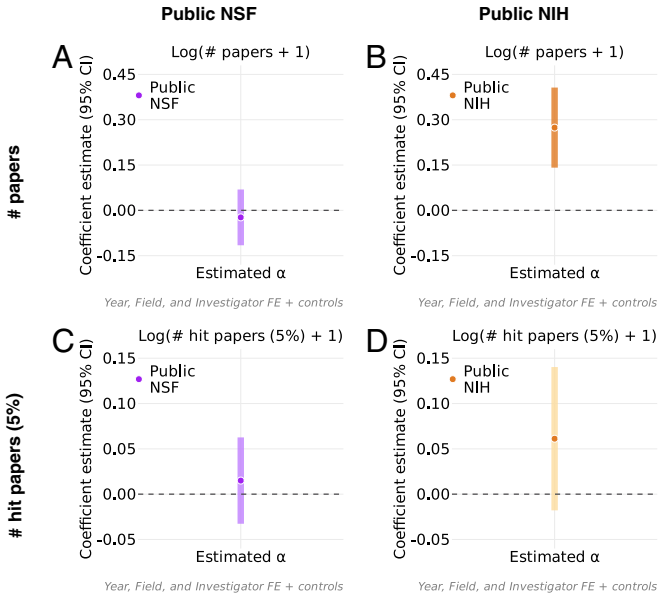

Fig. 4. LLM use and US federal research funding outputs. (*A* and *B*) Regression estimates relating grant-level LLM use (α) to the total number of resulting publications for NSF (*A*) and NIH (*B*) grants. (*C* and *D*) Corresponding estimates for high-impact outputs, where a "hit" paper is defined as one whose citations fall within the top 5% of all papers published worldwide in the same year and field. All regressions include grant start year, field, and investigator fixed effects, as well as controls for funding amount. Points indicate coefficient estimates, and bars denote 95% CIs.

of longer-run output (*SI Appendix*, section S2.2 and Figs. S3–S5). Nonetheless, because high-impact work may unfold over longer horizons, the relationship between LLM involvement and long-run outcomes remains an important open question for future work.

These productivity patterns both align with and qualify recent evidence of large output gains following LLM diffusion in manuscript production (16). While we observe a positive association between LLM involvement and publication output for NIH-funded grants, the magnitude of this association is smaller than the large increases documented in preprint production and is strongly agency-dependent, with no detectable relationship at NSF. In the federal funding context, where projects are selected through peer review and outputs reflect execution constraints and portfolio objectives rather than writing speed alone, LLM adoption appears to translate into additional publications primarily at NIH, and predominantly among nonhit papers, while showing no detectable relationship with publication output at NSF.

Robustness and Sensitivity Checks

We conduct a series of robustness and sensitivity checks across numerous dimensions. First, to better characterize proposals and awards with higher LLM involvement, we follow prior work (16) and compute the Flesch Reading Ease score for each abstract as a measure of writing complexity (*SI Appendix*, section S2.3). Consistent with prior work (16), we find that higher LLM involvement is positively associated with greater writing complexity in both proposals and awards (*SI Appendix*, Fig. S6).

Second, recent work has shown that the use of promotional language in grant proposals is positively associated with funding outputs and related proposal characteristics (14). To assess the extent to which our results may be driven by promotional words, we conduct a robustness analysis that explicitly removes promotional words at all stages of our text processing pipeline. The resulting LLM-use measures remain highly similar to the original estimates across all datasets, indicating that our results are robust to the exclusion of promotional words (*SI Appendix*, section S2.4 and Fig. S7).

Third, in the main analysis of semantic distinctiveness, we define the comparison set using grants funded in the preceding year. We redefine the comparison set using funded grants starting in the prior 2 y, finding the results are robust (*SI Appendix*, section S2.5 and Fig. S8).

Fourth, while our primary measure of semantic distinctiveness relies on cosine distances between embedding vectors, we also compute distances using L2 metrics and obtain consistent results (*SI Appendix*, section S2.6 and Fig. S9). In addition, to ensure that our findings are not driven by the percentile normalization, we replicate the main analysis using raw cosine distances without the within-year percentile transformation and find highly consistent results (*SI Appendix*, section S2.6 and Fig. S10).

Fifth, in our primary specifications, we use ordinary least squares regressions. We re-estimate proposal success using logistic regressions and publication outputs using negative binomial regressions and continue to observe the same associations between α and outcomes (*SI Appendix*, section S2.7 and Figs. S11 and S12).

Sixth, we further examine robustness by field using public award data where sample sizes are sufficiently large to support field-level analyses and find broadly consistent results (*SI Appendix*, section S2.8 and Figs. S13–S18). We also conduct analogous robustness checks by subagency using the public award data and again find broadly consistent patterns (*SI Appendix*, section S2.9 and Figs. S19–S24).

Seventh, to ensure that our findings are not driven by the idiosyncrasies of a single language model, we replicate the entire detection and analysis pipeline using Claude, Gemini, and Llama as alternative reference LLMs. The main analyses yield broadly consistent results across model families (*SI Appendix*, section S2.10 and Figs. S25–S29).

Eighth, because the detection method relies on word-frequency statistics, we systematically evaluate its sensitivity to preprocessing choices by varying the tokenization scheme, stopword handling, and punctuation normalization across nine alternative pipelines. The resulting α estimates remain highly correlated with the baseline, and the main results are broadly consistent under alternative pipelines (*SI Appendix*, section S2.11 and Figs. S30–S33).

Ninth, because AI-related concepts have become increasingly prevalent in scientific research (23), grant proposals may naturally contain more AI-method terminology over time, independent of LLM involvement in writing. To rule out the possibility that our estimator captures topic-driven AI vocabulary growth rather than LLM-driven shifts, we remove all AI-related terms identified in prior work (23) at every stage of the estimation pipeline and re-estimate α. The resulting measures remain highly correlated with the original estimates, and all main regression results are robust to this exclusion (*SI Appendix*, section S2.12 and Figs. S34–S37).

Tenth, to increase the interpretability of the detection method, we apply the generalized word-shift graph framework (41) to decompose the aggregate divergence into per-word contribution scores. This decomposition reveals that the LLM signal is driven by a concentrated set of evaluative and elaborative terms, while human-written text is characterized by simpler, more functional

Downloaded from https://www.pnas.org by NORTHWESTERN UNIVERSITY; SERIALS DEPARTMENT on November 18, 2026 from IP address 165.124.85.138.

vocabulary (*SI Appendix*, section S2.13 and Figs. S38 and S39). Our results are further robust to varying text-length thresholds, single- versus multi-investigator, and standardizing predictors (*SI Appendix*, sections S2.14–S2.16, Figs. S40–S49, and Tables S13–S15).

Discussion

Taken together, our findings reveal that the rapid diffusion of LLMs is already reshaping key stages of the US federal research funding pipeline, from proposal submission to downstream scientific output. By combining confidential proposal submissions with the full population of NSF and NIH awards, we provide systematic evidence that LLM adoption is associated with i) how scientific ideas are positioned relative to recently funded work, ii) whether proposals clear the funding threshold, and iii) the publication output that funded projects go on to produce. These findings speak directly to the two open questions raised at the outset. First, the early effects of generative AI in science may arise less from accelerating the execution of research than from reshaping how ideas are articulated, packaged, and aligned with prevailing norms, while the harder bottlenecks of project design, coordination, and implementation remain: The publication output gains we observe are modest and context-dependent, absent at NSF and concentrated in nonhit papers at NIH, in contrast to the large productivity increases documented at the manuscript level. Second, whereas prior work has largely studied the finished outputs of science, our findings show that generative AI already operates upstream, at the idea-generation stage, where scientific ideas are first articulated and compete for public investment. Together, these results reposition generative AI not merely as a tool that changes how scientific text is produced, but as a force that is beginning to reshape the articulation and allocation of ideas in the federal research funding system.

A central concern in science policy is maintaining a diverse and exploratory research portfolio, one that supports both cumulative progress and the pursuit of unconventional ideas. Across both agencies and both stages (proposal submission and award funding), higher LLM involvement is consistently associated with lower semantic distinctiveness: Proposals with greater LLM traces are positioned closer to recently funded work within the same agency. Importantly, our analysis suggests that the convergence we observe is not simply a byproduct of LLM-induced surface-level rewriting, but reflects shifts in the substantive positioning of proposals and awards. We further find that investigators with greater LLM involvement have, so far, largely stayed within their existing lines of inquiry: Proposals with greater LLM involvement are not systematically more distant from an investigator's own prior proposals within the same agency (*SI Appendix*, section S2.17 and Fig. S50). In other words, rather than moving investigators away from their established research trajectories, greater LLM involvement appears to be associated with proposals that more closely resemble what the agencies have funded in the past. A portfolio that is closer to recent funding patterns may reflect improved clarity, tighter alignment with reviewer expectations, or lower transaction costs in articulating a fundable project. But it also implies reduced exploration in the idea landscape, which matters for public funders explicitly tasked with sustaining high-variance discovery, with implications for long-run impact and sustainability of science.

Importantly, the downstream consequences of LLM adoption are agency-dependent. In the confidential submission data, LLM involvement appears unrelated to proposal success at NSF but positively associated with funding at NIH. In the public award data, we similarly detect no relationship between LLM involvement and publication output at NSF, whereas at NIH, higher LLM involvement predicts greater publication output, though concentrated in nonhit papers rather than the most highly cited work. Notably, these agency-dependent patterns do not appear to be explained solely by differences in field composition. For example, we restrict the comparison to a common field—Biomedical & Clinical Sciences—across agencies and extend it to include the Department of Defense (DOD). Within this shared field, we continue to observe distinct relationships between LLM involvement and downstream outputs across agencies (*SI Appendix*, section S2.18 and Fig. S51). Overall, these results suggest that the observed differences likely reflect an interaction between disciplinary context and agency-specific institutional structure, rather than a simple NSF-NIH contrast or a pure field effect. One potential interpretation is that NIH funding and review norms may more strongly reward incremental, executable projects that yield multiple publications, and LLM-assisted drafting may help proposals conform to those established templates. The contrast also raises questions regarding specific features of agency review, such as NIH's study section system, the role of preliminary data and demonstrated feasibility in R01 review, and differences in how agencies weight innovation versus feasibility, and whether these agency-specific features may have played a role. We emphasize, however, that these institutional accounts are speculative: Our analysis documents agency-dependent associations but does not test the mechanisms behind them, and we caution against reading the agency contrast as evidence of specific channels. In addition, publication counts capture a larger share of NIH outputs than NSF outputs, which often include software, datasets, instrumentation, and training. It is also possible that high-impact work unfolds over longer horizons than our follow-up window (*SI Appendix*, section S2.2 and Figs. S3–S5). Disentangling these possibilities and, more generally, explaining why LLM adoption translates into selection and output effects in some contexts but not others, is a key direction for future research.

This context dependence is especially striking when contrasted with recent large-scale evidence that LLM diffusion has been accompanied by sharp increases in manuscript production (16). Taken together, the manuscript setting and the funding-to-output setting suggest a simple but consequential hypothesis: A meaningful share of the productivity gains associated with LLM adoption may operate through accelerating communication—the speed and ease with which researchers articulate, package, and iterate on ideas—rather than accelerating execution, which remains constrained by project design, coordination, data collection, and other bottlenecks that LLM-enabled writing does not directly resolve. This interpretation is necessarily tentative, given limitations of publication-based output measures and the time required for high-impact work to accrue. Nonetheless, the contrast highlights an important implication for science policy: The benefits of generative AI may accrue first and most strongly where writing is the binding constraint, whereas their effects on the composition and ultimate impact of publicly funded research also hinge on institutional incentives and execution constraints.

Our within-investigator, across-grant design adds a distinct perspective to a growing literature on how AI adoption is reshaping scientific production. Existing work has documented that LLM adoption is associated with changes in productivity and referencing behavior in manuscripts and published papers,

Downloaded from https://www.pnas.org by NORTHWESTERN UNIVERSITY; SERIALS DEPARTMENT on November 18, 2026 from IP address 165.124.85.138.

and more broadly that AI-augmented research can contract the collective focus of science (16, 26). But these analyses largely rely on comparisons across researchers, across papers, or across AI versus non-AI work. By contrast, our design leverages variation in LLM involvement across a given investigator's own proposals and awards, allowing us to examine whether more intensive LLM use within the same investigator is associated with differences in semantic positioning, proposal success, and downstream output. At the same time, this design has limitations. While within-investigator comparisons hold constant time-invariant differences across investigators, they cannot eliminate time-varying strategic choices in when and how investigators use LLMs. For example, investigators may rely more heavily on LLMs for proposals that are already more incremental, executable, or aligned with agency priorities; in that case, higher LLM involvement may partly reflect the type of project being pursued rather than an LLM-induced shift toward the center of the funding portfolio. The relationships we report should therefore be interpreted as conditional correlations within investigators, not causal effects. In addition, the confidential proposal data come from two large US R1 universities. Although these data provide rare access to funded, unfunded, and pending proposals at the point of submission, two institutions cannot capture the full diversity of US research settings. Broadening institutional coverage remains an important direction for future work.

Beyond the analytical design, our study helps address a persistent blind spot in the literature: how generative AI enters the funding pipeline at the moment when ideas are first articulated for review. Because unfunded proposals are rarely observable at scale, most prior work has had to infer behavior from public records of funded projects or published papers. By directly observing confidential submissions alongside the population of funded awards, we can separate proposal preparation from postaward editing and examine selection outcomes, not just the characteristics of funded projects. It is important to note that our measurement strategy—textual traces of LLM involvement in proposal abstracts—captures only one dimension of how LLMs may be used in proposal preparation. Accordingly, our estimates should be interpreted as measuring how LLMs shape the articulation and positioning of ideas in proposal narratives, not the full scope of LLM use in research development.

Finally, because federal grants deploy public funds, the rise of LLMs in proposals has implications not only for scientific efficiency but also for accountability and trust. If LLM adoption systematically pulls proposals toward safer, more easily justifiable ideas, it could quietly tilt public portfolios toward exploitation even as agencies emphasize the value of high-risk, high-reward exploratory research. At the same time, heavy reliance on LLMs in documents intended to reflect investigators' original ideas raises questions about authorship, responsibility, and the integrity of the funding process. By quantifying LLM involvement at scale and documenting its associations with idea positioning, selection, and downstream output, our study provides an empirical foundation for evidence-informed policy, ranging from disclosure norms and reviewer training to mechanisms that protect high-variance exploration and preserve portfolio diversity as generative AI becomes increasingly embedded in scientific practice.

Materials and Methods

Datasets. This study integrates four complementary datasets capturing distinct stages of the US federal research funding pipeline at NSF and NIH. Two datasets consist of confidential proposal submissions obtained directly from universities (D1–D2), while two other datasets are constructed from publicly released award records (D3–D4). The data collection protocol on private NSF and NIH proposals from two US universities was reviewed and approved by the Institutional Review Board of Northwestern University (IRB No. STU00222200). Together, these data enable us to characterize the evolution of LLM use in federal funding and to examine how it relates to the positioning of scientific ideas, their selection for funding, and their translation into publicly supported research output.

D1 and D2 comprise confidential NSF and NIH proposal submissions collected from two large US R1 research universities, under approved IRB protocols. Data from the first university were obtained in May 2025, and data from the second university were obtained in November 2025. These data include the full population of proposals submitted to NSF (D1) and NIH (D2) with request start dates between 2021 and 2025, covering funded, unfunded, and pending proposals. The inclusion of unfunded proposals enables analyses that are not possible using public data alone. In total, D1 contains 1.6K NSF proposal submissions, and D2 contains 4.1K NIH proposal submissions over the study period (2021–2025).

For each proposal, the datasets include the proposal abstract text, which serves as the primary textual input for estimating LLM involvement. In addition, the data include proposal-level metadata used in the empirical analyses, including the funding agency (NSF or NIH), request start year, requested funding amount, funding outcome (funded, unfunded, or pending), and investigator names. We link investigators to Dimensions using names and affiliations and use Dimensions researcher IDs as unique investigator identifiers. One university provides principal investigator (PI) names only, whereas the other provides both PI and co-PI names. We assign scientific fields to each proposal by identifying, among grants funded by the same agency, the grant whose abstract has the smallest cosine distance to the proposal abstract based on SPECTER2 embeddings, and assigning the corresponding grant fields from Dimensions (42) to the proposal. Field assignments use Dimensions' main (level-0) field classification (e.g., Information and Computing Sciences, Biological Sciences), which defines 22 fields. These variables allow us to estimate proposal-level LLM involvement, relate it to semantic distinctiveness, and examine proposal success while controlling for funding amount and investigator, field, and year fixed effects.

D3 and D4 consist of the full population of publicly released NSF and NIH funded awards constructed from Dimensions. Dimensions (https://www.dimensions.ai/) is a large-scale, integrated research information platform that uniquely enables systematic linking of research grants to their resulting publications through funding acknowledgment data, making it particularly well suited for studying downstream scientific outputs of funded grants. D3 includes NSF awards with start dates between 2021 and 2025, and D4 includes NIH awards with start dates over the same period, aligning the observation window with the private proposal data. The award data were downloaded from Dimensions in November 2025. In total, D3 contains 57K NSF awards, and D4 contains 74K NIH awards.

For each funded award, the datasets include the award abstract text, which serves as the primary textual input for estimating LLM involvement at the award stage. In addition, the data include award-level metadata used in the empirical analyses, including the funding agency (NSF or NIH), award start year, awarded funding amount, scientific fields, and investigator names and identifiers. Both PI and co-PI identifiers are provided directly by Dimensions. Scientific fields are also directly obtained from Dimensions using its main (level-0) field classification system (e.g., Information and Computing Sciences, Biological Sciences), in which Dimensions defines 22 fields.

To examine downstream scientific output, we link funded awards in D3 and D4 to their resulting publications in Dimensions, including both preprints and published articles. When a paper has both a preprint and a published version, we retain only the published version to avoid double counting. Using these linked records, we construct measures of total publication output and high-impact output based on field- and year-normalized citation percentiles,

Downloaded from https://www.pnas.org by NORTHWESTERN UNIVERSITY; SERIALS DEPARTMENT on November 18, 2026 from IP address 165.124.85.138.

where citations are measured as cumulative citations through November 2025. These measures allow us to relate award-level LLM involvement to semantic distinctiveness relative to recently funded work within each agency and to assess downstream research outputs while controlling for investigator, field, and year fixed effects.

The combination of private proposal submissions (D1–D2) and public award records (D3–D4) enables a unified analysis across multiple stages of the funding process. The private data capture idea articulation and selection at submission, including proposals that were never funded, while the public data capture postselection outcomes, including funded project descriptions and subsequent publications.

Data, Materials, and Software Availability. D1 and D2, internal confidential grant proposal datasets from two US R1 universities, require a data use agreement with the universities; interested researchers should contact the authors for further information. D3 and D4 are bibliometric datasets from Dimensions (42). Those who are interested in raw data from Dimensions should contact Digital Science directly. The deidentified data necessary to reproduce main plots and statistical analyses are freely available at https://northwestern-cssi.github.io/llmfunding/.

ACKNOWLEDGMENTS. We thank Yian Yin for thoughtful comments. We thank all members of the Center for Science of Science and Innovation at Northwestern University for helpful discussions, and Alyse Freilich for her careful editing and valuable feedback. This work is supported by the NSF under Award Number 2404035. The authors used large language models (GPT-5 and Claude Opus 4) to improve the language, style, and readability of the text in this manuscript. The authors are fully responsible for the content.

Author affiliations: [a]Center for Science of Science and Innovation, Kellogg School of Management, Northwestern University, Evanston, IL 60208; [b]Ryan Institute on Complexity, Kellogg School of Management, Northwestern University, Evanston, IL 60208; [c]Northwestern Innovation Institute, Kellogg School of Management, Northwestern University, Evanston, IL 60208; [d]Management and Organizations Department, Kellogg School of Management, Northwestern University, Evanston, IL 60208; and [e]Industrial Engineering and Management Sciences Department, McCormick School of Engineering, Northwestern University, Evanston, IL 60208

1. V. Bush, *Science, the Endless Frontier: A Report to the President on a Program for Postwar Scientific Research* (U.S. Government Printing Office, Washington, DC, 1945).
2. P. Stephan, *How Economics Shapes Science* (Harvard University Press, 2012).
3. D. Li, P. Azoulay, B. N. Sampat, The applied value of public investments in biomedical research. *Science* **356**, 78–81 (2017).
4. E. Galkina Cleary, J. M. Beierlein, N. S. Khanuja, L. M. McNamee, F. D. Ledley, Contribution of NIH funding to new drug approvals 2010–2016. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2329–2334 (2018).
5. P. Azoulay, J. S. Graff Zivin, D. Li, B. N. Sampat, Public R&D investments and private-sector patenting: Evidence from NIH funding rules. *Rev. Econ. Stud.* **86**, 117–152 (2019).
6. L. Fleming, H. Greene, G. Li, M. Marx, D. Yao, Government-funded research increasingly fuels innovation. *Science* **364**, 1139–1141 (2019).
7. Y. Yin, Y. Dong, K. Wang, D. Wang, B. F. Jones, Public use and public funding of science. *Nat. Hum. Behav.* **6**, 1344–1350 (2022).
8. P. Azoulay, M. Clancy, D. Li, B. N. Sampat, What if NIH had been 40% smaller? *Science* **389**, 1303–1305 (2025).
9. A. C. Furnas, N. Fishman, L. Rosenstiel, D. Wang, Partisan disparities in the funding of science in the United States. *Science* **389**, 1195–1200 (2025).
10. Y. Wang *et al.*, Funding the frontier: Visualizing the broad impact of science and science funding. arXiv [Preprint] (2025). https://arxiv.org/abs/2509.16323 (Accessed 19 September 2025).
11. Y. Yin, Y. Wang, J. A. Evans, D. Wang, Quantifying the dynamics of failure across science, startups and security. *Nature* **575**, 190–194 (2019).
12. Y. Wang, B. F. Jones, D. Wang, Early-career setback and future career impact. *Nat. Commun.* **10**, 4331 (2019).
13. D. Li, L. Agha, Big names or big ideas: Do peer-review panels select the best science proposals? *Science* **348**, 434–438 (2015).
14. H. Peng, H. S. Qiu, H. B. Fosse, B. Uzzi, Promotional language and the adoption of innovative ideas in science. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2320066121 (2024).
15. W. Liang *et al.*, Quantifying large language model usage in scientific papers. *Nat. Hum. Behav.* **9**, 2599–2609 (2025).
16. K. Kusumegi *et al.*, Scientific production in the era of large language models. *Science* **390**, 1240–1243 (2025).
17. W. Liang *et al.*, "Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews" in *International Conference on Machine Learning (ICML)*, R. Salakhutdinov et al., Eds. (Proceedings of Machine Learning Research, 2024).
18. W. Liang *et al.*, The widespread adoption of large language model-assisted writing across society. *Patterns* **6**, 101366 (2025).
19. D. Kobak, R. González-Márquez, E. Á. Horvát, J. Lause, Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Sci. Adv.* **11**, eadt3813 (2025).
20. J. Liu, Y. He, Z. Zheng, Y. Bu, C. Ni, AI-assisted writing is growing fastest among non-English-speaking and less established scientists. arXiv [Preprint] (2025). https://arxiv.org/abs/2511.15872 (Accessed 19 November 2025).
21. H. Bao, M. Sun, M. Teplitskiy, Where there's a will there's a way: ChatGPT is used more for science in countries where it is prohibited. *Quant. Sci. Stud.* **6**, 716–731 (2025).
22. H. Wang *et al.*, Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
23. J. Gao, D. Wang, Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nat. Hum. Behav.* **8**, 2281–2292 (2024).
24. B. Jabarian, A. Imas, Artificial writing and automated detection. National Bureau of Economic Research [Preprint] (2025). https://www.nber.org/papers/w34223 (Accessed 30 September 2025).
25. K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, J. Zou, The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716–723 (2025).
26. Q. Hao, F. Xu, Y. Li, J. Evans, Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature* **649**, 1237–1243 (2026).
27. M. R. Frank, D. Wang, M. Cebrian, I. Rahwan, The evolution of citation graphs in artificial intelligence research. *Nat. Mach. Intell.* **1**, 79–85 (2019).
28. B. F. Jones, The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *Rev. Econ. Stud.* **76**, 283–317 (2009).
29. R. Hill *et al.*, The pivot penalty in research. *Nature* **642**, 999–1006 (2025).
30. L. Liu, N. Dehmamy, J. Chown, C. L. Giles, D. Wang, Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nat. Commun.* **12**, 5392 (2021).
31. G. Tripodi *et al.*, Tenure and research trajectories. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2500322122 (2025).
32. E. Shao *et al.*, SciSciGPT: Advancing human-AI collaboration in the science of science. *Nat. Comput. Sci.* **6**, 301–315 (2025).
33. D. Wang, AI agents are 'aeroplanes for the mind': Five ways to ensure that scientists are responsible pilots. *Nature* **651**, 32–34 (2026).
34. C. A. Bail, Can generative AI improve social science? *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2314021121 (2024).
35. S. Musslick *et al.*, Automating the practice of science: Opportunities, challenges, and implications. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401238121 (2025).
36. A. R. Doshi, O. P. Hauser, Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**, eadn5290 (2024).
37. J. G. March, Exploration and exploitation in organizational learning. *Organ. Sci.* **2**, 71–87 (1991).
38. B. Jones, Artificial intelligence in research and development. National Bureau of Economic Research [Preprint] (2025). https://www.nber.org/papers/w34312 (Accessed 31 October 2025).
39. E. Scharfmann, M. Marx, L. Fleming, Pasteur's quadrant researchers bring novelty, impact to publishing, and patenting. *Science* **390**, 891–893 (2025).
40. A. Singh *et al.*, "A multi-format benchmark for scientific document representations" in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, 2023), pp. 5548–5566.
41. R. J. Gallagher *et al.*, Generalized word shift graphs: A method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Sci.* **10**, 4 (2021).
42. C. Herzog, D. Hook, S. Konkiel, Dimensions: Bringing down barriers between scientometricians and data. *Quant. Sci. Stud.* **1**, 387–395 (2020).

Downloaded from https://www.pnas.org by NORTHWESTERN UNIVERSITY; SERIALS DEPARTMENT on November 18, 2026 from IP address 165.124.85.138.