nature portfolio

Peer Review File

Artificial Intelligence Tools Expand Scientists' Impact but Contract Science's Focus

Corresponding Author: Professor James Evans

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
Summary of Key Results

This paper presents a large-scale empirical analysis of the effects of AI adoption on scientific research and career trajectories. The key findings highlight an accelerated uptake of AI tools in scientific fields, a significant increase in individual productivity and impact for AI-adopting researchers, and a corresponding contraction in the diversity of scientific exploration. However, the study also finds that AI-driven research narrows the breadth of scientific focus, reducing knowledge extent and decreasing scientist engagement with follow-on research. According to the authors, these results underscore a fundamental paradox: AI amplifies individual success but may restrict the collective expansion of scientific knowledge.

Originality and Significance

The study provides a novel perspective on AI's role in shaping the landscape of scientific research. While prior works have examined AI's impact on specific scientific domains, this research systematically quantifies AI's influence across multiple disciplines, using an extensive dataset of 67.9 million papers. The combination of machine learning techniques and citation analysis presents a rigorous and comprehensive albeit non-experimental approach. The finding that AI research may be leading to "collective hill-climbing" rather than fostering diverse exploration is a critical contribution, aligning with prior concerns about research homogenization. The work is timely and relevant, addressing concerns about AI-driven research in the context of rapidly evolving technological advancements.

Data & Methodology

The study employs an extensive dataset (OpenAlex, covering 109 million papers) and utilizes BERT-based language models to identify AI-assisted research. The two-stage fine-tuning approach ensures an acceptable level of accuracy (F1-score = 0.876) in detecting AI-related publications. The validation process, which includes expert annotation with very high inter-rater agreement (Fleiss' Kappa = 0.96), further enhances the credibility of the methodology. The study effectively combines bibliometric analysis, machine learning classification, and statistical validation, making it a robust and reliable examination of AI's impact on science, within the limits of such methodologies.

Appropriate Use of Statistics and Treatment of Uncertainties

The authors appropriately employ statistical analyses, including t-tests, chi-square tests, and median tests, ensuring the validity of their conclusions. The use of bootstrapping to balance datasets further strengthens the robustness of the findings. The study acknowledges and addresses potential biases in AI-paper identification and citation-based measures.

Conclusions: Robustness, Validity, and Reliability

The paper's conclusions are well-supported by the presented data and findings, albeit they underplay the role of causal

overdetermination (see below). The findings are consistent across six major scientific disciplines, reinforcing the reliability of the results. The study's implications for science policy and future research directions are well-articulated, making the conclusions both relevant and actionable.

Suggested Improvements
1. Clearer Definition of "AI Papers" and Identification Criteria: While the study uses BERT models to classify AI papers, the exact criteria for identifying AI research are not explicitly defined. It would be beneficial to include a clearer definition early in the paper, specifying whether AI-assisted research includes papers merely using AI for analysis or those developing AI methodologies.

2. Addressing Overdetermined Findings: The claim that "AI research attracts more attention from academia, and AI-adopting scientists achieve higher scholarly productivity and impact" should be reframed. It is unclear whether AI itself drives these advantages or whether researchers who are already positioned for higher productivity are more likely to adopt AI. The study should consider alternative explanations and acknowledge potential self-selection biases, as well as methodological limitations. The same critique applies to other findings as well.

3. Considering Confounding Factors in Research Concentration: The finding that "AI in science has become more concentrated around specific hot topics" and that "AI-assisted research focuses on a narrower scope" might also be influenced by confounding factors. Certain topics may inherently lend themselves more to AI application, making them appear disproportionately represented. The study should discuss whether external factors, such as topicality, data availability or funding priorities, also contribute to this trend.

References: Appropriate Credit to Previous Work

The paper appropriately situates itself within the existing literature, citing key studies on AI in science, knowledge diffusion, and citation patterns. References to foundational AI research, bibliometric studies, and science-of-science literature provide strong grounding.

Clarity and Context

The paper is well-written, with a clear abstract that effectively summarizes the key findings. The introduction provides strong contextual background, making the research questions explicit. The conclusions succinctly distill the findings and their broader implications. However, a more structured discussion of limitations and potential future research directions would improve the clarity of the argument.

Final Assessment

This study makes a significant contribution to understanding the impact of AI on scientific research, all the while being timely and novel. Its findings will make a valuable resource for policymakers, researchers, and academic institutions. While the paper effectively demonstrates AI's dual role in expanding individual impact and contracting scientific focus, addressing concerns about definitional clarity, overdetermination of findings, and potential confounding factors are necessary for transparency and for appropriately situating the analysis. Overall, this is a compelling and insightful paper that advances the conversation on AI's role in shaping the future of science.

Referee #2

(Remarks to the Author)
Artificial Intelligence Expands Scientists' Impact but Contracts Science's Focus
2025-01-00126A.

26 March 2025
• Key results: Please summarise what you consider to be the outstanding features of the work.

This paper focuses on the potential influences of AI on science and scientists, as well as the potential conflict between individual and collective benefits. The researchers conclude that there is an accelerated adoption of AI among scientists but that there are less topics being studied and less interaction. They identify a paradox as being the movement towards rich areas of data, but that there is work to automate existing fields but not generate new fields.

• Validity: Does the manuscript have flaws which should prohibit its publication? If so, please provide details.

I have some concerns about the overall timing of the analysis. My overall impression of this manuscript is that it is somewhat premature.

• Originality and significance: If the conclusions are not original, please provide relevant references. On a more subjective note, do you feel that the results presented are of immediate interest to many people in your own discipline, and/or to people from several disciplines?

The work is original, with the authors undertaking an investigation of a very large number of scientific papers.

The conclusions are presented from an interesting perspective as a paradox.

The reported results were of immediate interest to me as a researcher who is curious about AI impact on the scientific community. They are at a high level of analysis and examine research across multiple areas. They suggest that we lack a detailed, dynamic understanding of AI on the entire scientific community.

• Data & methodology: Please comment on the validity of the approach, quality of the data and quality of presentation. Please note that we expect our reviewers to review all data, including any extended data and supplementary information. Is the reporting of data and methodology sufficiently detailed and transparent to enable reproducing the results?

I have some concerns regarding the method. I note that LLMs are recent and fundamentally different from machine learning methods.

Figure 1 (d) is confusing to me. The graph generally shows increases in the share of papers with AI. In approximately 2010, the big data era started. This provided input for machine learning research, which increased. Machine learning is a valid tool, based on statistical algorithms. It is different from LLMs, which is text based.

Prior to the emergence of Chat-GPT in late 2022, researchers were not using LLMs. Mass adoption of use, generally, takes time, even though LLMs were adopted quite rapidly. A paper can be submitted, but it takes time to go through the review process and be published.

When I examine this figure, I see that the increase in machine learning research (which started around 2015) has started to taper off, especially in materials science and chemistry. Although increasing in the other areas, the rate seems to be lessening. Therefore, this implies to me that paper is somewhat premature. If something is truly making an impact, it would not be tapering off as such.

In figure 1E, the trends were already going almost straight up. They started rising after 2015 and, although they appear to be increasing rapidly after 2023, the momentum started beforehand. Therefore, it is not clear that this trend can be attributed to AI.

There are also many other factors, in general, that result in the increase in publications at certain points in time. Collaboration with researchers worldwide has increasingly occurred, supported, in part, by document sharing and other collaborative tools, as well as publicly available data (e.g., the human genome mapping). Some journals, with which I am familiar, have increased their page count over the past 10 years or have published additional papers and supporting material online. Thus, I have a concern that there are other factors that might be influencing the results.

• Appropriate use of statistics and treatment of uncertainties: All error bars should be defined in the corresponding figure legends; please comment if that's not the case. Please include in your report a specific comment on the appropriateness of any statistical tests, and the accuracy of the description of any error bars and probability values.

The Fleiss Kappa of 0.9 seems very high and indicates almost perfect agreement. They are measuring whether a paper is an AI or not an AI paper, which would explain the high value. I think this could be made clearer. Then, BERT model has a 13% error, given an F1 score of 0.876. Then, this can be interpreted as being below humans. Is this what is intended?

At some points, I found it confusing as to whether the authors were referring to papers on AI or using AI.

The figures are appropriated labelled, although the amount of discussion seems excessive, given the descriptions in the main paper.

• Conclusions: Do you find that the conclusions and data interpretation are robust, valid and reliable?

The conclusions are ok, but I have some concerns with how they were arrived at. They seem somewhat premature. LLMs have not been around for very long and there is a lead time before publications using them appear.

• Suggested improvements: Please list additional experiments or data that could help strengthening the work in a revision.

I would like to know more about the citations. Citations to AI papers are higher than to non-AI papers, even though there are vastly more non-AI papers, but AI is currently a topic with high interest, especially since the introduction of large language models. Furthermore, usually citations to recently published papers take some time to accumulate.

• References: Does this manuscript reference previous literature appropriately? If not, what references should be included or excluded?

I do not have any further suggestions.

• Clarity and context: Is the abstract clear, accessible? Are abstract, introduction and conclusions appropriate?

It is generally clear what the authors are trying to do. There were some small places where I didn't entirely understand what the authors were trying to convey. For example, it was not clear to me how the research question "aligns with the promise

and peril of AI in scientific research." I mention that instance because of the importance of the statement of a research question.

• Inflammatory material: Does the manuscript contain any language that is inappropriate or potentially libelous?

No, it does not.

Springer Nature is committed to diversity, equity and inclusion; please raise any concerns that may in your view have an impact on this commitment.

I do not have any concerns.

• Please indicate any particular part of the manuscript, data, or analyses that you feel is outside the scope of your expertise, or that you were unable to assess fully.

I do not have a great understanding of the OpenAlex database, which seems to have started at Microsoft and now being supported in Europe. I have not used BERT, although I appreciate the mission of language models in general.

I am not an expert in statistics.

Referee #3

(Remarks to the Author)
I read with great interest the fascinating manuscript by Hao et al. The topic -- the impact of AI on scientific research -- is extremely timely and the results are confirmatory of my expectations and fears. This is actually a concern for me because of confirmatory bias. I thus want to be extra careful with what is the evidence provided to the reader.

My starting concern is that the text suffers, to some extent, from too much hype. This is due to the lack of clarification for what is meant by AI. Both scientists and the lay public interpret the term AI currently as meaning almost exclusively **generative AI**. Personally, I do **not** believe that generative AI has changed how we live or how most people work (second line of introduction). I am also rather concerned about poorly designed experiments that have claimed that generative AI can "facilitate the distillation and dissemination of scientific findings."

I strongly feel that the manuscript would benefit from an introduction that is more explicit about what the authors mean by AI. They study publications from as early as 1980, clearly at a time during which there was not generative AI.

Another concern relates to how AI papers are identified. This is described in the first section of the Results and in the methods. Some results are then presented in Fig. 1. I have to confess that I find the panels a-c in Fig 1 useless. One of my main questions in how well the identification model works for **different time periods**. AI has meant many different things over the 40 years studied by the authors. If AI papers for some stage of AI research are not well identified this is likely to alter findings.

As an aside, I am particularly unimpressed by Fig 1c. This is the kind of figure used to support something that is just confirmatory bias. More interesting would be to see an example of where the model makes a mistake.

Continuing with Fig 1, a linear plot is not appropriate for these data. If the growth appears to be exponential then the plot should use a log scale on the y-axis. and the growth estimates should be show as fits to straight regions in such a plot. And error bars would be useful too. The same exact issue arises for Fig. 2d. In that case, it is clear that the exponential fit is terrible for the case of young scientists not using AI. And, again, there should be error bars on the parameter estimates.

The analysis of the citations is also problematic. The distribution of citations to papers is very right skewed, making the average a rather unreliable statistic. More concerning is the **apparent** mixture of all types of articles in the analysis. Review articles have very different citation patterns from original research papers or from editorial pieces. I would recommend disaggregating the analysis and focusing on papers reporting on original research.

The authors also refer to 1.7 million researchers. The overwhelming majority of these individuals is highly unlikely to have remained in research careers. Thus, it is not clear what they have to teach us about the impact of the use of AI on career progression. Ioannidis' team has discussed a set of continuing publishing scholar, which would qualify as the worldwide core of active researchers. That number if remember correctly is on the order of 0.2 million << 1.7 million.

The most interesting result of the manuscript is reported in Fig. 3. (Please note some spelling mistakes throughout the manuscript such as 'centriod') It very much supports my own bias. But again I have concerns. Let us assume that the 768 dimension embedding of the articles is reliable and accurate. What the authors do then is use t-SNE embedding of the data and then calculate a distance between papers based on that embedding. This procedure **cannot** produce a reliable estimate. While t-SNE (or UMAP) can maintain distances approximately accurate for **close neighbors**, it does for general

pairs of distant data points. This means that the calculation of a cluster radius cannot be reliably made.

Finally concerning Fig. 4. I am not sure I follow what is being plotted in Figs. 4a and 4b. I imagine that in 4b, the y-axis is the EG defined in Eq. (7). But what is the y-axis in 4a? Fig. 4c is redundant, I just want to know what the Gini coefficient is with error bars. Fig. 4d suffers from the problem with the definition of distance between papers discussed for Fig. 3.

In summary, this is a very interesting manuscript, but one for which I would need to gain greater confidence in the robustness of the analyses.

Referee #4

(Remarks to the Author)
Thank you for the opportunity to review, “Artificial Intelligence Expands Scientists’ Impact but Contracts Science’s Focus”. This manuscript highlights some of the intended and unintended consequences of AI adoption by scientists. For full disclosure, I think this is very important work and I like this manuscript a lot. My comments are to help the authors make the biggest contribution possible.
Key Results: The authors show that AI adoption by scientists increases their publication rates (in a field), their paper’s citations, and accelerates their promotion schedules. However, at the field level, it narrows the range of topics studied in the field. The authors look at mechanisms for these effects and further find that AI causes a winnowing of attention on a smaller set of topics.
Validity: After reading the manuscript many times, I feel that it is very compelling. I have been thinking about whether/how the effects could be driven by mechanisms other than the ones presented by the authors. One area that could be ruled out more eloquently is whether the non-AI research that is being minimized or squeezed out should be squeezed out. That is, is more diversity of topics beneficial? (I do believe it is, but the question remains of whether some topics need to die out in a field and perhaps AI is accelerating this, versus whether it is pushing out fruitful areas).
Originality and Significance: I believe that this paper is really important. The world is adopting AI and it is speeding things up, but we haven’t fully grappled with the 2nd degree consequences of such AI progress. This paper implies that AI causes a faster switch to exploitation from exploration which is narrowing a scientific field’s focus, potentially limiting future innovation.
Data and Methodology: (see next subheading)
Appropriate Use of Statistics and Treatment of Uncertainties: My empirical expertise is not directly related to the analyses presented in this paper (particularly around the construction of the variables). Nothing jumps out at me as problematic in this domain. One note is that the error bars are not defined in your explanation on Figure 2a/b, Figure 3.
Conclusions: I think the authors present a very strong case for their conclusions. I think they could benefit by making the argument about why fewer areas of inquiring are worse for the field (which I believe is a problem but it could be better justified). I have more points below that did not fit into the reviewer template.
Suggested Improvements:
1. Curious about the phrasing of the research question (last sentence of page 1). In the question itself is the word “beneficial”. Why is it the “beneficial adoption” instead of simply “adoption”? Beneficial to whom? You of course find that it is beneficial for individual scientists but again, that is a conclusion you have made stemming from your RQ.
2. Curious about your findings in Figure 2a regarding citations of AI papers from 20+ years ago. Given there were fewer (AI) papers back then, and most papers need to cite something “historical” to situate their work in a broader conversation. I do not know if it is possible for you to identity foundational citations versus “throw away” citations that everyone uses more colloquially (this has an important problem vs. here is how we build off of X). Would you see the same pattern of citations for other topics that started small and then became the most prominent topic/method in a given field. Is this unique to AI citations versus other things that started small and became big?
3. 2 years to being “out” in your analyses. In your analysis regarding whether scientists exit, are your results robust to looking at lengths longer than 2 years. For instance, if someone goes on maternity leave and hence experiences a gap in publication due to life circumstances, can they come back or do most people fall off at the 2 year mark?
4. I think this paper does a whole lot, but I am curious about the people who are leaving the field at higher rates. Do you have any demographic information on these people? Again, this could be fodder for Future Research but does the adoption of AI affect all groups of scientists in a uniform manner?
5. On page 6, you state “…scientific field shrinks to focus attention on areas most amenable to AI research, such as those with an abundance of data”. Do you specifically tests for this assumption? Or is this an inference? I wonder whether this effect is based on the availability of data or idiosyncratic preferences of the scientists who chose the starting point (and then experienced the Matthew Effect).
6. I found your explanation of Figure 4d very confusing. The explanation could be revised (I am sorry that I cannot offer a path forward because I am not sure what you are saying).

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
I thank the authors for their thorough and extensive revision that addressed my remarks.
I have no further comments and recommend publication.

Giovanni Colavizza

Referee #2

(Remarks to the Author)

Thank you for the revised version of this paper. Obviously, the authors have put in considerable work to reworking the paper. However, I am still having problems understanding the paper. Also, I am still grappling with the potential contributions and whether there is enough mature evidence. I do agree that there is a great deal of research that involves the use of AI and it could be interesting to assess the impact. For example, machine learning is used in a wide variety of applications.

I have the following comments on the paper.

Like one other reviewers, I am still concerned that the paper suffers from too much hype. I also still have the nagging concern that generative AI is still a relatively new concept and that we don't have enough evidence to draw strong conclusions about its impact. The title is appealing and eye-catching, but I think it goes too far. The paper, in my opinion, is more of an exploratory analysis of research papers in different areas of research that use AI. I know this probably sounds far more toned down than the authors would like.

I am not convinced that the evidence provided by the authors is enough to conclude that LLMs have indeed transformed scientific writing. Many of the writings I have seen from genAI are not written well. Of course, as these tools advance, they probably will have more of an impact. Reference 23, for example, concludes: Our findings show a sharp increase in the estimated fraction of LLM-modified content in academic writing beginning about five months after the release of ChatGPT, with the fastest growth observed in Computer Science papers." Is that enough justification? This is a preprint that is mentioning something that starts 5 months after introduction.

I find that the paper suffers from presentation problems. Even the response document I found a bit overwhelming.

As a small comment, in my opinion, "we believe" does not belong in academic writing, in my opinion.

The research question is: How does the adoption of AI by individual scientists influence the collective character and dynamic exploration of science as a whole? The phrase "promise and peril" doesn't seem to fit. Further, I don't think that is what you do. Is there a discussion on the promise? Is there a discussion on the perils? Is the conclusion (the paradox you identify) the peril? I scanned for both words and don't see them throughout. Do you really explore the "collective character and dynamic exploration" and, if so, I think you should be explicit about how you do do.

I think you have a good statement of your contribution: "to reveal that the adoption of AI leads to a notable acceleration in the production and visibility of science produced by those scientists who incorporate AI."

"Hot topics" does not seem very formal for a journal such as Nature.

This sentence is awkward: "When the two papers happen to work on similar topics, however, the lack of engagement leads to mutual unawareness and overlap." Further, isn't some overlap good (replicability)?

In the last sentence just before the discussion, I did not understand what you mean by "knowledge extent" so I needed to look it up. Perhaps just indicate exactly what it means earlier in the paper.

What is collective hill-climbing? I am not familiar with that phrase.

Is it fair to say that mutual unawareness leads to redundant research? Probably. But isn't this also true, to some extent, generally? Researchers, especially if they are new to a topic, might overlook relevant research? Here, I think you are just positioning it as a bigger, and more systematic, problem.

I am curious about your use of the term innovation. I think of it in a much different sense; for example, in terms of creating a novel technology. Is the implication that every research paper is an innovation?

Cognitive extent is another term that was not familiar to me, so I needed to look up the reference. Maybe it would help if you could define it briefly, even if in a footnote. I like, for example, the way you defined paper families.

The EG formula is not intuitively obvious to me.

Please check the KE equation.

Also, you might want to define field problems.

Finally, some more discussion of the implications might be helpful. Isn't it good to expand the impact? Is there a problem with less focus?

Thank you for your work.

Referee #3

(Remarks to the Author)
I would like to thank the authors for the extraordinary amount of work that they have put into this revision. I feel that the manuscript is now stronger. I still have some questions and suggestions that I believe could strengthen the manuscript further. In particular, I am still not fully convinced by the manner in which KE is calculated and believe that this needs to be strengthened.

############################
Figure 1:
Line graphs would be more appropriate for panels b and f. Otherwise, best practice is to show bar graphs starting at zero.

While I agree that adding exponential fit line to the data in panels d and e is unnecessary (but thanks of plots in SI), I still think that using a log-scale on y-axis would make the figures easier to read.

In panel b, it is not clear to me why they authors use Fleiss' Kappa for comparison of expert classification but no for comparison of model to ground truth. Using Fleiss' Kappa for both would obliviate the need for two different y-axis and enable oranges to oranges comparison (average F1 score could still be given in text).

############################
Figure 2:
The caption and text quote over 2 million researchers for whom the authors could track their careers. This seems like an awful lot of people for one to be able to track accurately.

US universities hire one faculty per department per year. let's say that gives us about 20 new hires per university per year. Times 50 years that is about 1000 new scientists. Times 100 universities research universities in the US that is maybe 100,000 established scientists. But the analysis is looking only at 6 disciplines. So, those numbers seems very excessive to me...

I do not know that current approaches to author disambiguation are that robust. Which brings me to my other concern: the authors also quote a 45% rate of transition to established scientists for the junior scientists who adopted AI methods. This rate seems extraordinarily high...

I thank the authors for all their new analysis in response to my concerns relating to panel a. I believe however, that the current choice for panel an artificially exacerbates the distinction between AI and no-AI papers. Indeed, Figs S8b-d suggest that the difference is not so marked. Only when you look at S8a do you see a very large difference. As you demonstrate in Fig 4 with the calculated Gini coefficients, most of the difference may be due to the very top of the citation list.

To avoid this exaggeration, I strongly recommend that you use the median instead of the mean in Figs 2a and 2b.

Concerning, panel c, I wonder whether using separate plots for out of academia and become established research would avoid some confusion. The point of those plots is to show significance of difference of the ratio to 1, but with the two data points per discipline one feels tempted to look for statistical significant of the difference between the two.

For panel d and SI figures, I would plot the survival function as that smooths the data without the need for binning.

Finally, 'established' is misspelled in some places.

############################
Figures 3 and 4:
I thank the authors for the clarification. I now see that t-SNE projection was only used for visualization.

Nonetheless, the authors must be aware of the issues with the reliability of calculation of distances for high-dimensionality vectors. As a sanity check (and sensitivity analysis), I would recommend using a PCA approach and identifying the number of significant PCs for the data and then calculating distances within that embedding.

Such as check is particularly critical because knowledge extent the basis for the very important analyses presented in Figs 3 and 4.

Note that 'centroid' is still misspelled in 3b.

Referee #4

(Remarks to the Author)
I was happy to review, “Artificial intelligence expands scientists’ impact but contracts science’s focus” for a second time. I was positively overwhelmed by the care and attention the authors put into revising this manuscript. The supplemental analyses I raised in the previous review were addressed, and in my opinion, add substantial depth to the paper. In my professional opinion, this article is greatly improved and ready for publication.

Key Results: In addition to what was presented in the first submission, the authors have performed several robustness checks to triangulate on their proposed mechanisms and rule of rival plausible alternatives.

Validity: No issues

Originality and Significance: This paper makes an important and timely contribution to the science behind science. My concern is that we are adopting AI without considering its 2nd to nth order effects, which is the core topic of this paper.

Data and Methodology: No issues

Appropriate use of statistics and treatment of uncertainties: No issues

Conclusions: The additional analyses strengthened this manuscript

Suggested Improvements: Very minor wording things here and there. For instance, the authors state “Among the massive papers in the OpenAlex dataset, we select 66,117,158…” would benefit from “quantity of” to signify number versus importance.

Version 3:

Reviewer comments:

Referee #2

(Remarks to the Author)
I still think that the title is 'overhyped.'

In general, I think that the abstract should not report numbers, but I will leave that to the editors.

(Remarks on code availability)
NA

Referee #3

(Remarks to the Author)
I truly applaud the authors for all the additional work undertaken to answers my questions and those of the other reviewers. I feel that procedures are not much clearer and their robustness better evaluated.

I am happy with all the answers to my questions. I think that the definition of 'established scientists' as those who have led one or more research projects in particular clarifies matters.

I leave two final suggestions for the authors:

First, I found the percentile exploration motivated by panel 2a extremely useful. I think there is more richness than in the reporting of the average, and would encourage the authors to show the results for the top 10% and top 1% in Fig. 2.

Second, In Figs. 5b,c, the visualizations would be stronger if they showed box plots instead of bar plots.

(Remarks on code availability)

Revision of manuscript in Nature #2025-01-00126A

We are grateful for the exceptionally detailed and constructive feedback from the *Nature* editor and all reviewers. The reviews demonstrate remarkable depth of engagement with our work, providing comprehensive methodological scrutiny that strengthened our AI paper identification approach, insightful questions about causality that led us to conduct match-and-comparison analyses, critical concerns about statistical robustness that prompted multiple alternative analytical approaches, valuable suggestions for alternative explanations that resulted in extensive factor analyses, and thoughtful considerations of demographic heterogeneity that enhanced our understanding of AI's differential impacts. Here, we summarize our changes:

Critical Methodological Improvements

Enhanced AI Paper Identification (Reviewers 1, 2, 3)
• Clarified definition: Explicitly distinguished AI-augmented research from AI development papers
• Era-specific validation: Demonstrated *F1*-scores >0.85 across ML, DL, and LLM eras with Fleiss' κ>0.93
• Extended coverage: Updated dataset through March 2025 with monthly temporal resolution
• Method transparency: Added detailed examples of correct/incorrect classifications with attention visualizations

Strengthened Causal Inference (Reviewer 1)
• Match-and-comparison analysis: Controlled for early-career trajectories, showing 24.45% higher citations for AI adopters
• Self-selection bias addressed: Demonstrated AI adoption itself drives advantages, not pre-existing researcher characteristics
• Alternative explanations ruled out: Systematic analysis of topicality, funding priorities, and original topic impact

Robust Statistical Framework (Reviewers 2, 3)
• Multiple citation indicators: Beyond averages, analyzed percentiles and proportions for right-skewed distributions
• Exponential fitting improvements: Log-scale plots with confidence intervals for career transition models
• Publication type controls: Separate analyses excluding review articles, maintaining consistent findings
• Core researcher validation: Results hold for scientists with 5+ and 10+ consecutive publication years

Major Analytical Enhancements

Comprehensive Factor Analysis (Reviewer 1)

• Data abundance correlation: Statistical confirmation that AI concentrates in data-rich areas (r=0.35-0.43)
• Topic diversity benefits: Demonstrated broader knowledge spaces don't reduce field impact or novelty
• Original impact neutrality: AI adoption shows no preference for initially high-impact topics ($|r|$<0.1)

Demographic and Temporal Robustness (Reviewer 4)
• Threshold sensitivity: Career transition results consistent across 2-, 3-, and 4-year dropout definitions
• Demographic heterogeneity: Revealed different AI benefits across demographic features including institution types, geographic regions, gender and race meriting further investigation
• LLM-era validation: All major findings replicated using only recent papers, confirming contemporary relevance

Enhanced Citation Context Analysis (Reviewer 4)
• Core vs. superficial citations: Distinguished foundational from "throwaway" citations using semantic analysis
• Technology comparison: Contrasted AI citation patterns with nanotechnology, revealing AI's sustained influence
• Temporal citation dynamics: Tracked how citation roles evolve over decades for AI papers

Thanks to the engaged Reviewers, the revised manuscript is substantially more rigorous and comprehensive than the original:

Methodological Rigor
• Tripled analytical depth with 28 new supplementary figures addressing reviewer concerns
• Eliminated confounding explanations through systematic alternative factor analyses
• Enhanced reproducibility with detailed methodological descriptions and validation procedures

Broader Scope and Implications
• Demographic insights: First analysis of AI's heterogeneous impact across researcher populations
• Policy relevance: Concrete evidence that data abundance drives AI concentration, informing research policy
• Temporal validation: Demonstrated findings hold from historical AI adoption through current LLM era

Scientific Rigor
• Limitations acknowledged: Highlight limits of strict causal identification in ours and related observational work and clarify how findings should be interpreted
• Causality established: Move beyond simple correlation to demonstrate AI adoption's causal effects on scientific careers

• Robustness confirmed: Results hold across multiple analytical approaches, thresholds, subpopulations, and temporal periods
• Scope clarified: Clear delineation of AI-augmented vs. AI-development research with concrete examples

The resulting paper provides unprecedented empirical evidence for AI's paradoxical effects on science—simultaneously expanding individual impact while contracting collective focus—backed by rigorous methodology that addresses every major concern raised in review. We believe that this thorough revision process has transformed our work into a definitive analysis worthy of *Nature*'s standards for groundbreaking scientific contributions.

Below, we include a point-by-point response to reviewers' comments, with our responses below and our revisions in red.

Referee #1

Summary of Key Results

This paper presents a large-scale empirical analysis of the effects of AI adoption on scientific research and career trajectories. The key findings highlight an accelerated uptake of AI tools in scientific fields, a significant increase in individual productivity and impact for AI-adopting researchers, and a corresponding contraction in the diversity of scientific exploration. However, the study also finds that AI-driven research narrows the breadth of scientific focus, reducing knowledge extent and decreasing scientist engagement with follow-on research. According to the authors, these results underscore a fundamental paradox: AI amplifies individual success but may restrict the collective expansion of scientific knowledge.

Response: We thank the Reviewer for your time reviewing our manuscript. In this revision cycle, we endeavor to address your exceptionally engaged and constructive comments. Here, we provide a brief summary of our main revisions below.

a. We clarify our definition of "AI papers", namely papers adopting AI to assist research in natural science fields. We also illustrate the most popular AI methods adopted in different disciplines and eras, providing a more comprehensive but also intuitive understanding of what is identified as AI in our analysis.
b. We conduct a match-and-comparison analysis considering the alternative explanation regarding self-selection bias, showing that adopting AI itself contributes to researchers' subsequent advantages in productivity and impact. We also reframe our findings in the main text, making them substantially more rigorous.
c. We analyze how external factors, including topicality, data abundance, and funding priorities, might influence the selectivity of AI adoption across different topics. Results indicate that data abundance is a major external factor in the disproportionate representation of AI across different topics, while topical specificity, popularity, and funding intensity are not. This provides valuable insight into how AI can be better leveraged to promote scientific development.
d. We have expanded the discussion with limitations and potential future research directions.

Additionally, considering the rapid development of AI, we updated the OpenAlex dataset (and associated Web of Science concordance) used in our revision. The coverage period has been extended from the original cutoff date of May 30, 2024, to March 31, 2025, ensuring that our analysis remains as up-to-date as possible. In the latest OpenAlex snapshot, the dataset maintainers made several corrections, including the removal of abstracts containing invalid or junk content and updates to authorship information. These changes result in adjustments to the numerical values reported in our analysis; while the original conclusions remain valid and unchanged.

The details of our revision are summarized as point-by-point responses to your comments as follows.

Suggested Improvements

1. Clearer Definition of "AI Papers" and Identification Criteria: While the study uses BERT models to classify AI papers, the exact criteria for identifying AI research are not explicitly defined. It would be beneficial to include a clearer definition early in the paper, specifying whether AI-assisted research includes papers merely using AI for analysis or those developing AI methodologies.

Response 1.1: We thank the Reviewer for this excellent suggestion. We regret not clearly defining what "AI papers" refer to. **In our analysis, we focus on papers adopting AI to assist research in natural science fields, rather than those developing AI methodologies themselves.** To identify these papers, we select six representative disciplines covering the vast majority of the natural sciences—biology, medicine, chemistry, physics, materials science, and geology, and analyze AI-assisted research within them.

To illustrate the most commonly adopted AI methods in natural science research, we list the top 15 AI methods across all disciplines for all eras in Fig. 1c. These top methods include support vector machines and principal component analysis from the ML era, as well as convolutional neural networks and generative adversarial networks, which represent important techniques of the DL era. Meanwhile, large language models, which have emerged in recent years, also rank among the most frequently utilized methods in the latest period. We also show in detail the top 10 most frequently used AI methods for each discipline and each AI era in Supplementary Tables S5–S11. These provide a more intuitive understanding of what is identified as AI in our analysis.

According to your suggestion and the above discussion, we make the following revisions to our manuscript:

a. We revise the *Introduction* in the *Main Text*, *Methods-M1. Dataset and Paper selection* and *Methods-M3. Design and fine-tune the language model for AI paper identification*, specifying that we focus on AI-assisted research using AI for analysis, rather than those developing AI itself.
b. We list the top 15 AI methods in the *Main Text*, *Fig. 1c*, illustrating the most commonly adopted AI methods in natural science research.
c. We show in detail the top 10 most frequently used AI methods for each discipline and each AI era in *Supplementary Tables S5–S11* and *Supplementary Notes 1.3-Profile of the identified AI papers*, providing an intuitive understanding of what is identified as AI in our analysis.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 2 in the *Main Text, Introduction*-**Scope of our research focus] … Notably, we do not focus on computer science or mathematics, fields that develop AI methodologies directly, but rather on papers that augment research in natural science fields by adopting AI techniques ranging from early machine learning algorithms to contemporary generative AI. Specifically, we select six representative disciplines that cover the vast

majority of natural science contributions—biology, medicine, chemistry, physics, materials science, and geology.

[**Page 3 in the *Main Text, Results*-**Increasing prevalence of AI adoption in science] … Counting all eras and disciplines collectively, the most commonly adopted AI methods in natural science research include support vector machines and principal component analysis from the ML era, and convolutional neural networks and generative adversarial networks from the DL era. The large language model, which has emerged in recent years, also ranks among the most frequently utilized methods (Fig. 1c and Supplementary Table S5-S11).

[**Page 8 in the *Methods*-**M1. Dataset and Paper selection] In this paper, we emphasize the adoption of AI methods in conventional natural science disciplines and exclude researches developing AI methodologies themselves, separating the influence of AI on science from AI's own invention and refinement. Therefore, we select biology, medicine, chemistry, physics, materials science, and geology as representatives of natural science disciplines, while we exclude computer science and mathematics, where most works introducing and developing AI methods are published…

[**Page 9 in the *Methods*-**M3. Design and fine-tune the language model for AI paper identification] … Finally, we utilize optimal ensemble models during both stages to identify all papers that use AI to support natural science research from the selected representative natural science disciplines.

[***Supplementary Notes 1.3*-**Profile of the identified AI papers] For adopted AI methods, we extracted phrases from the abstracts of identified AI papers and calculated the frequency of phrases related to specific AI techniques. We identified and listed the top 10 most frequently used AI methods for each discipline and AI era (Supplementary Tables S5–S11). It is worth noting that in earlier years, due to the limited number of AI papers, the number of commonly used AI methods may be fewer than ten. The results show that during the ML era, the most frequently applied AI methods in natural science research included Artificial Neural Networks (ANN), Principal Component Analysis (PCA), and Support Vector Machines (SVM). In the DL era, Convolutional Neural Networks (CNN), a landmark of the time, dominated the landscape, while traditional methods such as SVM continued to see widespread adoption. Since the beginning of the LLM era in 2023, Large Language Model (LLM) has increasingly appeared among the most commonly used AI methods, with their frequency rankings steadily rising across disciplines. The evolution of AI methods in natural science research reflects the development and transformation of AI technologies over the past decades.

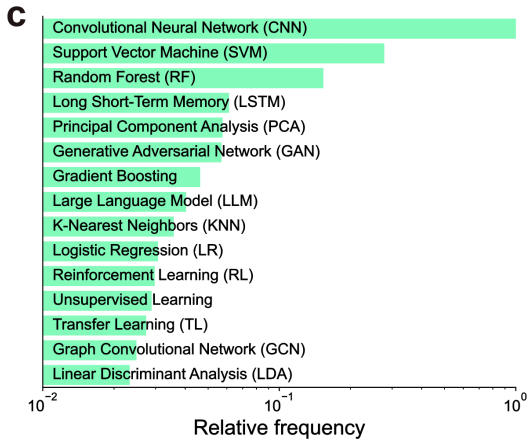


Figure 1. Increasing prevalence of AI adoption in science. (c) Relative adoption frequency of the top 15 AI methods across all disciplines for all selected AI development eras.

Supplementary Table S5. The most frequently adopted AI methods in all selected disciplines in different AI development eras.

Era	Year	Top AI methods									
		1	2	3	4	5	6	7	8	9	10
ML	[1980,1990]	Expert System	Artificial Neural Network (ANN)	Knowledge Base	Connectionist Network (CN)	Principal Component Analysis (PCA)	Hebbian Learning	Maximum Likelihood Classifier	Logistic Regression (LR)	Generalized Additive Model	Nonlinear Regression
	[1991,2000]	Artificial Neural Network (ANN)	Principal Component Analysis (PCA)	Expert System	Knowledge Base	Radial Basis Functions (RBF)	Hidden Markov Model (HMM)	Unsupervised Learning	Linear Discriminant Analysis (LDA)	Maximum Likelihood Classifier	Markov Chain Monte Carlo (MCMC)
	[2001,2005]	Artificial Neural Network (ANN)	Support Vector Machine (SVM)	Principal Component Analysis (PCA)	Hidden Markov Model (HMM)	Linear Discriminant Analysis (LDA)	Markov Chain Monte Carlo (MCMC)	Radial Basis Functions (RBF)	Unsupervised Learning	Knowledge Base	Bsyesian Network
	[2006,2010]	Support Vector Machine (SVM)	Artificial Neural Network (ANN)	Principal Component Analysis (PCA)	Linear Discriminant Analysis (LDA)	Hidden Markov Model (HMM)	Markov Chain Monte Carlo (MCMC)	Radial Basis Functions (RBF)	Non-negative Matrix Factorization (NMF)	Unsupervised Learning	Conditional Random Field (CRF)
	[2011,2015]	Support Vector Machine (SVM)	Artificial Neural Network (ANN)	Principal Component Analysis (PCA)	Markov Chain Monte Carlo (MCMC)	Linear Discriminant Analysis (LDA)	Hidden Markov Model (HMM)	Conditional Random Field (CRF)	Unsupervised Learning	Non-negative Matrix Factorization (NMF)	Radial Basis Functions (RBF)
DL	[2016,2020]	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Random Forest (RF)	Principal Component Analysis (PCA)	Generative Adversarial Network (GAN)	Long Short-term Memory (LSTM)	Decision Tree (DT)	Unsupervised Learning	Reinforcement Learning (RL)	Linear Discriminant Analysis (LDA)
	[2021,2022]	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Random Forest (RF)	Transfer Learning (TL)	Long Short-term Memory (LSTM)	Generative Adversarial Network (GAN)	Gradient Boosting	K-nearest Neighbor (KNN)	Graph Convolutional Network (GCN)	Principal Component Analysis (PCA)
LLM	2023	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Random Forest (RF)	Long Short-term Memory (LSTM)	Gradient Boosting	Large Language Model (LLM)	Generative Adversarial Network (GAN)	Transfer Learning (TL)	K-nearest Neighbor (KNN)	Graph Convolutional Network (GCN)
	2024	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Random Forest (RF)	Large Language Model (LLM)	Gradient Boosting	Long Short-term Memory (LSTM)	Generative Adversarial Network (GAN)	K-nearest Neighbor (KNN)	Shapley Additive Explanations (SHAP)	Logistic Regression (LR)
	2025	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Large Language Model (LLM)	Random Forest (RF)	Shapley Additive Explanations (SHAP)	Long Short-term Memory (LSTM)	Gradient Boosting	Generative Adversarial Network (GAN)	Graph Neural Network (GNN)	Logistic Regression (LR)

Supplementary Table S6-11. The most frequently adopted AI methods in each disciplines in different AI development eras. Please see the full tables in Supplementary Materials.

2. Addressing Overdetermined Findings: The claim that "AI research attracts more attention from academia, and AI-adopting scientists achieve higher scholarly productivity and impact" should be reframed. It is unclear whether AI itself drives these advantages or whether researchers who are already positioned for higher productivity are more likely to adopt AI. The study should consider alternative explanations and acknowledge potential self-selection biases, as well as methodological limitations. The same critique applies to other findings as well.

Response 1.2: We appreciate the Reviewer for pointing out some of our outsized and uncritical claims, as also the many ways in which self-selection becomes possible. To address the possibility of overdetermined findings and better characterize the naturalistic situation of the apparent impact and drift of AI in science, we reframe the manuscript from the following the perspectives:

First, we follow the Reviewer in adopting a **balanced, evidence-based tone that acknowledges both achievements and limitations** rather than making promotional claims about AI in science. We ground our discussion in specific, verifiable examples—AlphaFold's Nobel Prize recognition, measurable improvements in laboratory throughput, and concrete advances in fusion control—rather than speculative promises about AI's transformative potential. Crucially, we now pair each positive development with corresponding concerns or caveats, such as noting that while large language models facilitate scientific writing, they also raise "concerns about weakened confidence in AI-generated content." We now adopt language that is deliberately measured, using terms like "beginning to transform" and "helped some chemists" rather than hyperbolic claims about revolutionary breakthroughs and universal adoption. By presenting AI as a tool with demonstrated but bounded capabilities—one that "circumvents" costs and "upscales" experiments rather than fundamentally reimagining science—we now establish a realistic foundation for our subsequent empirical analysis, positioning the work as a dispassionate examination of observable effects rather than a validation of promotional narrative.

Second, we conduct extended analysis to **strengthen the causal identification** in the case you mentioned. Specifically, it is true that alternative explanations exist for the observation "AI research attracts more attention from academia, and AI-adopting scientists achieve higher scholarly productivity and impact". Namely, are these advantages driven by AI adoption itself, or do scientists who are already on a trajectory toward greater productivity and impact simply more likely adopt AI. To investigate this, we conducted a match-and-comparison analysis of scientists with similar early-career productivity and impact trajectories but who differ in their subsequent AI adoption behavior (Supplementary Fig. S16). For example, we filter 11,019 scientists who began adopting AI in the third year of their careers and match them with 1,926 scientists who exhibited comparable annual citation counts during their first three years but never adopted AI. The comparison reveals that starting from the third year, when the former group began adopting AI, a divergence in annual citation counts emerges between the two groups. By the tenth year of their careers, scientists who adopted AI in their third year have 24.45% higher annual citations and 9.61% higher productivity than their

non-AI-adopting counterparts with similar early trajectories. The same pattern holds when analyzing a second cohort: 9,837 scientists who adopted AI in the fifth year of their careers were compared with peers who had similar citation and productivity levels in their first five years but never adopted AI. Taken together, these findings suggest that for scientists with comparable early-career positions, adopting AI itself contributes to their subsequent advantages in productivity and impact.

Third, we **weaken causal claims** to make our statements more rigorous and add a new concluding paragraph to the discussion that highlights study limitations in line with Reviewer suggestions. We believe that this new paragraph strengthens the paper's credibility by acknowledging the inherent limitations of observational research and tempering what could otherwise appear as overdetermined causal claims. By explicitly stating "we cannot fully identify the causal linkage between AI adoption and scientific impact," we acknowledge some of our necessarily correlational findings, preventing readers from overinterpreting the results. We also acknowledge related methodological constraints associated with selection, such as missing "subtle and unmentioned forms of AI use" along with domain-specific limitations to reveal a more nuanced understanding of the complexity inherent in studying AI's impact on science "in the wild". We believe that these articulations enhance the paper's scholarly rigor by positioning our findings in conformity with rigorous epistemic standards. We couple this with a forward-looking discussion about expanding AI's sensory and experimental capacities to suggest a constructive research agenda that builds on our insights without overstating their definitive character.

According to your suggestion and the above discussion, we make the following revisions to our manuscript:

a. We reframe the *Introduction* to adopt a balanced, evidence-based tone that equally acknowledges achievements and limitations rather than making promotional claims about AI in science.
b. We add the match-and-comparison analysis in *Supplementary Fig. S16* and *Supplementary Notes 2.2- The effect of AI on individual scientists' careers*, demonstrating with evidence that adopting AI contributes to subsequent advantages in researchers' productivity and impact.
c. We reframe overdetermined findings in the *Main Text*, *Results-AI enlarges the impact of individual papers and enhances the career of individual scientists*, and *Results-AI adoption is associated with a contraction in knowledge focus among scientific fields*, making the statements more rigorous.
d. We added a concluding paragraph to the *Discussion* section that highlights methodological limitations of the study, and balanced discovered insights with articulation of a future research program that will be required to resolve uncertainty and the conflict between individual and collective scientific incentives.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 2 in the *Main Text, Introduction*-**AI fuels both hope and concern for its influence on science] ... Major investments in predictive and generative AI have catalyzed society-level debates over the future of AI at home and in the workplace. Perhaps more than any other domain, AI tools have become deeply entwined with the process of knowledge production, yielding findings that attract disproportionate attention in various scientific fields…. Models improved via deep reinforcement learning have become tuned to contain complex fusion reactions and discovered new, hardware-optimized forms to matrix multiplication that recursively accelerate deep learning itself. Autonomous laboratory systems driven by ChatGPT have helped some chemists and material scientists upscale the number of adaptive high-throughput experiments. Moreover, recent breakthroughs in large language models are beginning to transform scientific writing, broadening participation and facilitating the distillation of findings, but also raising concerns about weakened confidence in AI-generated content."

[**Page 3 in the *Main Text, Results*-**AI enlarges the impact of individual papers and enhances the career of individual scientists] … Furthermore, when controlling for and comparing scientists with similar early-career positions, the enhanced productivity and impact still hold (Supplementary Fig. S16). This suggests that, after accounting for potential selection-biases among researchers with different original achievements that may influence their choice of AI adoption, AI itself contributes to the observed advantages.

[**Page 4 in the *Main Text, Results*-**AI enlarges the impact of individual papers and enhances the career of individual scientists] … This indicates that AI-adopted scientists are associated with increased opportunities to lead research teams and reduced risks of dropping out from academia, thereby experiencing accelerated career transitions from junior to established scientists.

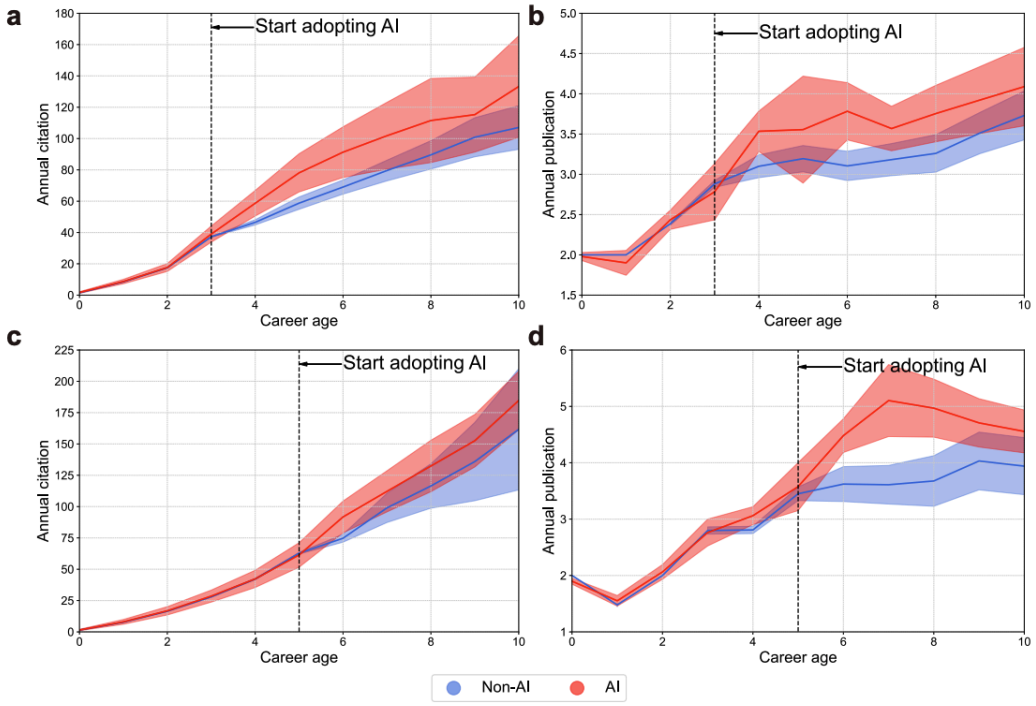
[**Page 5 in the *Main Text, Results*-**AI enlarges the impact of individual papers and enhances the career of individual scientists] … Collectively, these findings suggest that AI research receives more attention from academia, and AI-adopting scientists are associated with higher scholarly productivity and impact.

[**Page 6 in the *Main Text, Results*-**AI adoption is associated with a contraction in knowledge focus among scientific fields] … Compared to conventional research, AI research is associated with a 4.63% contracted median collective knowledge extent across science, which is consistent across all six disciplines…

[**Page 8 in the *Main Text, Discussion*-**study limitations and a future research program articulated to grow credible certainty about findings and remediate unintended consequences] … While our analysis provides new insight into AI's impact on science, clear limitations remain. Our identification approach, though validated by experts, misses subtle and unmentioned forms of AI use, and our focus on natural sciences excludes important domains where AI adoption patterns may differ. Moreover, despite consistently suggestive evidence, we cannot fully identify the causal linkage between AI adoption and scientific impact. Nevertheless, our findings demonstrate that currently attributed uses of

AI in science primarily augment cognitive tasks through data processing and pattern recognition. Looking forward, these findings illuminate a critical and expansive pathway for AI development in science. To preserve collective exploration in an era of AI use, we will need to reimagine AI systems that expand not only cognitive capacity but also sensory and experimental capacity, enabling scientists to search, select, and gather new types of data from previously inaccessible domains rather than merely optimizing analysis of standing data. The history of major discoveries has been most consistently linked with new views on nature. Expanding the scope of AI's deployment in science will be required for sustained innovation and to stimulate new fields rather than merely automating existing ones.

[***Supplementary Notes 2.2-***The effect of AI on individual scientists' careers] As shown in the main text, scientists who adopt AI tend to achieve higher scholarly productivity and impact. However, an alternative explanation for this observation is whether these advantages are driven by AI adoption itself, or whether scientists who are already on a trajectory toward greater productivity and impact are simply more likely to adopt AI. To investigate this, we conducted a match-and-comparison analysis of scientists with similar early-career productivity and impact trajectories, but who differ in their subsequent AI adoption behavior (Supplementary Fig. S16). Specifically, we filter 11,019 scientists who began adopting AI in the third year of their careers and match them with 1,926 scientists who exhibited comparable annual citation counts during their first three years but never adopted AI. The comparison reveals that starting from the third year, when the former group began adopting AI, a divergence in annual citation counts emerges between the two groups. By the tenth year of their careers, the scientists who adopted AI in their third year have 24.45% higher annual citations than their non-AI-adopting counterparts with similar early trajectories. Similarly, their annual productivity in the tenth year is 9.61% higher than the matched group of non-AI-adopters with comparable early-career productivity. The same pattern holds when analyzing a second cohort: 9,837 scientists who adopted AI in the fifth year of their careers were compared with peers who had similar citation and productivity levels in their first five years but never adopted AI. Taken together, these findings suggest that for scientists with comparable early-career positions, adopting AI itself contributes to their subsequent advantages in productivity and impact.



Supplementary Figure S16. Match-and-comparison analysis on the role of AI in driving the advantages of scientific productivity and scholarly impact. (a) Comparison between scientists who began adopting AI in the third year of their careers and matched counterparts who exhibited comparable annual citation counts during their first three years but never adopted AI. **(b)** Comparison between scientists who began adopting AI in the third year of their careers and matched counterparts who exhibited comparable annual productivity during their first three years but never adopted AI. **(c)** Comparison between scientists who began adopting AI in the fifth year of their careers and matched counterparts who exhibited comparable annual citation counts during their first three years but never adopted AI. **(d)** Comparison between scientists who began adopting AI in the fifth year of their careers and matched counterparts who exhibited comparable annual productivity during their first three years but never adopted AI. The results suggest that for scientists with comparable early-career positions, adopting AI itself contributes to their subsequent advantages in productivity and scholarly impact. For all panels, 99% CIs are shown as shaded envelopes.

3. Considering Confounding Factors in Research Concentration: The finding that "AI in science has become more concentrated around specific hot topics" and that "AI-assisted research focuses on a narrower scope" might also be influenced by confounding factors. Certain topics may inherently lend themselves more to AI application, making them appear disproportionately represented. The study should discuss whether external factors, such as topicality, data availability or funding priorities, also contribute to this trend.

Response 1.3: We thank the Reviewer for this suggestion. We agree that a variety of external factors could make certain topics more amenable to AI augmentation (and automation), disproportionately representing them among AI-supported work. Following your suggestion, we examine how the external factors of topicality, data abundance, and funding priority might influence the selectivity of AI adoption across different topics. Moreover, we also examine the factor of original, pre-AI topical impact.

For topicality, certain topics may inherently lend themselves more to AI applications, potentially making them more likely (and frequently) represented in AI-assisted research.

To investigate this possibility, we analyzed the topical correlations between AI and non-AI research (Supplementary Fig. S22). Based on citation relationships between AI and non-AI papers, we found that AI research across various fields is more frequently cited by non-AI studies than by AI studies themselves. This suggests that topics in AI research influence non-AI research rather than forming isolated, self-referential clusters. In this way, we find no obvious difference in topic selection between AI and non-AI research, indicating that inherent topicality does not account for the observed narrowing focus of AI-augmented research.

For original impact, it is possible that AI tends to concentrate on more fruitful areas rather than topics with lower original impact. To evaluate this, we categorize our selected papers into 1,883 topics according to the OpenAlex taxonomy. We calculate the average citation count and disruption score[1-2] of non-AI papers within each topic, obtaining proxies for the topic's original influence and novelty. We then examined, across topics, the correlation between degree of AI penetration and original influence and novelty across different AI development eras, where the absolute values of Pearson's r remain below 0.1 in all cases (Supplementary Fig. S23). This suggests that AI adoption does not selectively favor topics based on their original impact. Instead, AI homogeneously penetrates topics with both high and low original influence.

For funding priority, funding agencies may tend to differentially support AI research in specific topical areas. To evaluate this, we extract 31,115,808 funding data mentions from 32,437 acknowledged funding agencies from the Web of Science dataset and categorize our selected papers into 1,883 topics as above. We calculate the deviation of the probability that AI papers are funded in each topic from the average level and obtain the indicator of funding priority to each topic. We then examine, across the topics, the correlation between AI penetration and funding priority across different AI development eras, where the absolute values of Pearson's r are below 0.1 in all cases (Supplementary Fig. S24). These results suggest that across topics and AI eras, there are heterogeneous funding priorities for different topics, but these priorities are not clearly correlated with the proportion of AI adoption, suggesting that policymakers' choices do not obviously affect AI adoption on selective topics.

For data abundance, it is possible that AI research may shrink its focus of attention to scientific areas with an abundance of data, which are most amenable to data-hungry AI algorithms. To evaluate the impact of data abundance on the selectivity of AI adoption across different topics, we extract and quantify the appearance frequency of data-related terms, such as "data" and "dataset", in the titles and abstracts of papers within each topic. We then normalize these frequencies to serve as an indicator of data abundance for each topic. Next, we examine, across topics, the correlation between degree of AI penetration and data abundance across different AI development eras (Supplementary Fig. S25). Results show that data abundance is diverse across topics, while across AI eras, data abundance is positively and significantly ($p < 0.001$ for all eras) correlated with the proportion of AI adoption, suggesting that AI is more likely to be adopted in topics with abundant data resources. Notably, the correlation of AI-use with frequency of mentioned

data resources rises with the size and intensity of AI-models from 0.24 in the machine learning era to 0.36 in the deep learning era to 0.43 in the large model era .

In general, these extended analyses reveal that *data abundance* is a major external factor influencing the selectivity of AI adoption across different topics, contributing to the observed disproportionate representation within the knowledge space and the contraction of scientific focus. This finding helps to explain the core mechanism underlying our claim that AI is increasingly associated with a narrowing of scientific attention. Moreover, as models have become larger with more learned parameters, data has come to play an increasingly influential role in determining where AI will be deployed and where it will not. Nevertheless, as we demonstrate in the analysis described in the paragraph below, the topical narrowing of AI papers cannot be explained by data abundance alone. Overall, this finding provides valuable insight into how AI can be better leveraged to promote sustainable and comprehensive scientific development. We explore these implications in the new last paragraph of the discussion. We are grateful to the Reviewer for encouraging us to explore the data prevalence factor, which our initial draft mentioned but did not directly estimate.

We further examined the consistency of our findings regarding the contracted knowledge extent of AI-augmented research across different topics. We conducted stratified analyses to control for the effects of the external factors discussed above, which may impact the finding as confounding or explanatory variables (Supplementary Fig. S26). Results show that when we control each external factor by partitioning them into 10 tiers based on magnitude, contraction in knowledge extent for AI-augmented research remains evident within each tier. Regardless of whether a topic is more or less likely to be selected for AI adoption, our findings hold consistent: existing AI-augmented research within the topic tends to cover a contracted knowledge space compared to non-AI research.

We add the external factor analysis in *Supplementary Fig. S22-S26* and *Supplementary Notes 3.2- Selectivity of AI adoption in different topics*. For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 6 in the *Main Text, Results-***AI adoption is associated with a contraction in knowledge focus among scientific fields] Generally, these results highlight an emerging conflict between individual and collective incentives to adopt AI in science, where scientists receive expanded personal reach and impact, but the knowledge extent of entire scientific fields tends to shrink and focus attention on a subset of topical areas. According to analyses on possible factors that may influence the selectivity of AI adoption across different topics, we find that factors like inherent topicality, original impact, and funding priority, remain almost unrelated to the disproportionate AI adoption (Supplementary Fig. S22-S24). In contrast, data availability appears to be a major impacting factor, where areas with an abundance of data are increasingly and disproportionately amenable to AI research, contributing to the observed concentration within knowledge space (Supplementary Fig. S25).

[***Supplementary Notes 3.2***-Selectivity of AI adoption in different topics] To gain a deeper understanding of our main findings that "AI in science has become more concentrated around specific hot topics" and that "AI-augmented research focuses on a narrower scope", we examine whether external factors might influence the selectivity of AI adoption across different topics, potentially leading to their disproportionate representation.

Topicality. Certain topics may inherently lend themselves more to AI applications, potentially making them more likely to be represented in AI-assist research. To investigate this possibility, we analyzed the topical correlations between AI and non-AI research (Supplementary Fig. S22). Based on citation relationships between AI and non-AI papers, we found that AI research across various fields is more frequently cited by non-AI studies than by AI studies themselves. This suggests that topics in AI research influence non-AI research rather than forming isolated, self-referential clusters of AI literature. There is no obvious difference in topic selection between AI and non-AI research, indicating that inherent topicality does not account for the disproportionate prevalence of AI-augmented research in different fields.

Original impact. Another possibility concerns whether the topics being squeezed out by AI are likely to be marginalized anyway. That is, does AI tend to concentrate on more fruitful areas rather than topics with lower prior and potential impact? To evaluate this, we categorize our selected papers into 1,883 topics according to the OpenAlex taxonomy. We calculate the average citation count and disruption score[1-2] of non-AI papers within each topic, obtaining proxies for the topic's original influence and novelty. We then examine, across topics, the correlation between the degree of AI penetration and the original influence and novelty across different AI development eras, where the absolute values of Pearson's r are below 0.1 in all cases (Supplementary Fig. S23). The results show that across topics and AI eras, neither original citation nor original disruption is obviously correlated with the proportion of AI adoption in that topic. This suggests that AI adoption does not selectively favor topics based on their original impact. Instead, AI homogeneously penetrates into topics with both high and low original influence.

Funding priority. Another potential factor that might influence the selectivity of AI adoption across different topics is funding priority, which means that funding agencies may tend to directly support AI research more in some specific research topics. To evaluate this, we merge information from the "grants" field in the OpenAlex dataset and acknowledgment texts in the Web of Science (WOS) dataset, obtaining 31,115,808 coded funding data acknowledgments from 32,437 funding agencies. By categorizing our selected papers into topics as mentioned above, we calculate the deviation of the probability that AI papers are funded in each topic from the average level and obtain the indicator of funding priority to the topics. We then examined, across topics, correlation between the degree of AI penetration and funding priority across different AI development eras, where the absolute values of Pearson's r are below 0.1 in all cases (Supplementary Fig. S24). Results show that across topics and AI eras, there are heterogeneous funding priorities, but funding priority is not obviously correlated with

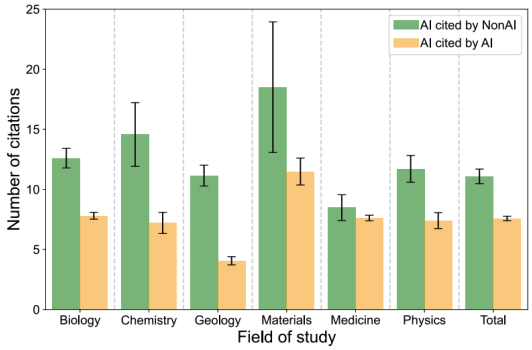
proportion of AI adoption, suggesting that policy-makers' choices do not obviously affect AI adoption on selective topics.

Data abundance. As discussed in the main text, the knowledge extent of entire scientific fields tends to shrink to focus attention on areas most amenable to AI research, such as those with an abundance of data. To directly evaluate the impact of data abundance on the selectivity of AI adoption across topics, we extract and quantify the appearance frequency of data-related terms, such as "data" and "dataset", in titles and abstracts of papers within each topic. We then normalize these frequencies to serve as an indicator of data abundance for each topic. We then examine, across topics, correlation between the degree of AI penetration and data abundance across different AI development eras (Supplementary Fig. S25). Results show that, as expected, data abundance is diverse across topics, while across AI eras, data abundance is positively and significantly ($p < 0.001$ for all eras) correlated with the proportion of AI adoption, suggesting that AI is more likely to be adopted in topics with abundant data resources. Notably, the correlation of AI-use with frequency of mentioned data resources rises with the size and intensity of AI-models from 0.24 in the machine learning era to 0.36 in the deep learning era to 0.43 in the large model era.

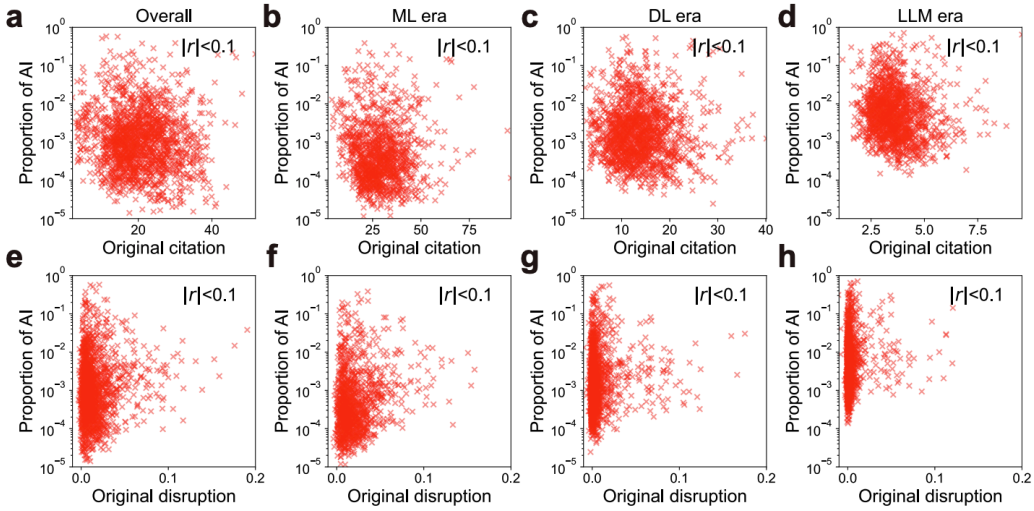
In general, data abundance is a major external factor influencing the selectivity of AI adoption across different topics, contributing to the observed disproportionate representation within knowledge space and the contraction of scientific focus. This finding elucidates factors underlying the heterogeneous adoption of AI across topics and provides valuable insights into how AI can be better leveraged to promote sustainable and comprehensive scientific development.

Given the selectivity of AI across topics, we further examined the consistency of our findings regarding the contracted knowledge extent of AI-augmented research in different topics. We conducted stratified analyses to control for the effects of the external factors discussed above, which may impact the finding as confounding variables (Supplementary Fig. S26). Results show that when we control each external factor by partitioning them into 10 tiers based on magnitude, contraction in knowledge extent for AI-augmented research remains evident within each tier. Regardless of whether a topic is more or less likely to be selected for AI adoption, our findings hold consistently: existing AI-augmented research within the topic tends to cover a contracted knowledge space compared to non-AI research.

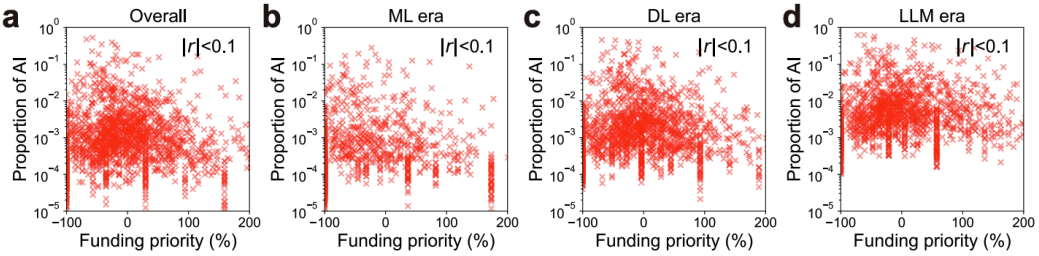
[1] Lingfei Wu, Dashun Wang, and James A Evans. "Large teams develop and small teams disrupt science and technology". In: Nature 566.7744 (2019), pp. 378–382.
[2] Michael Park, Erin Leahey, and Russell J Funk. "Papers and patents are becoming less disruptive over time". In: Nature 613.7942 (2023), pp. 138–144.



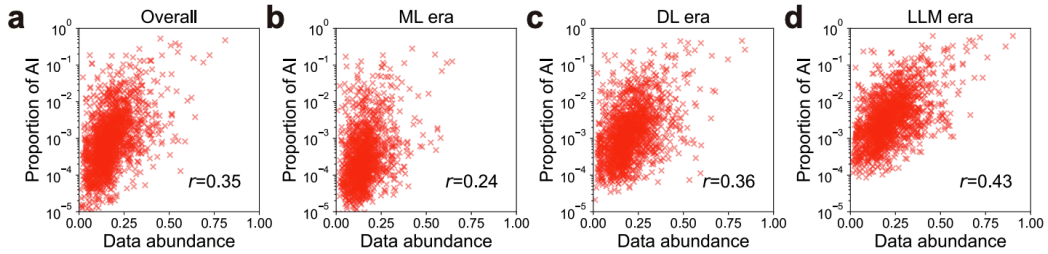
Supplementary Figure S22. Selectivity of AI adoption in different topics regarding topicality itself. The citation relationships between AI and non-AI papers show that AI research across various fields is more frequently cited by non-AI studies than by AI studies themselves. This suggests that topics in AI research influence non-AI research rather than forming isolated, self-referential clusters of AI literature. There tends to be no obvious difference in topic selection between AI and non-AI research, indicating that inherent topicality does not account for the observed more narrow focus of AI-augmented research. For all panels, 99% CIs are shown as error bars.



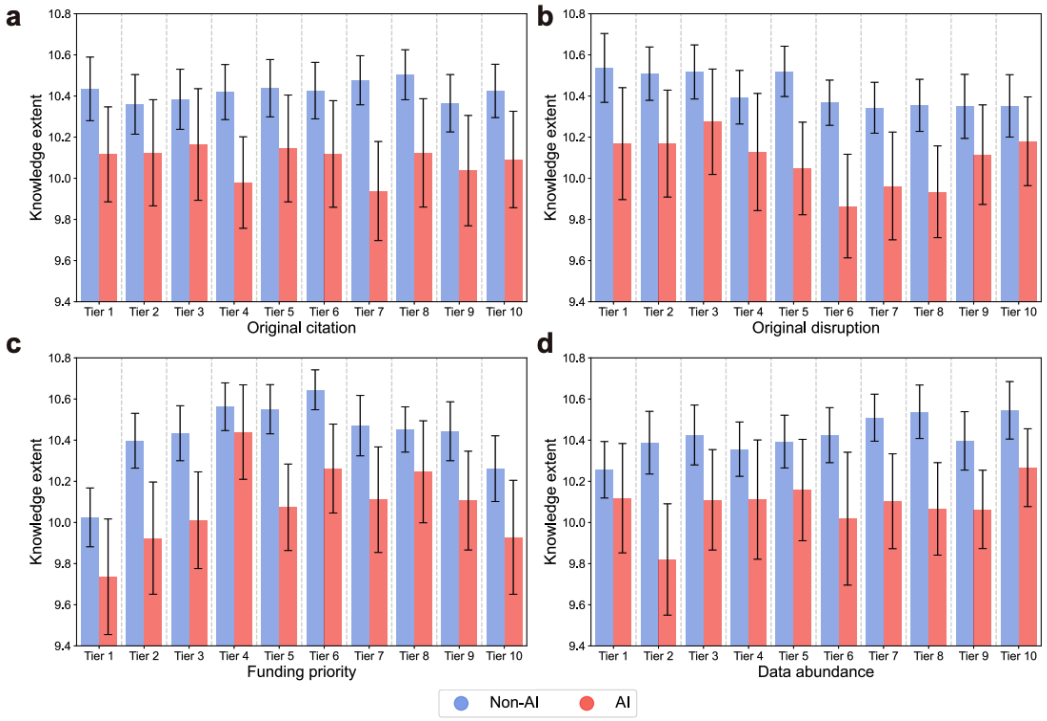
Supplementary Figure S23. Selectivity of AI adoption in different topics regarding the topics' original impact. In all panels, each scatter shows the degree of AI penetration in a topic and the original citation count or disruption score within that topic. The corresponding Pearson's r is indicated correspondingly in each panel. Results show that across topics and AI eras, neither original citation nor original disruption is obviously correlated with the proportion of AI adoption, suggesting that AI adoption does not selectively favor topics based on original impact.



Supplementary Figure S24. Selectivity of AI adoption in different topics regarding funding priority across the topics. In all panels, each scatter shows the degree of AI penetration in a topic and funding priority to that topic. The corresponding Pearson's r is indicated correspondingly in each panel. Results show that across topics and AI eras, there are heterogeneous funding priorities for different topics, but funding priority is not obviously correlated with the proportion of AI adoption, suggesting that AI adoption does not selectively favor topics prioritized by funding agencies.



Supplementary Figure S25. Selectivity of AI adoption in different topics regarding data abundance in the topics. In all panels, each scatter shows the degree of AI penetration in a topic and the data abundance in that topic. The corresponding Pearson's r is indicated correspondingly in each panel. The results show that across topics and AI eras, the data abundance is positively correlated with the proportion of AI adoption, suggesting that AI is more likely to be adopted in topics with abundant data resources. Notably, the correlation of AI-use with frequency of mentioned data resources rises with the size and intensity of AI-models.



Supplementary Figure S26. Stratified analyses controlling for the effects of potential confounding variables on the contracted knowledge extent of AI-augmented research. The results show that when dividing each external factor into 10 tiers based on magnitude, the contraction in knowledge extent for AI-augmented research remains evident within each tier. Therefore, regardless of whether a topic is more or less likely to be selected for AI adoption, existing AI-augmented research within the topic consistently covers a contracted knowledge space compared to non-AI research. For all panels, 99% CIs are shown as error bars.

Clarity and Context

The paper is well-written, with a clear abstract that effectively summarizes the key findings. The introduction provides a strong contextual background, making the research questions explicit. The conclusions succinctly distill the findings and their broader implications. However, a more structured discussion of limitations and potential future research directions would improve the clarity of the argument.

Response 1.4: We thank the Reviewer for this suggestion. As detailed above, we have expanded the discussion with limitations and potential future research directions and adjusted its structure following this guidance. We believe that these articulations enhance the paper's scholarly rigor by positioning our findings in conformity with rigorous epistemic standards. We couple this with a forward-looking discussion about expanding AI's sensory and experimental capacities to suggest a constructive research agenda that builds on our insights without overstating their definitive character.

According to your suggestion and the above discussion, we make the following revision to our manuscript:

a. We added a concluding paragraph to the *Discussion* section that highlights methodological limitations of the study, and balanced discovered insights with articulation of a future research program that will be required to resolve uncertainty and the conflict between individual and collective scientific incentives.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 8 in the *Main Text, Discussion*-**study limitations and a future research program articulated to grow credible certainty about findings and remediate unintended consequences] … While our analysis provides new insight into AI's impact on science, clear limitations remain. Our identification approach, though validated by experts, misses subtle and unmentioned forms of AI use, and our focus on natural sciences excludes important domains where AI adoption patterns may differ. Moreover, despite consistently suggestive evidence, we cannot fully identify the causal linkage between AI adoption and scientific impact. Nevertheless, our findings demonstrate that currently attributed uses of AI in science primarily augment cognitive tasks through data processing and pattern recognition. Looking forward, these findings illuminate a critical and expansive pathway for AI development in science. To preserve collective exploration in an era of AI use, we will need to reimagine AI systems that expand not only cognitive capacity but also sensory and experimental capacity, enabling scientists to search, select, and gather new types of data from previously inaccessible domains rather than merely optimizing analysis of standing data. The history of major discoveries has been most consistently linked with new views on nature. Expanding the scope of AI's deployment in science will be required for sustained innovation and to stimulate new fields rather than merely automating existing ones.

Final Assessment

This study makes a significant contribution to understanding the impact of AI on scientific research, all the while being timely and novel. Its findings will make a valuable resource for policymakers, researchers, and academic institutions. While the paper effectively demonstrates AI's dual role in expanding individual impact and contracting scientific focus, addressing concerns about definitional clarity, overdetermination of findings, and potential confounding factors are necessary for transparency and for

appropriately situating the analysis. Overall, this is a compelling and insightful paper that advances the conversation on AI's role in shaping the future of science.

Response: We are delighted that the Reviewer finds our work scientifically relevant and significant, especially as AI methods are increasingly adopted across the scientific enterprise. With these valuable suggestions, we believe our manuscript is substantially improved by addressing the mentioned concerns.

We sincerely appreciate the Reviewer for the valuable time and effort spent engaging with our manuscript. We believe that our manuscript has substantially improved in clarity and robustness across this revision cycle.

Referee #2

Key results

This paper focuses on the potential influences of AI on science and scientists, as well as the potential conflict between individual and collective benefits. The researchers conclude that there is an accelerated adoption of AI among scientists but that there are less topics being studied and less interaction. They identify a paradox as being the movement towards rich areas of data, but that there is work to automate existing fields but not generate new fields.

Validity

I have some concerns about the overall timing of the analysis. My overall impression of this manuscript is that it is somewhat premature.

Response: We thank the Reviewer for the time and effort reviewing our manuscript and engaging with our research. In this revision cycle, we endeavor to address all of the comments and concerns so kindly provided. First, we include multiple improvements according to your concerns regarding the overall timing of our analysis.

a. For problems regarding the trends of AI growth shown in our Fig. 1d and Fig. 1e, we identify how these issues are primarily due to the relatively low temporal resolution used in our initial statistics. To address this, we increase the temporal resolution in our statistics from one year to one month, allowing for more accurate statistics of AI's growth trend across different eras.
b. Considering that LLMs have not been around for very long, which leads to limited data for analysis, we update the OpenAlex dataset to the latest version, ensuring that our analysis covers as much recent use and impact as possible. We also replicate our results with only the subset of papers published during the LLM era. Consistent results in such separate, exploratory analysis verify the robustness of our general findings over decades of AI development through the era of LLMs.

Moreover, we also revise the manuscript following other detailed suggestions you provided.

c. We clarify our definition of "AI papers", namely papers adopting AI to assist research in natural science fields. We also illustrate top AI methods adopted in different disciplines and eras, providing a more intuitive understanding of what is identified as AI in our analysis.
d. We provide more detailed explanations of our identification accuracy evaluation, making the procedure and the involved measurements in human expert labeling and BERT accuracy evaluation clearer.
e. We provide more information about the citation, including its accumulation over time and multiple alternative statistical indicators beyond the average number.
f. We revise the introduction to better convey our research question. We also polish the captions of our figures, making the discussion more concise.

With respect to the timing of the article, we believe that these changes increase the maturity of AI's use in science (by a year), with even stronger effects. For example,

AI-assisted science during the LLM era is nearly twice as correlated with mentions of data use than in the earlier machine learning (ML) era. Our hope is that early findings about AI's influence on scientific careers and scientific ideas can allow scientists and science policy-makers (e.g., funders) to steer the use of AI toward collective scientific benefit, rather than waiting for AI's influence on science to become definitively settled by inertia. In short, we seek to survey AI's use and influence early in order to inform beneficial collective action.

We detail our revisions point-by-point in response to the Reviewer's comments as follows.

Data & methodology

I have some concerns regarding the method. I note that LLMs are recent and fundamentally different from machine learning methods.

Figure 1 (d) is confusing to me. The graph generally shows increases in the share of papers with AI. In approximately 2010, the big data era started. This provided input for machine learning research, which increased. Machine learning is a valid tool, based on statistical algorithms. It is different from LLMs, which is text based.

Prior to the emergence of Chat-GPT in late 2022, researchers were not using LLMs. Mass adoption of use, generally, takes time, even though LLMs were adopted quite rapidly. A paper can be submitted, but it takes time to go through the review process and be published.

When I examine this figure, I see that the increase in machine learning research (which started around 2015) has started to taper off, especially in materials science and chemistry. Although increasing in the other areas, the rate seems to be lessening. Therefore, this implies to me that the paper is somewhat premature. If something is truly making an impact, it would not be tapering off as such.

In figure 1E, the trends were already going almost straight up. They started rising after 2015 and, although they appear to be increasing rapidly after 2023, the momentum started beforehand. Therefore, it is not clear that this trend can be attributed to AI.

Response 2.1: We thank the Reviewer for this comment. We agree that the original trends of AI growth shown in Fig. 1d and Fig. 1e raise questions. Upon further analysis, we identified that these issues are primarily due to the relatively low temporal resolution used in our initial statistics. Specifically, due to the rapid growth of AI adoption in recent years, there are noticeable changes in the proportion month by month, where aggregating the data into yearly intervals introduces substantial inaccuracy. To address this, we increase the temporal resolution in Fig. 1d and Fig. 1e from one year to one month, allowing for much more accurate statistics of AI's growth trend. In addition, we update the OpenAlex dataset to the latest version, extending the coverage period from the original cutoff date of May 30, 2024, to March 31, 2025, ensuring that our analysis remains maximally up-to-date. The results show that the proportion of AI in all

disciplines keeps growing, indicating an increasing prevalence and rapid development of AI in science.

Beyond the general upward trend of AI adoption in science, we further investigate how the recent emergence of LLMs influences this trajectory. We calculate the monthly increase in proportion of AI papers and researchers across disciplines in recent years (Supplementary Fig. S7). Results show that following the release of ChatGPT in December 2022, which we regard as the beginning of the LLM era, growth rates in the proportion of papers and researchers initially remain consistent with prior trends. A period of time later, however, these growth rates begin to exhibit a marked acceleration across disciplines, providing evidence for the impact of LLM advancement. These patterns align with the intuitive expectation: the mass adoption of LLMs in natural science research requires time, and there is an inherent delay for LLM-based research to pass through peer review and reach publication.

Here, we also note that in the latest OpenAlex snapshot, the dataset maintainers made several corrections, including the removal of abstracts containing invalid or junk content and updates to higher-resolution authorship information. These changes result in adjustments to the numerical values reported in our analysis; however, all original conclusions remain valid and unaffected.

According to the Reviewer's suggestion and above discussion, we made the following revisions to our manuscript:

a. We extend the coverage period of our dataset to March 31, 2025, and increase the temporal resolution in the *Main Text*, *Fig. 1d* and *Fig. 1e* from one year to one month. This provides a more accurate representation of AI's growth trend.
b. We add analysis on how the recent emergence of LLMs influences the increasing trend of AI in *Supplementary Notes 1.4-Increasing prevalence of AI adoption in science* and *Supplementary Fig. S7*, where the growth trend of AI adoption aligns with the emergence of LLMs.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 2 in the *Main Text, Introduction***] … We conduct a large-scale quantitative analysis of the impact of AI on scientists and science, covering 41,298,433 research papers spanning from 1980 to 2025 in the OpenAlex dataset…

[***Supplementary Notes 1.4***-Increasing prevalence of AI adoption in science] Beyond the general upward trend of AI adoption in science, we further investigate how the recent emergence of LLMs influences this trajectory. We calculate the monthly increase in the proportion of AI papers and researchers across disciplines in recent years (Supplementary Fig. S7). Results show that following the release of ChatGPT in December 2022, which we regard as the beginning of the LLM era, growth rates in the proportion of papers and researchers initially remain consistent with prior trends. A period of time later, however, these growth rates begin to exhibit marked acceleration across disciplines, providing evidence for the impact of LLM advancement. Meanwhile, it also aligns with intuitive

expectation: **the mass adoption of LLMs in natural science research requires time, and there is an inherent delay for LLM-based research to pass through peer review and achieve publication.**

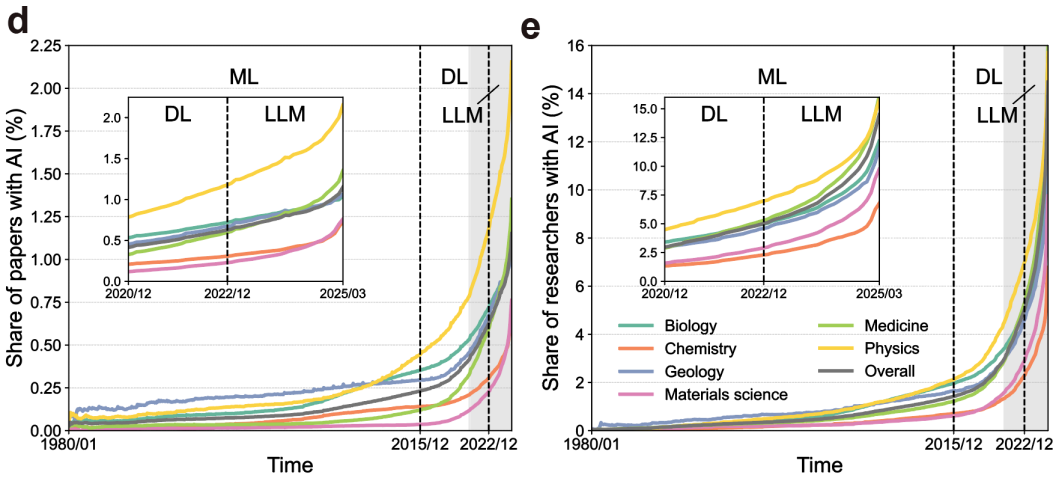
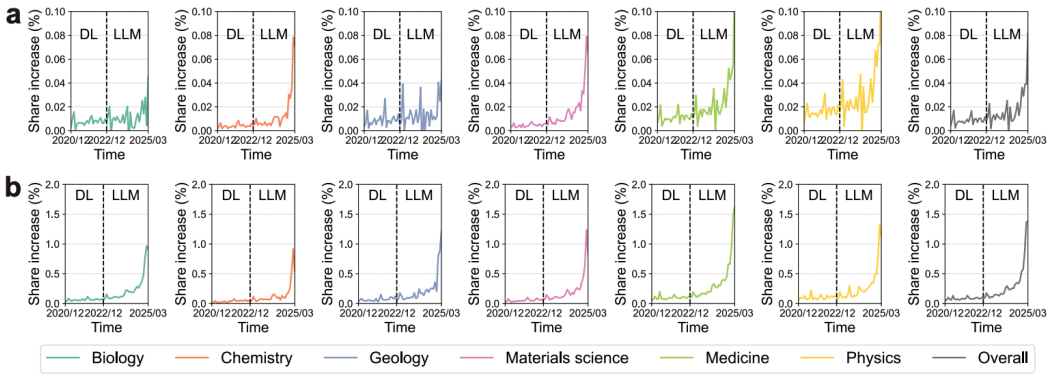


Figure 1. Increasing prevalence of AI adoption in science. (d)-(e) The growth of AI-augmented papers (d, $n = 41,298,433$) and AI-adopted researchers (e, $n = 5,377,346$) across the eras of ML, DL, and LLM between 1980 and 2025 in the selected scientific disciplines.



Supplementary Figure S7. Monthly increase in the proportion of AI (a) papers and (b) researchers across disciplines in recent years. Following the release of ChatGPT in December 2022, growth rates in the proportion of papers and researchers initially remain consistent with prior trends. A period of time later, these growth rates begin to exhibit a marked acceleration across disciplines, providing evidence for the impact of LLM advancement and use.

There are also many other factors, in general, that result in the increase in publications at certain points in time. Collaboration with researchers worldwide has increasingly occurred, supported, in part, by document sharing and other collaborative tools, as well as publicly available data (e.g., the human genome mapping). Some journals, with which I am familiar, have increased their page count over the past 10 years or have published additional papers and supporting material online. Thus, I have a concern that there are other factors that might be influencing the results.

Response 2.2: We thank the Reviewer for raising these important considerations. It is true that the number of academic publications has been increasing over time[1]. To

illustrate the rising prevalence and rapid development of AI in science with normalization to the total number of publications, we show the relative proportion of AI papers and AI researchers among all papers and researchers at each time point in Fig. 1d and Fig. 1e. Despite the increase in the total number of publications, these relative shares keep growing. For better clarification, we revise the corresponding text in the manuscript.

[**Page 3 in the *Main Text, Results*-**Increasing prevalence of AI adoption in science] … Statistically, despite the overall rise in number of papers published annually across all disciplines[1], the share of AI-augmented papers surged by 10.70 (geology, $Z = 348.60$, $p < 0.001$ and $df = 1$ in Cochran-Armitage test) to 51.89 (biology, $Z = 1388.70$, $p < 0.001$ and $df = 1$ in Cochran-Armitage test) times from 1980 to 2025 (Fig. 1d). Similarly, the proportion of researchers adopting AI has grown even more rapidly, from 135.46 times in geology ($Z = 546.81$, $p < 0.001$ and $df = 1$ in Cochran-Armitage test) to 362.16 in physics ($Z = 2237.51$, $p < 0.001$ and $df = 1$ in Cochran-Armitage test) (Fig. 1e)…

[1] Johan SG Chu and James A Evans. "Slowed canonical progress in large fields of science". In: Proceedings of the National Academy of Sciences 118.41 (2021), e2021636118.

Appropriate use of statistics and treatment of uncertainties

The Fleiss's Kappa of 0.9 seems very high and indicates almost perfect agreement. They are measuring whether a paper is an AI or not an AI paper, which would explain the high value. I think this could be made clearer. Then, BERT model has a 13% error, given an F1 score of 0.876. Then, this can be interpreted as being below humans. Is this what is intended?

Response 2.3: We thank the Reviewer for their engagement and this important question. Given papers processed by the BERT model, our identification accuracy evaluation includes two steps. In the first step, we recruited a team of human experts to label the results. We use Fleiss's Kappa to measure the agreement between human experts. The Reviewer is correct that the task of judging whether a paper is AI-augmented or not is relatively clear, enabling us to achieve high agreement. When agreement is high, we believe that the experts' labels are sufficiently reliable to regard them as ground truth. In the second step, we compare the BERT classification results with these ground truths and obtain an F1-score of 0.875 (value adjusted as we update the dataset). This suggests that the BERT model's classifications are sufficiently accurate to justify their use for automatically classifying the full corpus of papers and support our subsequent analysis.

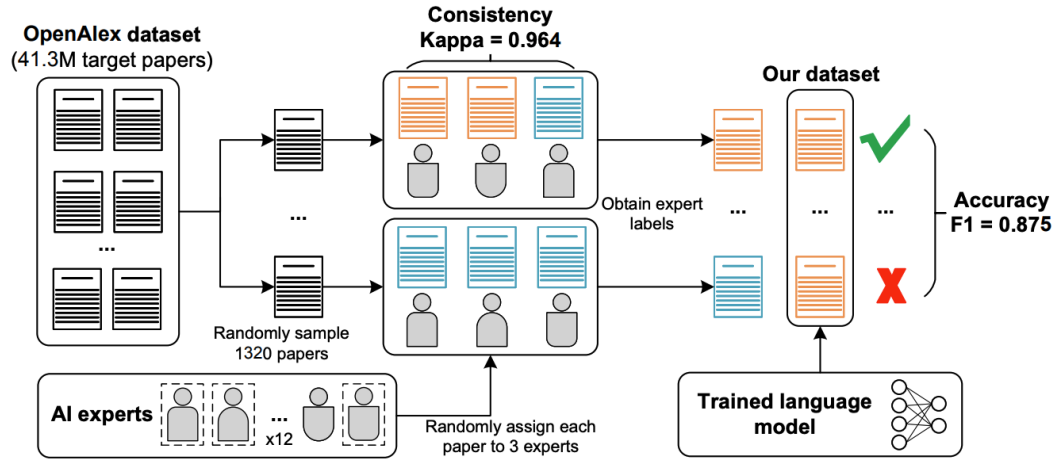
In brief, we use Fleiss's Kappa to determine whether expert labels are reliable enough to be used as ground truth and then use the F1-score to determine whether BERT classification results are sufficiently accurate compared to ground truth. We now also propagate our error assessments by cascading the two steps and evaluate identification accuracy rather than comparing performance between the BERT model and human experts. We also note that an F1-score of 0.875 does not necessarily correspond to a 13% error rate. Our precision is 0.897 and recall is 0.854, resulting in

$$F1 = 2\frac{precision \times recall}{precision + recall} = 2\times\frac{0.897\times0.854}{0.897+0.854} = 0.875$$

To make the procedure of identification accuracy evaluation clearer, we have revised the corresponding text in the manuscript and add an illustration in Extended Data Fig. 2. For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 3 in the *Main Text, Results*-**Increasing prevalence of AI adoption in science] To evaluate the accuracy of our identification, we recruited a team of human experts to validate the results (Methods M4 and Extended Data Fig. 2). These experts demonstrate strong consensus across their independent annotation of papers arbitrarily sampled from the six disciplines mentioned above, achieving an average Fleiss' Kappa of 0.964. The BERT model attains an F1-score of 0.875 when evaluated against this expert-labeled data…



Extended Data Fig. 2. Procedure of accuracy evaluation via expert evaluation. We randomly sample 1320 papers and delegate three experts to scrutinize the identification results for each paper. We then draw the final expert label of each paper from the three experts according to the principle of the minority obeying the majority and validate the result of the language model with it. Results indicate strong consistency among experts and high accuracy with our identification results.

At some points, I found it confusing as to whether the authors were referring to papers on AI or using AI.

Response 2.4: We thank the Reviewer for this excellent suggestion. We regret not clearly defining what "AI papers" refer to. **In our analysis, we focus on papers adopting AI to assist research in natural science fields, rather than those developing AI methodologies themselves.** To identify these papers, we exclude the disciplines of computer science and mathematics, where papers on the development of AI itself are published. We select six representative disciplines covering the vast majority of the natural sciences—biology, medicine, chemistry, physics, materials science, and geology, and analyze AI-assisted research within them.

To illustrate the most commonly adopted AI methods in natural science research, we list the top 15 AI methods across all disciplines for all eras in Fig. 1c. These top methods include support vector machines and principal component analysis from the ML era, as well as convolutional neural networks and generative adversarial networks, which represent important techniques of the DL era. Meanwhile, large language models, which have emerged in recent years, also rank among the most frequently utilized methods in the latest period. We also show in detail the top 10 most frequently used AI methods for each discipline and each AI era in Supplementary Tables S5–S11. These provide a more intuitive understanding of what is identified as AI in our analysis.

According to your suggestion and the above discussion, we make the following revisions to our manuscript:

d. We revise the *Introduction* in the *Main Text*, *Methods-M1. Dataset and Paper selection* and *Methods-M3. Design and fine-tune the language model for AI paper identification*, specifying that we focus on AI-assisted research using AI for analysis, rather than those developing AI itself.

e. We list the top 15 AI methods in the *Main Text*, *Fig. 1c*, illustrating the most commonly adopted AI methods in natural science research.
f. We show in detail the top 10 most frequently used AI methods for each discipline and each AI era in *Supplementary Tables S5–S11* and *Supplementary Notes 1.3-Profile of the identified AI papers*, providing an intuitive understanding of what is identified as AI in our analysis.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 2 in the *Main Text, Introduction***] … Notably, we do not focus on work that develops AI methodologies themselves, but rather on papers that augment research in natural science fields by adopting AI techniques ranging from early machine learning algorithms to contemporary generative AI. Specifically, we select six representative disciplines that cover the vast majority of the natural science contributions—biology, medicine, chemistry, physics, materials science, and geology—while we exclude computer science and mathematics in order to exclude the development of AI itself…

[**Page 3 in the *Main Text, Results-***Increasing prevalence of AI adoption in science] … Counting all eras and disciplines collectively, the most commonly adopted AI methods in natural science research include support vector machines and principal component analysis from the ML era, and convolutional neural networks and generative adversarial networks from the DL era. The large language model, which has emerged in recent years, also ranks among the most frequently utilized methods (Fig. 1c and Supplementary Table S5-S11).

[**Page 8 in the *Methods-***M1. Dataset and Paper selection] In this paper, we emphasize the adoption of AI methods in conventional natural science disciplines and exclude research developing AI methodologies themselves, separating the influence of AI on science from AI's own invention and refinement. Therefore, we select biology, medicine, chemistry, physics, materials science, and geology as representatives of natural science disciplines, while we exclude computer science and mathematics, where most works introducing and developing AI methods are published…

[**Page 9 in the *Methods-***M3. Design and fine-tune the language model for AI paper identification] … Finally, we utilize optimal ensemble models during both stages to identify all papers that use AI to support natural science research from the selected representative natural science disciplines.

[***Supplementary Notes 1.3-***Profile of the identified AI papers] For adopted AI methods, we extracted phrases from the abstracts of identified AI papers and calculated the frequency of phrases related to specific AI techniques. We identified and listed the top 10 most frequently used AI methods for each discipline and AI era (Supplementary Tables S5–S11). It is worth noting that in earlier years, due to the limited number of AI papers, the number of commonly used AI methods may be fewer than ten. The results show that during the ML era, the most frequently applied AI methods in natural science research included Artificial Neural Networks (ANN), Principal Component Analysis (PCA), and

Support Vector Machines (SVM). In the DL era, Convolutional Neural Networks (CNN), a landmark of the time, dominated the landscape, while traditional methods such as SVM continued to see widespread adoption. Since the beginning of the LLM era in 2023, Large Language Model (LLM) has increasingly appeared among the most commonly used AI methods, with their frequency rankings steadily rising across disciplines. The evolution of AI methods in natural science research reflects the development and transformation of AI technologies over the past decades.

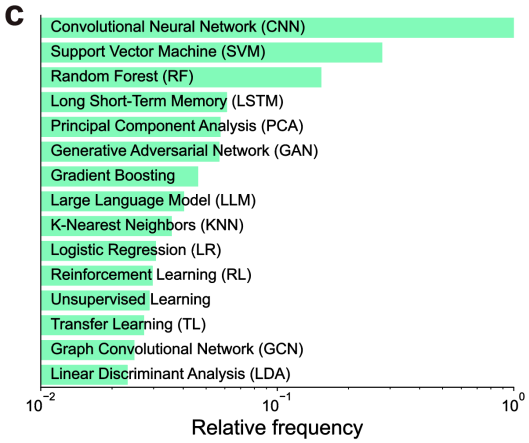


Figure 1. Increasing prevalence of AI adoption in science. (c) Relative adoption frequency of the top 15 AI methods across all disciplines for all selected AI development eras.

Supplementary Table S5. The most frequently adopted AI methods in all selected disciplines in different AI development eras.

Era	Year	Top AI methods									
		1	2	3	4	5	6	7	8	9	10
ML	[1980,1990]	Expert System	Artificial Neural Network (ANN)	Knowledge Base	Connectionist Network (CN)	Principal Component Analysis (PCA)	Hebbian Learning	Maximum Likelihood Classifier	Logistic Regression (LR)	Generalized Additive Model	Nonlinear Regression
	[1991,2000]	Artificial Neural Network (ANN)	Principal Component Analysis (PCA)	Expert System	Knowledge Base	Radial Basis Functions (RBF)	Hidden Markov Model (HMM)	Unsupervised Learning	Linear Discriminant Analysis (LDA)	Maximum Likelihood Classifier	Markov Chain Monte Carlo (MCMC)
	[2001,2005]	Artificial Neural Network (ANN)	Support Vector Machine (SVM)	Principal Component Analysis (PCA)	Hidden Markov Model (HMM)	Linear Discriminant Analysis (LDA)	Markov Chain Monte Carlo (MCMC)	Radial Basis Functions (RBF)	Unsupervised Learning	Knowledge Base	Bsyesian Network
	[2006,2010]	Support Vector Machine (SVM)	Artificial Neural Network (ANN)	Principal Component Analysis (PCA)	Linear Discriminant Analysis (LDA)	Hidden Markov Model (HMM)	Markov Chain Monte Carlo (MCMC)	Radial Basis Functions (RBF)	Non-negative Matrix Factorization (NMF)	Unsupervised Learning	Conditional Random Field (CRF)
	[2011,2015]	Support Vector Machine (SVM)	Artificial Neural Network (ANN)	Principal Component Analysis (PCA)	Markov Chain Monte Carlo (MCMC)	Linear Discriminant Analysis (LDA)	Hidden Markov Model (HMM)	Conditional Random Field (CRF)	Unsupervised Learning	Non-negative Matrix Factorization (NMF)	Radial Basis Functions (RBF)
DL	[2016,2020]	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Random Forest (RF)	Principal Component Analysis (PCA)	Generative Adversarial Network (GAN)	Long Short-term Memory (LSTM)	Decision Tree (DT)	Unsupervised Learning	Reinforcement Learning (RL)	Linear Discriminant Analysis (LDA)
	[2021,2022]	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Random Forest (RF)	Transfer Learning (TL)	Long Short-term Memory (LSTM)	Generative Adversarial Network (GAN)	Gradient Boosting	K-nearest Neighbor (KNN)	Graph Convolutional Network (GCN)	Principal Component Analysis (PCA)
LLM	2023	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Random Forest (RF)	Long Short-term Memory (LSTM)	Gradient Boosting	Large Language Model (LLM)	Generative Adversarial Network (GAN)	Transfer Learning (TL)	K-nearest Neighbor (KNN)	Graph Convolutional Network (GCN)
	2024	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Random Forest (RF)	Large Language Model (LLM)	Gradient Boosting	Long Short-term Memory (LSTM)	Generative Adversarial Network (GAN)	K-nearest Neighbor (KNN)	Shapley Additive Explanations (SHAP)	Logistic Regression (LR)
	2025	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	Large Language Model (LLM)	Random Forest (RF)	Shapley Additive Explanations (SHAP)	Long Short-term Memory (LSTM)	Gradient Boosting	Generative Adversarial Network (GAN)	Graph Neural Network (GNN)	Logistic Regression (LR)

Supplementary Table S6-11. The most frequently adopted AI methods in each discipline in different AI development eras. Please see the full tables in Supplementary Materials.

The figures are appropriately labelled, although the amount of discussion seems excessive, given the descriptions in the main paper.

Response 2.5: We thank the Reviewer for this comment. We have polished the captions of our figures, making the discussion more concise and less redundant. For your convenience, we quote the revised texts below.

Revision:

[**Page 2 in the *Main Text, Figure 1 Caption***] … **(b)** Accuracy evaluation of our identification results by human experts. For samples spanning three eras of AI, experts reached consensus with Fleiss' Kappa ≥ 0.93. Our model identification results have strong accuracy in validation against expert-labeled data with an F1-score ≥ 0.85... **(d)-(f)** The growth of AI-augmented research papers (d, $n = 41,298,433$) and AI-adopted researchers (e, $n = 5,377,346$) across the eras of ML, DL, and LLM

between 1980 and 2025 in selected scientific disciplines, where 99% CIs of monthly growth rates for AI papers and researchers are shown as error bars.

[**Page 4 in the *Main Text, Figure 2 Caption***] **(a)** Annual citations after publication of AI and non-AI papers ($n = 27,405,011$), where AI papers attract more citations. **(b)** Average annual citations of researchers adopting AI and their counterparts without AI ($n = 5,377,346$), where researchers adopting AI garner 4.84 times more citations than their counterparts without AI… **(d)** Probability density distribution of the transition time for junior scientists to become established ($n = 2,282,029$). The probability density distributions can be well-fit with exponential distributions, where junior scientists adopting AI become established approximately four years earlier than their counterparts without AI.

[**Page 5 in the *Main Text, Figure 3 Caption***] … **(b)** For visualization, we use the t-SNE algorithm to flatten the high-dimensional embeddings of a small random batch of 10,000 papers, half of which are AI papers and half are non-AI papers, into a 2-D plot. As shown by the solid arrows and circular boundaries, the knowledge extent of AI papers is smaller across the entirety of natural science, and AI papers are more clustered in knowledge space, indicating more concentration on specific problems. **(c)** Knowledge extent of AI and non-AI papers in each field ($n = 6$ disciplines $\times 2000$ samples), where AI research focuses on a more contracted knowledge space. **(d)** Knowledge entropy of AI and non-AI papers in each field ($n = 6$ disciplines $\times 2000$ samples), where AI research has a lower entropy.

[**Page 7 in the *Main Text, Figure 4 Caption***] **(a)** Knowledge extent of individual AI and non-AI paper families, i.e., an original paper and its cumulative citations ($n = 27,405,011$), where the knowledge space of individual AI paper families is broader and grows faster. **(b)** Engagement among papers that cite AI vs. non-AI papers ($n = 23,342,516$), where there are fewer follow-on interactions among papers that cite the same original paper in AI research... **(c)** Distribution of citations to AI vs. non-AI papers ($n = 27,405,011$), where AI papers tend to concentrate more on a smaller number of top papers.

Conclusions

The conclusions are ok, but I have some concerns with how they were arrived at. They seem somewhat premature. LLMs have not been around for very long and there is a lead time before publications using them appear.

Response 2.6: We thank the Reviewer for this consideration. It is undoubtably true that LLMs have not been around for very long, and there is a lead time before publications using them appear, making the data available for analysis limited. As we mentioned above, we update the OpenAlex dataset to the latest version, extending the coverage period from the original cutoff date of May 30, 2024, to March 31, 2025, which contains nearly 3 years of LLM application, ensuring that our analysis covers as much recent use and impact as possible. Also, to verify the robustness of our general findings over

decades of AI development to the specific era of LLM, we replicated our findings with only the subset of papers published during the LLM era. As we show in Supplementary Fig. S26-S28, the results in the LLM era are consistent with those from prior decades of AI development.

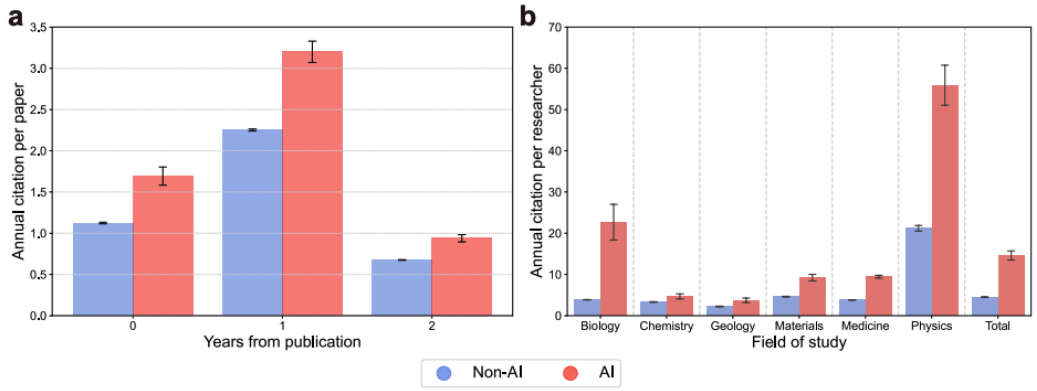
This separate analysis of the LLM era could exhibit limitations due to the lack of data caused by the relatively short history of LLMs. For example, we are not able to replicate the analysis regarding the career development of individual scientists due to the limited time span for detecting scientists' career role transitions. Besides, some results in this separate analysis manifest less significance than the corresponding results in the general findings with more data, but this does not affect the qualitative conclusions. Moreover, some findings are stronger in the LLM era, such as the correlation between data mentions and AI, as algorithms grow increasingly data hungry. More validation could necessarily be achieved as LLM develops over a longer period of time and thus produces more abundant data. Nevertheless, we believe in the value of this early analysis—years into the use of LLMs and associated large models in other domains—in order for science to be made aware of these trends, which remain consistent with others more than two decades in the making.

We add the above separate, exploratory analysis in *Supplementary Fig. S26-S28* and *Supplementary Notes 4.1-Robustness of results in the LLM era*. For your convenience, we quote the revised texts and figures below.

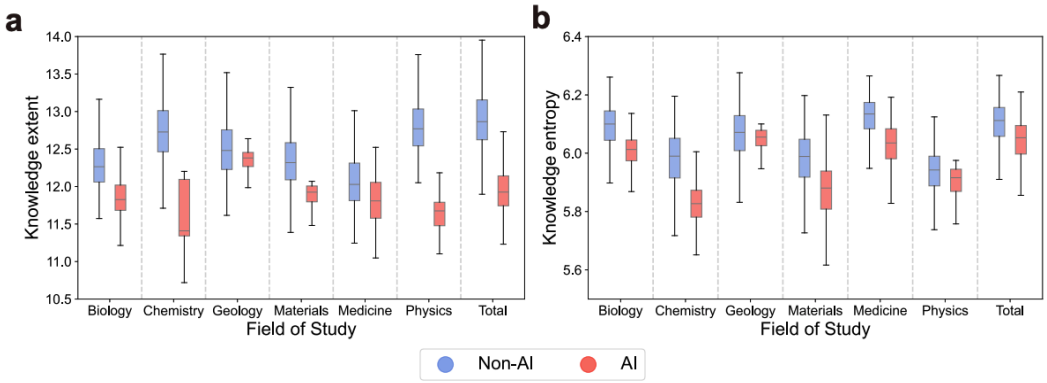
Revision:

[***Supplementary Notes 4.1-*** Robustness of results in the LLM era] Because LLMs have not been around for very long and there remains a lead time before publications using them appear, the available data regarding LLMs are necessarily limited for analysis. To verify the robustness of our general findings over decades of AI development to the specific era of LLM, we replicated our findings with only the subset of papers published during the LLM era. Of the 41,298,433 publications in our OpenAlex and Web of Science analyses that cover six disciplines, 4,797,614 are published in the LLM era. As we show in Supplementary Fig. S26-S28, all results in the LLM era are consistent with those from prior decades of AI development.

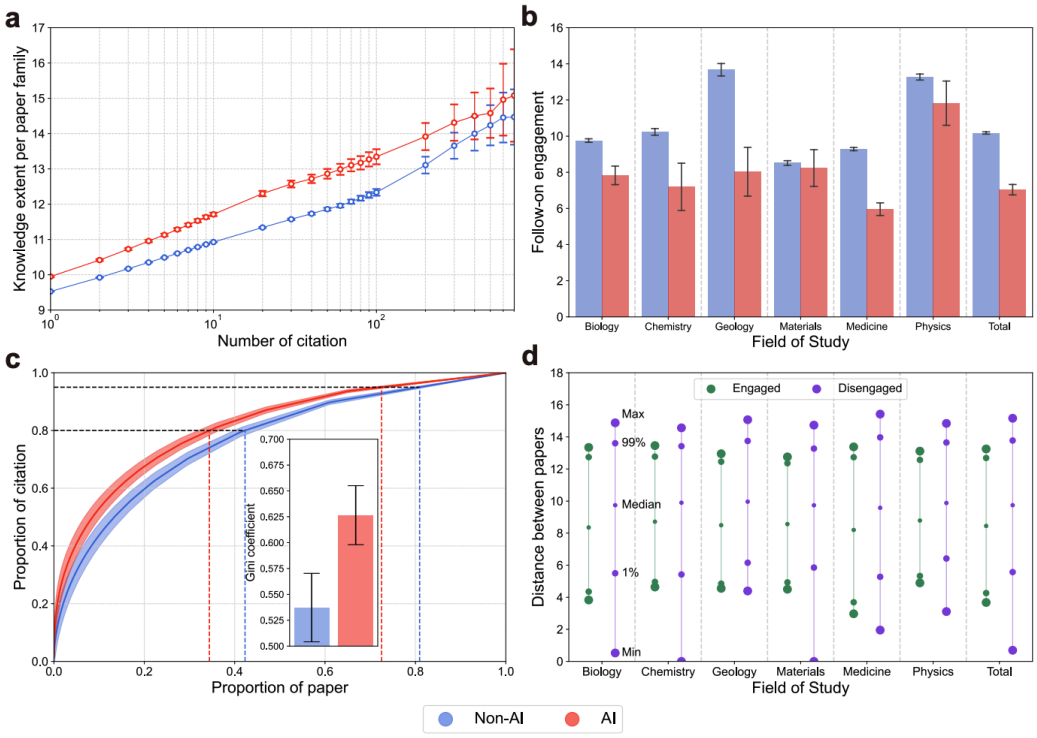
This separate analysis of AI from the LLM era could exhibit limitations due to the lack of data caused by the relatively short history of LLMs. One major point is that our method of detecting scientists' career role transitions requires scientists to maintain a certain length of publication history, which is not feasible separately in the short LLM era. Therefore, we are not able to replicate the analysis regarding the career development of individual scientists. Besides, some results in this separate analysis may appear less significant than the corresponding results in the general findings, in part from the smaller sample size, but this does not affect any qualitative conclusions. Additional validation should be achieved as LLM develops over a longer period of time and thus produces more abundant data.



Supplementary Figure S26. Replication of "AI enlarges the impact of papers" with only the subset of papers published during the LLM era. (a) Annual citations after publication of AI and non-AI papers. **(b)** Average annual citations of researchers adopting AI and their counterparts without AI. For all panels, 99% CIs are shown as error bars. Results in the LLM era are consistent with those obtained over the full time span featured in the main text.



Supplementary Figure S27. Replication of "AI adoption is associated with a contraction in knowledge extent within and across scientific fields" with only the subset of papers published during the LLM era. (a) Knowledge extent of AI and non-AI papers in each field. **(b)** The knowledge entropy of AI and non-AI papers in each field. We show 1.5 times the inter-quartile range (IQR) from the box as whiskers. Results in the LLM era are consistent with those over the full time span detailed in the main text.



Supplementary Figure S28. Replication of "over-concentration of AI research results in redundant innovation" with only the subset of papers published during the LLM era. (a) Knowledge extent of individual AI and non-AI paper families. **(b)** Engagement among papers that cite AI vs. non-AI papers. For the above panels, 99% CIs are shown as error bars. **(c)** Distribution of citations to AI vs. non-AI papers. **(d)** Distribution of distances between paper pairs that cite the same papers, with or without citing each other—engaged versus disengaged. Results in the LLM era are consistent with those on the full time span featured in the main text.

Suggested improvements

I would like to know more about the citations. Citations to AI papers are higher than to non-AI papers, even though there are vastly more non-AI papers, but AI is currently a topic with high interest, especially since the introduction of large language models. Furthermore, usually citations to recently published papers take some time to accumulate.

Response 2.7: We thank the Reviewer for this question. Considering the vastly different total number of AI and non-AI papers, as well as the time needed from paper publication to citation accumulation, in Fig. 2a, we present the **number of annual citations** received by **each paper** from year of publication until decades later, namely the "velocity" of citation. As our results show, citation velocity generally increases over the first three years after publication, peaks in years 3 to 5, and then gradually declines. Notably, annual citations to AI papers are consistently higher than those to non-AI counterparts. Here, our measurement of annual citations per paper has already been normalized to the different total number of AI and non-AI papers, and taken into account the time for recently published papers to accumulate citations.

To provide more information about the accumulation of citations over time, in addition to the "velocity" aforementioned, we also compare the "acceleration" of citation, namely

the year-over-year change in annual citation counts, between AI and non-AI papers (Supplementary Fig. S9). Results reveal that during the initial years after publication, AI papers exhibit a faster acceleration in citation growth. In the subsequent period, after reaching peak citation "velocity", AI papers exhibit a more rapid deceleration in annual citations. Throughout the entire citation lifecycle, citation "velocity" of AI papers consistently remains higher than for non-AI papers.

Considering that citation distributions are empirically right-skewed[1-2], we adopt multiple statistical indicators beyond average annual citation count to ensure more robust conclusions (Supplementary Fig. S8). For the right tail of the distribution, namely papers with high citation counts, we observe that the top 1st and 5th-percentile annual citations of AI papers exceed those of non-AI papers by 152.39% and 48.27%, respectively. For the left tail of the distribution, namely papers with low citation counts, the proportion of AI papers receiving fewer than three and fewer than five citations each year is 2.49% and 2.46% lower. These suggest that AI papers not only tend to receive higher citations but are also less likely to become low-impact publications, indicating the enhanced visibility and academic attention garnered by AI research.

Moreover, we examine the influence of AI and non-AI papers from the perspective of disruption[3-4] (Supplementary Fig. S15). Despite higher citation counts for AI papers, we find that their disruption scores are somewhat lower. This suggests that while AI papers tend to influence a larger number of subsequent studies, the higher impact tends to be primarily developmental rather than disruptive, which means that they contribute to advancing (and completing) existing fields rather than initiating new ones. This observation is consistent with our finding that AI contracts science's focus.

According to your suggestion and the above discussion, we make the following revisions to our manuscript:

a. We add the analysis regarding the "acceleration" of citations in *Supplementary Fig. S9* and *Supplementary Notes 2.1-The effect of AI on individual papers*, providing more information about the accumulation of citations over time.
b. We add multiple alternative statistical indicators to compare citations between AI and non-AI papers in *Supplementary Fig. S8* and *Supplementary Notes 2.1-The effect of AI on individual papers*, taking into account the right-skewed distribution of citation.
c. We add the comparison between AI and non-AI papers from the perspective of disruption, a measurement about the character of research paper influence other than citation, in *Supplementary Fig. S15* and *Supplementary Notes 2.1-The effect of AI on individual papers*, providing a more comprehensive understanding about the influence of AI papers.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 3 in the *Main Text, Results-***AI enlarges the impact of individual papers and enhances the career of individual scientists] … In addition to higher annual average citations, the higher scientific impact of AI-augmented papers is also reflected by

multiple alternative statistical indicators about top and bottom annual citation count (Supplementary Fig. S8).

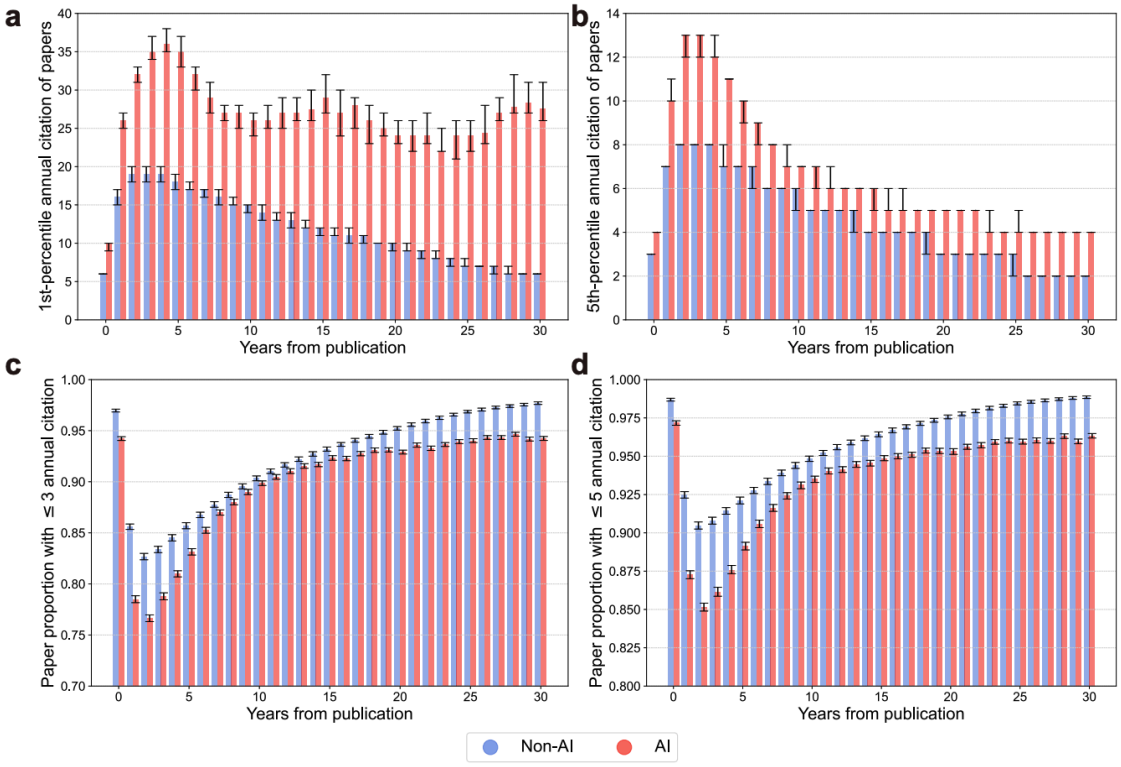
[***Supplementary Notes 2.1-***The effect of AI on individual papers] As demonstrated in the main text, AI-related papers, on average, receive higher annual citation counts than non-AI papers. Given that citation distributions are empirically right-skewed[1-2], we adopt multiple statistical indicators beyond the average annual citation counts to ensure more robust conclusions (Supplementary Fig. S8). For the right tail of the distribution, namely papers with high citation counts, we observe that from year of publication and across subsequent decades, citations for the top 1st and 5th-percentile of AI papers exceed those of non-AI papers by 152.39% and 48.27%, respectively. For the left tail of the distribution, namely papers with low citation counts, we find that, over the same time period, the proportion of AI papers receiving fewer than three and fewer than five citations each year is 2.49% and 2.46% lower, respectively, than non-AI papers. Taken together, these findings suggest that AI papers not only tend to receive higher citations but are also less likely to become low-impact publications, indicating the enhanced visibility and academic attention garnered by AI research.

[***Supplementary Notes 2.1-***The effect of AI on individual papers] In addition to the "velocity" of citation, shown above as the number of citations received per year after publication, we also compare the "acceleration" of citations, namely the year-over-year change in annual citation counts, between AI and non-AI papers (Supplementary Fig. S9). Results reveal that during the initial post-publication years when annual citations are generally increasing, AI papers exhibit a faster acceleration in citation growth. In the subsequent period, after reaching peak citation "velocity", AI papers also exhibit a more rapid deceleration in annual citations. Over the longer term, the changing patterns of annual citations to AI and non-AI papers converge, with no significant differences observed. Nevertheless, throughout the entire citation lifecycle, the citation "velocity" of AI papers consistently remains higher than that of non-AI papers.

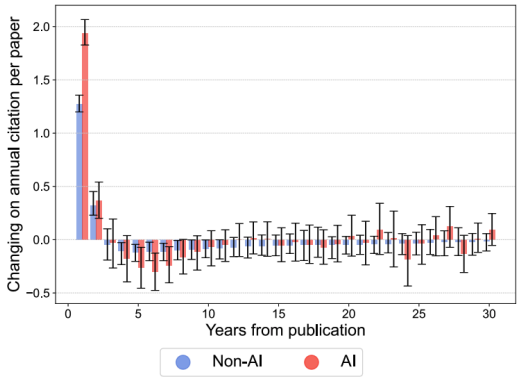
[***Supplementary Notes 2.1-***The effect of AI on individual papers] … we examine the influence of AI and non-AI papers from the perspective of disruption[3-4] (Supplementary Fig. S15). Despite higher citation counts for AI papers, we find that their disruption scores are lower. This suggests that while AI papers tend to influence a larger number of subsequent studies, the higher impact is primarily developmental rather than disruptive, which means that they contribute to advancing (and completing) existing fields rather than initiating new ones. This observation is consistent with our later finding that AI contracts science's focus.

[1] Sidney Redner. "How popular is your paper? An empirical study of the citation distribution". In: The European Physical Journal B-Condensed Matter and Complex Systems 4.2 (1998), pp. 131–134.
[2] Per O Seglen. "The skewness of science". In: Journal of the American society for information science 43.9 (1992), pp. 628–638.
[3] Lingfei Wu, Dashun Wang, and James A Evans. "Large teams develop and small teams disrupt science and technology". In: Nature 566.7744 (2019), pp. 378–382.

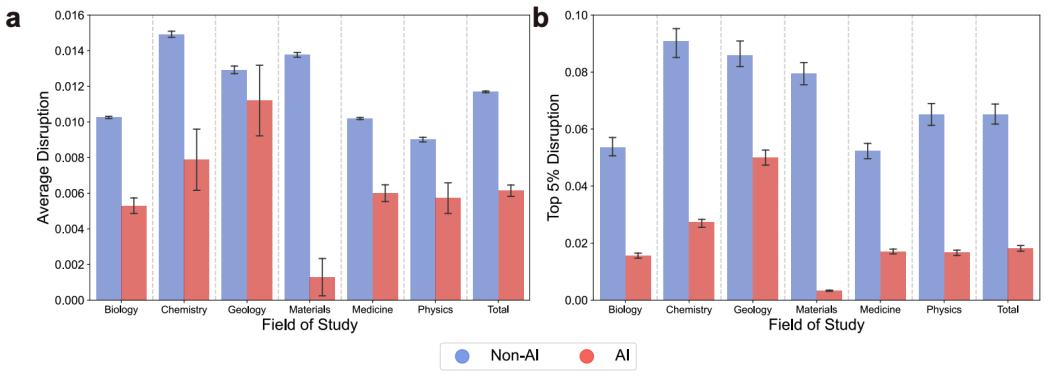
[4] Michael Park, Erin Leahey, and Russell J Funk. “Papers and patents are becoming less disruptive over time”. In: Nature 613.7942 (2023), pp. 138–144.



Supplementary Figure S8. Alternative statistical indicators for the annual citation comparison between AI and non-AI papers. (a) Top 1st-percentile annual citation for AI and non-AI papers from year of publication. **(b)** Top 5th-percentile annual citation for AI and non-AI papers from year of publication. **(c)** Proportion of AI and non-AI papers receiving fewer than three citations each year from year of publication. **(d)** Proportion of AI and non-AI papers receiving fewer than five citations each year from year of publication. For all panels, 99% CIs are shown as error bars.



Supplementary Figure S9. Comparison of “acceleration” of citation, namely year-over-year change in annual citation counts, between AI and non-AI papers. AI papers experience a faster acceleration in citation growth during the initial post-publication years, and exhibit a more rapid deceleration in annual citations after reaching peak annual citations. Throughout the entire citation lifecycle, the annual citation of AI papers consistently remains higher than that of non-AI papers. For all panels, 99% CIs are shown as error bars.



Supplementary Figure S15. Comparison of disruption between AI and non-AI papers. (a) Average disruption of AI and non-AI papers in each field. **(b)** Top 5th-percentile disruption of AI and non-AI papers in each field. For all panels, 99% CIs are shown as error bars.

Clarity and context

It is generally clear what the authors are trying to do. There were some small places where I didn't entirely understand what the authors were trying to convey. For example, it was not clear to me how the research question "aligns with the promise and peril of AI in scientific research." I mention that instance because of the importance of the statement of a research question.

Response 2.8: We thank the Reviewer for this constructive question and suggestion. We see that the word "align" was confusing. By exploring the correlates of AI adoption for scientists and science as a whole, we seek to characterize *both* the promise and peril of AI use for science. We have revised the introduction to make this more clear and quote the revised text below.

Revision:

[**Page 2 in the *Main Text, Introduction***] Here we seek to explore the promise and peril of AI in scientific research by posing the question: How does the adoption of AI by individual scientists influence the collective character and dynamic exploration of science as a whole?

Other comments

I do not have a great understanding of the OpenAlex database, which seems to have started at Microsoft and is now being supported in Europe. I have not used BERT, although I appreciate the mission of language models in general.

Response 2.9: We thank the Reviewer for this comment. We regret not providing enough details about the dataset and methods used. OpenAlex is a scientific research database supported by non-profit organizations in Europe. As you suggest, the database is built upon the foundation of Microsoft Academic Graph (MAG). Since Microsoft shut down the service of MAG in 2021, OpenAlex has continued to be updated to provide a sustainable global resource for research information. MAG and OpenAlex were partly

developed to add emerging publication types (e.g., conference papers, particularly important in fields like computer science) to the open analysis of scientific production. We also analyze the subset of OpenAlex publications in the Web of Science, a more established but narrower collection, to demonstrate the consistency of our findings with alternative databases. We add a brief introduction to OpenAlex in *Methods-M1. Dataset and Paper selection*.

Moreover, we thank you for recognizing the utilization of BERT. BERT is a powerful pre-trained language model recognized for its text representation and contextual understanding capabilities. Many studies leverage BERT as a foundation, fine-tuning it to create specialized models for automatically processing scientific texts for a wide variety of tasks[1-3]. In our work, we utilize BERT to identify AI-augmented papers from massive research corpora. We add relevant references in the main text.

Revision:

[**Page 8 in the *Methods*-**M1. Dataset and Paper selection] In this paper, we conduct our major analyses based on OpenAlex. OpenAlex is a scientific research database built upon the foundation of the Microsoft Academic Graph (MAG). Supported by non-profit organizations, OpenAlex is continuously updated, providing a sustainable global resource for research information. As of March 2025, OpenAlex contains 265.7M research papers, along with related data about citation, author, institution, etc.

[**Page 3 in the *Main Text, Results-*** Increasing prevalence of AI adoption in science] To identify AI papers in various fields across eras, we fine-tune BERT, an established language model[1-3], on articles published in explicitly AI-oriented scientific journals and conferences for automatically extracting and interpreting information from context.

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text". In: EMNLP/IJCNLP (1). Association for Computational Linguistics, 2019, pp. 3613–3618.

[2] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. "SPECTER: Document-level Representation Learning using Citation-informed Transformers". In: ACL. Association for Computational Linguistics, 2020, pp. 2270–2282.

[3] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. "SciRepEval: A Multi-Format Benchmark for Scientific Document Representations". In: EMNLP. Association for Computational Linguistics, 2023, pp. 5548–5566.

We sincerely appreciate the Reviewer for the valuable time and effort spent engaging with our manuscript. We believe that our manuscript has substantially improved in clarity, robustness, and insight across this revision cycle thanks, in large part, to this helpful review.

Referee #3

I read with great interest the fascinating manuscript by Hao et al. The topic -- the impact of AI on scientific research -- is extremely timely and the results are confirmatory of my expectations and fears. This is actually a concern for me because of confirmatory bias. I thus want to be extra careful with what is the evidence provided to the reader.

Response: We thank the Reviewer for this positive feedback on our manuscript, and share your expressed concern about the critical importance of accuracy. In our revision, we endeavor to address all the comments you so kindly provided. We believe that our revision provides readers with more convincing evidence in support of our stated findings. Here, we provide a brief summary of our main revisions below.

a. We revise the introduction, making it more explicit about what we regard as AI over the past decades.
b. We improve the description of our AI paper identification procedure and identification accuracy evaluation. We evaluate how well the identification model works for different time periods, providing this assurance of reliability for our subsequent analysis. We also improve the examples of our identification results, providing an intuitive understanding of the identification results.
c. We improve the representation of exponential curves, where we use a log-scale *y*-axis and estimate the parameters which fit to straight (log-scaled) regions of the plot. We also add error bars to show confidence intervals from parameter estimations.
d. We add robustness analysis to the effect of AI adoption on individual papers, where we use multiple statistical indicators beyond average annual citation count considering the right-skewed distribution of citations. We also distinguish special publications, like review articles, and replicate our analyses exclusively on original research papers.
e. We add robustness analysis to the effect of AI adoption on individual researchers, where we select a small proportion of core researchers with years of continuous publication records and replicate our analyses exclusively on them.
f. We improve explanations of our analysis regarding the extent of knowledge contraction. We clarify that our distance calculation is conducted within the native, high-dimensional embedding space without t-SNE projection. We also revise the representation and description of Fig. 4 to provide readers with improved understanding.

The details of our revision are summarized as a point-by-point response to your comments as follows.

My starting concern is that the text suffers, to some extent, from too much hype. This is due to the lack of clarification for what is meant by AI. Both scientists and the lay public interpret the term AI currently as meaning almost exclusively **generative AI**. Personally, I do **not** believe that generative AI has changed how we live or how most people work (second line of introduction). I am also rather concerned about poorly

designed experiments that have claimed that generative AI can "facilitate the distillation and dissemination of scientific findings."

I strongly feel that the manuscript would benefit from an introduction that is more explicit about what the authors mean by AI. They study publications from as early as 1980, clearly at a time during which there was no generative AI.

Response 3.1: We thank the Reviewer for this comment and this valuable suggestion. You are right that in our context, the term "AI" encompasses conventional machine learning approaches from the 1980s, subsequent deep learning techniques, and more recent generative AI advancements such as large language models (LLMs). Therefore, it would be inappropriate to only use terms related to "generative AI" as a general description in our introduction. To clarify, we have revised our introduction. We mention both **predictive and generative** AI with corresponding examples, and explicitly clarify what we mean by "AI" in the context of this research. We also revised our original simple affirmation of generative AI into a **two-sided discussion**, including its impact on both facilitating the distillation of scientific findings and raising concerns about weakened confidence in AI-generated content.

For your convenience, we quote the revised text below.

Revision:

[**Page 1 in the *Main Text, Introduction*-**Scope of our research study and a focus on a range of machine learning and AI methods.] Artificial intelligence (AI) has made significant strides in recent decades, promising to impact myriad aspects of society, including education, healthcare, and industry. Major investments in predictive and generative AI have catalyzed society-level debates over the future of AI at home and in the workplace. Perhaps more than any other domain, AI tools have become deeply entwined with the process of knowledge production, yielding findings that attract disproportionate attention in various scientific fields1 . For example, AlphaFold learns known protein structures to accurately predict the unexplored ones, circumventing the capital and human cost of conventional structural inference and recently granted a 2024 Nobel Prize. Models improved via deep reinforcement learning have become tuned to contain complex fusion reactions and discovered new, hardware-optimized forms to matrix multiplication that recursively accelerate deep learning itself. Autonomous laboratory systems driven by ChatGPT have helped some chemists and material scientists upscale the number of adaptive high-throughput experiments. Moreover, recent breakthroughs in large language models are beginning to transform scientific writing, broadening participation and facilitating the distillation of findings, but also raising concerns about weakened confidence in AI-generated content.

[**Page 2 in the *Main Text, Introduction*-**Scope of our research study and a focus on a range of machine learning and AI methods.] Notably, we do not focus on work that develops AI methodologies themselves, but rather on papers that augment research in natural science fields by adopting AI techniques ranging from early machine learning

algorithms to contemporary generative AI. Specifically, we select six representative disciplines…

Another concern relates to how AI papers are identified. This is described in the first section of the Results and in the methods. Some results are then presented in Fig. 1. I have to confess that I find the panels a-c in Fig 1 useless. One of my main questions in how well the identification model works for **different time periods**. AI has meant many different things over the 40 years studied by the authors. If AI papers for some stage of AI research are not well identified this is likely to alter findings.

Response 3.2: We thank the Reviewer for these comments. We agree that accurately identifying AI related papers across different eras of AI technologies represents the critical basis of our analysis. We use a pre-trained BERT model and design a two-stage fine-tuning process to adapt it to the AI paper identification task, which reaches reliable performance to support our subsequent analyses. In Fig. 1, we illustrate the fine-tuning process and our accuracy evaluation of this identification model.

In Fig. 1a, we show the increasing performance of AI paper identification during the two-stage fine-tuning of BERT pre-trained models and mark selected optimal models in both stages that we ensembled and used to identify all selected papers. Additionally, we add a flowchart in Extended Data Fig. 1, detailing our model structure and two-stage fine-tuning procedure. Taken together, we provide a clearer explanation of our model fine-tuning design, where the rising and converging performance in Fig. 1a demonstrates the effectiveness of our training process.

In Fig. 1b, we agree that how well the identification model works for different time periods is critically important for the reliability of our subsequent analysis. Following your suggestion, we group the 1,320 samples used to evaluate the identification results according to their publication dates (see Supplementary Table S2). We then calculated the identification accuracy separately for each AI era and show the results in Fig. 1b. Our results exhibit an overall Fleiss' Kappa of 0.964[1], and each of the three eras of AI has a Fleiss' Kappa greater than 0.93 (Supplementary Table S3), indicating strong consensus across the independent annotation of distinct experts. Taking the expert labels as ground truth and validating the identification results of our BERT model against it, our model reaches an overall F1-score of 0.875, and the F1-scores of the three eras of AI all exceed 0.85 (Supplementary Table S4). This indicates the high quality of our AI paper identification, which lays a robust foundation for our subsequent analysis.

In Fig. 1c, we remove the example of AI paper identification from the main text as you suggest. In compensation, we provide examples of both correct and incorrect identifications in the Supplementary Materials. Please refer to our response to the next comment for further details.

According to your suggestion and the above discussion, we make the following revisions to our manuscript:

a. We revise the *Main Text, Results-Increasing prevalence of AI adoption in science* describing Fig. 1a, and add a flowchart in *Extended Data Fig. 1*, providing clearer descriptions of our model structure and two-stage fine-tuning procedure.
b. We revise *Fig. 1b* and add *Supplementary Table S2-S4*, showing that the identification is accurate across all eras of AI.

For your convenience, we quote the revised text and figures below.

Revision:

[**Page 3 in the *Main Text, Results-*Increasing prevalence of AI adoption in science]** …These experts demonstrate strong consensus across their independent annotation of papers sampled at random from the six disciplines mentioned above, achieving an average Fleiss' Kappa of 0.964[1]. The BERT model attains an F1-score of 0.875 when evaluated against this expert-labeled data.

[**Page 3 in the *Main Text, Results-*Increasing prevalence of AI adoption in science]** … Meanwhile, the strong consensus among experts and high quality for identification is consistent across samples from different eras of AI, confirming the reliability of our identification accuracy and laying a robust foundation for our subsequent analysis (Fig. 1b and Supplementary Table S1-S4)…

[***Supplementary Notes 1.2-*Identification accuracy evaluation]** … We sample 2 groups of papers at random from each of the 6 disciplines, resulting in 12 paper groups in total, where samples in each group span all three eras of AI (Supplementary Table S2). We assign three different experts to independently label each sampled paper and measure the consistency among experts based on Fleiss' Kappa coefficient. Results exhibit an overall Fleiss' Kappa of 0.964, and each of the three eras of AI has a Fleiss' Kappa greater than 0.93 (Supplementary Table S3), indicating strong consensus across the independent annotation of distinct experts. Taking the expert labels as ground truth and validating the identification results of our BERT model against it, our model reaches an overall F1-score of 0.875, and the F1-scores of the three eras of AI are all greater than 0.85 (Supplementary Table S4). This indicates the consistent high quality of our AI paper identification, which lays a robust foundation for our subsequent analysis.

[1] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". In: Biometrics (1977), pp. 159–174.

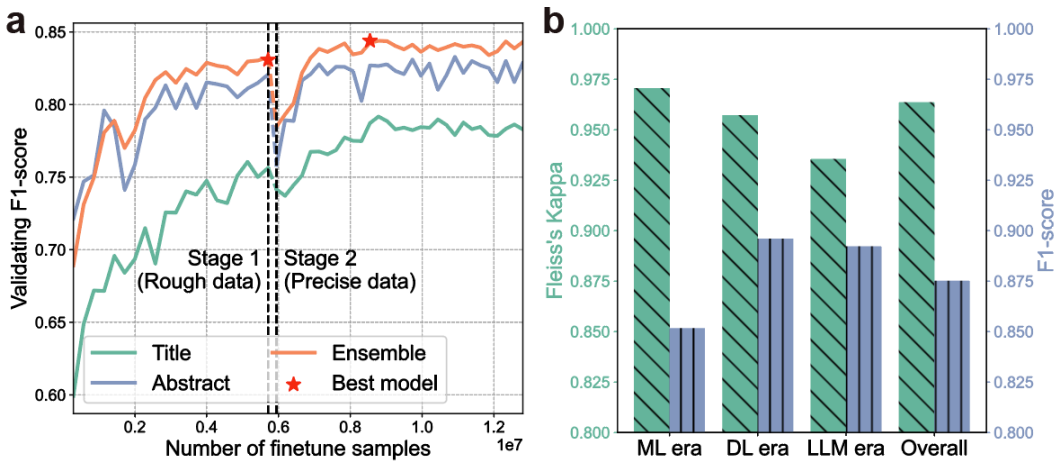
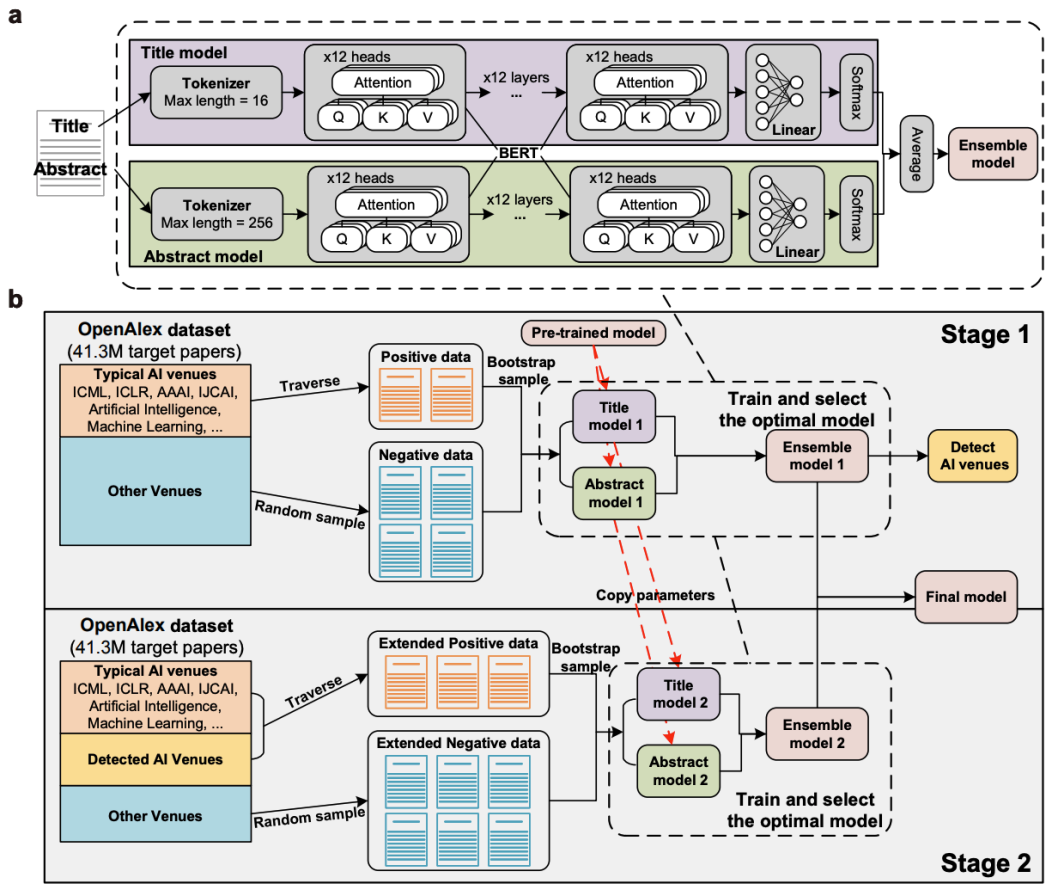


Figure 1. Increasing prevalence of AI adoption in science. (a) Increasing performance of AI paper identification during the two-stage fine-tuning of the BERT pre-trained models, where we use rough training data in Stage 1 to evolve precise assessments in Stage 2. We independently train two models based on titles and abstracts, respectively, and then integrate them into an ensemble that selects the optimal models during both stages (red stars) to identify all selected papers. **(b)** Accuracy evaluation of our identification results by human experts. For samples spanning three eras of AI, experts reached consensus with Fleiss' Kappa≥ 0.93. Our model identification results have strong accuracy in validation against expert-labeled data with an F1-score≥ 0.85.



Extended Data Fig. 1. Illustration for the method of identifying AI usage in research papers with fine-tuned language models. (a) Structure of our deployed language model, which consists of the tokenizer, the core BERT

model, and the linear layer. **(b)** Procedure of the two-stage model fine-tuning process, where we design specific approaches for constructing positive and negative data at each stage.

Supplementary Table S2. Number of papers sampled for expert annotation in identification accuracy evaluation.

Group	ML era	DL era	LLM era	Overall
Biology 1	59	33	18	110
Biology 2	57	39	14	110
Chemistry 1	60	32	18	110
Chemistry 2	69	28	13	110
Geology 1	69	25	16	110
Geology 2	55	40	15	110
Materials science 1	65	29	16	110
Materials science 2	63	32	15	110
Medicine 1	60	33	17	110
Medicine 2	54	37	19	110
Physics 1	66	29	15	110
Physics 2	60	36	14	110
Overall	737	393	190	1320

Supplementary Table S3. Fleiss' Kappa of expert annotation results in identification accuracy evaluation.

Group	ML era	DL era	LLM era	Overall
Biology 1	0.955	0.943	0.926	0.950
Biology 2	0.951	0.951	1.000	0.963
Chemistry 1	1.000	0.952	0.924	0.976
Chemistry 2	0.981	0.940	0.889	0.963
Geology 1	0.921	0.941	1.000	0.939
Geology 2	0.976	1.000	0.909	0.975
Materials science 1	1.000	0.893	1.000	0.976
Materials science 2	1.000	1.000	0.909	0.988
Medicine 1	0.976	0.898	0.921	0.952
Medicine 2	0.914	0.961	0.921	0.939
Physics 1	0.960	1.000	0.909	0.964
Physics 2	1.000	0.962	0.899	0.976
Overall	0.971	0.957	0.935	0.964

Supplementary Table S4. F1-score of BERT identification results against expert labels in identification accuracy evaluation.

Group	ML era	DL era	LLM era	Overall
Biology 1	0.923	0.920	0.889	0.917
Biology 2	0.878	0.897	0.857	0.885
Chemistry 1	0.809	0.857	0.900	0.844
Chemistry 2	0.892	0.923	0.857	0.898
Geology 1	0.800	0.812	0.941	0.826
Geology 2	0.783	0.912	0.875	0.857
Materials science 1	0.857	0.895	0.941	0.881
Materials science 2	0.800	0.923	0.875	0.855
Medicine 1	0.978	0.913	0.875	0.935
Medicine 2	0.882	0.917	0.917	0.906
Physics 1	0.825	0.833	0.875	0.835
Physics 2	0.816	0.909	0.875	0.862
Overall	0.852	0.896	0.892	0.875

As an aside, I am particularly unimpressed by Fig 1c. This is the kind of figure used to support something that is just confirmatory bias. More interesting would be to see an example of where the model makes a mistake.

Response 3.3: We thank the Reviewer for this instructive comment. We agree that only a single example of correct AI paper identification may strengthen confirmatory bias. Therefore, we remove the original example of AI-augmented paper identification from the main text. In replacement, we provide more examples of both correctly and incorrectly classified papers to expand the original Fig. 1c and include them in the Supplementary Materials.

First, we provide multiple correct examples from different eras of AI to provide rationale and explainability for our identification results. We visualize the average attention strengths within our title and abstract BERT models (Supplementary Fig. S2). In examples from different AI eras, our model consistently allocates substantial attention to AI-related terms. These illustrate how our identification model correctly interprets and accurately identifies various AI-related contents from papers published in different eras.

Second, we present an example in which the model makes a mistake to gain a deeper understanding of the identification results (Supplementary Fig. S3). The sample paper, published in Geology in 2005[5], is incorrectly classified by the model as AI-related, but human experts confirm that the article does not involve AI methodology. As we illustrate, the model assigns substantial attention to terms like “representing” and “deep”, which are commonly associated with AI research. In the context of this article, however, “representing” refers to general expression or depiction, which is unrelated to representation learning in AI, and “deep” is used to describe the extent of uncertainty rather than indicating a deep neural network or similar AI concepts. This example illustrates a potential edge case where the identification model may produce misclassifications. Nevertheless, given the high F1-score observed in the evaluation of

identification accuracy, such cases appear rare, and their effect on our subsequent analyses minimal.

We add an external factor analysis in *Supplementary Fig. S2-S3* and *Supplementary Notes 1.2-Identification accuracy evaluation*. For your convenience, we quote the revised texts and figures below.

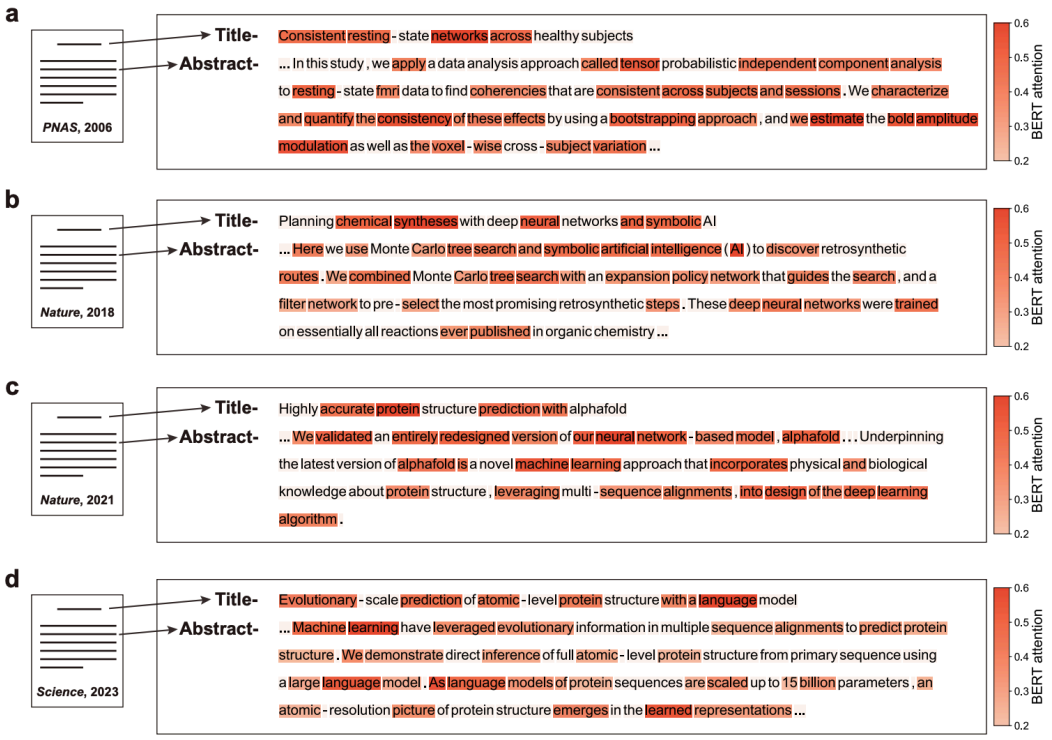
Revision:

[**Page 3 in the *Main Text, Results*-**Increasing prevalence of AI adoption in science] … To provide a rationale and explainability for our identification results, we visualize attention strengths in the BERT model with examples, where the model allocates substantial attention to terms such as "neural network" and "large language model", illustrating how the model correctly interprets and accurately identifies AI-related contents from papers published in different eras of AI development (Supplementary Fig. S2-S3)…

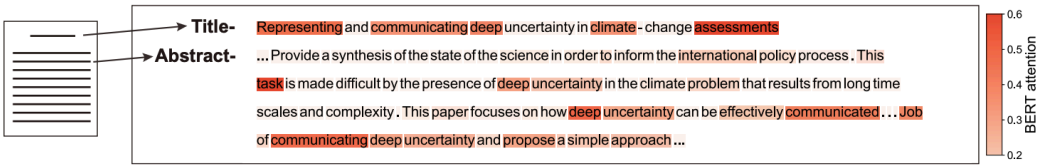
[***Supplementary Notes 1.2-***Identification accuracy evaluation] To provide rationale and explainability for our identification results, we offer multiple identification examples from different eras of AI and visualize the average attention strengths within our title and abstract BERT models. Our example for the ML era is a medicine paper with AI published in *PNAS* 2006[1], where our model allocates substantial attention to terms such as "independent component analysis" (Supplementary Fig. S2a). In the DL era, our first example is a chemistry paper with AI published in *Nature* 2018[2], and our second is a biology paper with AI published in *Nature* 2021[3], where our model allocates substantial attention to terms such as "neural network" (Supplementary Fig. S2bc). Our example of the LLM era is a biology paper using AI and published in *Science* 2023[4], where the model allocates substantial attention to terms such as "large language model" (Supplementary Fig. S2d). These illustrate how our identification model correctly interprets and accurately identifies various AI-related contents from papers published in different eras of AI.

Furthermore, to gain a deeper understanding of our identification results, we present an example in which the model makes a mistake (Supplementary Fig. S3). The sample paper, published in *Géoscience* in 2005[5], is incorrectly classified by the model as AI-related, but human experts confirm that the article does not involve AI methodologies. As we illustrate, the model assigns substantial attention to terms like "representing" and "deep", which are commonly associated with AI research. In the context of this article, however, "representing" refers to general expression or depiction, which is unrelated to representation learning in AI, and "deep" is used to describe the extent of uncertainty rather than indicating a deep neural network or similar AI concepts. This example illustrates an edge case where the identification model may produce misclassifications. Nevertheless, given the high F1-score observed in our evaluation of identification accuracy, such cases appear to be rare, and the effect on our subsequent analyses minimal.

[1] Jessica S Damoiseaux, Serge ARB Rombouts, Frederik Barkhof, Philip Scheltens, Cornelis J Stam, Stephen M Smith, and Christian F Beckmann. “Consistent resting-state networks across healthy subjects”. In: Proceedings of the national academy of sciences 103.37 (2006), pp. 13848–13853.
[2] Marwin HS Segler, Mike Preuss, and Mark P Waller. “Planning chemical syntheses with deep neural networks and symbolic AI”. In: Nature 555.7698 (2018), pp. 604–610.
[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. “Highly accurate protein structure prediction with AlphaFold”. In: Nature 596.7873 (2021), pp. 583–589.
[4] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. “Evolutionary-scale prediction of atomic-level protein structure with a language model”. In: Science 379.6637 (2023), pp. 1123–1130.
[5] Milind Kandlikar, James Risbey, and Suraje Dessai. “Representing and communicating deep uncertainty in climate change assessments”. In: Comptes Rendus. Géoscience 337.4 (2005), pp. 443–455.



Supplementary Figure S2. Correct examples of AI paper identification across different eras of AI development. **(a)** Example from the ML era, a medicine paper with AI published in *PNAS* 2006. **(b)** Example from the DL era, a chemistry paper with AI published in *Nature* 2018. **(c)** Example from the DL era, a biology paper with AI published in *Nature* 2021. **(d)** Example from the LLM era, a biology paper with AI published in *Science* 2023.



Supplementary Figure S3. An incorrect example of AI paper identification. The example is a geology paper without AI published in 2006, which is mistaken to be AI-related by the identification model.

Continuing with Fig 1, a linear plot is not appropriate for these data. If the growth appears to be exponential then the plot should use a log scale on the y-axis. and the growth estimates should be shown as fits to straight regions in such a plot. And error bars would be useful too.

Response 3.4: We thank the Reviewer for this helpful suggestion. We agree that a log-scale plot would be better suited to represent exponential growth. We also agree that linear fitting in the log-scale plot can benefit the growth estimates a lot. Considering the limited space in the Main Text for Fig. 1d and Fig. 1e, however, we felt that displaying increasing trends for all disciplines along with linear fittings in the log-scale plot might make the figures appear overly cluttered. As a result, we retain the linear scale *y*-axis in the Main Text for brevity. We add error bars in Fig. 1f to indicate confidence intervals (CIs) for the average monthly growth rates of AI papers and researchers. Additionally, considering the rapid development of AI, we update the OpenAlex dataset used in our revision. The coverage period is now extended from the original cutoff date of May 30, 2024, to March 31, 2025, ensuring that our analysis remains as up-to-date as possible. This update results in adjustments to the numerical values reported in our analysis, but the original conclusions remain valid and unaffected.

To detail the exponential growth trend of AI in each discipline and estimate the growth rate, we separately plot the proportion of papers and researchers adopting AI in each discipline with a log-scale *y*-axis and estimate the growth rate with exponential fitting in each era in Supplementary Fig. S5-S6. As the results show, the proportion of papers and researchers adopting AI fit straight regions in the log-scale plot, indicating an exponential growth trend. Meanwhile, the growth rate of AI papers and AI researchers in all disciplines increased progressively from the ML era to the DL and LLM eras, as estimated by linear fittings in log-scale. The growth rate estimation is consistent with average monthly growth rates in Fig 1f. Such an increasing trend underscores the increasing prevalence and fast development of AI in science.

We add the detailed log-scale plot and growth rate estimation in *Supplementary Fig. S5-S6* and *Supplementary Notes 1.4-Increasing prevalence of AI adoption in science*. For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 3 in the *Main Text, Results*-**Increasing prevalence of AI adoption in science] … Meanwhile, growth rates for AI-augmented papers and researchers have accelerated across the three eras (Fig. 1f and Supplementary Fig. S5-S6)…

[***Supplementary Notes 1.4-***Increasing prevalence of AI adoption in science] In the main text, the proportion of papers and researchers adopting AI exhibit exponential growth over the past decades. To further illustrate and analyze this upward trend, we separately plot the proportion of papers and researchers adopting AI in each discipline with a log-scale y-axis and estimate the growth rate with exponential fitting in each era (Supplementary Fig. S5-S6). As results show, the proportion of papers and researchers adopting AI fit to straight regions in the log-scale plot, indicating an exponential growth trend. Meanwhile, the growth rate of AI papers and AI researchers in all disciplines increased progressively from the ML to the DL and LLM eras, as estimated by the exponential fitting, which underscores the increasing prevalence and fast development of AI in science.

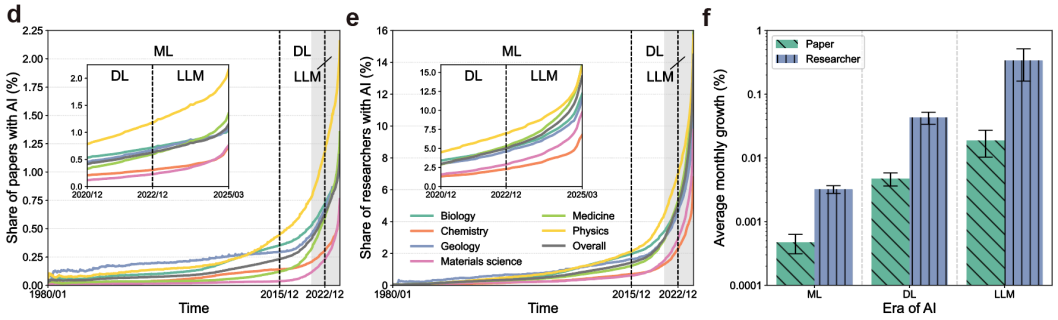
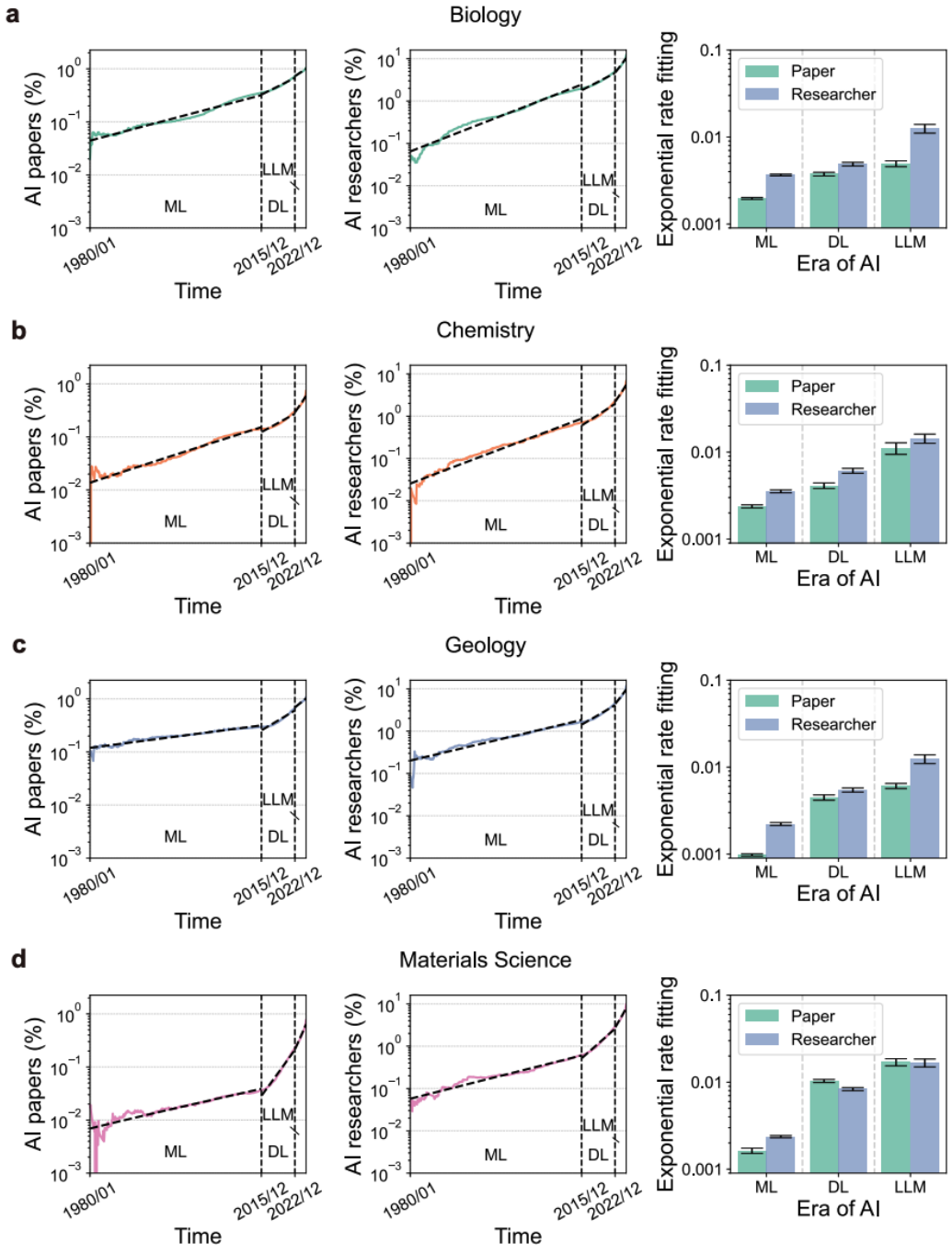
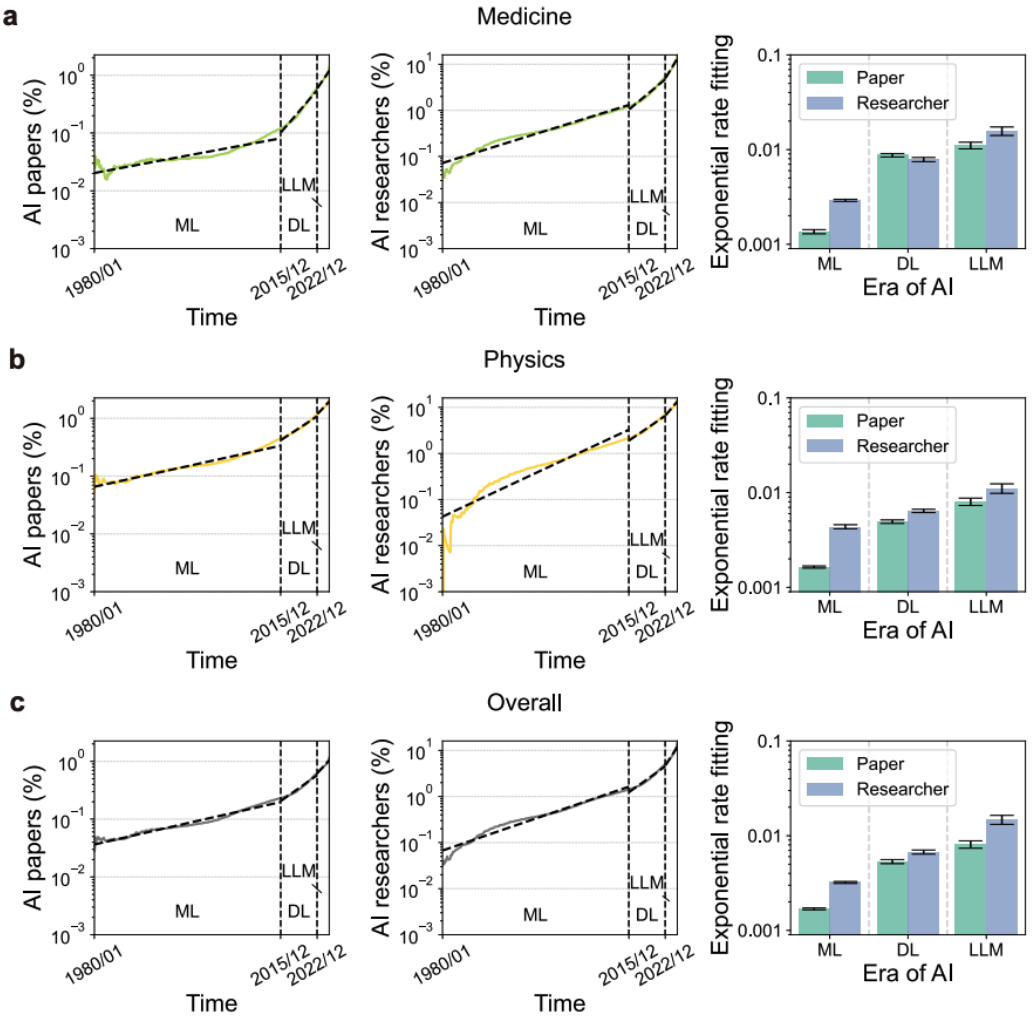


Figure 1. Increasing prevalence of AI adoption in science. (d)-(f) The growth of AI research papers (d, $n = 41,298,433$) and AI researchers (e, $n = 5,377,346$) over the eras of ML, DL, and LLM between 1980 and 2025 in selected scientific disciplines, where 99% CIs of monthly growth rates for AI papers and researchers are shown as error bars.



Supplementary Figure S5. Proportion of papers and researchers adopting AI in each discipline with log-scale *y*-axis and estimated growth rate with exponential fitting for each era. Proportion of papers and researchers adopting AI fit to straight regions in the log-scale plot, indicating high precision of the exponential fitting. Meanwhile, the growth rate of AI papers and AI researchers in all disciplines increased progressively from the ML to the DL and LLM eras, as estimated by the exponential fitting. For estimated growth rates, 99% CIs are shown as error bars.



Supplementary Figure S6. (Continued from Supplementary Fig. S5) Proportion of papers and researchers adopting AI in each discipline with log-scale *y*-axis and estimated growth rate with exponential fitting for each era. Proportion of papers and researchers adopting AI fit to straight regions in the log-scale plot, indicating high precision of the exponential fitting. Meanwhile, the growth rate of AI papers and AI researchers in all disciplines increased progressively from the ML to the DL and LLM eras, as estimated by the exponential fitting. For estimated growth rates, 99% CIs are shown as error bars.

The same exact issue arises for Fig. 2d. In that case, it is clear that the exponential fit is terrible for the case of young scientists not using AI. And, again, there should be error bars on the parameter estimates.

Response 3.5: We thank the Reviewer for pointing this out. We apologize for the original exponential fitting in Fig. 2d. According to our stochastic process model of junior scientists' career development (Methods M7), the transition time to established scientist should follow an exponential distribution, regardless of whether the scientist adopts AI or not. Upon careful examination, however, we find that the inconsistency between the original results and our mathematical model was due to the presence of some uncorrected data in the original OpenAlex dataset, which led to errors in our statistical analysis. To address this, we now update the dataset to the latest version, in which the maintainers

have corrected various issues, including the removal of abstracts with invalid or junk content along with updates to authorship information. We also improve our data processing code by adding filters to exclude invalid records, thereby preventing such errors.

Following your suggestion, we change the *y*-axis to log scale. The revised probability distributions now exhibit the expected exponential shape, fitting to linear regions in the log-scale plots. The linear fittings yield $R^2 \geq 0.95$. Additionally, we add error bars to our parameter estimations. The updated results indicate that AI adoption enables junior scientists to lead research teams and transition to established researchers approximately three years earlier than their non-AI counterparts.

We revise *Fig. 2d* and *Extended Data Fig. 7*. For your convenience, we quote the revised texts and figures below.

Revision:

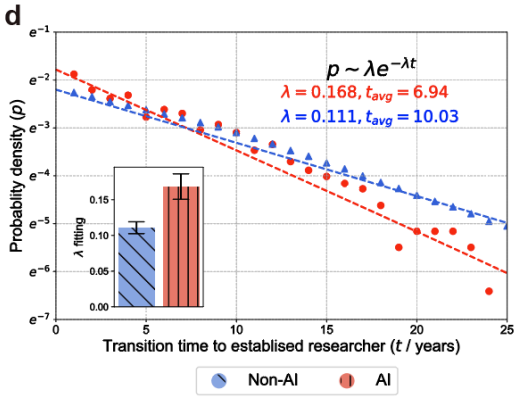
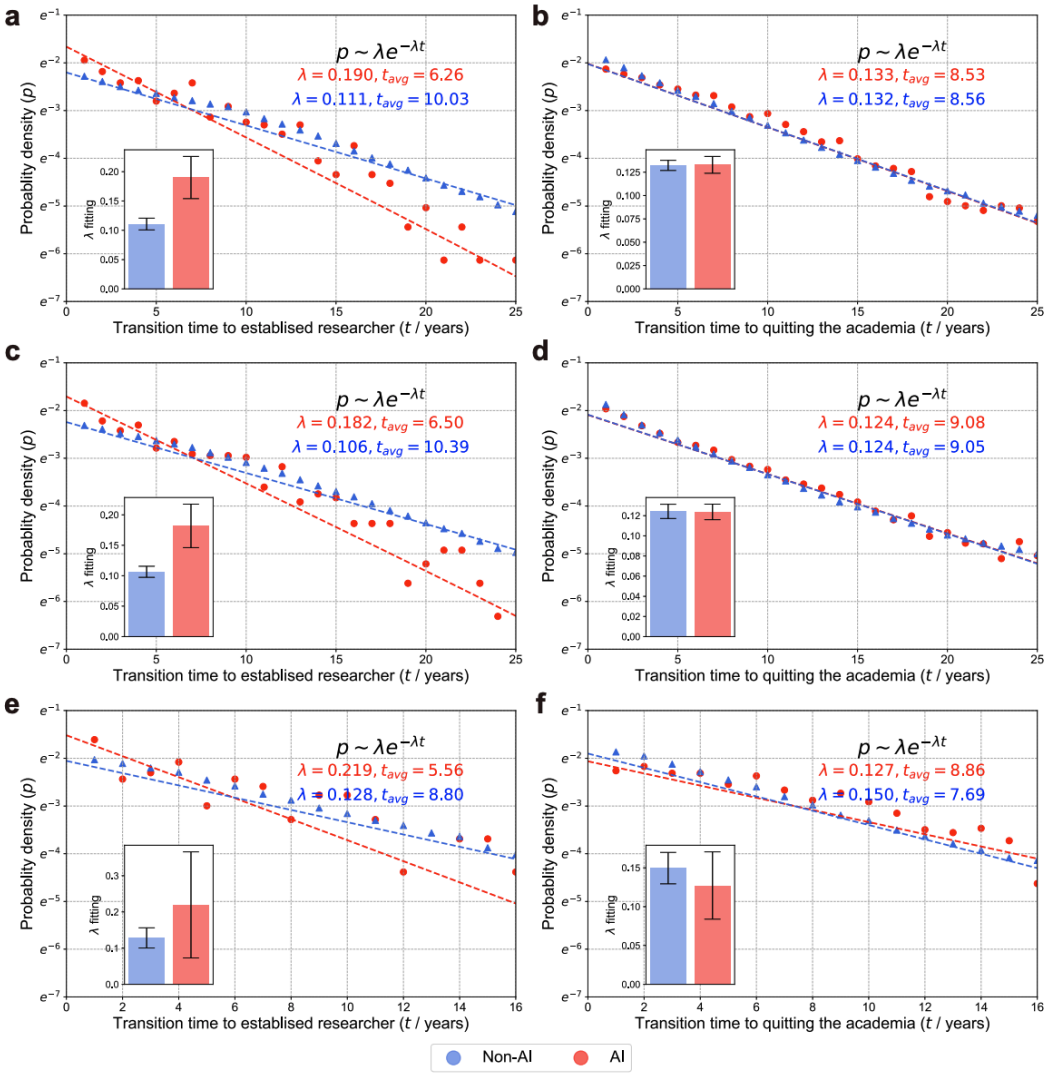


Figure 2. AI enlarges paper impact and enhances researcher careers. (d) Probability density distribution of the transition time for junior scientists to become established ($n = 2,282,029$). The probability density distributions can be well-fit with exponential distributions, where junior scientists adopting AI become established approximately four years earlier than their counterparts without AI. For all panels, 99% CIs are shown as error bars.



Extended Data Fig. 7. Model fitting the role transition time of junior scientists. (a) (c) (e) Probability density distribution of the transition time to become established scientists for junior scientists in biology, medicine, and physics. **(b) (d) (f)** Probability density distribution of the transition time to leave academia for junior scientists in biology, medicine, and physics. All probability density distributions can be well-fit with exponential distributions, where the expected time for junior scientists to become established is approximately 4 years shorter for those who adopt AI, while the expected time for junior scientists to abandon academia is somewhat longer for those who adopt AI. Results indicate that AI not only provides junior scientists opportunities to become established scientists at a younger age, but also reduces the risk of their exiting academia early. For all panels, 99% CIs are shown as error bars.

The analysis of the citations is also problematic. The distribution of citations to papers is very right skewed, making the average a rather unreliable statistic. More concerning is the **apparent** mixture of all types of articles in the analysis. Review articles have very different citation patterns from original research papers or from editorial pieces. I would recommend disaggregating the analysis and focusing on papers reporting on original research.

Response 3.6: We thank the Reviewer for this suggestion. Citation distributions are empirically right-skewed[1-2], as you note. Considering this, we adopt multiple statistical

indicators beyond the average annual citation count to ensure more robust conclusions (Supplementary Fig. S8). For the right tail of the distribution, namely papers with high citation counts, we observe that the top 1st and 5th-percentile of annual citations for AI papers exceed those for non-AI papers by 152.39% and 48.27%, respectively. For the left tail of the distribution, namely papers with low citation counts, the proportion of AI papers receiving fewer than three and fewer than five citations each year is 2.49% and 2.46% lower. These results suggest that AI papers not only tend to receive higher citations but are also less likely to become low-impact publications, indicating the enhanced visibility and academic attention garnered by AI research, which is consistent with statistics in the main text.

We also agree that special types of publications, like review articles and editorial pieces, tend to have very different citation patterns from original research. To verify the robustness of our finding that AI papers garner higher academic impact, we distinguish these special publications from those reporting original research and replicate our analyses exclusively on the latter (Supplementary Fig. S10-S11). By combining the publication type classification provided in the OpenAlex dataset with keyword matching in article titles, such as "review" or "survey", we identified these special types of publications. They account for 9.26% of the papers included in our original analysis. On the subset excluding them, AI papers still receive higher visibility and academic attention than their non-AI counterparts across the aforementioned statistical indicators.

Following your suggestions and the above discussion, we make two important revisions to our manuscript:

a. We add multiple alternative statistical indicators to compare the citations between AI and non-AI papers in *Supplementary Fig. S8* and *Supplementary Notes 2.1-The effect of AI on individual papers*, taking into account the right-skewed distribution of citation.
b. We add analysis excluding review articles that only focus on original research papers in *Supplementary Fig. S10-S11* and *Supplementary Notes 2.1-The effect of AI on individual papers*, enhancing the robustness of our findings.

For your convenience, we quote the revised texts and figures below.

Revision:

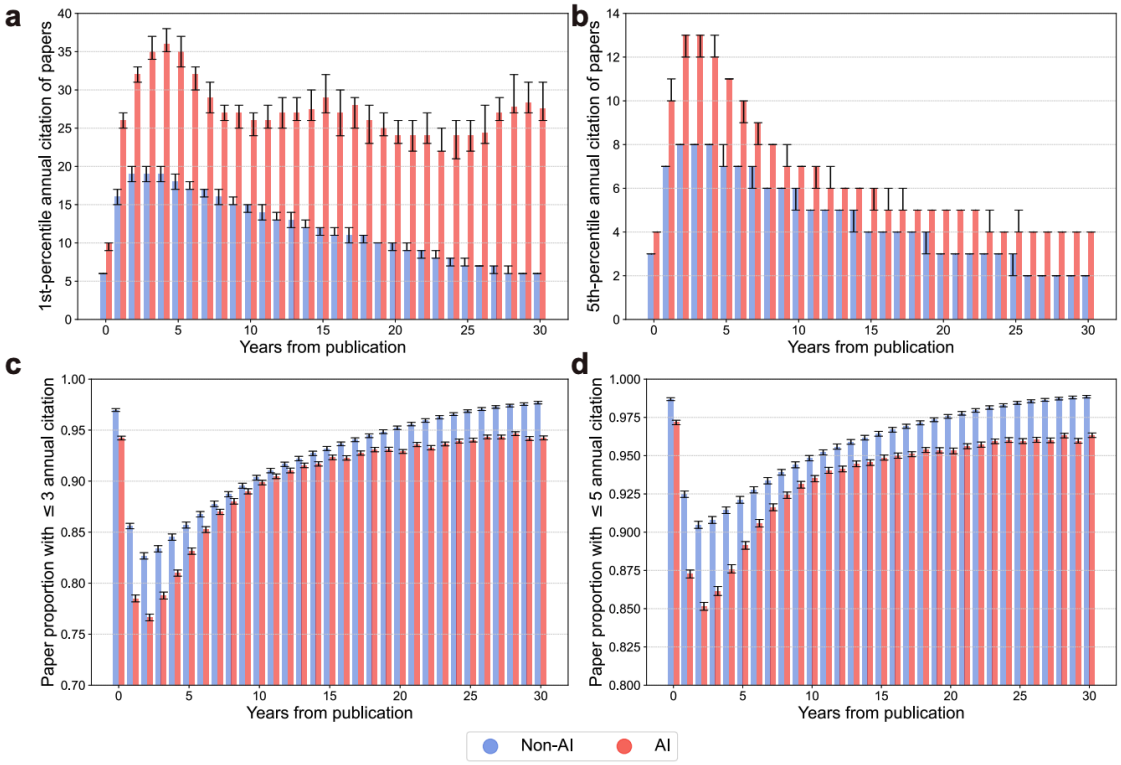
[**Page 3 in the *Main Text, Results-***AI enlarges the impact of individual papers and enhances the career of individual scientists] … In addition to higher annual average citations, the higher scientific impact of AI-augmented papers is also reflected by multiple alternative statistical indicators. (Supplementary Fig. S8).

[***Supplementary Notes 2.1-***The effect of AI on individual papers] As demonstrated in the main text, AI-related papers, on average, receive higher annual citation counts than non-AI papers. Given that citation distributions are empirically right-skewed[1-2], we adopt multiple statistical indicators beyond the average annual citation counts to ensure more robust conclusions (Supplementary Fig. S8). For the right tail of the distribution, namely papers with high citation counts, we observe that from year of publication and across subsequent decades, citations for the top 1st and 5th-percentile of AI papers exceed those

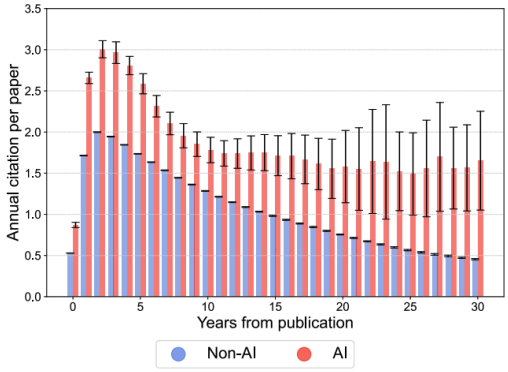
of non-AI papers by 152.39% and 48.27%, respectively. For the left tail of the distribution, namely papers with low citation counts, we find that, over the same time period, the proportion of AI papers receiving fewer than three and fewer than five citations each year is 2.49% and 2.46% lower, respectively, than non-AI papers. Taken together, these findings suggest that AI papers not only tend to receive higher citations but are also less likely to become low-impact publications, indicating the enhanced visibility and academic attention garnered by AI research.

[***Supplementary Notes 2.1-***The effect of AI on individual papers] Empirically, review articles, editorial pieces, and other types of special publications tend to exhibit markedly different citation patterns compared to original research papers. To verify the robustness of our finding that AI papers garner higher academic impact, we distinguish these special publications from those reporting original research and replicate our analyses exclusively on the latter (Supplementary Fig. S10-S11). Specifically, we filter publications labeled as "article" in the OpenAlex dataset, thereby excluding journals dedicated to reviews as well as publication types such as letters, editorials, and erratum. To further eliminate a small number of review articles that may still be included in non-review journals, we additionally filter all papers with "review" or "survey" in their titles. As a result, these special pieces account for 9.26% of the 27,405,011 publications with intact reference records from our original analysis. After excluding them, we obtained a refined dataset of 24,867,012 original research papers. On this subset, AI papers still receive higher visibility and academic attention than their non-AI counterparts across the previously introduced statistical indicators.

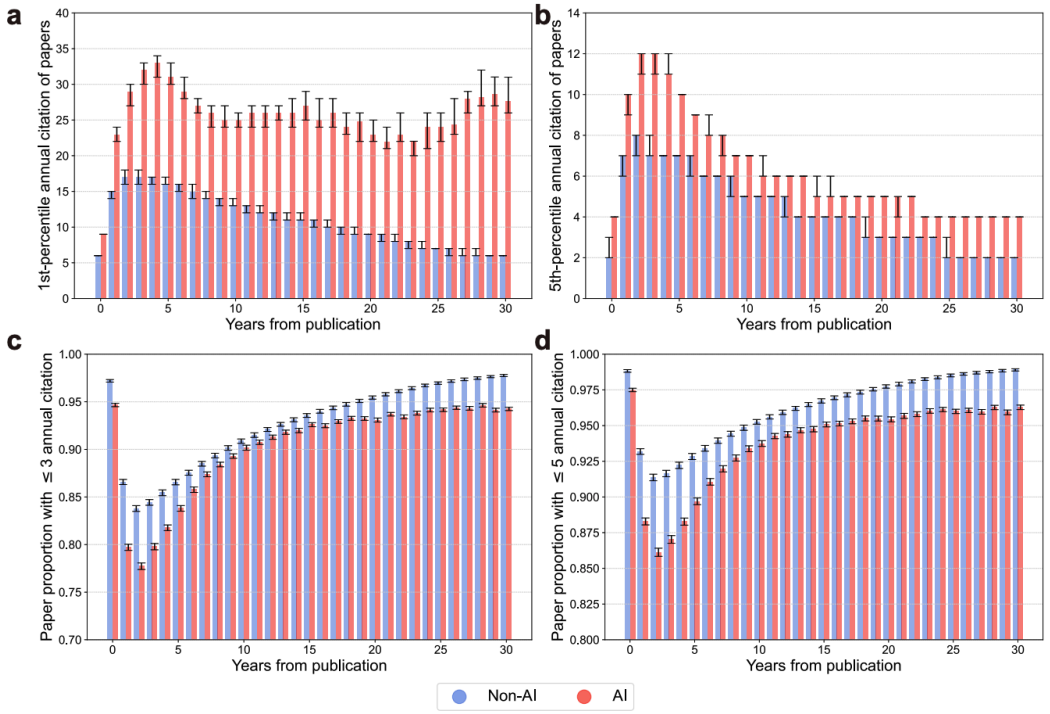
[1] Sidney Redner. "How popular is your paper? An empirical study of the citation distribution". In: The European Physical Journal B-Condensed Matter and Complex Systems 4.2 (1998), pp. 131–134.
[2] Per O Seglen. "The skewness of science". In: Journal of the American society for information science 43.9 (1992), pp. 628–638.



Supplementary Figure S8. Alternative statistical indicators for the annual citation comparison between AI and non-AI papers. (a) Top 1st-percentile of annual citation for AI and non-AI papers from year of publication. **(b)** Top 5th-percentile of annual citation for AI and non-AI papers from year of publication. **(c)** Proportion of AI and non-AI papers receiving fewer than three citations each year from year of publication. **(d)** Proportion of AI and non-AI papers receiving fewer than five citations each year from year of publication. For all panels, 99% CIs are shown as error bars.



Supplementary Figure S10. Annual citations after the publication of AI and non-AI original research papers, excluding review articles, editorial pieces and other special publications. Results show that AI papers attract more citations, indicating higher academic impact than papers without AI. Results are consistent with the overall statistics in the main text. 99% CIs are shown as error bars.



Supplementary Figure S11. Alternative statistical indicators for the annual citation comparison between AI and non-AI original research papers, excluding review articles, editorial pieces and other special publications. (a) Top 1st-percentile annual citation for AI and non-AI papers from year of publication. **(b)** Top 5th-percentile annual citation for AI and non-AI papers from year of publication. **(c)** Proportion of AI and non-AI papers receiving fewer than three citations each year from year of publication. **(d)** Proportion of AI and non-AI papers receiving fewer than five citations each year from year of publication. Results are consistent with the overall statistics in the main text. For all panels, 99% CIs are shown as error bars.

The authors also refer to 1.7 million researchers. The overwhelming majority of these individuals is highly unlikely to have remained in research careers. Thus, it is not clear what they have to teach us about the impact of the use of AI on career progression. Ioannidis' team has discussed a set of continuing publishing scholars, which would qualify as the worldwide core of active researchers. That number if remembered correctly is on the order of 0.2 million << 1.7 million.

Response 3.7: We thank the Reviewer for this sharp comment. In Fig. 2c, we now analyze the impact of AI adoption on researcher dropout, that is, whether a researcher becomes a "continuing publishing scholar." Following your suggestion, we add a separate analysis to examine the effect of AI adoption on core researchers with multi-year continuous publication records.

To evaluate this, we identified 1,495,265 researchers with uninterrupted publication records over at least five consecutive years and 525,716 researchers over at least ten consecutive years, accounting for 52.30% and 18.39% of all researchers in our dataset, respectively (Supplementary Fig. S17). The proportion of researchers with uninterrupted publication records over at least ten consecutive years is close to the proportion of "core researchers" you mentioned. As the results show, among those with at least five

consecutive years of publications, researchers adopting AI published 2.40 times more papers and received 3.88 times more citations per year than their non-AI counterparts, with consistent patterns observed across disciplines. Similar results hold among researchers with at least ten consecutive years of publications. These findings confirm the positive impact of AI adoption on the career progression of both continuously publishing core scientists and the broader population of normal scientists.

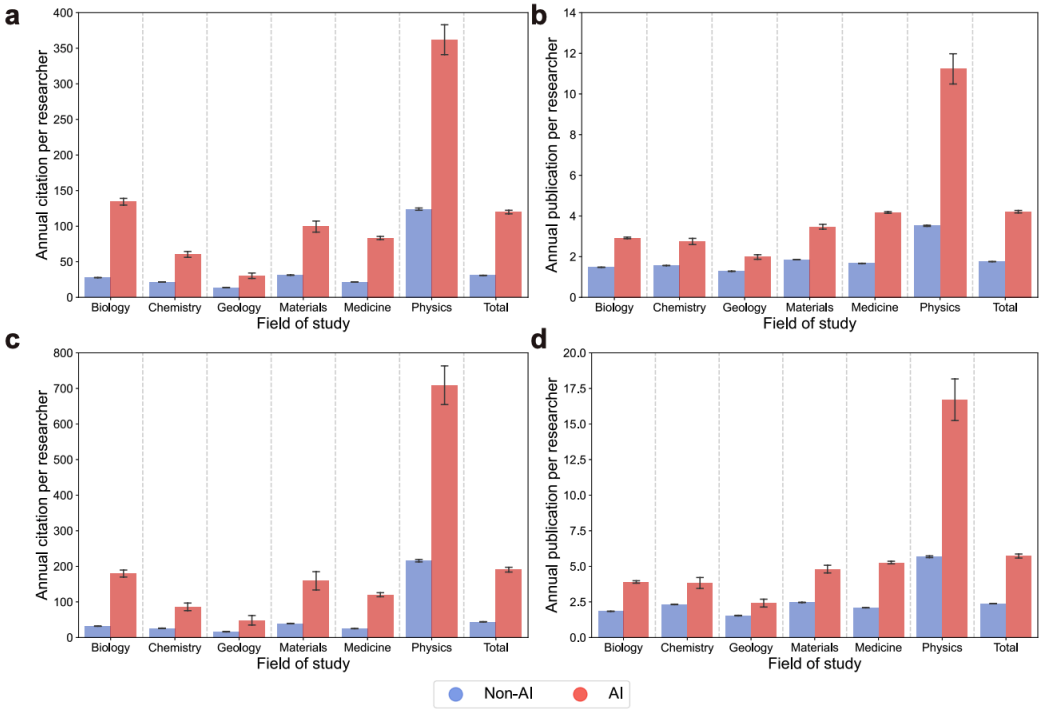
We add the analysis on researchers with continuous publication records in *Supplementary Fig. S17* and *Supplementary Notes 2.2- The effect of AI on individual scientists' careers*. For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 3 in the *Main Text, Results-***AI enlarges the impact of individual papers and enhances the career of individual scientists] … On average, researchers adopting AI annually publish 3.02 times more papers ($t \geq 47.18$, $p < 0.001$ and $df > 10^3$ in t-test on any discipline) and garner 4.84 times more citations ($t \geq 30.32$, $p < 0.001$ and $df > 10^3$ in t-test on any discipline) compared with those not adopting AI, with consistency across disciplines and robustness for core researchers with multi-year continuous publication records[1] (Supplementary Fig. S17).

[***Supplementary Notes 2.2-***The effect of AI on individual scientists' careers] In our main analysis, we include 2.3 million scientists with complete career trajectories and show that those adopting AI are more productive and receive more citations. Prior research suggests that only a relatively small subset of scientists continue publishing over long periods, however, forming what has been recognized as the global core of active researchers[1]. To examine whether our findings hold for this persistent group, we identified 1,495,265 researchers with uninterrupted publication records over at least five consecutive years and 525,716 researchers over at least ten consecutive years, accounting for 52.30% and 18.39% of all researchers in our dataset, respectively (Supplementary Fig. S17). We then compared the productivity and citation impact of AI-adopting versus non-AI-adopting researchers within these subsets. Among those with at least five consecutive years of publications, researchers adopting AI published 2.40 times more papers and received 3.88 times more citations per year than their non-AI counterparts, with consistent patterns observed across disciplines. Similarly, for researchers with at least ten consecutive years of publications, who are more productive and receive more citations than researchers with five consecutive years of publications, AI adopters published 2.41 times more papers and received 4.31 times more citations annually. These results further confirm the positive impact of AI adoption on the career progression of both the continuously publishing core scientists and broader population of normal scientists.

[1] John PA Ioannidis, Kevin W Boyack, and Richard Klavans. "Estimates of the continuously publishing core in the scientific workforce". In: PloS one 9.7 (2014), e101698.



Supplementary Figure S17. The impact of AI on the productivity and citation of researchers with multi-year continuous publication records. (a) Comparison of annual citations between researchers adopting AI and their counterparts without AI among those with at least five consecutive years of publications. **(b)** Comparison of annual publications between researchers adopting AI and their counterparts without AI among those with at least five consecutive years of publications. **(c)** Comparison of annual citations between researchers adopting AI and their counterparts without AI among those with at least ten consecutive years of publications. **(d)** Comparison of annual publications between researchers adopting AI and their counterparts without AI among those with at least ten consecutive years of publications. These results confirm the positive impact of AI adoption on the career progression of continuously publishing core scientists. For all panels, 99% CIs are shown as error bars.

The most interesting result of the manuscript is reported in Fig. 3. (Please note some spelling mistakes throughout the manuscript such as 'centriod') It very much supports my own bias. But again, I have concerns. Let us assume that the 768-dimension embedding of the articles is reliable and accurate. What the authors do then is use t-SNE embedding of the data and then calculate a distance between papers based on that embedding. This procedure **cannot** produce a reliable estimate. While t-SNE (or UMAP) can maintain distances approximately accurate for **close neighbors**, it does for general pairs of distant data points. This means that the calculation of a cluster radius cannot be reliably made.

Response 3.8: We thank the Reviewer for this comment and the opportunity to clarify our procedure. We apologize for causing this misunderstanding in the brevity of our desrcription. In fact, the only place where we use t-SNE to reduce the dimension is the visualizations in Fig. 3b. In all other results regarding the distance between papers, we calculate distances based on the original 768-dimensional embeddings. Given the strong capability of the advanced scientific text embedding model we adopted[1], we believe the obtained distances and cluster radius support our subsequent analysis with reliability.

To better clarify this, we make the following revisions to our manuscript:

a. We revise the *Caption* of *Figure 3*, making it clear that the t-SNE dimensional reduction is only applied for visualization.
b. We revise the section regarding *AI adoption is associated with a contraction in knowledge focus among scientific fields* in the *Main Text*, pointing out that the distances are calculated in the 768-dimensional embedding space.
c. We revise *Method M8. Measure the knowledge extent of papers*, making the mathematical formulas more rigorous and clear.

Meanwhile, we thank the Reviewer for carefully pointing out our spelling problems. We have corrected the errors and double-checked the entire paper. For your convenience, we quote the revised texts below.

Revision:

[**Page 5 in the *Main Text, Figure 3 Caption***] … **(b)** For visualization, we use the t-SNE algorithm to flatten the high-dimensional embeddings of a small random batch of $10,000$ papers, half of which are AI papers and half are non-AI papers, into a 2-D plot.

[**Page 6 in the *Main Text, Results-***AI adoption is associated with a contraction in knowledge focus among scientific fields] … We employ SPECTER 2.0, a specialized text embedding model pre-trained on a large scientific literature corpus and fine-tuned with citation information[1] to project research articles onto this 768-dimensional embedding space of science… Within the high-dimensional embedding space, we first design the measurement of knowledge extent, which represents the coverage topical ground of AI and non-AI papers by comparably sized samples in each given domain[2-3] (Fig. 3b and Methods M8).

[**Page 11 in the *Methods-***M8. Measure the knowledge extent of papers] To assess the knowledge extent of a set of research papers within their high-dimensional embeddings

$$\{e_1, e_2, ..., e_n\}, e_i \in R^{768},$$

we first compute the centroid as the mean of their vector locations:

$$c = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

[1] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. "SciRepEval: A Multi-Format Benchmark for Scientific Document Representations". In: EMNLP. Association for Computational Linguistics, 2023, pp. 5548–5566.

[2] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojevic, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. "Science of science". In: Science 359.6379 (2018), eaao0185.

[3] Staša Milojevic. "Quantifying the cognitive extent of science". In: Journal of Informetrics 9.4 (2015), pp. 962–973.

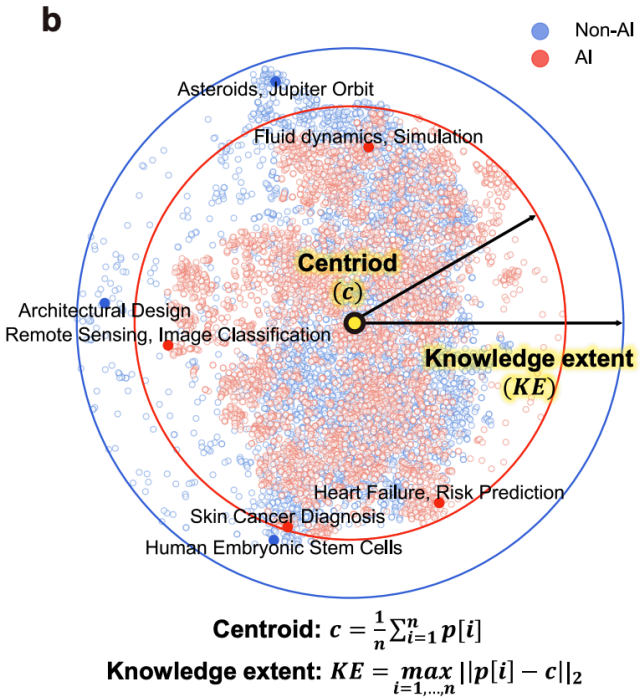


Figure 3. AI adoption is associated with a contraction in knowledge extent within and across scientific fields. (b) For visualization, we use the t-SNE algorithm to flatten the high-dimensional embeddings of a small random batch of 10,000 papers, half of which are AI papers and half are non-AI papers, into a 2-D plot. As shown by the solid arrows and circular boundaries, the knowledge extent of AI papers is smaller across the entirety of natural science, and AI papers are more clustered in knowledge space, indicating more concentration on specific problems.

Finally concerning Fig. 4. I am not sure I follow what is being plotted in Figs. 4a and 4b. I imagine that in 4b, the y-axis is the EG defined in Eq. (7). But what is the y-axis in 4a? Fig. 4c is redundant, I just want to know the Gini coefficient with error bars. Fig. 4d suffers from the problem with the definition of distance between papers discussed for Fig. 3.

Response 3.9: We thank the Reviewer for these comments. We apologize for not explaining Fig. 4 clearly enough. In Fig. 4, we further examine the observed conflict between the growing influence of individual papers and researchers and the narrowing of domain knowledge within AI research. Here, we provide more detailed clarifications.

In Fig. 4a, we plot the knowledge extent of "paper families", namely a focal paper and its follow-on citations, in the *y*-axis, which measures the size of the space covered by research derived from each original paper. We add a mathematical definition of this measurement in Methods M9. Results show that the knowledge extent of AI papers' citation families is larger than non-AI papers', indicating that the contraction of knowledge space in AI research is not attributable to the narrowing of knowledge space that can be derived from each original work of research.

In Fig. 4b, you are correct that we plot follow-on engagement (EG), namely how frequently citations of the same original paper cite one another, which is defined in

Methods M10. Results show that AI papers tend to only concentrate on the original paper, rather than forming dense interactions among each other.

In Fig. 4c, we further examine the concentration of AI papers in the Matthew effect in citations. In this figure, we show the cumulative distribution of citations for AI papers and non-AI papers. This plot helps to intuitively understand the contraction of citations, where a small number of superstar papers in AI research dominate the field, with 22.20% of the top papers receiving 80% of the citations and the top 54.14% receiving 95% of citations. As you suggested, we also add the Gini coefficient with error bars. Through statistical tests, we verify that the Gini coefficient of citations for AI papers is significantly higher than for non-AI papers, signaling a disparity in recognition.

In Fig. 4d. we find that, on average, a pair of disengaged papers indeed focuses on less relevant topics and lies farther apart. In some cases, however, when the two papers happen to work on similar topics, lack of engagement leads to mutual unawareness. As a result, they are likely to lie very close to each other, producing overlapping research. This reveals that the observed concentration leads to more overlapping ideas and redundant innovations linked to a contraction in knowledge extent. As we mentioned above, we only use *t*-SNE to reduce the dimension for the visualizations in Fig. 3b. In Fig. 4d, we calculate distances within the original 768-dimensional embedding space.

To make Fig. 4 more accessible for readers, we make the following revisions to our manuscript:

a. We revise the section regarding *Over-concentration of AI research and redundant innovation* in the *Main Text*, clarifying the measurement of knowledge extent among "paper families" in Fig. 4a.
b. We add *Methods-M9. Measure the knowledge extent of paper families*, providing a mathematical definition for the knowledge extent of "paper families".
c. We revised *Fig. 4c* and *Extended Data Fig. 10*, illustrating Gini coefficients with error bars.
d. We revise the section regarding *Over-concentration of AI research and redundant innovation* in the *Main Text*, clarifying that the distances between papers are calculated in the uncompressed 768-dimensional embedding space.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 6 in the *Main Text, Results*-**Over-concentration of AI research and redundant innovation] … We first examine the knowledge extent of "paper families", i.e., a focal paper and its follow-on citations, which measures the size of the space covered by research derived from each original paper (Fig. 4a and Methods M9). Results show that the knowledge extent of AI papers' citation families is on average 3.46% more expanded compared to non-AI papers' ($t \geq 1.91, p \leq 0.1$ and $df > 10^3$ in t-test on 30 out of 32 pairs of data). Therefore, the contraction of knowledge space in AI research is not attributable to the narrowing of knowledge space that can be derived from each original work of research.

[**Page 6 in the *Main Text, Results*-**Over-concentration of AI research and redundant innovation] … Further evidence of this concentration is found in the Matthew effect among AI paper citations across different fields (Fig. 4c and Extended Data Fig. 10). In AI research, a small number of superstar papers dominate the field, with 22.20% of top papers receiving 80% of the citations and the top 54.14% receiving 95% of citations. This unequal distribution leads to a GINI coefficient of 0.754 in citation patterns surrounding AI research, higher than 0.690 for non-AI papers ($t = 27.86$, $p < 0.001$ and $df = 198$ in t-test), signaling a disparity in recognition.

[**Page 6 in the *Main Text, Results*-**Over-concentration of AI research and redundant innovation] … We examine distances between these pairs of papers within the 768-dimensional vector space (Fig. 4d).

[**Page 11 in the *Methods*-**M9. Measuring the knowledge extent of paper families] To measure how much knowledge space can be derived from each original research, we calculate the knowledge extent of "paper families", i.e., a focal paper and its follow-on citations. Focusing on an original research paper ϕ, which corresponds to a high-dimensional embedding $e_{\phi} \in R^{768}$, we extract all n_{ϕ} research papers that cite this original paper. These citing papers are sorted chronologically by publication date, from earliest to most recent. The corresponding high-dimensional embeddings of these sorted papers are:

$$\left\{e_{\phi}^{1}, e_{\phi}^{2}, \ldots, e_{\phi}^{n_{\phi}}\right\},\ e_{\phi}^{i} \in R^{768}.$$

Thereby, we calculate knowledge extent covered by the "paper family" consists of the original paper φ and the first n follow-on papers citing it ($1 \leq n \leq n_{\phi}$) as:

$$KE_{\phi}[n] = \{\| e_{\phi} - e_{\phi}^{i} \|_{2}\}.$$

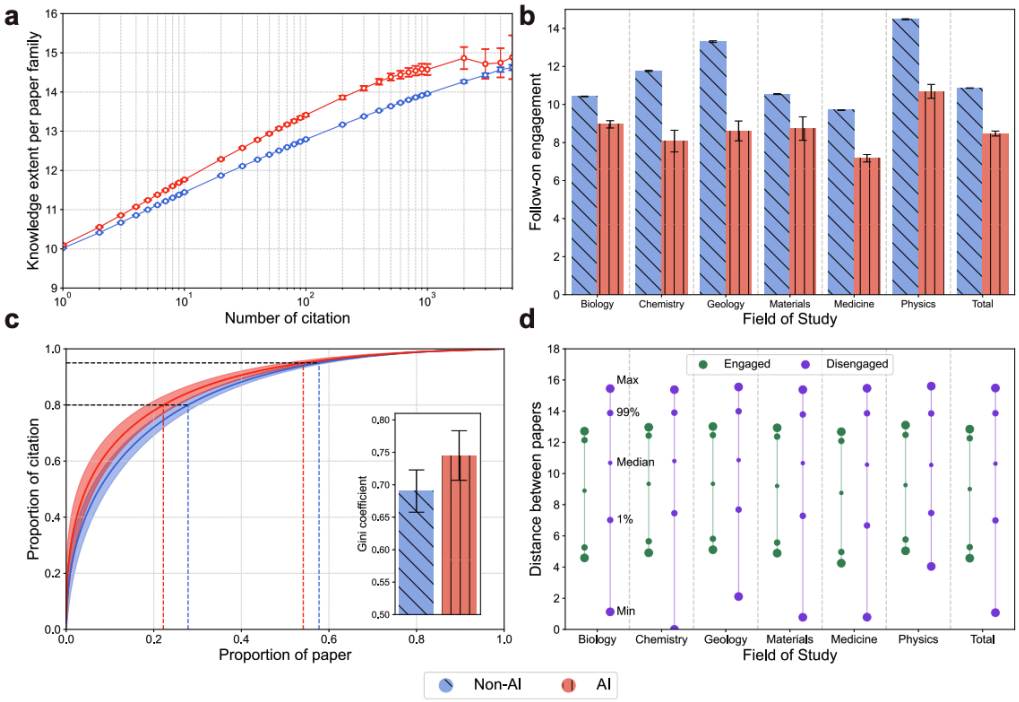
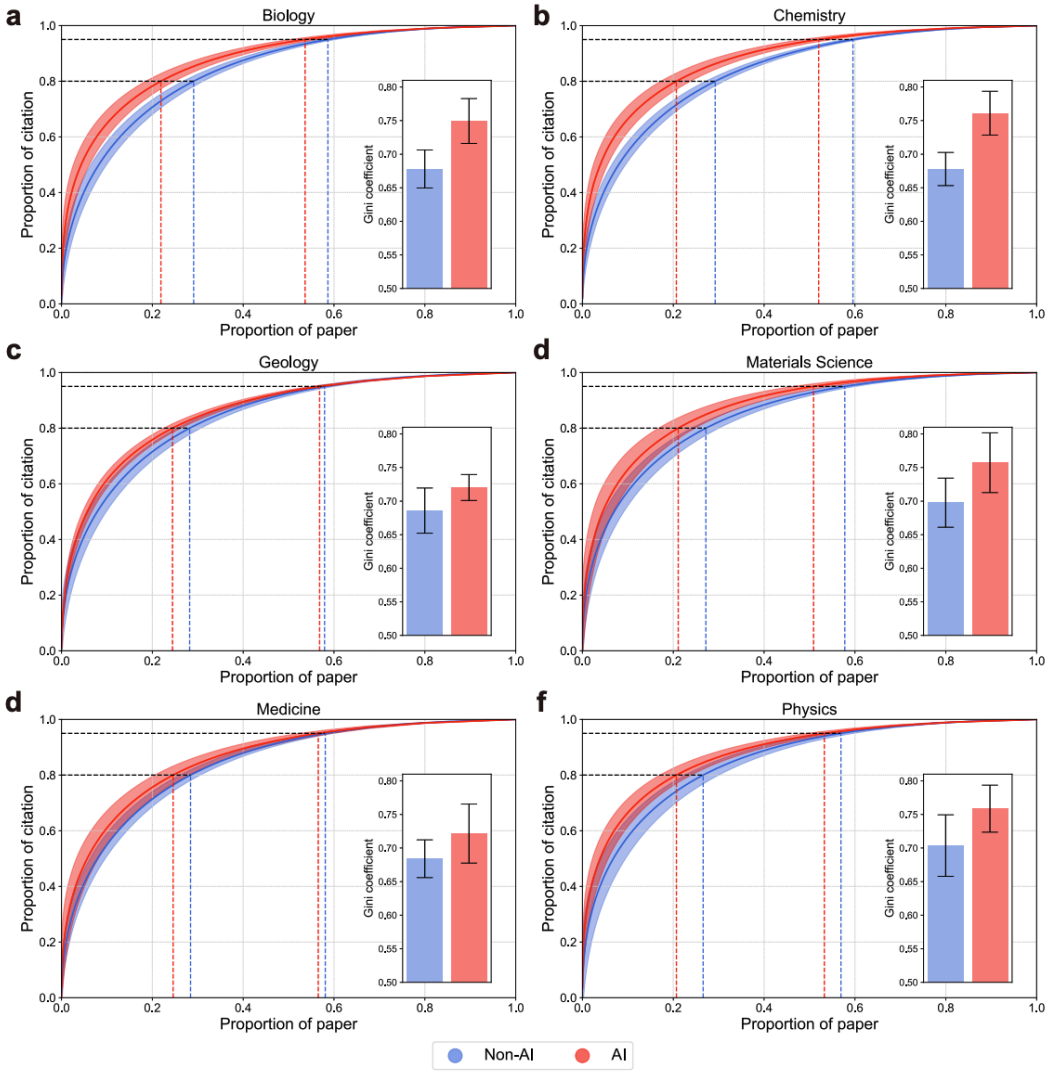


Figure 4. Over-concentration of AI research results in redundant innovation. (a) Knowledge extent of individual AI and non-AI paper families, i.e., an original paper and its cumulative citations ($n = 27,405,011$), where the knowledge space of individual AI paper families is broader and grows faster. (b) Engagement among the papers that cite AI vs. non-AI papers ($n = 23,342,516$), where there are fewer follow-on interactions among papers that cite the same original paper in AI research. For the above panels, 99% CIs are shown as error bars. (c) The distribution of citations to AI vs. non-AI papers ($n = 27,405,011$), where AI papers tend to concentrate more on a smaller number of top papers. (d) The distribution of distances between paper pairs that cite the same papers, with or without citing each other—engaged versus disengaged ($n = 590,325,130$ sampled paper pairs). Results show that for papers not engaged with each other, the median distance is larger, but the minimum distance is smaller, indicating a higher probability of overlap with each other in knowledge space.



Extended Data Fig. 10. The Matthew effect in citations to AI and non-AI papers. In AI research, a small number of superstar papers dominate the field, with approximately 20% of top papers receiving 80% of citations and 50% receiving 95%. This unequal distribution leads to a higher GINI coefficient in citation patterns surrounding AI research. Such disparity in the recognition of AI papers is consistent across all fields examined.

In summary, this is a very interesting manuscript, but one for which I would need to gain greater confidence in the robustness of the analyses.

Response: We thank the Reviewer for this encouraging and thoughtful feedback. With your valuable suggestions, we believe the robustness of our analyses has become substantially improved.

We sincerely appreciate the Reviewer for the valuable time and effort spent engaging with our manuscript. We believe that our manuscript has markedly improved in clarity, robustness, and insight across this revision cycle, in large part thanks to this review.

Referee #4

Thank you for the opportunity to review, "Artificial Intelligence Expands Scientists' Impact but Contracts Science's Focus". This manuscript highlights some of the intended and unintended consequences of AI adoption by scientists. For full disclosure, I think this is very important work and I like this manuscript a lot. My comments are to help the authors make the biggest contribution possible.

Response: We thank the Reviewer for this encouraging feedback on our manuscript. In this revision cycle, we endeavor to address all the comments you kindly provided. We believe that our manuscript is substantially improved with your insightful suggestions. Here, we provide a brief summary of our main revisions below.

a. We analyze the benefit of topic diversity in scientific research and illustrate that AI adoption does not selectively favor fruitful topics. Instead, AI homogeneously penetrates into topics with both high and low original influence.
b. We look into the semantic role of the citations within the context and analyze how the roles change over time. We also add a comparison of the pattern of citations between AI and other technologies that started small and became large, revealing similarities and differences.
c. We examine the robustness of our findings to the setting of the threshold for researcher dropout detection. We also explore the most appropriate threshold considering comprehensive factors and adjust the threshold to 3 years.
d. We add analysis on the relationship between the researchers' demographic information and their dropout, providing preliminary evidence of the heterogeneous impact of AI on scientific careers across different demographic groups.
e. We add a statistical analysis to examine our inference that "scientific field shrinks to focus attention on areas most amenable to AI research, such as those with an abundance of data", illustrating that data abundance is a major external factor influencing the selectivity of AI adoption across different topics.
f. We polish the texts and captions in our manuscript, making the descriptions more rigorous.

Additionally, considering the rapid development of AI, we updated the OpenAlex dataset used in our revision. The coverage period has been extended from the original cutoff date of May 30, 2024, to March 31, 2025, ensuring that our analysis remains as up-to-date as possible. In the latest OpenAlex snapshot, the dataset maintainers made several corrections, including the removal of abstracts containing invalid or junk content and updates to authorship information. These changes result in adjustments to the numerical values reported in our analysis; however, all original conclusions remain valid and unaffected.

The details of our revision are summarized as a point-by-point response to your comments as follows.

Validity

After reading the manuscript many times, I feel that it is very compelling. I have been thinking about whether/how the effects could be driven by mechanisms other than the ones presented by the authors. One area that could be ruled out more eloquently is whether the non-AI research that is being minimized or squeezed out should be squeezed out. That is, is more diversity of topics beneficial? (I do believe it is, but the question remains of whether some topics need to die out in a field and perhaps AI is accelerating this, versus whether it is pushing out fruitful areas).

Response 4.1: We thank the Reviewer for this interesting conjecture. First, we examine the most basic aspect: Is a broader knowledge space, namely a greater diversity of research topics, indeed beneficial to the overall advancement of scientific knowledge? To examine this, we categorized our selected papers into 252 distinct sub-fields and computed, for each sub-field, the average citation count and disruption score[1-2]—which are widely applied measurements for scientific impact in science of science research—across different eras of AI (Supplementary Fig. S21). Results show that, across different sub-fields and eras of AI, neither citation (as a proxy for impact) nor disruption (as a proxy for novelty) is obviously correlated with the sub-field's knowledge extent. This indicates that higher topic diversity within a subfield does not diffuse the sub-field's intellectual energy and thus reduces its scholarly impact or innovative capacity. Therefore, maintaining a broader diversity of research topics appears to be beneficial: it does not hinder a sub-field's performance; rather, it may offer more opportunities for advancing the scientific frontier and provide researchers with a wider array of investigation choices, ultimately benefiting the general state of knowledge.

Subsequently, we investigate whether the topics squeezed out by AI should indeed be marginalized. In other words, does AI tend to concentrate on more fruitful areas rather than topics with lower impact? To evaluate this, we categorize our selected papers into 1,883 topics according to the OpenAlex taxonomy. We calculate the average citation count and disruption score[1-2] for non-AI papers within each topic, obtaining proxies for the topic's original influence and novelty. We then examine, across the topics, the correlation between the degree of AI penetration and the original influence and novelty across different AI development eras, where the absolute values of Pearson's r are below 0.1 in all cases (Supplementary Fig. S23). Results show that across topics and AI eras, neither the original citation nor the original disruption is obviously correlated with the proportion of AI adoption for that topic. This suggests that AI adoption does not selectively favor topics based on their original impact or promise. Instead, AI homogeneously penetrates into topics with both high original influence, namely popular or apparently promising areas, and low original influence, namely unappreciated areas or those in a state of decline.

According to your suggestion and the above discussion, we make the following revisions to our manuscript:

a. We add the analysis regarding the benefit of topic diversity in *Supplementary Fig. S21* and *Supplementary Notes 3.1- The benefit of topic diversity*.

b. We add analysis regarding the selectivity of AI adoption in topics with different original impacts in *Supplementary Fig. S23* and *Supplementary Notes 3.2- Selectivity of AI adoption in different topics*, enhancing the robustness of our findings.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 6 in the *Main Text, Results-***AI adoption is associated with a contraction in knowledge focus among scientific fields] Generally, these results highlight an emerging conflict between individual and collective incentives to adopt AI in science, where scientists receive expanded personal reach and impact, but the knowledge extent of entire scientific fields tends to shrink and focus attention on a subset of topical areas. According to analyses on possible factors that may influence the selectivity of AI adoption across different topics, we find that factors like inherent topicality, original impact, and funding priority, remain almost unrelated to the disproportionate AI adoption (Supplementary Fig. S22-S24)...

[***Supplementary Notes 3.1***-The benefit of topic diversity] Given our observation of the contracted knowledge extent of AI-augmented research, a natural question arises: Is a broader knowledge space, namely a greater diversity of research topics, indeed beneficial to the overall advancement of scientific knowledge? To examine this, we categorize our selected papers into 252 distinct sub-fields and compute, for each sub-field, the average citation count and disruption score[1-2] across different eras of AI (Supplementary Fig. S21). Results show that, across different sub-fields and eras of AI, neither citation (as a proxy for impact) nor disruption (as a proxy for novelty) is obviously correlated with the sub-field's knowledge extent. Specifically, the absolute values of Pearson's r are below 0.1 in almost all cases. This indicates that higher topic diversity within a subfield does not dissipate the sub-field's intellectual energy and reduce its scholarly impact or innovative capacity. Therefore, maintaining a broader diversity of research topics does not hinder a sub-field's performance; rather, it likely offers more opportunities for advancing the scientific frontier and provides researchers with a wider array of investigation choices, ultimately benefiting the state of collective knowledge.

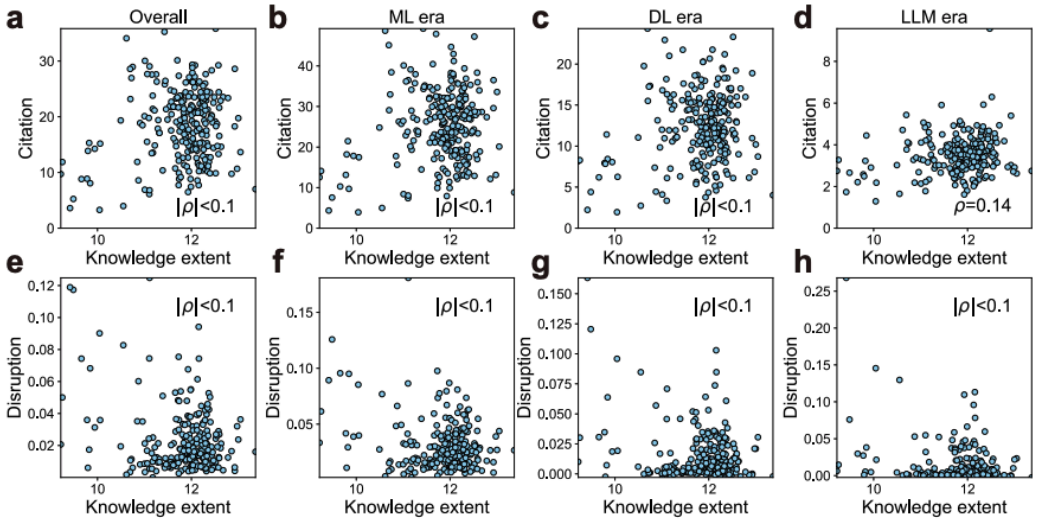
[***Supplementary Notes 3.2***-Selectivity of AI adoption in different topics] To gain a deeper understanding of our main findings that "AI in science has become more concentrated around specific hot topics" and that "AI-augmented research focuses on a narrower scope", we examine whether external factors might influence the selectivity of AI adoption across different topics, potentially leading to their disproportionate representation.

Original impact. Another possibility concerns whether the topics being squeezed out by AI are likely to be marginalized anyway. That is, does AI tend to concentrate on more fruitful areas rather than topics with lower prior and potential impact? To evaluate this, we categorize our selected papers into 1,883 topics according to the OpenAlex taxonomy. We calculate the average citation count and disruption score[1-2] of non-AI papers within each topic, obtaining proxies for the topic's original influence and novelty. We then examine, across topics, the correlation between the degree of AI penetration and the

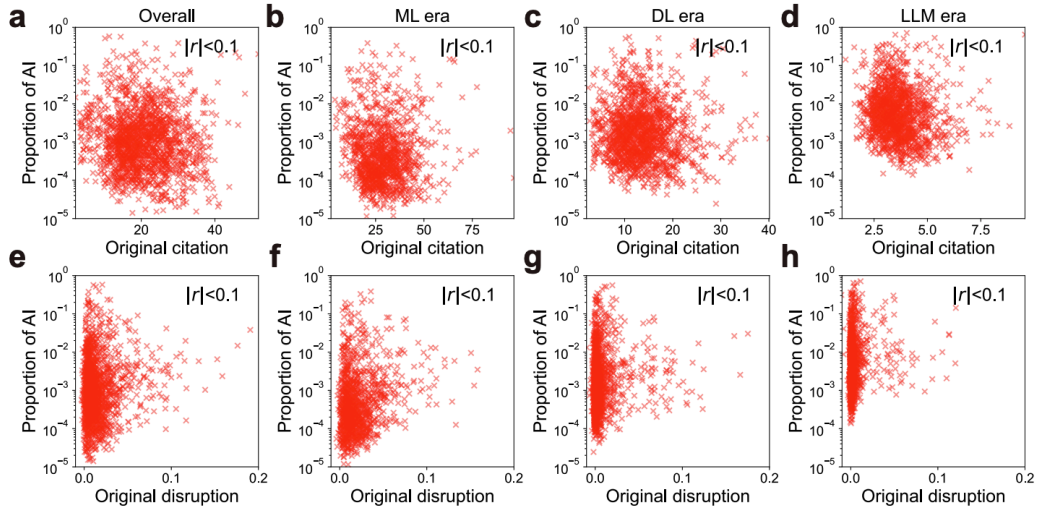
original influence and novelty across different AI development eras, where the absolute values of Pearson's r are below 0.1 in all cases (Supplementary Fig. S23). The results show that across topics and AI eras, neither original citation nor original disruption is obviously correlated with the proportion of AI adoption in that topic. This suggests that AI adoption does not selectively favor topics based on their original impact. Instead, AI homogeneously penetrates into topics with both high and low original influence.

[1] Lingfei Wu, Dashun Wang, and James A Evans. "Large teams develop and small teams disrupt science and technology". In: Nature 566.7744 (2019), pp. 378–382.

[2] Michael Park, Erin Leahey, and Russell J Funk. "Papers and patents are becoming less disruptive over time". In: Nature 613.7942 (2023), pp. 138–144.



Supplementary Figure S21. Correlation analysis between knowledge extent and citation or disruption of sub-fields. For all panels, each scatter represents a sub-field, and the corresponding Pearson's r is indicated correspondingly. Results show that, across sub-fields and eras of AI development, neither citation impact (as a proxy for impact) nor disruption (as a proxy for novelty) is obviously correlated with the sub-field's knowledge extent.



Supplementary Figure S23. Selectivity of AI adoption in different topics associated with the topics' original impact. In all panels, each scatter shows the degree of AI penetration in a topic and the original citation count or

disruption score within that topic. The corresponding Pearson's r is indicated in each panel. Results show that across topics and AI eras, neither the original citation nor the original disruption is obviously correlated with the proportion of AI adoption, suggesting that AI adoption does not selectively favor topics based on differential original impact.

Appropriate Use of Statistics and Treatment of Uncertainties

My empirical expertise is not directly related to the analyses presented in this paper (particularly around the construction of the variables). Nothing jumps out at me as problematic in this domain. One note is that the error bars are not defined in your explanation on Figure 2a/b, Figure 3.

Response 4.2: We thank the Reviewer for this comment. In the revised paper, we use error bars to show 99% CIs for the statistics in all panels of Figure 2. In Figure 3, we show 1.5 times the inter-quartile range (IQR) from the box as whiskers in panels (c) and (d). To make our definitions clearer, we have polished the captions of our figures. For your convenience, we quote the revised text below.

Revision:

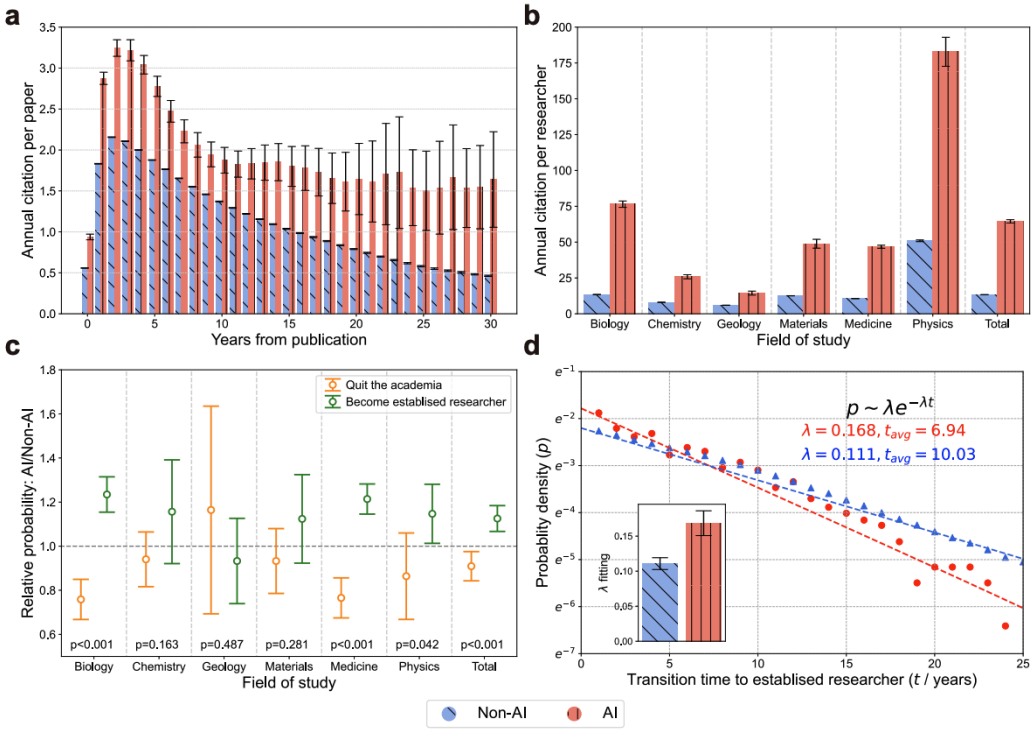


Figure 2. AI enlarges paper impact and enhances researcher careers. ... For all panels, 99% CIs are shown as error bars.

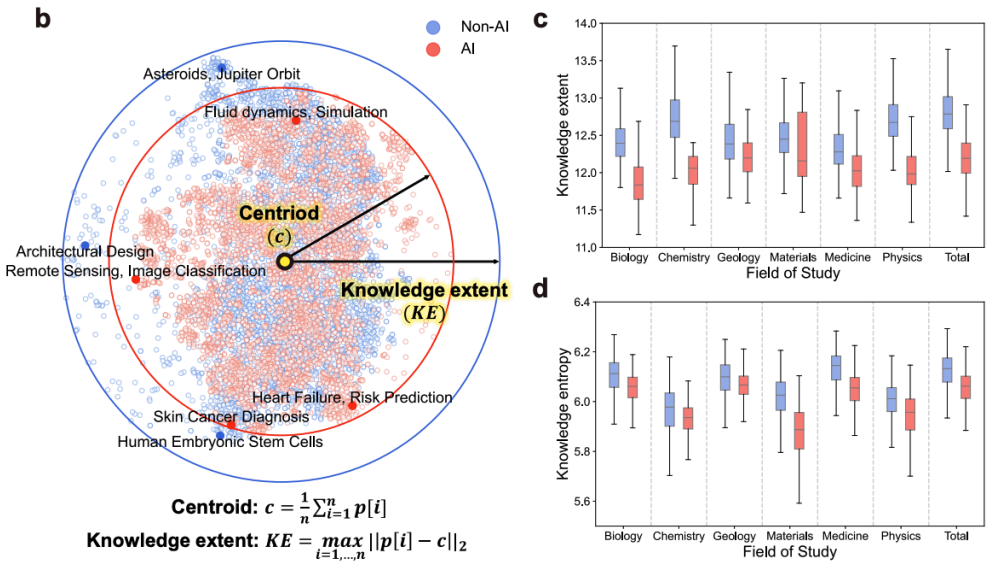


Figure 3. AI adoption is associated with a contraction in knowledge extent within and across scientific fields. … For boxplots in panels (c) and (d), we show 1.5 times the inter-quartile range (IQR) from the box as whiskers.

Suggested Improvements

1. Curious about the phrasing of the research question (last sentence of page 1). In the question itself is the word "beneficial". Why is it "beneficial adoption" instead of simply "adoption"? Beneficial to whom? You of course find that it is beneficial for individual scientists but again, that is a conclusion you have made stemming from your RQ.

Response 4.3: We thank the Reviewer for this question. We agree with this unnecessary ambiguity in the research question and have removed that misleading word. As you mentioned, this "benefit" would be actually illustrated later in our analyses for individual scientists. For your convenience, we quote the revised text below.

Revision:

[**Page 2 in the *Main Text, Introduction*-**Research design] Aligning the promise and peril of AI in scientific research, here we pose the question: How does the adoption of AI by individual scientists influence the collective character and dynamic exploration of science as a whole?

2. Curious about your findings in Figure 2a regarding citations of AI papers from 20+ years ago. Given there were fewer (AI) papers back then, and most papers need to cite something "historical" to situate their work in a broader conversation. I do not know if it is possible for you to identify foundational citations versus "throw away" citations that everyone uses more colloquially (this has an important problem vs. here is how we build off of X). Would you see the same pattern of citations for other topics that started small and then became the most prominent topic/method in a given field. Is this unique to AI citations versus other things that started small and became big?

Response 4.4: We thank the Reviewer for this very interesting comment and line of inquiry. We now explore the semantic role of citations within context and examine how that role changes over time to avoid misattributing a change in convention and scholarly manners to one of outsized influence.

To investigate this, we use the criteria and methodology proposed in our previous work for distinguishing between core and superficial citations[1], and compute the times that AI and non-AI papers are cited as core citations in each year following publication (Supplementary Fig. S12). Results show that the proportion of AI papers to be cited as core citations tends to decrease over time, which aligns with the intuitive expectation that older papers are more likely to be cited for historical context rather than as active intellectual foundations. Nevertheless, foundational influence can still be observed in decades-old papers.

For example, two recent studies published in 2020 and 2021, which used AI techniques to predict diabetes[2-3], both cite a 1988 paper on the ADAP learning algorithm[4] as a foundational reference. Both the newer papers are built upon the old algorithm to design more advanced ensemble models. Moreover, in terms of absolute counts of being core citations, AI papers still surpass non-AI papers by 98.70%, reflecting the heightened academic attention attracted by AI research.

We also explore whether citations to AI are unique as compared with other things that started small and became large. To investigate this, we conducted a comparative analysis using Nanotechnology as a case study (Supplementary Fig. S13). We identified Nanotechnology-related research by detecting the presence of the word "nano" in the titles of all publications. As shown, Nanotechnology was sparsely studied in its early stages but began to gain widespread attention across disciplines such as biology, chemistry, and physics starting in the 1990s. We observe that Nanotechnology exhibits a citation pattern partially similar to AI: Nanotechnology papers are still cited as core citations many years after publication, although the proportion of being such core citations declines over time. In contrast to AI, however, both total number of citations and frequency of being core citations for Nanotechnology papers eventually decline to levels indistinguishable from non-Nanotechnology papers, suggesting that nanotechnology does not exhibit the same sustained outsized academic influence characterizing AI.

According to your suggestion and the above discussion, we make the following revisions to our manuscript:

a. We add analysis regarding the semantic role of citations within context in *Supplementary Fig. S12* and *Supplementary Notes 2.1- The effect of AI on individual papers*, demonstrating the changing role of citations over time.
b. We compare citation patterns for AI and Nanotechnology in *Supplementary Fig. S13* and *Supplementary Notes 2.1- The effect of AI on individual papers*, revealing similarities and differences between AI and other fields that started small and became large.

For your convenience, we quote the revised texts and figures below.

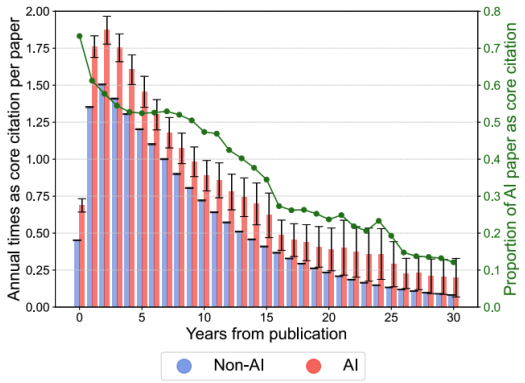
Revision:

[***Supplementary Notes 2.1-***The effect of AI on individual papers] In our analysis of annual citations for AI and non-AI papers, we find that both types of papers continue to be cited even 20 or more years after publication. Given that fewer papers, especially AI-related ones, were published several decades ago and that most scholarly works cite "historical" papers to situate their research within a broader context, it is important to understand whether referenced papers are being cited as foundational citations or simply "throwaway" citations that the scholarly crowd uses more colloquially. To address this, we follow an established methodology for distinguishing between core and superficial citations[1], and we compute the times AI and non-AI papers are cited as core citations in each year following publication (Supplementary Fig. S12). Results show that the proportion of AI papers to be cited as core citations tends to decrease over time, which aligns with the intuitive expectation that older papers are more likely to be cited for historical context rather than as active intellectual foundations. Nevertheless, foundational influence can still be observed in decades-old papers. For example, two recent studies published in 2020 and 2021, which used AI techniques to predict diabetes[2-3], both cite a 1988 paper on the ADAP learning algorithm[4] as a foundational reference. Both newer papers build upon the prior algorithm to design more advanced ensemble models. Moreover, in terms of absolute counts of being core citations, AI papers still surpass non-AI papers by 98.70%, reflecting the heightened academic attention attracted by AI research.

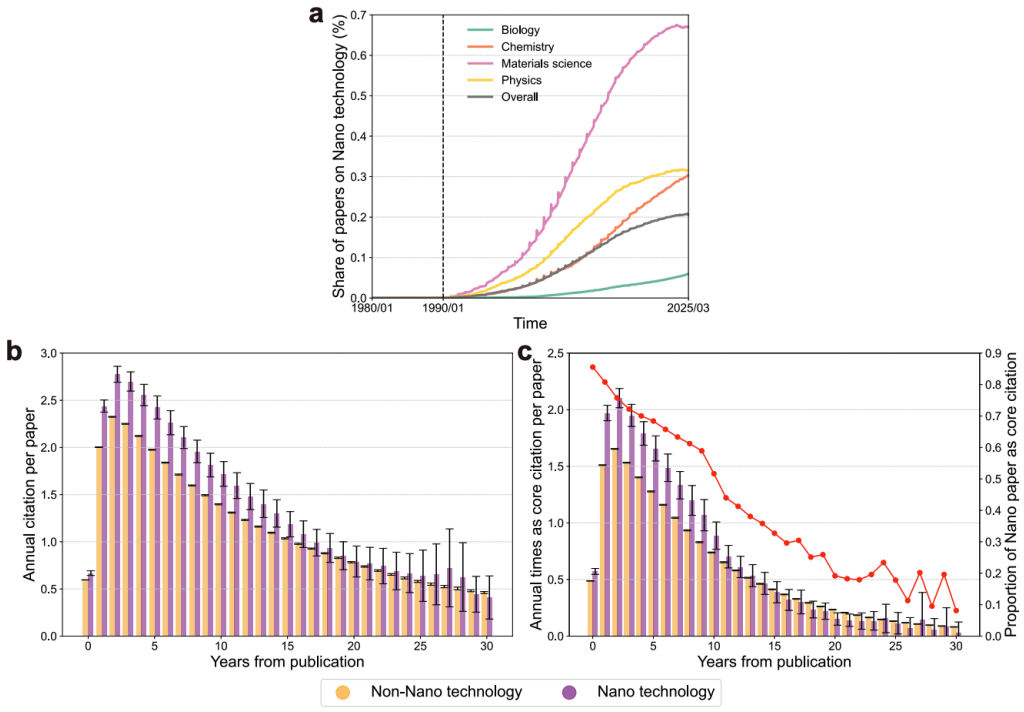
[***Supplementary Notes 2.1-***The effect of AI on individual papers] AI represents a technology that started small and then became the most prominent method in multiple fields. As such, AI papers exhibit a distinctive citation pattern compared to non-AI papers. To investigate whether this citation pattern is unique to AI or also characteristic of other emerging technologies that started small and then became widely appreciated and used, we conduct a comparative analysis using Nanotechnology as case study (Supplementary Fig. S13). We identified Nanotechnology-related research by detecting the presence of the word "nano" in the titles of all publications. As shown, Nanotechnology was sparsely studied in its early stages but began to gain widespread attention across disciplines such as biology, chemistry, and physics beginning in the 1990s. We observe that Nanotechnology exhibits a citation pattern partially similar to AI: Nanotechnology papers are still cited as core citations many years following publication, although the proportion of being such core citations declines over time. In contrast to AI, however, both total number of citations and the frequency of being core citations for Nanotechnology papers eventually decline to levels indistinguishable from non-Nanotechnology papers, suggesting that nanotechnology does not exhibit the same sustained higher academic influence as AI.

[1] Qianyue Hao, Jingyang Fan, Fengli Xu, Jian Yuan, and Yong Li. "HLM-Cite: Hybrid Language Model Workflow for Text-based Scientific Citation Prediction". In: NeurIPS. 2024.

[2] Md Kamrul Hasan, Md Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. “Diabetes prediction using ensembling of different machine learning classifiers”. In: IEEE Access 8 (2020), pp. 76516–76531.
[3] Saloni Kumari, Deepika Kumar, and Mamta Mittal. “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier”. In: International Journal of Cognitive Computing in Engineering 2 (2021), pp. 40– 46.
[4] Jack W Smith, James E Everhart, William C Dickson, William C Knowler, and Robert Scott Johannes. “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus”. In: Proceedings of the annual symposium on computer application in medical care. 1988, p. 261.



Supplementary Figure S12. Distinguishing between core and superficial citations. Results show that the proportion of AI papers cited as core citations tends to decrease over time (green line), while foundational influence can still be observed in decades-old papers. In terms of the absolute count of instances being among the core citations in future papers, AI papers still outperform non-AI papers (red and blue bars), reflecting their enhanced academic impact. 99% CIs are shown as error bars.



Supplementary Figure S13. Comparison between citation patterns of Nanotechnology and non-Nanotechnology papers. (a) Growth of Nanotechnology over time in different disciplines. **(b)** Annual citations after publication of Nanotechnology and non-Nanotechnology papers. **(c)** Annual times of being core citation following publication of Nanotechnology and non-Nanotechnology papers. Nanotechnology papers are still cited as core citations many years after publication, with the proportion of being among the core citations declining over time. In contrast to AI, both total number of citations and the frequency of being core citations for Nanotechnology papers eventually decline to levels indistinguishable from non-Nanotechnology papers. For panels (b) and (c), 99% CIs are shown as error bars.

3. 2 years to being "out" in your analyses. In your analysis regarding whether scientists exist, are your results robust to looking at lengths longer than 2 years. For instance, if someone goes on maternity leave and hence experiences a gap in publication due to life circumstances, can they come back or do most people fall off at the 2 year mark?

Response 4.5: We thank the Reviewer for this astute question. To ensure the robustness of our findings to the setting of this threshold, we switch to 3 or 4 years and replicate our results. As we illustrate in Supplementary Fig. S18, results are consistent for different thresholds, underscoring the role of AI in bringing about increased opportunities for junior scientists to lead research teams and reducing their risk of leaving academia.

More importantly, we explore the most appropriate threshold used for detecting researcher dropout. We analyzed the distribution of gap year durations in researchers' careers (Supplementary Fig. S19). Here, a period of gap years refers to a temporary interruption in a researcher's publication activity, followed by a subsequent resumption of publishing. Results show that 44.67% of all gap periods lasted only a single year, and 76.94% lasted no more than three years. Indicating that a three-year threshold can appropriately capture the majority of cases where researchers temporarily paused and then resumed publication activity. Moreover, previous researchers also use the threshold of three years in similar statistics[1]. A threshold too short would misclassify many researchers with temporary gaps as dropouts, while a threshold too long would result in a large number of researchers being classified as having uncertain status and excluded from analysis and reducing the sample size. Accounting for these considerations, we adjust the threshold to three years as a balanced and robust choice. All related original results are updated.

We add analysis on the choice and robustness of this threshold in *Supplementary Fig. S18-S19* and *Supplementary Notes 2.2- The effect of AI on individual scientists' careers*. For your convenience, we quote the revised texts and figures below.

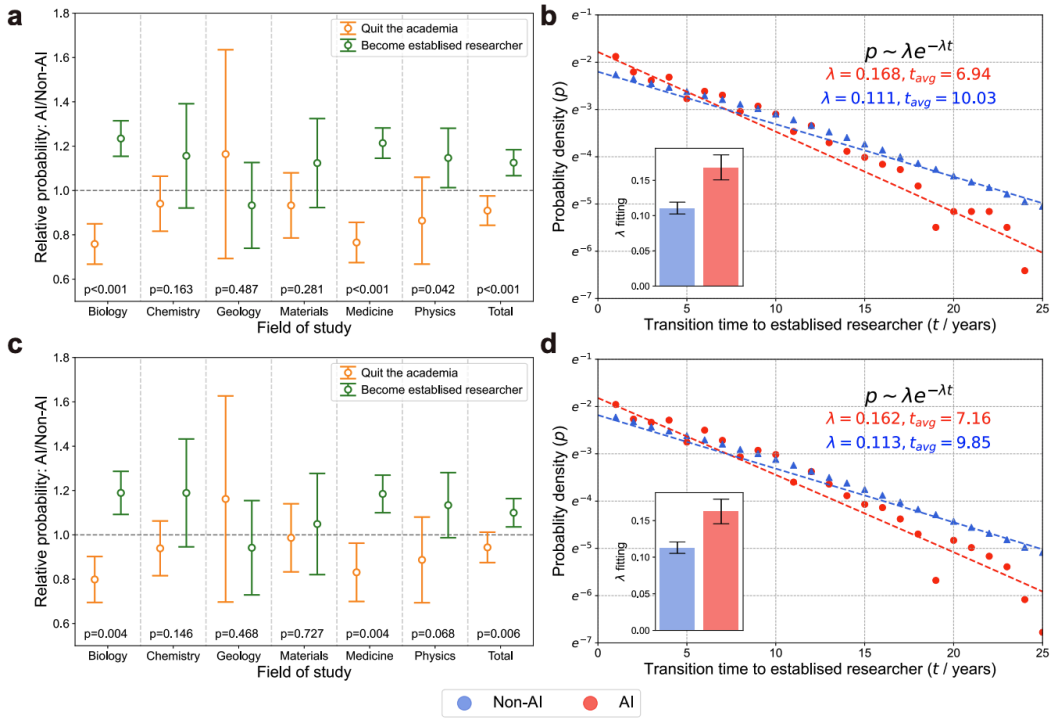
Revision:

[**Page 10 in the *Methods*-**M6. Detect scientists' career role transition] Therefore, we follow the settings in previous research[1] to use a threshold of 3 years and regard scientists who have no more publications after 2022 as having exited academia, while those who still publish papers after 2022 are considered to have an unclear ultimate status and are excluded from analysis.

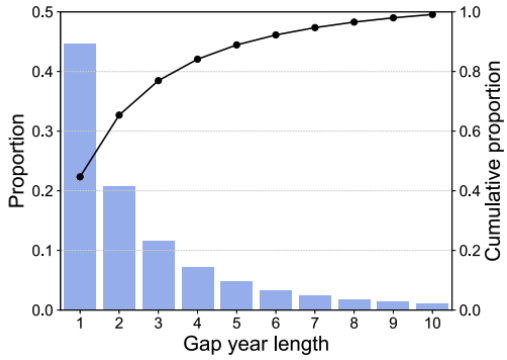
[***Supplementary Notes 2.2-***The effect of AI on individual scientists' careers] In our career transition model of researchers, we set a threshold of 3 years and regard scientists who

have no more publications after 2022 as having exited academia, where the threshold setting is consistent with previous research[1]. To ensure the robustness of our findings, we switch to different thresholds in detecting the dropout of researchers and replicate our results (Supplementary Fig. S18). As we illustrate, when using different dropout thresholds of 2 or 4 years, the probability for AI-adopted junior scientists to transition to established scientists is consistently higher than for their counterparts who do not adopt AI, and the anticipated transition time to becoming established scientists is shorter for AI-adopted junior scientists compared to their counterparts. These robust results underscore the role of AI in bringing about increased opportunities for junior scientists to lead research teams and reducing the risks of their leaving academia. To assess the appropriateness of our threshold for detecting researcher dropout, we analyzed the distribution of gap year durations in researcher careers (Supplementary Fig. S19). Here, a period of gap years refers to a temporary interruption in a researcher's publication activity, followed by a subsequent resumption of publishing. Results show that 44.67% of all gap periods lasted only one year, and 76.94% lasted no more than three years. Therefore, when choosing a three-year threshold for identifying dropout events, it can appropriately capture the majority of cases where researchers temporarily paused and then resumed publication activity. Additionally, as we illustrate, both shorter and longer thresholds yield similar overall results. Nevertheless, a threshold that is too short would misclassify many researchers with temporary gaps as dropouts, while a threshold too long would result in a large number of researchers being classified as having uncertain status and excluded from analysis, thereby reducing the sample size. Taking these considerations into account, we adopt a three-year threshold as a balanced and robust choice.

[1] Staša Milojevic, Filippo Radicchi, and John P Walsh. "Changing demographics of scientific careers: The rise of the ´ temporary workforce". In: Proceedings of the National Academy of Sciences 115.50 (2018), pp. 12616–12623.



Supplementary Figure S18. Impact of AI adoption on researchers' careers with different thresholds in detecting the dropout of researchers. (a) (c) The probability of two role transitions, where we show the quotients between junior scientists adopting AI and their counterparts without AI. **(b) (d)** Probability density distribution of the transition time for junior scientists to become established. In (a) and (b), the threshold for detecting researcher dropout is 2 years. In (c) and (d), the threshold for detecting researcher dropout is 4 years. Results are consistent for both shorter and longer thresholds. For all panels, 99% CIs are shown as error bars.



Supplementary Figure S19. The distribution of gap year durations in researcher careers. A period of gap years refers to a temporary interruption in a researcher's publication activity, followed by a subsequent resumption of publishing. Results show that 44.67% of all gap periods lasted only one year, and 76.94% lasted no more than three years.

4. I think this paper does a whole lot, but I am curious about the people who are leaving the field at higher rates. Do you have any demographic information on these people? Again, this could be fodder for Future Research but does the adoption of AI affect all groups of scientists in a uniform manner?

Response 4.6: We thank the Reviewer for this thoughtful question. It would indeed be insightful to look into the relationship between the researchers' demographic information, AI use, and dropout. We conduct some additional analyses to explore this (Supplementary Fig. S20).

First, we extract preliminary demographic information of researchers based on the institution to which they are affiliated. On the one hand, we categorize institutions by type and find that researchers affiliated with educational institutions exhibit the lowest dropout rates, although the differences in dropout rates across institution types are relatively modest. Notably, across all institutional types, researchers who adopt AI manifest similarly reduced dropout probabilities. On the other hand, we categorize institutions by geographic region, where researchers based in Africa and South America have higher dropout probabilities compared to those in Europe, Asia-Pacific, and North America. Furthermore, while AI adoption is associated with reduced dropout rates for researchers in regions such as Asia-Pacific and North America, the benefit is far less pronounced for those in Africa and South America.

Second, we follow established methods in prior work and infer gender and ethnicity information based on author names[1]. Results show that White and Asian researchers have lower dropout probabilities compared to their Hispanic and Black counterparts. Meanwhile, AI adoption is associated with reduced dropout rates for White and Asian researchers, while the benefit is much less pronounced for Black researchers. We also observe heterogeneity in the effect between male and female researchers. Male researchers tend to have lower original dropout probabilities and receive greater benefit from AI adoption. In contrast, female researchers exhibit higher original dropout probabilities and gain relatively less from adopting AI. These findings provide preliminary evidence regarding the heterogeneous impact of AI on scientific careers across different demographic groups, suggesting that the benefits of AI are unequally distributed, but the causes and consequences of this heterogeneity merit further investigation.

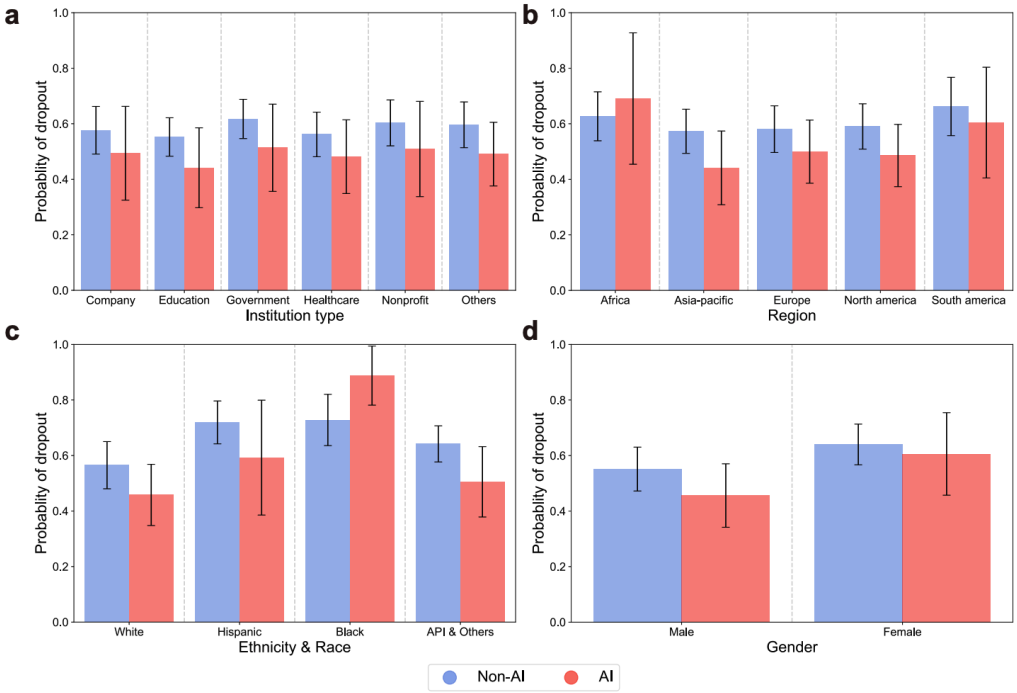
We add the analysis of researchers' demographic information and their dropout in *Supplementary Fig. S20* and *Supplementary Notes 2.2- The effect of AI on individual scientists' careers*. For your convenience, we quote the revised texts and figures below.

Revision:

[***Supplementary Notes 2.2-***The effect of AI on individual scientists' careers] To better understand whether the adoption of AI affects all groups of scientists uniformly, we provide demographic information on the people who are leaving the field (Supplementary Fig. S20). First, we utilized the "institution" field in the OpenAlex dataset to determine researchers' affiliations and categorized them into demographic groups based on

institutional attributes. On the one hand, we categorize institutions by type, such as companies, educational institutions, government agencies, etc. Researchers affiliated with educational institutions exhibit the lowest dropout rates, although differences in dropout rates across institution types are relatively modest. Notably, across all institutional types, researchers who adopt AI show similarly reduced dropout probabilities. On the other hand, we categorize institutions by geographic region, including Africa, Europe, Asia-Pacific, etc. We find that researchers based in Africa and South America have higher dropout probabilities compared with those in Europe, Asia-Pacific, and North America. Furthermore, while AI adoption is associated with reduced dropout rates for researchers in regions such as Asia-Pacific and North America, the benefit is far less pronounced for those in Africa and South America. Second, we follow established methods in prior work and infer gender and ethnicity information based on author names[1]. The results show that White and API (Asian or Pacific Islander) researchers have lower dropout probabilities compared to their Hispanic and Black counterparts. Meanwhile, AI adoption is associated with reduced dropout rates for White and API researchers, while the benefit is much less pronounced for Black researchers. We also observe heterogeneity in the effect between male and female researchers. Male researchers tend to have lower original dropout probabilities and receive greater benefit from AI adoption. In contrast, female researchers exhibit higher original dropout probabilities and gain relatively less from adopting AI. These findings provide preliminary evidence regarding the heterogeneous impact of AI on scientific careers across different demographic groups, suggesting that the benefits of AI are unequally distributed, but the causes and consequences of this heterogeneity merit further investigation.

[1] Jeffrey W Lockhart, Molly M King, and Christin Munsch. “Name-based demographic inference and the unequal distribution of misrecognition”. In: Nature Human Behaviour 7.7 (2023), pp. 1084–1095.



Supplementary Figure S20. Impact of AI adoption on research careers across different demographic groups. (a) Researchers affiliated with institutions of different types. **(b)** Researchers affiliated with institutions within different geographic regions. (c) Researchers of different ethnicity and race. (d) Researchers of different genders. Results show the heterogeneous impact of AI on scientific careers across different demographic groups. Results show the heterogeneous impact of AI on scientific careers across different demographic groups. For all panels, 99% CIs are shown as error bars.

5. On page 6, you state "…scientific field shrinks to focus attention on areas most amenable to AI research, such as those with an abundance of data". Do you specifically test for this assumption? Or is this an inference? I wonder whether this effect is based on the availability of data or idiosyncratic preferences of the scientists who chose the starting point (and then experienced the Matthew Effect).

Response 4.7: We thank the Reviewer for this question. In the original manuscript, we infer that scientific fields tend to shrink and focus attention on areas with an abundance of data, which may be most amenable to AI research. Here, we add statistical analysis to estimate the strength and significance of this inference.

To evaluate the impact of data abundance on the selectivity of AI adoption across different topics, we categorize our selected papers into 1,883 topics according to the OpenAlex taxonomy. We extract and quantify the appearance frequency of data-related terms, such as "data" and "dataset", in titles and abstracts of papers within each topic. We then normalize these frequencies to serve as an indicator of data abundance for each topic. Next, we examine, across topics, the correlation between degree of AI penetration and data abundance across different AI development eras (Supplementary Fig. S25). Results show that data abundance is diverse across topics, while across AI eras, data abundance is positively correlated with the proportion of AI adoption, suggesting that AI is more likely to be adopted in topics with abundant data resources.

Result reveals that data abundance appears to be a major external factor influencing the selectivity of AI adoption across different topics, contributing to the observed disproportionate representation within knowledge space and the contraction of scientific focus. This finding elucidates a critical factor underlying the heterogeneous adoption of AI across topics and provides valuable insight into how AI can be better leveraged to promote sustainable and comprehensive scientific development by not only consuming, but intelligently selecting, curating, and generating relevant data.

We add this extended analysis in *Supplementary Fig. S25* and *Supplementary Notes 3.2-Selectivity of AI adoption in different topics*. For your convenience, we quote the revised texts and figures below.

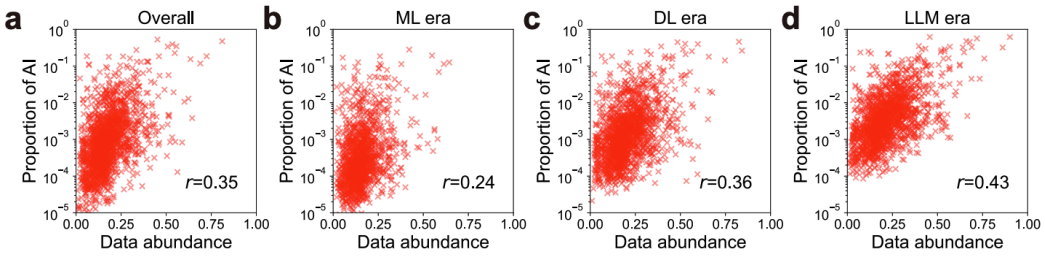
Revision:

[**Page 6 in the *Main Text, Results-***AI adoption is associated with a contraction in knowledge focus among scientific fields] Generally, these results highlight an emerging conflict between individual and collective incentives to adopt AI in science, where scientists receive expanded personal reach and impact, but the knowledge extent of entire scientific fields tends to shrink to focus attention on specific areas… Data availability turns out to be a major impacting factor, where areas with an abundance of data are increasingly and disproportionately amenable to AI research, contributing to the observed concentration within knowledge space (Supplementary Fig. S25).

[***Supplementary Notes 3.2***-Selectivity of AI adoption in different topics] To gain a deeper understanding of our main findings that "AI in science has become more concentrated around specific hot topics" and that "AI-augmented research focuses on a narrower scope", we examine whether external factors might influence the selectivity of AI adoption across different topics, potentially leading to their disproportionate representation.

Data abundance. As discussed in the main text, the knowledge extent of entire scientific fields tends to shrink to focus attention on areas most amenable to AI research, such as those with an abundance of data. To directly evaluate the impact of data abundance on the selectivity of AI adoption across topics, we extract and quantify the appearance frequency of data-related terms, such as "data" and "dataset", in titles and abstracts of papers within each topic. We then normalize these frequencies to serve as an indicator of data abundance for each topic. We then examine, across topics, correlation between the degree of AI penetration and data abundance across different AI development eras (Supplementary Fig. S25). Results show that, as expected, data abundance is diverse across topics, while across AI eras, data abundance is positively and significantly ($p < 0.001$ for all eras) correlated with the proportion of AI adoption, suggesting that AI is more likely to be adopted in topics with abundant data resources. Notably, the correlation of AI-use with frequency of mentioned data resources rises with the size and intensity of AI-models from 0.24 in the machine learning era to 0.36 in the deep learning era to 0.43 in the large model era.

In general, data abundance is a major external factor influencing the selectivity of AI adoption across different topics, contributing to the observed disproportionate representation within knowledge space and the contraction of scientific focus. This finding elucidates factors underlying the heterogeneous adoption of AI across topics and provides valuable insights into how AI can be better leveraged to promote sustainable and comprehensive scientific development.



Supplementary Figure S25. Selectivity of AI adoption in different topics regarding data abundance in the topics. In all panels, each scatter shows the degree of AI penetration in a topic and the data abundance in that topic. The corresponding Pearson's r is indicated correspondingly in each panel. The results show that across topics and AI eras, the data abundance is positively correlated with the proportion of AI adoption, suggesting that AI is more likely to be adopted in topics with abundant data resources. Notably, the correlation of AI-use with frequency of mentioned data resources rises with the size and intensity of AI-models.

6. I found your explanation of Figure 4d very confusing. The explanation could be revised (I am sorry that I cannot offer a path forward because I am not sure what you are saying).

Response 4.8: We thank the Reviewer for this important catch. We regret not more clearly explaining how the sample points in Fig. 4d were obtained and how the relationships between different presented statistics should be interpreted. To address this, we have improved the relevant section, making the explanation of Fig. 4d much more clear.

We first summarize the ideas in Fig. 4 a-c. Here, we intend to further explore the observed conflict between the growing influence of individual AI-enhanced papers and researchers and the narrowing of domain knowledge within AI research. In Fig. 4a, we found that the contraction of knowledge space in AI research is not attributable to a narrowing of knowledge attributable to each work of original research. To further investigate, in Fig. 4b and Fig. 4c, we subsequently reveal that AI papers tend to concentrate on the original paper, rather than engaging with and building upon each other.

In Fig. 4d. We sample 590,325,130 pairs of papers, where each pair cites the same original work. Among these, 51,723,984 pairs not only cite the same original work but also cite each other (engaged), while the remaining pairs do not cite each other (disengaged). We examine distances between these pairs of papers. Results show that, on average, a pair of disengaged papers indeed focuses on less relevant topics and lies farther apart. In some cases, however, when the pair of papers happen to work on similar topics, the lack of engagement leads to mutual unawareness. As a result, they likely lie very close to one another, producing overlapping research. Collectively, these results

reveal that greater concentration in knowledge extend leads to more overlapping ideas and redundant research activity.

To simplify these dynamics for straightforward comprehension, we have revised our corresponding explanations. For your convenience, we quote the revised texts and figures below.

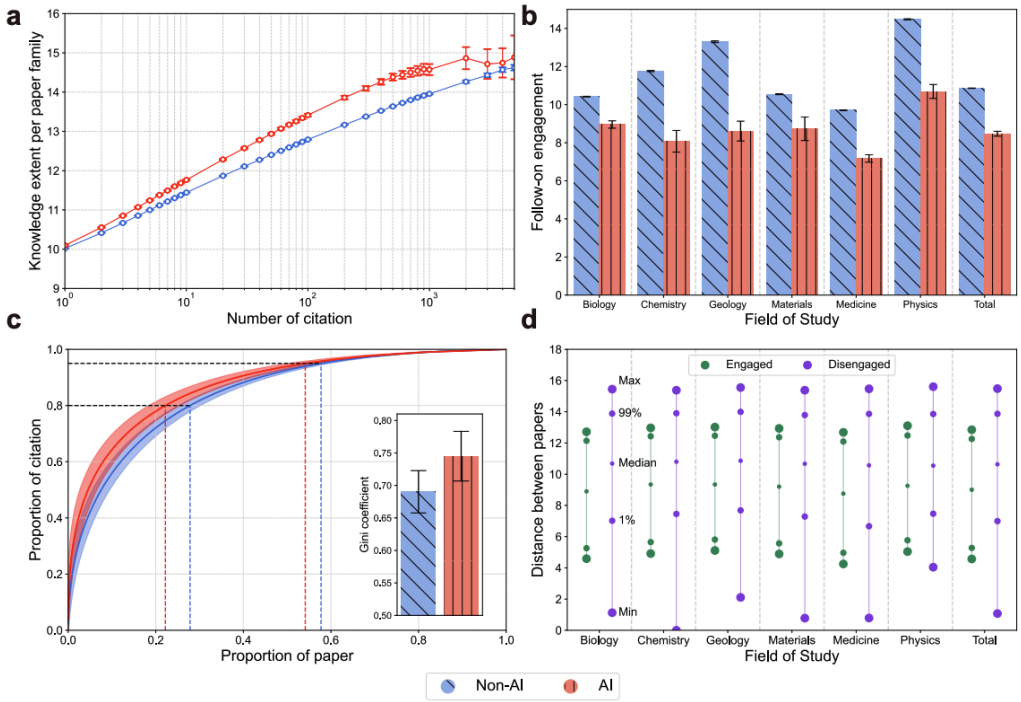


Figure 4. Over-concentration of AI research results in redundant innovation. (d) Distribution of distances between paper pairs that cite the same prior research, with or without citing one another—engaged versus disengaged ($n = 590,325,130$ sampled paper pairs). Results show that for papers not engaged with each other, the median distance is larger, but the minimum distance is smaller, indicating a higher probability of overlapping in knowledge space.

Revision:

[**Page 6 in the *Main Text, Results*-**Over-concentration of AI research and redundant innovation] To further analyze the impact of reduced follow-on engagement, we sample $590,325,130$ pairs of papers, where each pair cites the same original work. Among these, $51,723,984$ pairs not only cite the same original work but also cite each other (engaged), while the remaining pairs do not cite each other (disengaged). We examine distances between these pairs of papers within the 768-dimensional vector space (Fig. 4d) and find that median distance between paper pairs disengaged from one another tends to be 18.11% larger than between paper pairs engaged with each other ($\chi^2 \geq 106$, $p < 0.001$ and $df = 1$ in median-test on all disciplines). In contrast, the closest disengaged paper pairs are 76.51% closer to one another than the closest paper pairs engaged with one another. Taken together, a pair of disengaged papers indeed focuses on less relevant topics and lies farther apart in the embedding space. When the two papers happen to work on similar topics, however, the lack of engagement leads to mutual

unawareness and overlap. As a result, they are likely to lie very close to each other, producing redundant research.

We sincerely appreciate the Reviewer for the valuable time and effort spent engaging with our manuscript. We believe that our manuscript has substantially improved in clarity and robustness across this revision cycle, including with the benefit of insights and curiosity from this review!

Revision of manuscript in Nature #2025-01-00126B

We are delighted to receive this thoughtful feedback on our revised paper from the *Nature* editor and all reviewers. The reviews demonstrate continued depth of familiarity and engagement with our work, providing excellent suggestions that we have used to substantially improve our paper.

Here, we summarize our changes:

Improved claims and statements **(Reviewers 2)**
• Difference between conventional ML and LLMs: Clearly articulate the limitations of the data and analyses in the era of LLMs and Generative AI and renamed the "LLM era" the "Generative AI" era following the Reviewer's framing for greater generality.
• More rigorous claims: Toned down the detailed claims to be more rigorous and precise.
• Wording and presentation: Polished presentation to better in line with the academic style and broad audience of *Nature*.

Improved technical details **(Reviewers 3)**
• Author disambiguation: Detailed the inputs and algorithms in the OpenAlex's Author Name Disambiguation (AND) system.
• Sample size: Clarified the analyzed author sample size according to our definition of "established researcher".
• KE equation: A sensitivity analysis on the reliability of distance calculations to validate and support distance-based knowledge extent measurements.
• Figure polishing: Updating figures ranging from the aspect of plot types and statistical choices to axis scales and more.

Below, we include a point-by-point response to reviewers' comments, with our responses below and our revisions in red.

Response to Referee#1's Comments

I thank the authors for their thorough and extensive revision that addressed my remarks. I have no further comments and recommend publication.

Response: We are pleased to hear that our revision has addressed your concerns. We thank the Reviewer for their valuable time spent reviewing our manuscript.

Response to Referee#2's Comments

Thank you for the revised version of this paper. Obviously, the authors have put in considerable work to reworking the paper. However, I am still having problems understanding the paper. Also, I am still grappling with the potential contributions and whether there is enough mature evidence. I do agree that there is a great deal of research that involves the use of AI and it could be interesting to assess the impact. For example, machine learning is used in a wide variety of applications.

Response: We thank the Reviewer for your time and effort reviewing our manuscript. We appreciate the acknowledgement of our revision efforts and are dedicated to further strengthening our paper following these additional suggestions.

a. We have clarified the difference between classic machine learning techniques and generative AI/LLMs. We have modified our claims accordingly, making it clear that due to limited data in the era of LLMs, our corresponding results serve as a starting point for further study as LLMs develop in science rather than a strong conclusion.
b. We have toned down some of the claims on the implications on contracted science focus per your suggestions, making it more rigorous rather than simply labeling it a negative impact.
c. We have thoroughly polished our wording and presentation, making it more appropriate as a formal academic research paper and more accessible and readable for the broad audience of *Nature*.

We detail our revisions point-by-point in response to your further comments as follows.

I am still concerned that the paper suffers from too much hype. I also still have the nagging concern that generative AI is still a relatively new concept and that we don't have enough evidence to draw strong conclusions about its impact. The title is appealing and eye-catching, but I think it goes too far. The paper, in my opinion, is more of an exploratory analysis of research papers in different areas of research that use AI. I know this probably sounds far more toned down than the authors would like.

Response 2.1: We thank the Reviewer for this comment. It is true that generative AI is still a relatively new concept, which may exhibit different features and impacts compared to classic machine learning techniques as its scientific use plays out further. Your framing feels more general than ours, and so we have changed the name of this last period from the LLM to the GenAI (or GAI)-era. Due to GenAI's relatively short history of development, it may be too early to have enough evidence to draw strong conclusions about its impact.

To verify the robustness of our general findings across the decades of AI development to the specific era of LLMs and GenAI, we have replicated our findings with only the subset of papers published during the GenAI era in the last round of revision. Here, we make it clearer that with the limited data currently available, results in the GenAI era only preliminarily show a phenomenon consistent with past eras, but serve as a starting point

for future research. As LLMs and GenAI evolve over longer periods of time, generate richer data, are used in more applications, and those applications exert their impact on science, additional evaluations should be undertaken, further revealing how GenAI impacts scientific development consistent with prior machine learning and deep learning techniques, or exerting new forces and different directions.

We agree that our paper is an exploratory analysis of research papers in different areas of research that use AI. We uncovered that the adoption of AI is associated with expanded impacts to individual scientists, but contracted focus to natural science as a whole. As the Reviewer thoughtfully points out here and later, however, we went too far in concluding that the contracted focus of science represented a universally negative effect (a "peril") against what we expected would be viewed as a positive effect (the "promise") of expanded individual impact. Following your suggestions, we have toned down some of the claims, describing the contracted science focus as an objectively measurable phenomena, rather than using diminutive or subjectively negative expressions like "peril" and "redundant". Please see detailed revisions for this point in responses to later comments.

For your convenience, we quote the revised texts below (and note that "LLM era" has been changed to "Generative AI", "GenAI", or "GAI" era throughout the text and figures).

Revision:

[**Page 2 in the *Main Text, Introduction***] … on papers that augment research in natural science fields by adopting AI, primarily covering decades involving development and deployment of conventional machine learning algorithms and also extending to a necessarily more preliminary analysis of the latest generative AI techniques.

[**Page 2 in the *Main Text, Introduction***] … We separate the periods in which AI was predominantly conventional machine learning, deep learning, and most recently generative designs including large language models. With abundant data-based evidence across decades of conventional machine learning and deep learning, we validate these AI-based measurements and use them to reveal that the adoption of AI leads to an amplifying effect on the career of individual scientists… Nevertheless, this effect corresponds with a contracted focus within collective science… Meanwhile, analyses using currently available data within the latest era of generative AI including large language models reveal a preliminary consistency with prior periods, providing a starting point for further study as generative AI develops over a longer period.

[***Supplementary Notes 4.1-*** Robustness of results in the generative AI era] This separate analysis of AI from the generative AI era could exhibit limitations due to the lack of data caused by the relatively short history of generative AI, including LLMs… Although reaching quantitatively consistent conclusions, with the limited data currently available, results in the generative AI era may appear less significant than corresponding results in our general findings and serve as a starting point for future research as these new models develop and are deployed in new ways. As large, generative models evolve over longer

periods of time and produce richer data, additional evaluations should be pursued, further revealing how generative AI impacts scientific development consistent with traditional machine learning techniques or representing new directions.

I am not convinced that the evidence provided by the authors is enough to conclude that LLMs have indeed transformed scientific writing. Many of the writings I have seen from genAI are not written well. Of course, as these tools advance, they probably will have more of an impact. Reference 23, for example, concludes: Our findings show a sharp increase in the estimated fraction of LLM-modified content in academic writing beginning about five months after the release of ChatGPT, with the fastest growth observed in Computer Science papers." Is that enough justification? This is a preprint that is mentioning something that starts 5 months after introduction.

Response 2.2: Thank you for this question. Currently, there is evidence that LLMs are increasingly used to assist scientific writing, but it is indeed difficult to assess the extent to which they have "transformed" the essence of scientific writing. To improve the presentation, we change to a more rigorous statement: "recent developments in large language models are making them increasingly incorporated in assisting scientific writing". Also, LLM generated contents are not ensured, which may raise concerns about quality in AI-generated content.

Regarding references for the statement, we appreciate your patience in double-checking the listed papers. Actually, the mentioned reference (#23 in the original version) is already published in the Conference on Language Modeling (COLM), a rising top conference focused on language models. We have updated the reference correspondingly. This research is collaboratively conducted by multiple established researchers, and have received nearly 150 citations within one year, which we consider to be good evidence for the increased utilization of LLMs in assisting scientific writing. Also, we add multiple other references published in different domains that discuss LLMs for scientific writing to better support our statement.

For your convenience, we quote the revised texts and references below.

Revision:

[**Page 1 in the *Main Text, Introduction***] … Moreover, recent developments in large language models are making them increasingly incorporated in assisting scientific writing[1–4], facilitating the distillation of scientific findings, but also raising concerns about weakened confidence in AI-generated content.

[1] Ali Salimi and Hady Saheb. "Large language models in ophthalmology scientific writing: ethical considerations blurred lines or not at all?" *American journal of ophthalmology* 254 (2023), pp. 177–181.

[2] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts,

Christopher D Manning, and James Y. Zou. "Mapping the Increasing Use of LLMs in Scientific Papers". *First Conference on Language Modeling*. 2024.
[3] Taesoon Hwang, Nishant Aggarwal, Pir Zarak Khan, Thomas Roberts, Amir Mahmood, Madlen M Griffiths, Nick Parsons, and Saboor Khan. "Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts". *PLoS One* 19.2 (2024), e0297701.
[4] Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. "Delving into LLM-assisted writing in biomedical publications through excess vocabulary". *Science Advances* 11.27 (2025), eadt3813.

The research question is: How does the adoption of AI by individual scientists influence the collective character and dynamic exploration of science as a whole? The phrase "promise and peril" doesn't seem to fit. Further, I don't think that is what you do. Is there a discussion on the promise? Is there a discussion on the perils? Is the conclusion (the paradox you identify) the peril? I scanned for both words and didn't see them throughout. Do you really explore "collective character and dynamic exploration" and, if so, I think you should be explicit about how you do it.

Response 2.3: We thank the Reviewer for this comment. As you suggest, the words "promise and peril" do not precisely fit what we have done. In this manuscript, we discuss the impact of AI on accelerating individual career advancement, which can be regarded as "promise" (especially for those whom it advances), but we did not have analyses and discussions that demonstrate any universal "peril" for science. Following your caution, the paradox we uncovered, namely an expansion of individual scientists' impact but a contraction in collective science's reach, is an objective description of the phenomenon, but should not necessarily be regarded with negative judgement and framed as "peril". Furthermore, we now see that the original words "character and dynamic" we used to articulate our research question were not specific enough.

To summarize what we undertake and accomplish more precisely and rigorously, our research is to explore the impact of AI in scientific research at different scales, spanning the levels of individual scientists' careers and the collective exploration of science. We have revised sentences in the "*Main Text*, *Introduction"* with this new articulation. We have also added sentences at the beginning of the "*Main Text*, *Results-AI adoption is associated with a contraction in knowledge focus among scientific fields"* to indicate the broadening of our analyses from the level of the individual scientist to that of collective science as a whole.

For your convenience, we quote the revised text below.

Revision:
[**Page 1 in the *Abstract***] … In this way, AI adoption in science presents a seeming paradox—an expansion of individual scientists' impact but a contraction in collective science's reach—as AI-augmented work moves collectively toward areas richest in data.

With reduced follow-on scientific engagement, AI appears to automate established fields rather than explore new ones, highlighting a tension between personal advancement and collective scientific progress.

[**Page 1 in the *Main Text, Introduction***] … Here we seek to explore the impact of AI in scientific research at different scales by posing the question: How does the adoption of AI influence individual scientists' careers and the collective exploration of science as a whole?

[**Page 2 in the *Main Text, Introduction***] … Nevertheless, this effect corresponds with a contracted focus within collective science. Measured with "knowledge extent", the "diameter" covered by a sampled batch of papers in vector space, AI-driven science covers less topical ground and is associated with a decrease in follow-on scientific engagement, suggesting that AI is currently more likely to focus on existing popular research problems rather than explore new ones.

[**Page 5 in the *Main Text, Results-***AI adoption is associated with a contraction in knowledge focus among scientific fields] … The accelerating use of AI in science and its impact on individual scientists raises questions about its influences across the entire scientific field. To evaluate how AI collectively impacts the frontiers of scientific exploration, we design a measurement to characterize the breadth of scholarly attention represented by a collection of research papers.

I find that the paper suffers from presentation problems. Even the response document I found a bit overwhelming.

Response 2.4: Thank you for this detailed comments and we regret the overwhelming presentation of the response document. We have polished the corresponding paragraphs following your suggestions. Details and the revised texts are as follow:

a. In my opinion, "we believe" does not belong in academic writing.

   We thank the Reviewer for this suggestion. We have checked through our manuscript, the supplementary information, as well as the response document, and now avoid using this construction.

b. "Hot topics" does not seem very formal for a journal such as *Nature*.

   We thank the Reviewer for this suggestion. We have replaced it with the phrase "popular research topics" to make it more suitable for readers of a formal academic journal.

   Revision:
   [**Page 6 in the *Main Text, Results-***Reduced follow-on engagement and more overlapping works in AI research] … This results in a star-like structure around

specific popular research topics, rather than a network of emergent and interconnected research projects.

[**Page 6 in the *Main Text, Results*-**Reduced follow-on engagement and more overlapping works in AI research] … These findings suggest that AI in science has become more concentrated around popular research topics that become "lonely crowds" with reduced interaction among papers.

[***Supplementary Notes 3.2-*** Selectivity of AI adoption across different topics] To gain a deeper understanding of our main findings that "AI in science has become more concentrated around some popular research topics"...

c. This sentence is awkward: "When the two papers happen to work on similar topics, however, the lack of engagement leads to mutual unawareness and overlap." Further, isn't some overlap good (replicability)?

We thank the Reviewer for this comment. It is true that more overlapping research may have a two-sided implication. On the one hand, some overlap on specific scientific research directions could benefit validation and replication, enhancing the robustness of existing discoveries. On the other hand, as the collective scientific discovery explores a vast and complex landscape, concentrating attention on a shrinking number of specific questions could increase the likelihood that science becomes fixed on local maxima rather than searching in a more broad, decoupled, and diverse way.

We have rephrased this sentence and added corresponding sentences in the *Main Text, Discussion*.

Revision:

[**Page 6 in the *Main Text, Results*-**Reduced follow-on engagement and more overlapping works in AI research] … Taken together, a pair of disengaged papers commonly focus on less related topics and lie farther apart in the embedding space. Occasionally, however, due to the lack of reciprocal engagement, it is possible that mutually-unaware papers lie very close to each other, which indicates more overlapping research.

[**Page 8 in the *Main Text, Discussions***] … It is true that more overlapping attention and a contracted focus may benefit scientific replication and extension, accelerating the emergence of solid and practical solutions to specific questions. Insofar as scientific discovery represents a vast and complex landscape, however, concentrating attention on the same developments may increase the likelihood that science becomes fixed on local maxima of scientific explanation and prediction rather than searching in a more broad, decoupled, and diverse way.

d. In the last sentence just before the discussion, I did not understand what you mean by "knowledge extent" so I needed to look it up. Perhaps just indicate exactly what it means earlier in the paper.

We thank the Reviewer for this suggestion. We now refer to the measurement of "knowledge extent" early in the *Introduction* with a brief description. We also add a more detailed explanation to the definition of "knowledge extent" where it first appears in the *Results*, with reference to the graphical demonstration (*Fig. 3b*) and mathematical formula (*Methods M8*) of "knowledge extent". For your convenience, we quote the revised text below.

Revision:

[**Page 2 in the *Main Text, Introduction***] … Measured with "knowledge extent", the "diameter" covered by a sampled batch of papers in vector space, AI-driven science spans less topical ground and is associated with a decrease in follow-on scientific engagement, suggesting that AI is currently more likely to focus on existing popular research problems rather than explore new ones.

[**Page 6 in the *Main Text, Results-***AI adoption is associated with a contraction in knowledge focus among scientific fields] Within the high-dimensional embedding space, we design the measurement of knowledge extent (KE), which is the "diameter" of vector space covered by a sampled batch of papers, which allows us to compare the coverage of topical ground between AI and non-AI papers in each given domain (Fig. 3b and Methods M8).

e. What is collective hill-climbing? I am not familiar with that phrase.

We thank the Reviewer for this question. In machine learning research, "hill climbing" typically refers to an optimization algorithm that seeks to find the best solution by iteratively making small improvements to a current solution, always moving in the direction that increases the objective function, like climbing uphill to reach the peak. By adding "collective" to the "hill-climbing" metaphor, we reference a situation in which researchers act like a group of climbers all scaling the same popular mountain from the same route. Because both the "path" of a known approach to the "peak" of an anticipated solution offer limited space, this collective rush leads to "crowding" and may discourage the search for other, potentially higher "mountains" representing new questions and answers.

To make this more explicit, we add a short-form of this explanation in a footnote.

Revision:

[**Page 7 in the *Main Text, Discussions***] … Further, with a greater concentration of collective attention to the same AI papers, the adoption of AI appears to induce authors to engage in collective hill-climbing[1], catalyzing solutions to known problems rather than creating new ones.

[1]The metaphor of "collective hill-climbing" describes a situation where researchers act like a group of climbers all scaling the same popular mountain from the same route. Because both the "path" of a known approach and the "peak" of an anticipated solution are constrained, this collective rush leads to "crowding" and may discourage the search for other, potentially higher "mountains" representing new questions and answers.

f. Is it fair to say that mutual unawareness leads to redundant research? Probably. But isn't this also true, to some extent, generally? Researchers, especially if they are new to a topic, might overlook relevant research? Here, I think you are just positioning it as a bigger, and more systematic, problem.

We thank the Reviewer for these questions and proposed solutions. We observed from our analyses and associated statistics that work that builds on AI research tends to engage less with each other. This increases the possibility of more overlapping research that lies very close to each other. As you suggest, however, the phrasing "mutual unawareness leads to redundant research" overstates our findings as a causality claim. Moreover, the word "redundant" oversteps by attaching labeling this development with a negative valence. As discussed above, more overlapping research may have two-sided implications, and it is not appropriate to oversimplify that overlapping research is simply "redundant".

To make it more precise, we have revised the corresponding claims as follows.

Revision:

[**Page 6 in the *Main Text, Results*-**Reduced follow-on engagement and more overlapping works in AI research] … Occasionally, however, due to the lack of reciprocal engagement, it is possible that mutually-unaware papers lie very close to each other, which indicates more overlapping research… These findings suggest that AI in science has become more concentrated around popular research topics that become "lonely crowds" with reduced interaction among papers, linking to more overlapping research and a contraction in knowledge extent and diversity across science.

g. I am curious about your use of the term innovation. I think of it in a much different sense; for example, in terms of creating a novel technology. Is the implication that every research paper is an innovation?

We thank the Reviewer for this question. Upon reflection, use of the term innovation here feels inappropriate. You are right that "innovation" is more like creating a novel technology, which should involve an overlap with research papers, but does not necessarily mean the same thing. Considering that our evidence is drawn from research papers, here we substitute the term "innovation" with "research work", corresponding to the precise content we have analyzed.

Revision:

[**Page 6 in the *Main Text, Results*-**Reduced follow-on engagement and more overlapping works in AI research] … This results in a star-like structure around specific popular research topics, rather than a network of emergent and interconnected research works.

[**Page 6 in the *Main Text, Results*-**Reduced follow-on engagement and more overlapping works in AI research] … This concentration leads to more overlapping research works linked to a contraction in knowledge extent and diversity across science.

[**Page 7 in the *Main Text, Discussions***] … In all fields, AI-augmented research focuses on a narrower scope of scientific topics and reduces the scientific engagement of follow-on research, leading to more overlapping research works that slows the expansion of knowledge.

[**Page 8 in the *Main Text, Discussions***] … Expanding the scope of AI's deployment in science will be required for sustained scientific research and to stimulate new fields rather than merely automate existing ones.

[***Supplementary Notes 1-*** Identifying artificial intelligence in scientific research] To thoroughly investigate how the adoption of AI impacts scientific research… features of identified AI papers along the trend of scientific research over decades.

h. Cognitive extent is another term that was not familiar to me, so I needed to look up the reference. Maybe it would help if you could define it briefly, even if in a footnote. I like, for example, the way you defined paper families.

We thank the Reviewer for this question. "Cognitive extent" is a concept we draw from previous work in the field of Science of Science. As it was not coined by us, we previously did not provide a definition to it, but we agree that it is not sufficiently familiar that an explanation is necessary to make our paper readable for the audience of *Nature* from broad fields beyond the Science of Science.

Per your suggestion, we add a brief explanation in the footnote.

Revision:

[**Page 11 in the *Methods*-**M8. Measure the knowledge extent of papers] To address this issue, we build on prior work about cognitive extent[1].

[1]Cognitive extent is a measure of the breadth of a scientific field's cognitive territory. It is quantified by the number of unique phrases, as a proxy for scientific concepts, found within a sampled batch of papers with given size.

[1] Staša Milojevic. "Quantifying the cognitive extent of science". *Journal of Informetrics* 9.4 (2015), pp. 962–973.

i. The EG formula is not intuitively obvious to me.

We apologize for not explaining the formula clearly. The metric of follow-on engagement (EG) is designed to quantify how frequently citations of the same original paper interact with each other. For an original paper with n citations, if everyone cites all other papers published before their own, there are at most

$$(n-1)+(n-2)+\ldots+2+1=\frac{n(n-1)}{2}$$

citations among these n citing papers. When citations among these n citing papers are relatively sparse, however, such that there are only k citations among them, the follow-on engagement metric is calculated as

$$EG=\frac{k}{\frac{n(n-1)}{2}}=\frac{2k}{n(n-1)}=\frac{2k}{n(n-1)}\times 100\ (\%).$$

In a nutshell, EG is the ratio between the actual (k) and maximum possible number of citations among the n papers citing the same original paper, measuring the degree to which the n papers engage with each other.

To provide more intuitive understanding, we have improved the explanation and detailed the formula in the manuscript, as follows.

Revision:

[**Page 12 in the *Methods*-**M10. Measure follow-on engagement among papers] To quantify how frequently citations of the same original paper interact with each other, we design a metric called follow-on engagement building on prior work[1]. For an original paper with n citations, there are at most $\frac{n(n-1)}{2}$ possible citations among these n citing papers if everyone cites all papers published earlier than their own. We then count how many times these n citing papers actually cite one another, denoted as k. Our metric for follow-on engagement is calculated as the ratio of actual to maximum possible citations:

$$EG=\frac{k}{\frac{n(n-1)}{2}}=\frac{2k}{n(n-1)}=\frac{2k}{n(n-1)}\times 100\ (\%).$$

[1] Peter McMahan and James Evans. "Ambiguity and engagement". *American Journal of Sociology* 124.3 (2018), pp. 860–912.

j. Please check the KE equation.

We thank the Reviewer for pointing out this. We apologize for the notation misalignment between the KE equations in Fig. 3b and Method 8. The knowledge extent is calculated as the maximum Euclidean distance from each paper embedding point to the centroid. We now update the notation in *Method 8*,

uniformly using the letter p plus the index to represent the embedding point of each paper, and using the letter c to represent the centroid of a group of papers.

Futhermore, we add a sensitivity check to the distance metric. We substitute all Euclidean distance into Cosine distance and replicate our analyses. As we show in *Supplementary Fig. 37*, all results with Cosine distance are consistent with the original ones with Euclidean distance in the main text. These results demonstrate that our distance calculations based on paper embeddings are robust, ensuring the reliability of our distance-based analyses.

Revision:

[**Page 11 in the *Methods*-**M8. Measure the knowledge extent of papers] To assess the knowledge extent of a set of research papers within their high-dimensional embeddings

$$\{p[1],\ p[2],\ \ldots,\ p[n]\},\ p[i] \in R^{768},$$

we first compute the centroid as the mean of their vector locations:

$$c = \frac{1}{n}\sum_{i=1}^{n} p[i].$$

Next, we compute the Euclidean distance from each embedding to the centroid, where the knowledge extent of the set of papers is defined as the maximum distance or "diameter" of the vector space covered:

$$KE = max_{1\le i\le n}||p[i] - c||_2.$$

[***Supplementary Notes 4.3*-**Robustness of distance calculations in the high-dimensional paper embedding space] …we also check the robustness regarding different distance measurements. We substitute all Euclidean distance into Cosine distance and replicate our analyses. As we show in Supplementary Fig. 37, all results with Cosine distance are consistent with the original ones with Euclidean distance in the main text.

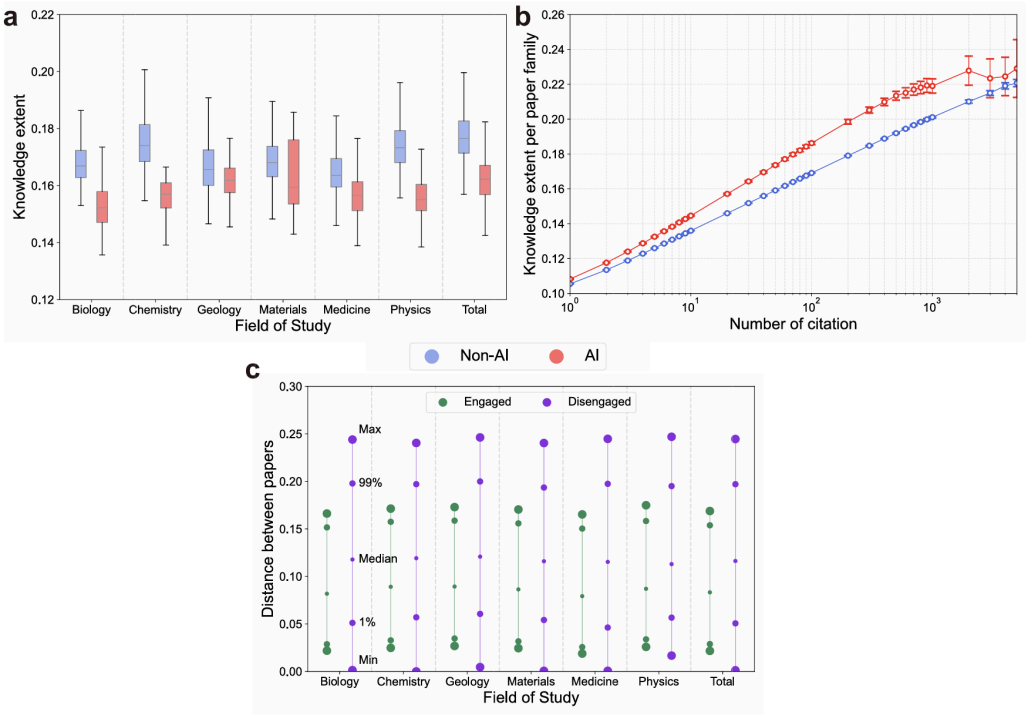


Figure S37. Replication of results related to high-dimensional distance calculations with Cosine distance. (a) Knowledge extent of AI and non-AI papers in each field. We show 1.5 times the inter-quartile range (IQR) from the box as whiskers. (b) Knowledge extent of individual AI and non-AI paper families. We show 99% CIs as error bars. (c) The distribution of distances between paper pairs that cite the same papers, with or without citing each other—engaged versus disengaged. Results with Cosine distance are consistent with those with Euclidean distance reported in the main text.

k. Also, you might want to define field problems.

We thank the Reviewer for this point. Actually, by "AI is more likely to solve critical field problems than catalyze new ones", we intended to express that AI is more likely to focus on existing popular problems across fields rather than create new problems, but not to define some specific problems for AI to solve.

To avoid misunderstanding, we have modified the expression, as follows:

Revision:

[**Page 2 in the *Main Text, Introduction***] … AI-driven science covers less topical ground and is associated with a decrease in follow-on scientific engagement, suggesting that AI is currently more likely to focus on existing popular research problems rather than explore new ones.

Finally, some more discussion of the implications might be helpful. Isn't it good to expand the impact? Is there a problem with less focus?

Response 2.5: We thank the Reviewer for this suggestion. It is indeed good for individual researchers to expand the academic impact to their worthy, AI-enhanced work. This can accelerate their personal career development and this benefit may be a driving force behind the accelerated rate of AI adoption. Regarding the contracted focus, we consider it as having a two-sided implication. For those fields garnering AI-driven attention, more overlapping focus may benefit scientific validation and replication, accelerating the emergence of solid, practical solutions. For scientific research as a whole, insofar as natural understanding represents a complex and rugged landscape, concentrating attention on specific questions may exacerbate the likelihood that science becomes fixed on local maxima rather than searching in a more broad, decoupled, and diverse way. We have added corresponding discussion as you suggest.

For your convenience, we quote the revised texts below.

Revision:

[**Page 7 in the *Main Text, Discussions***] … We find that individual scientists are increasingly rewarded with expanded academic impact and accelerated career development for incorporating AI assistance in research across these waves and in all natural science research fields we studied… This substantial academic benefit may be a driving force behind the accelerated rate of AI adoption.

[**Page 8 in the *Main Text, Discussions***] … It is true that more overlapping attention and a contracted focus may benefit scientific replication and extension, accelerating the emergence of solid and practical solutions to specific questions. Insofar as scientific discovery represents a vast and complex landscape, however, concentrating attention on the same developments may increase the likelihood that science becomes fixed on local maxima of scientific explanation and prediction rather than searching in a more broad, decoupled, and diverse way.

We sincerely appreciate the Reviewer for sharing your valuable time and effort through engaging with us and our manuscript. We have carefully revised our manuscript in this revision cycle thanks, in large part, to this helpful review.

Response to Referee#3's Comments

I would like to thank the authors for the extraordinary amount of work that they have put into this revision. I feel that the manuscript is now stronger. I still have some questions and suggestions that I believe could strengthen the manuscript further.

Response: We thank the Reviewer for this positive and constructive feedback on our revision. Most importantly, we are glad to see suggestions on the technical details, which are helping us refine and improve the manuscript. Here, we provide a brief summary of our main revisions below.

a. We have added a detailed description of the author name disambiguation methodology and clarified the resulting author sample size.
b. We have added a sensitivity analysis to validate the reliability of our distance calculations for high-dimensional vectors. This analysis supports our subsequent distance-based measurements, such as knowledge extent.
c. We have clarified the details for using two different metrics, Fleiss' Kappa and the F1-Score, in our identification accuracy evaluation.
d. Following suggestions, we have polished our figures from the aspect of plot types, statistical choices, axis scales, etc., enhancing their clarity and rigor.

We detail our revisions point-by-point in response to the Reviewer's further comments as follows.

Figure 1

Line graphs would be more appropriate for panels b and f. Otherwise, best practice is to show bar graphs starting at zero.

While I agree that adding exponential fit lines to the data in panels d and e is unnecessary (but thanks for plots in SI), I still think that using a log-scale on y-axis would make the figures easier to read.

Response 3.1: We thank the Reviewer for these suggestions on improving the format of our figures. We have now changed panels b and f into line graphs to make them more attuned to our results. We also shift to using a log-scale on the *y*-axis in panels d and e. This better utilizes the space and makes the figures easier to read.

For your convenience, we paste the revised figures below.

Revision:

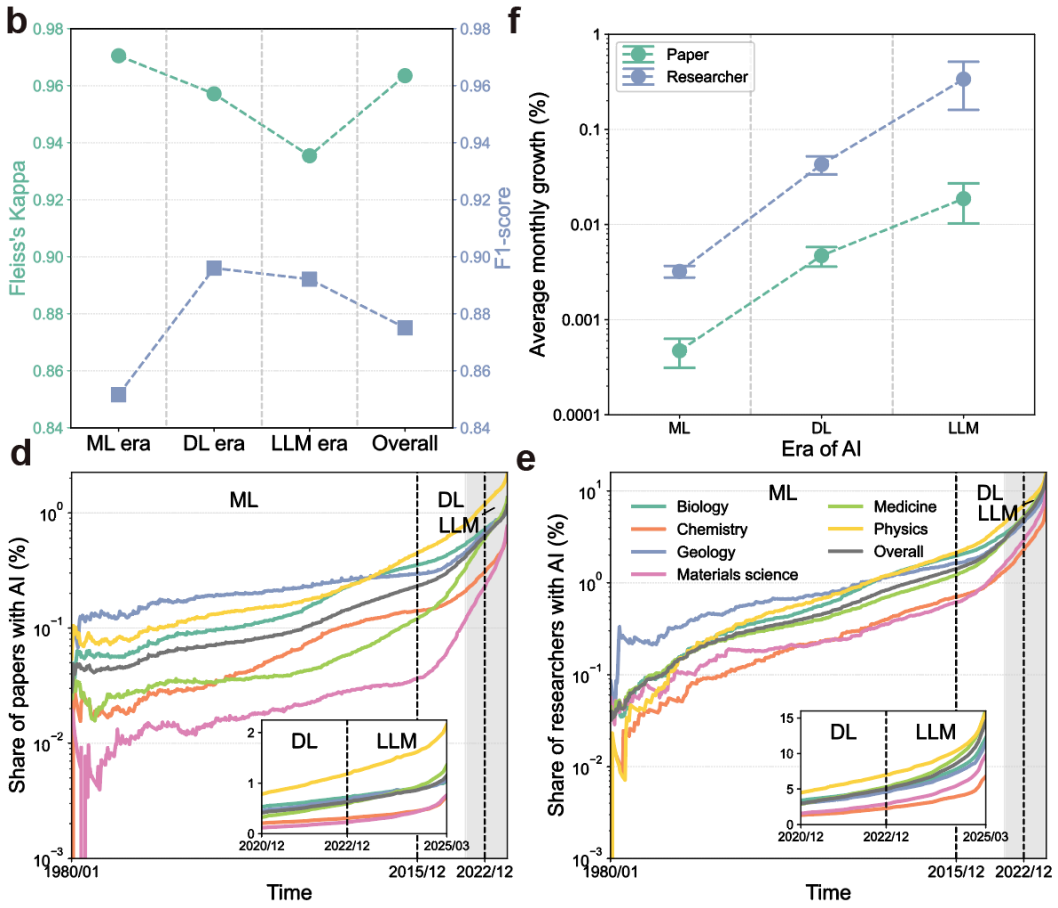


Figure 1. Increasing prevalence of AI adoption in science. (b) Accuracy evaluation of our identification results by human experts. For samples spanning three eras of AI, experts reached consensus with Fleiss' Kappa (κ) $\geq$ 0.93. Our model identification results have strong accuracy in validation against expert-labeled data with an F1-score $\geq$ 0.85. (d)-(e) The growth of AI-augmented papers (d, n = 41,298,433) and AI-adopted researchers (e, n =5,377,346) across the eras of ML, DL, and GLLM between 1980 and 2025 in selected scientific disciplines. The y-axes are set to log-scale. (f) MThe monthly growth rates for AI papers and researchers across the eras of ML, DL, and GLLM across all selected disciplines, where 99% CIs are shown as error bars.

In panel b, it is not clear to me why the authors use Fleiss' Kappa for comparison of expert classification but not for comparison of model to ground truth. Using Fleiss' Kappa for both would obliviate the need for two different y-axis and enable oranges to oranges comparison (average F1 score could still be given in text).

Response 3.2: We thank the Reviewer for this question. Here, we use different measurements in the two steps of our identification accuracy evaluation because of their distinct attributes.

In the first step, we evaluate the inter-rater reliability among our human experts. For each sample, we collected labels from three independent experts. In this context, the experts' labels are treated as equivalent; no single label is considered ground truth over others. Our objective is to measure the level of consensus among them. This is an unsupervised task, for which Fleiss' Kappa is the standard metric. Fleiss' Kappa (κ) is a statistical

measure of inter-rater reliability that generalizes Scott's π from two to any number of raters, given categorical ratings and a fixed number of items. Fleiss κ can be interpreted as articulating the extent to which the observed agreement among raters exceeds what would be expected if all raters made their ratings at random. A high Fleiss' κ score confirms that expert-generated labels are consistent and reliable.

In the second step, we evaluate the performance of our BERT model against expert labels. This relationship is asymmetrical: expert labels serve as the ground truth, and the model's predictions are what we seek to validate. This is a supervised task. We therefore employ the F1-Score, which is a standard metric for assessing the accuracy of a model in a supervised setting. A high F1-Score demonstrates that our BERT model is reliable for our subsequent large-scale classification and analysis.

To make the procedure and reason for using corresponding measurements more clear, we made the following revisions to our manuscript:

a. We revise the *Main Texts*, *Results-Increasing prevalence of AI adoption in science* in the *Main Text*, *Methods-M4. Scrutinize our identification results by disciplinary experts*, specifying the unsupervised and supervised settings for using different measurements.
b. We revise *Extended Data Fig.2*, illustrating the different attributes required by these two steps of measurement.

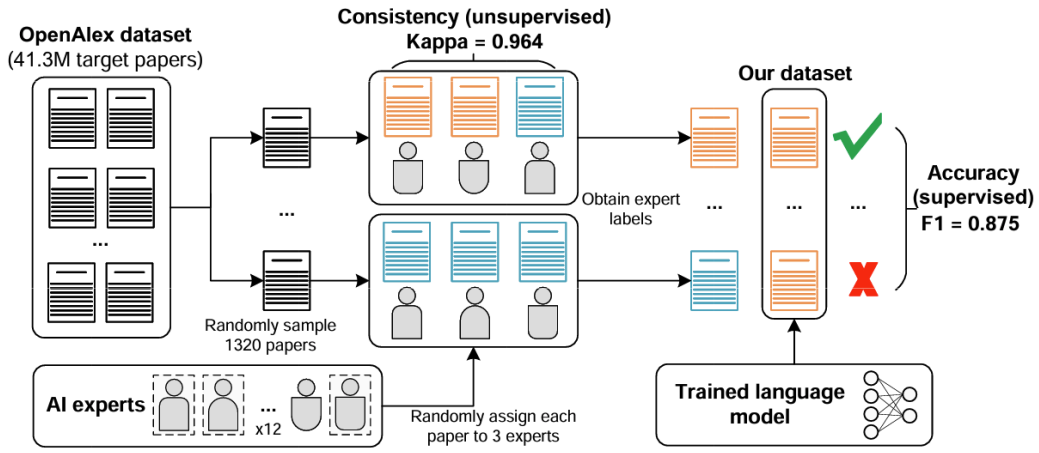
For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 2 in the *Main Text, Results-***Increasing prevalence of AI adoption in science] To evaluate the accuracy of our identification, we recruited a team of human experts to validate these results (Methods M4 and Extended Data Fig. 2). The experts demonstrate strong consensus across their independent annotation of papers sampled at random from the six disciplines mentioned above, achieving an average Fleiss' Kappa (κ) of 0.964[1-2]. The BERT model attains an F1-score of 0.875 in an evaluation that uses the expert labels as ground truth.

[**Page 9 in the *Methods-***M4. Scrutinize our identification results by disciplinary experts] … In this way, each paper is repeatedly labeled by three distinct experts, and we evaluate the consistency among these experts based on Fleiss' Kappa coefficient (κ)[1-2], which is an unsupervised measurement for assessing the agreement between independent raters. Having confirmed consensus among our experts, we draw the final expert label of each paper from the three experts according to the principle of the minority obeying the majority. We regard the expert labels as ground truth and validate the result of our BERT model against it with the F1-Score, which is a supervised accuracy measurement of accuracy.

[1] Joseph L Fleiss. "Measuring nominal scale agreement among many raters." *Psychological bulletin* 76.5 (1971), p. 378.
[2] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". *Biometrics* (1977), pp. 159–174.



Extended Data Fig.2. Procedure of accuracy evaluation via expert evaluation. We randomly sample 1320 papers and delegate three experts to scrutinize the identification results for each paper. We then draw the final expert label of each paper from the three experts according to the principle of the minority obeying the majority and validate the result of the language model with it. Results indicate strong consistency among experts and high accuracy with our identification results.

Figure 2

The caption and text quote over 2 million researchers for whom the authors could track their careers. This seems like an awful lot of people for one to be able to track accurately.

US universities hire one faculty per department per year. Let's say that gives us about 20 new hires per university per year. Times 50 years, that is about 1000 new scientists. Times 100 research universities in the US that are maybe 100,000 established scientists. But the analysis is looking only at 6 disciplines. So, those numbers seem very excessive to me.

I do not know that current approaches to author disambiguation are that robust. Which brings me to my other concern: the authors also quote a 45% rate of transition to established scientists for the junior scientists who adopted AI methods. This rate seems extraordinarily high.

Response 3.3: We thank the Reviewer for this detailed observation and comment.

First, regarding the technical aspect of author identification, tracking over 2 million authors is feasible given the structure and organization of the data. We rely on OpenAlex's Author Name Disambiguation (AND) system, which utilizes an XGBoost model to distinguish authors based on their institutions, co-authors, citations, ORCIDs, and other features, assigning a unique ID to each author. We do not need to perform the author disambiguation ourselves; we track a large number of authors simply by using OpenAlex's author IDs. Other studies have also conducted analyses involving a comparable or even larger number of authors. For example, a recent study on the "pivot penalty" published in *Nature* involved up to 8.32 million author observations[2].

Second, it seems that we have a misalignment on the definition of "established researcher". Following the definition used in prior studies[1], a researcher who has served as last author on at least one paper is regarded as "established". Therefore, our cohort includes far more than just faculty members working as principal investigators (PIs) at top 100 research universities in the U.S.; it also encompasses researchers with many other diverse roles. We consider that our original phrasing, "lead a research team" might have caused misunderstanding, so we have revised it to "lead a research project", which better aligns with our operational definition of having served as last author on at least one paper.

In fact, our study involves researchers from 56,497 institutions globally. Among them, 130,972 individuals are from the top 100 U.S. universities, of whom 51,443 are established researchers. This figure is largely consistent with your estimation considering that we only focus on six disciplines. The average number of researchers in other institutions is smaller than in the top 100 U.S. universities. For instance, all 1,393 universities in the U.S. account for a total of 298,321 researchers in our dataset, while all 14,837 U.S. institutions combined account for 536,488. Based on these statistics, our total number of global researchers seems reasonable.

Furthermore, based on the exact definition we used, the bar is not so high to become an "established researcher". Many PhD students (junior researchers) in the "AI for Science" field may take on roles such as postdocs or engineers in educational, commercial, or governmental institutions after graduation. In these positions, it is entirely possible for them to lead a research project and thus become a researcher who leads a research project. Therefore, reaching a 45% transition rate to established researchers becomes plausible.

According to your suggestion and above discussion, we made the following revisions to our manuscript:

a. We revised the definition of "established researcher" in the *Main Texts Results-AI enlarges the impact of individual papers and enhances the career of individual scientists* and *Methods-M5. Determine the project leader of papers* to make it more precise and clear.
b. We add a brief introduction in *Methods-M6. Detect scientists' career role transition* on how we can track a large number of authors simply by using unique author IDs from OpenAlex's Author Name Disambiguation (AND) system.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 4 in the *Main Text, Results-*** AI enlarges the impact of individual papers and enhances the career of individual scientists] To identify the implications of AI adoption on scientist's career development, we classify the scientists into "junior" and "established", where junior scientists are defined as newcomers who have not yet led a research project, whereas established scientists refer to those who have led one or more research projects (Methods M5 and Extended Data Fig. 5).

[**Page 9 in the *Methods*-**M5. Determine the project leader of papers] Here, we define the project leader as the last author of a research paper, in alignment with conventions established by previous studies[1]. To ensure that in most papers, the last authors represent the project leader, we examine the fraction of papers that list authors following alphabetical order.

[**Page 10 in the *Methods*-**M6. Detect scientists' career role transition] The OpenAlex dataset incorporates a well-designed Author Name Disambiguation (AND) mechanism, which utilizes an XGBoost model to predict the likelihood that two authors are the same based on features like their institutions, co-authors, and citations, and then applies a custom, ORCID-anchored clustering process to group their works, assigning a unique ID for each author. Simply utilizing unique IDs, we are able to track a large number of authors at the same time[2], where we depict an individual scientist's career trajectory using a role transition model (Extended Data Fig.4a) and extract the role transition trajectories for scientists.

[1] Vedran Sekara, Pierre Deville, Sebastian E Ahnert, Albert-László Barabási, Roberta Sinatra, and Sune Lehmann. "The chaperone effect in scientific publishing". *Proceedings of the National Academy of Sciences* 115.50 (2018), pp. 12603 12607.
[2] RyanHill, Yian Yin, Carolyn Stein, Xizhao Wang, Dashun Wang, and Benjamin F Jones. "The pivot penalty in research". *Nature* (2025), pp. 1–8.

I thank the authors for all their new analysis in response to my concerns relating to panel a. I believe however, that the current choice for panel a artificially exacerbates the distinction between AI and no-AI papers. Indeed, Figs S8b-d suggest that the difference is not so marked. Only when you look at S8a do you see a very large difference. As you demonstrate in Fig 4 with the calculated Gini coefficients, most of the difference may be due to the very top of the citation list.

To avoid this exaggeration, I strongly recommend that you use the median instead of the mean in Figs 2a and 2b.

Response 3.4: Thank you for this comment. It is true that the distribution of citations is right-skewed, and average citation is more impacted by a small number of highly-cited outlier papers than by the large number of papers with normal citation counts. As you note, our current results—including the high Gini coefficients of citation in Fig. 4c and the marked difference in top 1% of annual citations in Fig. S8a—all provide evidence for this.

To avoid the exaggeration caused by top-cited papers, using the median instead of the mean might represent a safe choice. As we try this, however, it does not yield ideal results.

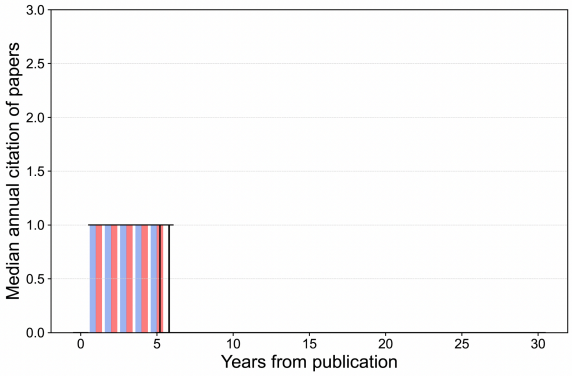


Figure Discussion 1. Median annual citations after publication of AI and non-AI papers.

Upon analysis, we found that this is because most papers receive only very few annual citations, a condition under which the median value loses resolution. For example, in the year of publication, 73.36% of Non-AI papers and 63.95% of AI papers received zero annual citations, causing the median for both groups to be 0. In the year following publication, papers with one annual citation spanned the 44.38%–67.26% percentile in the Non-AI cohort and the 32.41%–57.50% percentile in the AI cohort, causing the median for both to be 1. In such cases, substantial differences exist in the overall annual citation patterns of Non-AI and AI papers (a difference of 5.63×10^8 total citations), but the median value fails to illustrate this distinction.

From another perspective, the proportion of highly cited papers is another important indicator of a domain's impact, which should be considered when comparing the impact of AI and Non-AI research. Therefore, we consider the mean value is still a useful statistic. Nevertheless, we agree that mean value alone is not enough for fully illustrating the difference between the citations of AI and Non-AI research. Considering the limited resolution of median value discussed above, we replicate the results in Figs 2a and 2b with various percentile statistics. As the results show, different percentile statistics consistently indicate that AI papers and researchers attract more citations. Meanwhile, as we analyzed above, lower percentile statistics may lose resolution and only result in 0 and 1 at some points. We have included the extended percentile statistics in *Supplementary Fig. S9*.

For your convenience, we paste the added figures below.

Revision:

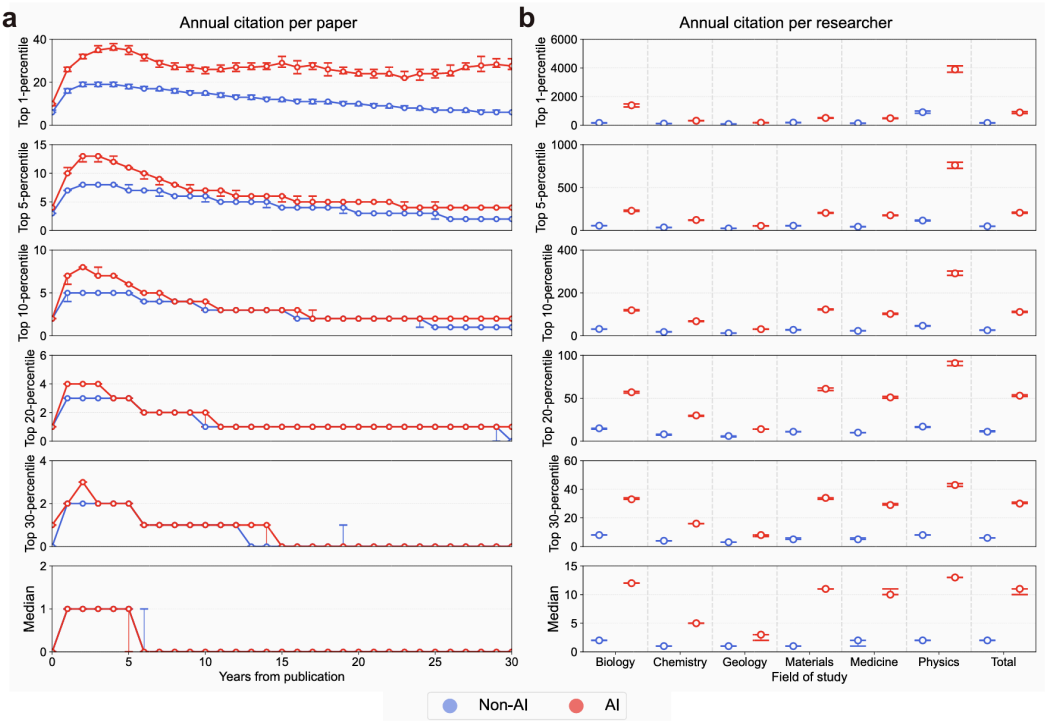


Figure S9. Different percentile statistics for the annual citation comparison between AI and non-AI papers. (a) Annual citations after publication of AI and non-AI papers. (b) Annual citations for researchers adopting AI and their counterparts without AI. Consistently indicated by different percentile statistics, AI papers and researchers attract more citations. For all panels, 99% CIs are shown as error bars.

Concerning panel c, I wonder whether using separate plots for out of academia and becoming established researchers would avoid some confusion. The point of those plots is to show significance of difference of the ratio to 1, but with the two data points per discipline one feels tempted to look for statistical significance of the difference between the two.

For panel d and SI figures, I would plot the survival function as that smooths the data without the need for binning.

Response 3.5: We thank the Reviewer for this helpful comment. For panel c, you correctly identify how a key finding involves AI researchers having a lower probability of leaving academia and a higher probability of becoming established researchers compared to their non-AI peers. To better emphasize this point, we revise the figure to present two separate plots as you suggest. This better shows how the ratio for each cohort (AI vs. non-AI) deviates from 1, rather than leading the reader's attention to a comparison between the two cohorts.

For panel d and similar results in SI figures, we appreciate the suggestion of graphing the survival function. For the raw probability distribution of

$$P(t) = \lambda e^{-\lambda t}, \; t > 0,$$

the corresponding survival function is

$$S(t) = 1 - \int_0^t P(u)du = e^{-\lambda t},\ t > 0.$$

As you point out, this transformation helps to smooth the data. We also found that our exponential fitting program was occasionally sensitive and inaccurate when fitting the original $P(t)$ formulation. In contrast, the simpler form of $S(t)$ is more robust and now ensures consistently accurate fitting results.

We have added the derivation of survival function into *Methods* and updated all figures related to the above points in the *Main Text*, *Extended Data Figures*, and *Supplementary Figures*. For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 10 in the *Methods*-**M7. Estimate the birth-death model for career development of junior scientists] … Then, the transition time for junior scientists to become established scientists or leave academia follows the exponential distribution:

$$P(t) = \lambda e^{-\lambda t},\ t > 0,$$

and the corresponding survival function is

$$S(t) = 1 - \int_0^t P(u)du = e^{-\lambda t},\ t > 0.$$

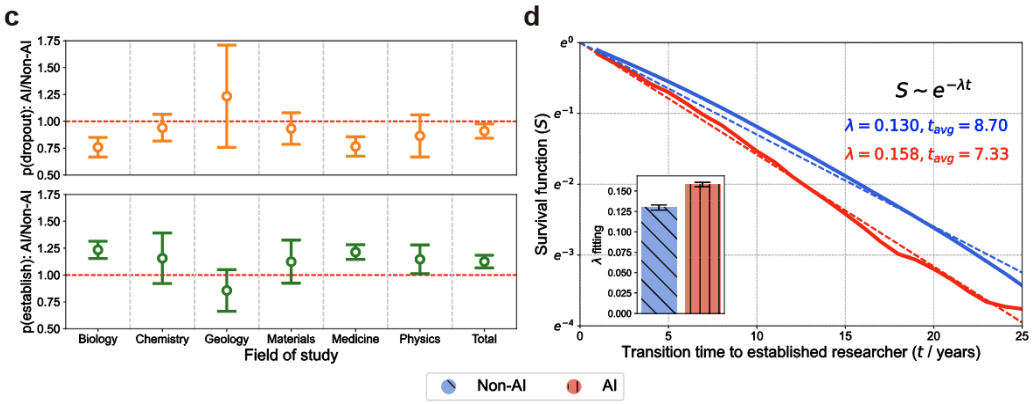
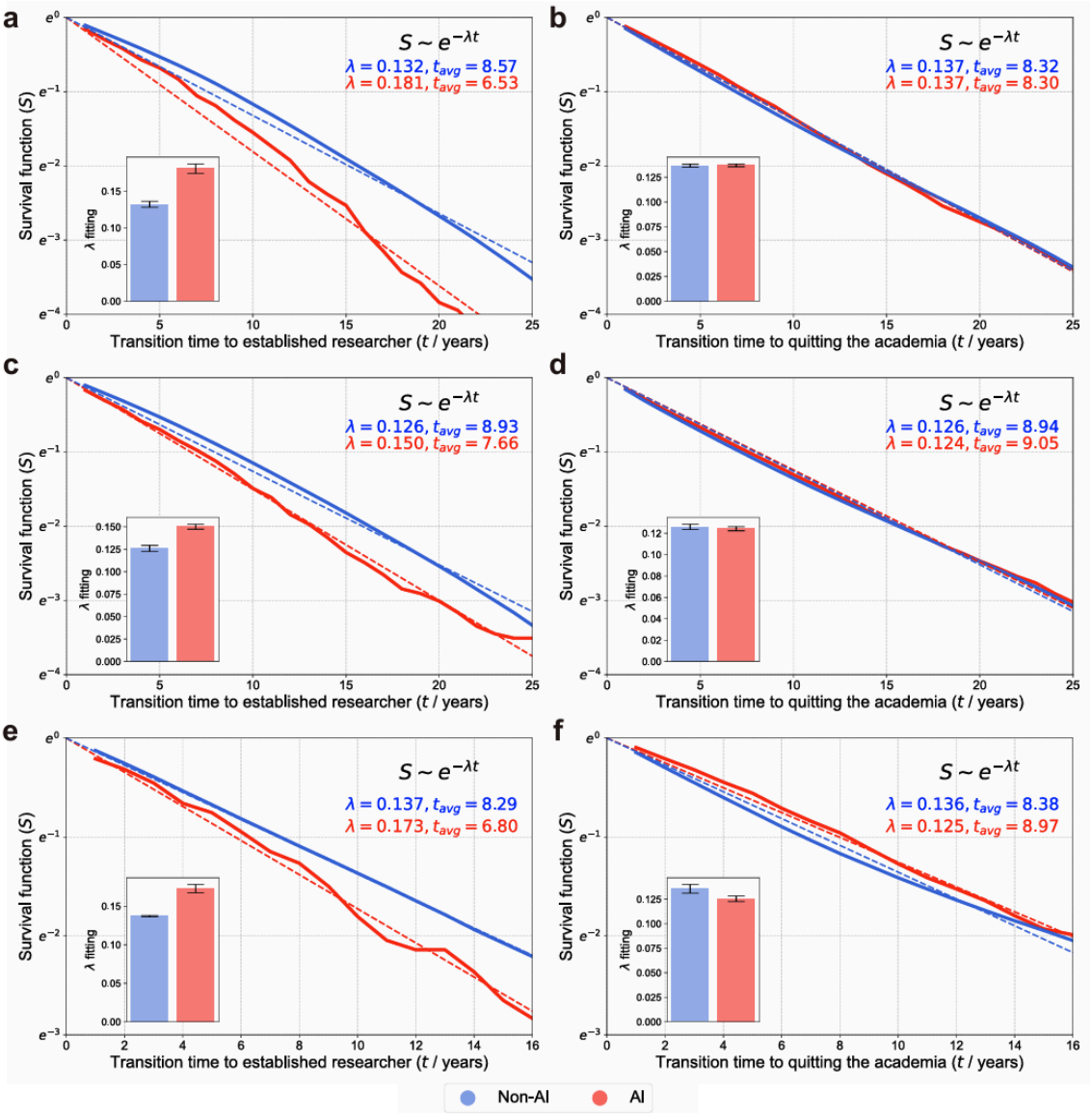


Figure 2. AI enlarges paper impact and enhances researcher careers. (c) The probability of two role transitions between junior scientists adopting AI and their counterparts without AI ($n = 6$ disciplines $\times$ 46 year observations). Junior scientists adopting AI have a higher probability of becoming established researchers and a lower probability of quitting academia compared with their counterparts without AI. (d) Survival function for the transition from junior to established researcher ($n = 2,282,029$). The survival function can be well-fit with exponential distributions, where junior scientists adopting AI become established earlier than their counterparts without AI.



Extended Data Fig. 7. Model fitting the role transition time of junior scientists. (a) (c) (e) Survival functions for the transition from junior to established researcher in biology, medicine, and physics. (b) (d) (f) Survival functions for the transition from junior researcher to leave academia in biology, medicine, and physics. All survival functions can be well-fit with exponential distributions, where the expected time for junior scientists to become established is shorter for those who adopt AI, while the expected time for junior scientists to abandon academia is similar or slightly longer for those who adopt AI. Results indicate that AI not only provides junior scientists opportunities to become established scientists at a younger age, but also reduces the risk of their exiting academia early. For all panels, 99% CIs are shown as error bars.

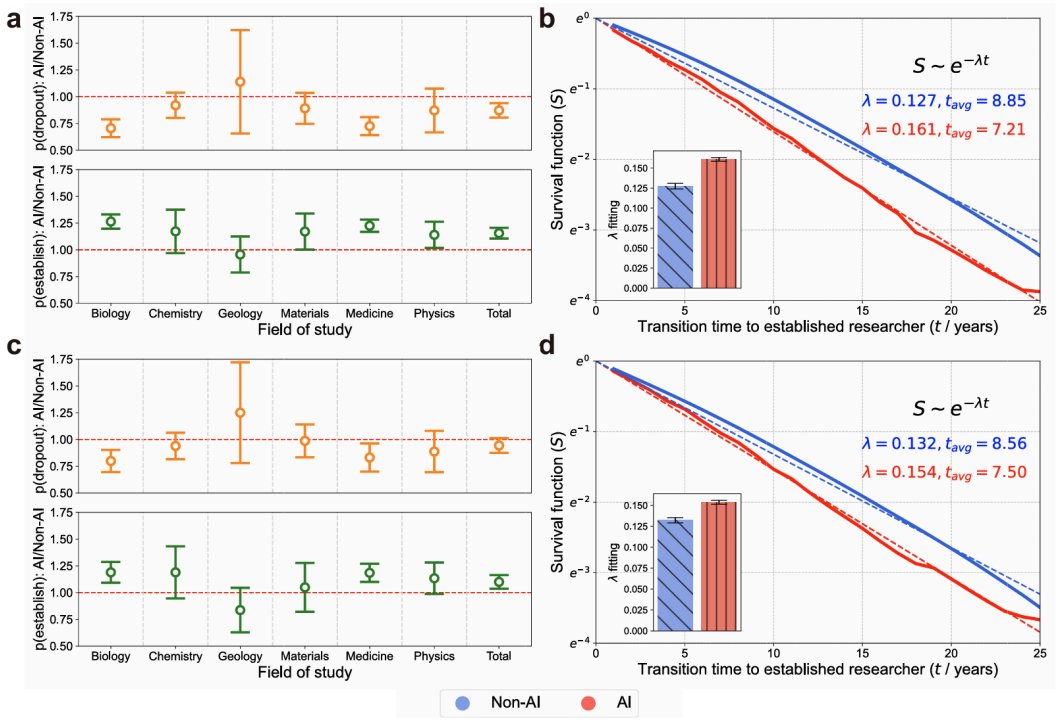


Figure S19. Impact of AI adoption on researchers' careers with different thresholds in detecting the dropout of researchers. (a) (c) The probability of two role transitions between junior scientists adopting AI and their counterparts without AI. (b) (d) Survival functions for the transition from junior to established researcher. In (a) and (b), the threshold for detecting researcher dropout is 2 years. In (c) and (d), the threshold for detecting researcher dropout is 4 years. Results are consistent for both shorter and longer thresholds. For all panels, 99% CIs are shown as error bars.

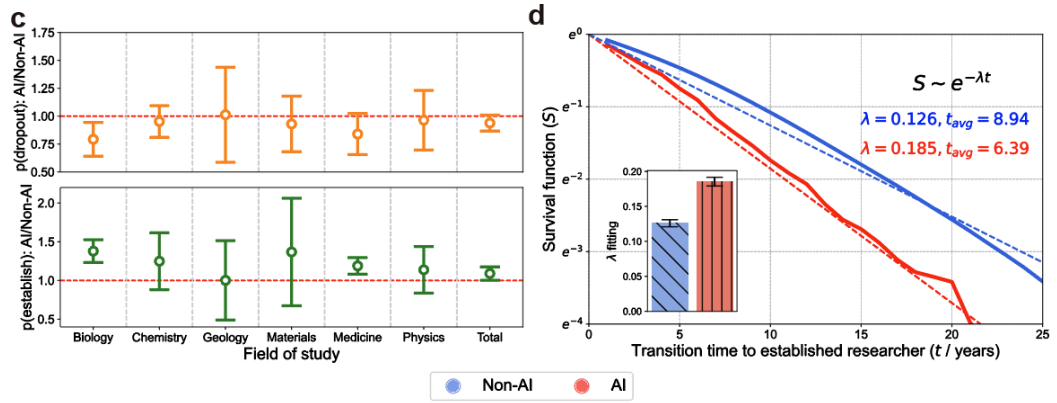


Figure S31. Replication of "AI enlarges the impact of papers and enhances the career of researchers" on the WOS dataset. (c) The probability of two role transitions between junior scientists adopting AI and their counterparts without AI. (d) Survival functions for the transition from junior to established researcher. For all panels, 99% CIs are shown as error bars. Results on the WOS dataset are consistent with those from the OpenAlex dataset featured in the main text.

Figures 3 and 4

I thank the authors for the clarification. I now see that t-SNE projection was only used for visualization.

Nonetheless, the authors must be aware of the issues with the reliability of calculation of distances for high-dimensionality vectors. As a sanity check (and sensitivity analysis), I would recommend using a PCA approach and identifying the number of significant PCs for the data and then calculating distances within that embedding.

Such a check is particularly critical because knowledge extent is the basis for the very important analyses presented in Figs 3 and 4.

Response 3.6: We thank the Reviewer for this suggestion. It is crucial to check the reliability of our distance calculation for high-dimensional vectors. Following your suggestion, we utilize a PCA approach. Specifically, we first identify the number of principal components required to explain 90% of the total variance in the original 768-dimensional embeddings, which turns out to be 135. We then project the original embeddings onto these principal components and recalculate the distances in this reduced-dimensional space to replicate our analysis. As we show in *Supplementary Fig. S34-S35*, all results in the reduced-dimensional space are consistent with those from the original embedding space reported in the main text.

In addition to replicating our analysis in reduced-dimensional space, we further test the sensitivity of our distance calculations from a more general perspective. Using the same PCA method, we reduce the original embeddings to 86, 135, and 197 dimensions, which correspond to principal components that explain 80%, 90%, and 95% of the total variance, respectively. We then randomly sample 1,000 NonAI-NonAI, 1,000 AI-AI, and 1,000 NonAI-AI paper pairs and calculate the distance for each pair within each of the reduced-dimensional spaces. As we now show in *Supplementary Fig. S36*, distances in the various reduced-dimensional spaces are highly correlated with those from the original space, with Pearson's r greater than 0.95 ($p < 0.001$ for all groups).

Besides the sensitivity to embedding dimensions, we also check the robustness of distance metric. We substitute all Euclidean distance into Cosine distance and replicate our analyses. As we show in *Supplementary Fig. 37*, all results with Cosine distance are consistent with the original ones with Euclidean distance reported in the main text. Taken together, these results demonstrate that our distance calculations based on paper embeddings are robust, ensuring the reliability of our distance-based analyses and suggesting the potential for future research to leverage these distance metrics for further analyses.

For your convenience, we quote the revised texts and figures below.

Revision:

[***Supplementary Notes 4.3-***Robustness of distance calculations in the high-dimensional paper embedding space] Because multiple important analyses in the main text are presented based on the distance calculated in the paper embedding space, it is crucial to ensure reliability of our distance calculations for high-dimensional vectors. To evaluate this, we utilize a principal components analysis (PCA) approach. We first identified the number of principal components required to explain 90% of the total variance in the

original 768-dimensional embeddings, which we determined to be 135. We then projected the original embeddings onto these principal components and recalculated distances in this reduced-dimensional space to replicate our analyses. As we show in Supplementary Fig. S34-S35, all results in the reduced-dimensional space are consistent with those from the original embedding space reported in the main text.

In addition to replicating our analysis in the reduced-dimensional space, we further tested the sensitivity of our distance calculations. Using the same PCA method, we reduce the original embeddings to 86, 135, and 197 dimensions, which correspond to the principal components that explain 80%, 90%, and 95% of the total variance, respectively. We then randomly sample 1,000 NonAI-NonAI, 1,000 AI-AI, and 1,000 NonAI-AI paper pairs and calculate the distance for each pair within each of the reduced-dimensional spaces. As we show in Supplementary Fig. S36, distances in the various reduced-dimensional spaces are highly correlated with those in the original space, with Pearson's r greater than 0.95 ($p < 0.001$ for all groups).

Besides the sensitivity to embedding dimensions, we also check the robustness regarding different distance measurements. We substitute all Euclidean distance into Cosine distance and replicate our analyses. As we show in Supplementary Fig. 37, all results with Cosine distance are consistent with the original ones with Euclidean distance in the main text. These results demonstrate the robustness of our distance calculations based on paper embeddings, ensuring the reliability of our distance-based analyses and suggesting the potential for future research to leverage these distance metrics for further analyses.

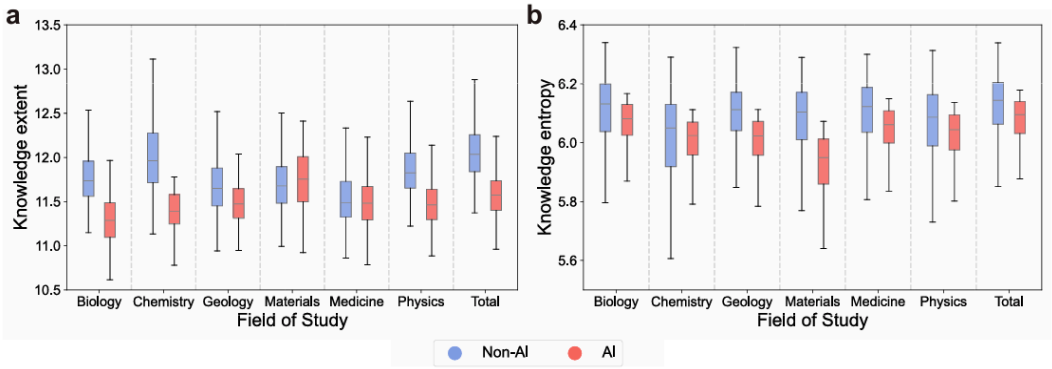


Figure S34. Replication of "AI usage is associated with a contraction in knowledge extent within and across scientific fields" in an embedding space after dimensional reduction using PCA. (a) Knowledge extent of AI and non-AI papers in each field. (b) Knowledge entropy of AI and non-AI papers in each field. We show 1.5 times the inter-quartile range (IQR) from the box as whiskers. Results in the reduced-dimensional space are consistent with those from the original embedding space reported in the main text.

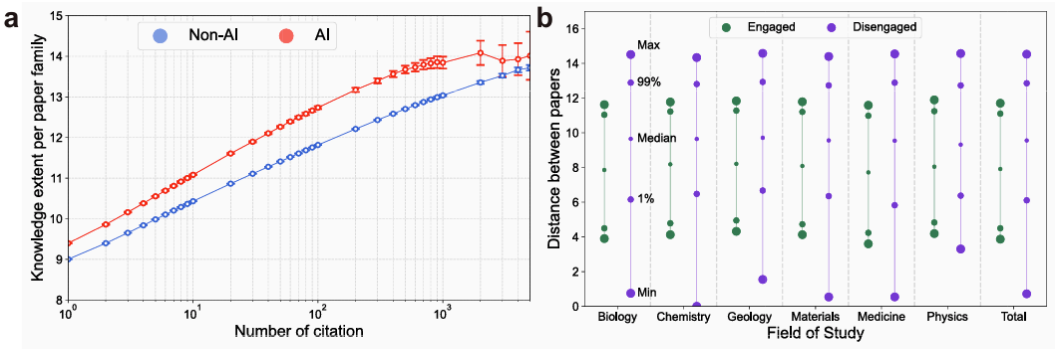


Figure S35. Replication of "reduced follow-on engagement and more overlapping works in AI research" in an embedding space after dimensional reduction using PCA. (a) Knowledge extent of individual AI and non-AI paper families. (b) The distribution of distances between paper pairs that cite the same papers, with or without citing each other—engaged versus disengaged. Results in the reduced-dimensional space are consistent with those from the original embedding space reported in the main text.

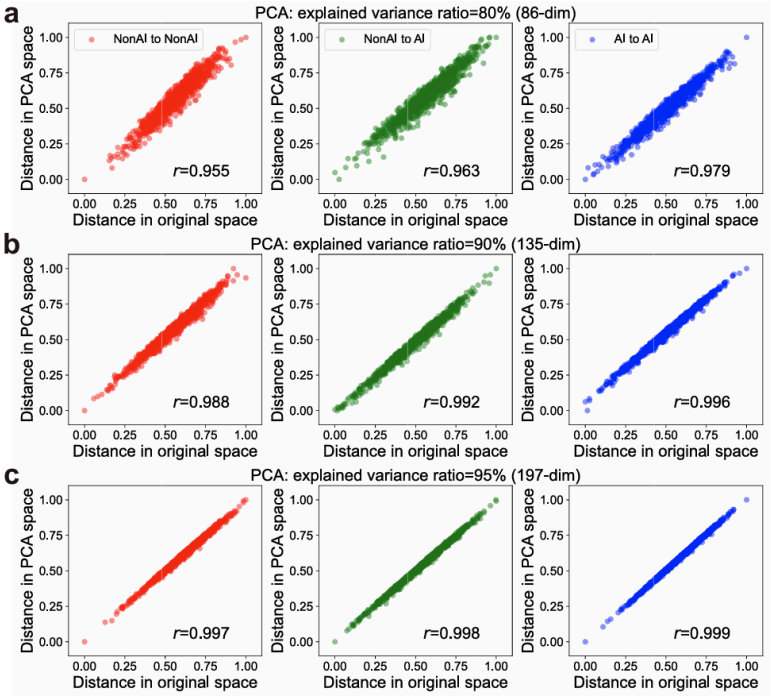


Figure S36. General sensitivity check on the distance calculations in high-dimensional paper embedding space. We reduce the original embeddings to 86, 135, and 197 dimensions with PCA, which correspond to the principal components that explain (a) 80%, (b) 90%, and (c) 95% of the total variance, respectively. We then randomly sampled 1,000 NonAI-NonAI, 1,000 AI-AI, and 1,000 NonAI-AI paper pairs and calculated the distance for each pair within each of the reduced-dimensional spaces. Results show that the distances in the various reduced-dimensional spaces are highly correlated with those in the original space, indicating the reliability of our paper embeddings for distance calculations.

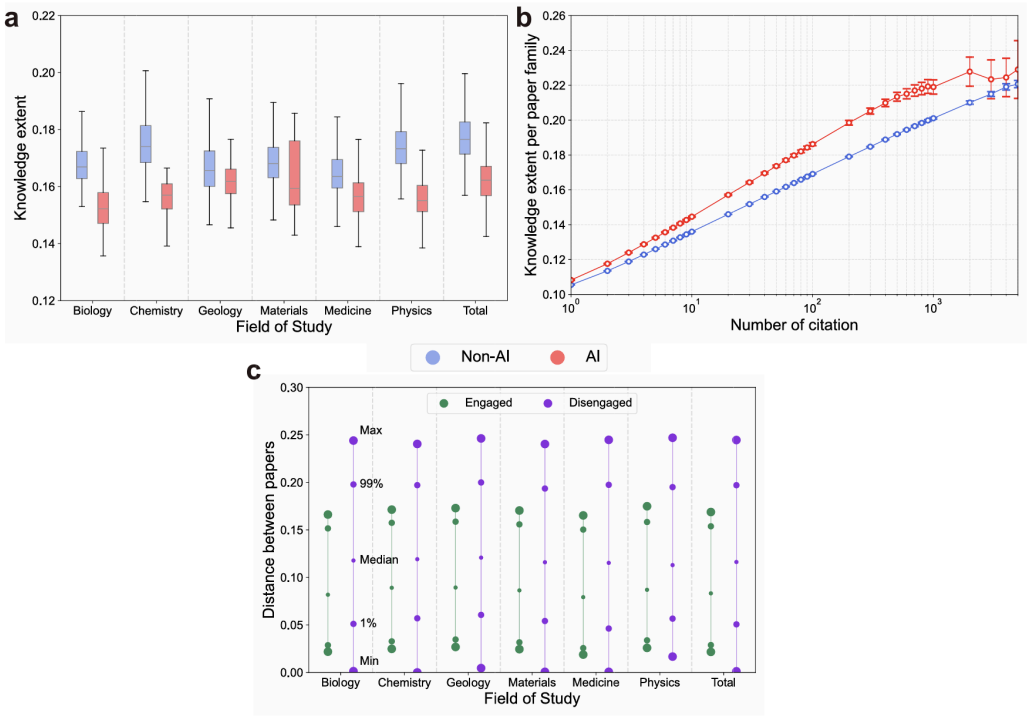


Figure S37. Replication of results related to high-dimensional distance calculations with Cosine distance. (a) Knowledge extent of AI and non-AI papers in each field. We show 1.5 times the inter-quartile range (IQR) from the box as whiskers. (b) Knowledge extent of individual AI and non-AI paper families. We show 99% CIs as error bars. (c) The distribution of distances between paper pairs that cite the same papers, with or without citing each other—engaged versus disengaged. Results with Cosine distance are consistent with those with Euclidean distance reported in the main text.

"Established" is misspelled in some places and note that "centroid" is still misspelled in 3b.

Response 3.7: We thank the Reviewer for their careful reading and apologize for the typos remaining figure typos. We have double-checked the complete passage again and corrected all potential mistakes.

For your convenience, we paste the revised figures below.

Revision:

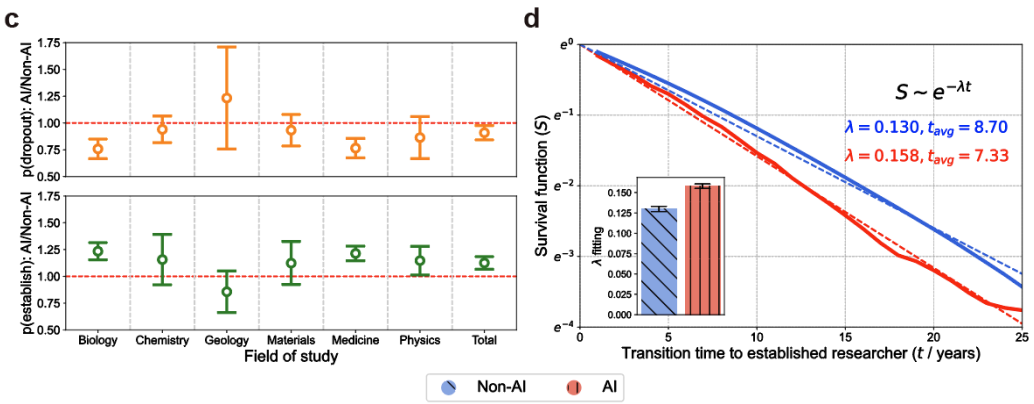


Figure 2. AI enlarges paper impact and enhances researcher careers. (c) The probability of two role transitions between junior scientists adopting AI and their counterparts without AI ($n = 6$ disciplines $\times$ 46 year observations). Junior scientists adopting AI have a higher probability of becoming established researchers and a lower probability of exiting academia compared with their counterparts without AI. (d) Survival functions for the transition from junior to established researcher ($n = 2,282,029$). The survival function can be well-fit with exponential distributions, where junior scientists adopting AI become established earlier than their counterparts without AI.

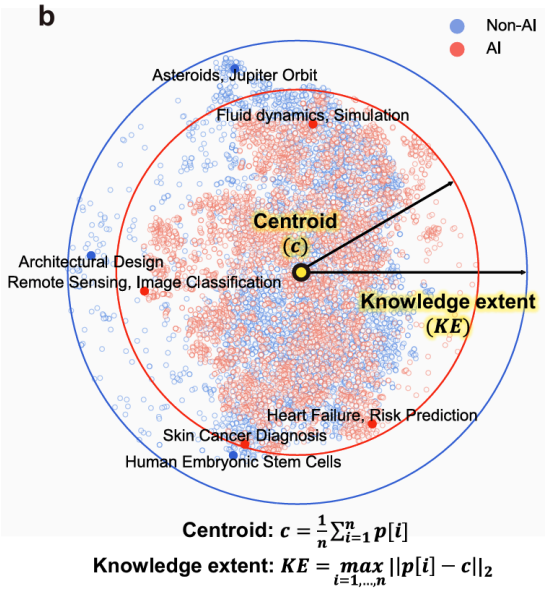


Figure 3. AI adoption is associated with a contraction in knowledge extent within and across scientific fields. (b) For visualization, we use the t-SNE algorithm to flatten the high-dimensional embeddings of a random batch of 10,000 papers, half of which are AI papers and half are non-AI papers, into a 2-D plot. As shown by the solid arrows and circular boundaries, the knowledge extent of AI papers (calculated in the unflatted space) is smaller across the entirety of natural science, and AI papers are more clustered in knowledge space, indicating more concentration on specific problems.

We sincerely appreciate the Reviewer's valuable time and effort spent engaging with our manuscript. We are dedicated to further improving the clarity and robustness of our manuscript in this revision cycle.

Response to Referee#4's Comments

I was happy to review, "Artificial intelligence expands scientists' impact but contracts science's focus" for a second time. I was positively overwhelmed by the care and attention the authors put into revising this manuscript. The supplemental analyses I raised in the previous review were addressed, and in my opinion, add substantial depth to the paper. In my professional opinion, this article is greatly improved and ready for publication.

Response: We sincerely appreciate this encouraging feedback on our manuscript. We thank the Reviewer for constructive suggestions across the revision cycle, which have substantially helped us in improving this paper toward the standards required of *Nature*.

Suggested Improvements: Very minor wording things here and there. For instance, the authors state "Among the massive papers in the OpenAlex dataset, we select 66,117,158…" would benefit from "quantity of" to signify number versus importance.

Response 4.1: We thank the Reviewer for this additional suggestion. Following it, and combining comments on presentations from other Reviewers, we have thoroughly polished the manuscript to make it more professional and readable for a broader audience.

For your convenience, we quote the revised texts and figures below.

Revision:

[**Page 8 in the *Methods*-**M1. Dataset and Paper selection] As of March 2025, OpenAlex contains 265.7M research papers, along with related data about citation, author, institution, etc. Among the massive quantity of papers in the OpenAlex dataset, we select 66,117,158 English research papers…

Revision of manuscript in Nature #2025-01-00126C

We are delighted to receive the decision of acceptance of our paper from the *Nature* editor and all reviewers. The review process provided excellent suggestions that we have used to substantially improve our paper. Here are our responses to the editorial comments and the remaining issues raised by the referees.

Response to Referee#2's Final Comments

I still think that the title is "overhyped".

Response: It is true that most people would assume "Artificial Intelligence" means genAI whereas the data in our paper is more skewed to ML techniques. Following the suggestions by the Editor, we have changed the original expression of "Artificial Intelligence" into the term "AI Intelligence Tools" to better express the nuance.

In general, I think that the abstract should not report numbers, but I will leave that to the editors.

Response: We have finalized the abstract per the Editor's suggestions.

Response to Referee#3's Final Comments

I truly applaud the authors for all the additional work undertaken to answer my questions and those of the other reviewers. I feel that procedures are now much clearer and their robustness better evaluated.

I am happy with all the answers to my questions. I think that the definition of 'established scientists' as those who have led one or more research projects in particular clarifies matters.

Response: We are pleased to hear that our revision has addressed the Reviewer's concerns. We thank the Reviewer for their valuable time spent reviewing our manuscript.

I leave two final suggestions for the authors:

First, I found the percentile exploration motivated by panel 2a extremely useful. I think there is more richness than in the reporting of the average, and would encourage the authors to show the results for the top 10% and top 1% in Fig. 2.

Response: Thanks for this suggestion. We have added the top 10% and top 1% annual citation alongside the original mean result in Fig. 2. This enhances the richness of our mean results. The revised figure is as follows:

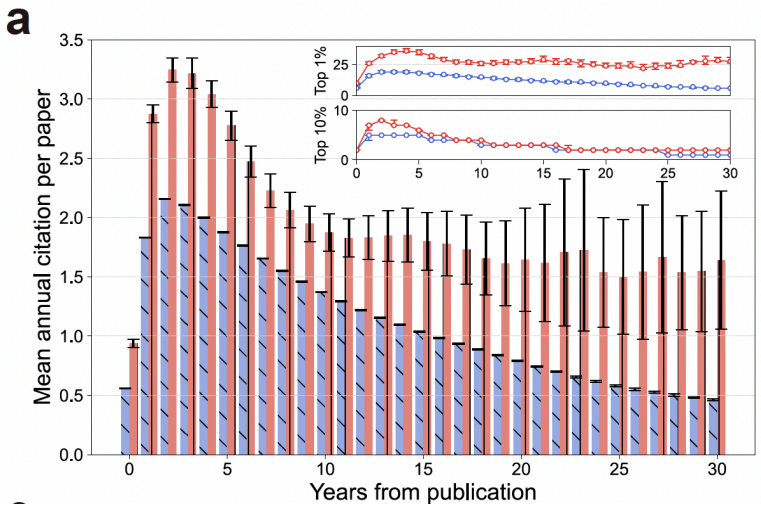


Figure 2. AI enlarges paper impact and enhances researcher careers. (a) Average (insets: top 1% and 10%) annual citations after publication of AI and non-AI papers ($p<0.001$, $n=27,405,011$), where AI papers attract more citations.

Second, In Figs. 5b, c, the visualizations would be stronger if they showed box plots instead of bar plots.

Response: We considered (and prototyped) this, but it didn't look as consistent with the other panels in the overall paper. As a result, we have retained the original bar structure.