The Rapid Growth of AI Foundation Model Usage in Science

Ana Trišović[*†], Alex Fogelson[†], Janakan Sivaloganathan, Neil Thompson[*]

MIT FutureTech, 32 Vassar St, Cambridge, MA 02139, USA.

[†]These authors share first authorship.

[*]Corresponding author: ana_tris@mit.edu, neil_t@mit.edu

We present the first large-scale analysis of AI foundation model *usage* in science – not just *citations* or *keywords*. We find that adoption has grown rapidly, at nearly-exponential rates, with the highest uptake in Linguistics, Computer Science, and Engineering. Vision models are the most used foundation models in science, although language models' share is growing. Open-weight models dominate. As AI builders increase the parameter counts of their models, scientists have followed suit but at a much slower rate: in 2013, the median foundation model built was 7.7x larger than the median one adopted in science, by 2024 this had jumped to 26x. We also present suggestive evidence that scientists' use of these smaller models may be limiting them from getting the full benefits of AI-enabled science, as papers that use larger models appear in higher-impact journals and accrue more citations.

At the forefront of the Artificial Intelligence (AI) revolution are foundation models: large pre-trained systems with a broad applicability that contrasts with earlier narrow, task-specific tools (*1, 2*). Although their development demands substantial investments in compute, data, and human capital for training, these models can thereafter be used for a wide range of downstream scientific tasks.

Foundation models are now proving transformative across diverse scientific domains (*3*). In

biology and bioinformatics, they are used for processing biological sequence data, enabling advancements in structure prediction, molecular property analysis, drug design, and single-cell omics research (*4–6*). In physics, they support scientific reasoning and particle event modeling (*7, 8*). Pretrained physics-informed models also enhance accuracy for climate modeling and fluid dynamics simulations in health, industry, and climate-related applications (*9–11*). Foundation models are also accelerating materials science, using large experimental databases to identify new materials and predict properties such as superconductivity and polymer structures (*10, 12, 13*).

With the widespread release of large language models (LLMs), usage has further expanded to tasks central to everyday scientific practice. LLMs assist with literature review and synthesis, manuscript drafting and editing, code generation and analytical pipeline construction, and even experimental design and interpretation (*14–16*). Increasingly integrated into routing workflows, LLMs act as versatile research assistants (*17*). While many uses of foundation models play a supportive role, others such as image segmentation, classification, and predictive modeling directly influence research findings, enabling machine-assisted discovery.

Understanding the prevalence of AI in scientific research is important for evaluating scientists' productivity (*18*), the pace of innovation (*19*), and the broader trajectory of economic growth (*20*). It also provides insight into how the scientific method is changing with the increasing use of AI, which has broader implications for the validity, transparency, and reproducibility of research findings (*10, 21–24*). The significant computational requirements of AI also seem to be deepening existing inequalities in access to research infrastructure (*10, 12, 23, 25–29*). Measuring and interpreting trends in foundation model use is therefore essential for informing science policy and funding strategies, as well as for anticipating the long-term economic and societal impacts of AI-driven discovery.

When measuring foundation model usage in science, choosing the wrong indicator can yield a distorted picture. Most previous broad-based studies of AI use citations, keyword-based methods, or abstract classification to identify AI adoption (*18, 30–35*). However, using citations as an indicator of adoption conflates meaningful use of a model (e.g., via API or fine-tuning) with papers that merely mention AI for context. Our analysis shows that, over the past five years, citation-based analysis *overestimates* true foundation model usage by 141%. Because text classification approaches, e.g. via LLMs, are imperfect, naively deploying them also creates mismeasurement. We find that using

frontier language models without adjustment would lead to a 15% overestimation of model adoption and an 86% overestimation of model customization over that same period. Given these challenges, it is perhaps not surprising that previous work has reported an enormous range of potential adoption rates, ranging from 0.75% by 2025 to already at 5.3% by 2021 (*18, 32*).

Related work has assessed foundation model usage through surveys of scientific practitioners, particularly for LLM usage (*36, 37*). That work finds high rates of usage by scientists for various science-related tasks. While an important measure of adoption in its own right, such surveys do not provide an indication of how central LLMs were to the scientific work. Our analysis, by contrast, provides a strong indicator of importance since "all elements necessary to allow interpretation and replication of the results" merit inclusion in a paper (*38*). That most scientific usage would fall below this threshold of importance is consistent with other work that shows relatively small impacts of LLMs on non-scientific productivity (*39, 40*).

To address these limitations, we build an analysis pipeline that directly examines how foundation models are being used in scientific research, and we do extensive data validation to adjust our measurement for the imperfections of language model classifications. Our analysis identifies early adopters, emerging trends, usage disparities, and examines how model domains, scale, and openness shape integration and collaboration across fields. To support this work, we introduce the *FutureTech AI in Science Database*, a comprehensive catalog of foundation models and their research adoption patterns.

Tracking Foundation Model Adoption in Science

Foundation Model Definition

A wide variety of definitions for foundation models have been proposed (*41, 42*), for example by the Stanford Institute for Human-Centered Artificial Intelligence, the European Union's AI Act, and the U.S. Department of Homeland Security (*2, 43, 44*). For our purposes, we focus on a definition that centers on AI's utility as a tool in science:

> A foundation model is a pretrained model whose initial development requires substantial investment in data, compute, and expertise, but which can be efficiently adapted and

reused across a wide range of downstream tasks. Its value lies not just in its architectural design, but in the broad utility enabled by the upfront training effort.

FutureTech AI Adoption in Science Database

To systematically track the use of foundation models in scientific research, we introduce the *FutureTech AI in Science Database* that links foundation models to the scientific publications which cite or apply them.[1] The database includes the following:

- **1,942 foundation models** released by 2024, manually curated according to our definition. Metadata includes (where available) model origin, openness, number of trainable parameters, and Semantic Scholar identifiers for 1,556 models with accompanying publications.
- **268,694 unique papers** citing these foundation models, with associated metadata including authorship, institutional affiliations, and country of origin. Relevant papers and full texts are identified using the Semantic Scholar Academic Graph (*45*).
- **736,161 in-text citations**, including three sentences of context, where foundation models are cited, each annotated with a citation-intent label indicating if and how the model is being applied, and a disambiguation key for cases involving multiple model versions (e.g., different sizes).

Together, these components enable rich cross-sectional analyses of foundation model use in science, spanning scientific fields, industry affiliations, geographic regions, and the specific ways in which the models are applied. To characterize the latter, we introduce a custom citation-intention classifier that categorizes references to foundation models into three distinct types:

- **Background Citation**: The model is referenced in a paper to provide context or background, such as in surveys discussing its significance, components, or comparisons with other models.
- **Model Usage**: The model is applied as a methodological tool without model weight modification, for example via local use or API for content generation, classification, or embeddings.

[1]Details on this dataset are described in the appendix.

- **Model Customization**: The model is used and further customized through fine-tuning or similar methods for domain-specific adaptation.

Throughout the text, we refer to *both* usage and customization as collectively representing **model adoption**.

Key Trends of Foundation Model Adoption Across Scientific Fields

Foundation model adoption is surging in science

Since 2015, the use and customization of foundation models has grown rapidly, at near-exponential rates (Figure 1A). During this period, overall adoption grew from 0.03% of publications to 1.4%. This includes both the usage of foundation models as-is, which grew from 0.03% of publications to 0.9% in 2024, and also the more technically-demanding customization of models, which rose from 0.001% of publications to 0.4%. Even these modest percentages translate to an enormous number of papers annually – we estimate that in 2023, 47,900 papers used a foundation model and 19,800 papers customized them. Such prolific use reveals that foundation models have already become an important tool in modern science.[2]

As of 2024, the scientific fields with the highest adoption rates were Linguistics (34%), followed by Computer Science (18%) and Engineering (4.6%). AI adoption has become pervasive, with 87% of scientific fields in our analysis reaching 1% adoption by 2024 (Table 1). Among fields with sufficiently rich data to analyze, Chemistry and Biology usage grew the fastest in recent years, with a three-year compound annual growth rate (CAGR) of 168% and 123% for usage of AI models, and 47% and 29% rate for customization, respectively (Table 1).[3] Except for Law (66%), all other fields demonstrated usage growth rate below 35%. Psychology recorded the highest growth in model customization (107%), while Environmental Science experienced a slight decline in model customization of -1.5%. In general, however, many scientists found it valuable to adapt models with their own training data, with 12% of all papers citing foundation models from 2015 to 2024 customizing models.

[2]Unsurprisingly, the mentions of AI foundation models as background context is science have also grown from 0.12% to 1.8% share of publications.

[3]See Appendix for field selection details.

	Model Adoption						Model Usage	Model Customization	
Field of Study	**Model Parameters Avg (Q1, Q3)**	**Top 3 Foundation Models**	**Model Age in Years Avg (Q1, Q3)**	**Total Unique Models Adopted**	**Open-Weight**	**Adoption by 2024**	**Usage Growth (3-yr CAGR)**	**Customizations Growth (3-yr CAGR)**	**Customizers (% of Adopters)**
Biology	105M (19M,589M)	AlphaFold (2) (41%) ResNet-152 (ImageNet) (7%) alphafold-multimer (4%)	4.5 (3.0, 6.0)	223	95%	2.8%	123%*	29%*	28%
Chemistry	108M (24M,482M)	AlphaFold (2) (65%) alphafold-multimer (4%) BERT-Large (2%)	3.8 (3.0, 4.0)	99	97%	1.4%	168%*	47%*	14%
Computer Science	121M (11M,1B)	ResNet-152 (ImageNet) (12%) BERT-Large (6%) VGG16 (6%)	4.9 (3.0, 7.0)	724	91%	18%	13%	13%	42%
Engineering	50M (7M,373M)	ResNet-152 (ImageNet) (14%) VGG16 (7%) AlexNet (4%)	5.7 (4.0, 7.0)	383	93%	4.6%	7.2%	11%	42%
Environmental Science	55M (11M,272M)	ResNet-152 (ImageNet) (13%) AlphaFold (2) (7%) VGG16 (7%)	5.6 (4.0, 7.0)	269	97%	2.4%	15%	-1.5%	42%
Law	504M (36M,7B)	BERT-Large (27%) RoBERTa-Large (12%) distilbert (5%)	3.8 (2.0, 5.0)	78	87%	3.3%	66%	27%	49%
Linguistics	635M (51M,8B)	BERT-Large (23%) RoBERTa-Large (7%) xlm-roberta (5%)	3.3 (2.0, 4.0)	297	87%	34%	16%	27%	43%
Materials Science	53M (7M,385M)	ResNet-152 (ImageNet) (14%) AlphaFold (2) (11%) VGG16 (6%)	6.1 (4.0, 8.0)	125	95%	1.1%	34%	19%	45%
Mathematics	93M (7M,1B)	ResNet-152 (ImageNet) (19%) VGG16 (9%) ResNet-110 (cifar-10) (5%)	5.2 (3.0, 7.0)	228	91%	1.3%	17%	21%	48%
Medicine	65M (10M,419M)	ResNet-152 (ImageNet) (15%) VGG16 (7%) AlphaFold (2) (6%)	5.7 (4.0, 8.0)	327	96%	1.5%	28%	18%	48%
Physics	95M (8M,1B)	ResNet-152 (ImageNet) (13%) VGG16 (6%) image-to-image_cgan (4%)	5.9 (4.0, 7.0)	156	84%	1.0%	18%	55%	48%
Psychology	376M (22M,6B)	BERT-Large (15%) ResNet-152 (ImageNet) (7%) VGG16 (6%)	5.0 (3.0, 6.0)	136	84%	1.0%	22%	107%	37%

Table 1: Adoption of foundation models in 2019-2024 publications, by scientific field. Model age is the difference between model and publication release dates. Asterisk indicates false-positive unadjusted data (details in Section 1.2.5).

In all scientific fields, researchers tend to use older, more established models that predate their publication by 3 or more years, with many lagging by 5 or 6 years. Notable exceptions are Linguistics, Law, and Chemistry, which use models which are 3.3, 3.8, and 3.8 years old, on average, respectively. Computer Science adopts the highest number of unique models (724), whereas Law adopts the fewest (78) (Table 1).

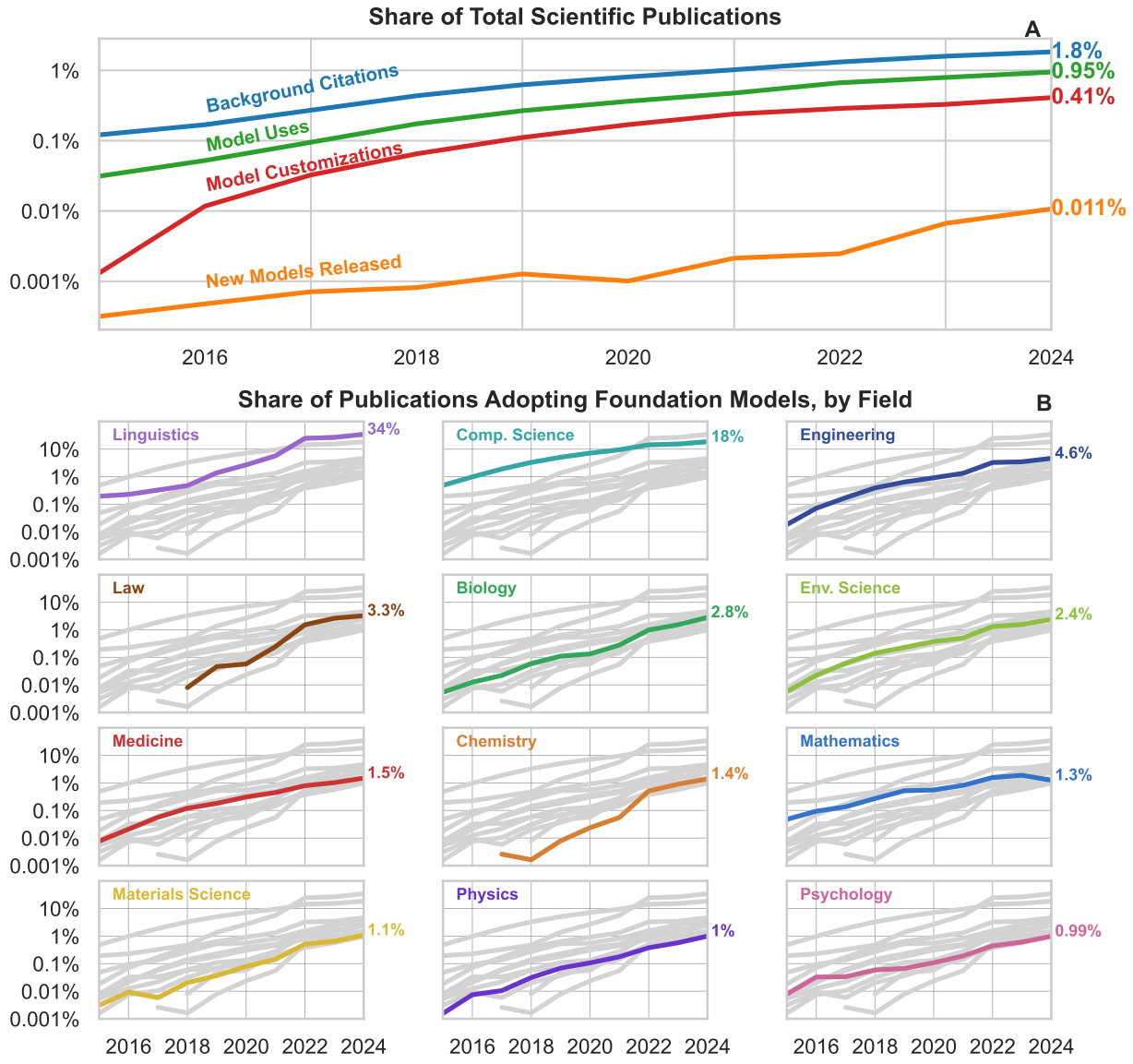


Figure 1: Trends in scientific engagement with foundation models: (A) Share of publications citing, using, customizing, or releasing foundational models. Publications with multiple types are categorized by their most technically demanding category (release, then customization, then use, then citation); (B) Share of publications adopting foundation models by field, where adoption is defined as use or customization. Grey lines represent the other academic fields.

Smaller models dominate adoption, large-scale models dominate development

Despite the enormous growth in the size of cutting-edge models (*46, 47*), scientists largely have stuck with small models. Even in 2024, nearly half of the models adopted had less than 100 million parameters and nearly three-quarters had less than 1 billion parameters (Figure 2A). By contrast, in that same year, nearly 75% of the foundation models being built had 1 billion or more parameters (Figure 2B).

For many years, the size of models being adopted by scientists increased only gradually. Growth accelerated after 2022, with average model sizes increasing 3.9x by 2024. As might be expected, language and multimodal models played a role in this growth (growing 5.0x and 11x, respectively). But other model types also grew significantly over this period, by 2.1x. Despite such growth, the adoption of the largest foundation models in science is limited. We estimate that by 2024 there had been fewer than 10,000 adoptions of 100+ billion parameter models (e.g. GPT4) across all of science (Figure 2C).

Up until 2018, adopters typically used larger models than builders, a trend largely driven by widespread use of VGG, ResNet and AlexNet models, which were large models for the time (*48*). Since 2018, scientists have continued the use of similar-sized models even as model builders built ever-larger systems. This has led to a widening gap between the size of foundation models being built and those being used (Figure 2D). Over the past five years, the gap in median model size between builders and adopters has grown from 1.8x to 26x. Since model capabilities grow with greater model size (*49–51*), this suggests that scientists are only partially reaping the benefits of the recent AI advances. A plausible explanation for this limited uptake of cutting-edge models is the cost and technical demands of working with larger systems, a concern frequently raised in policy discussions (*27, 52*). Despite this broad trend, there is evidence that some scientific fields, particularly Psychology and Linguistics, have adopted larger models more quickly, suggesting that modern LLMs have particular incremental value for these fields (Figure 2E).

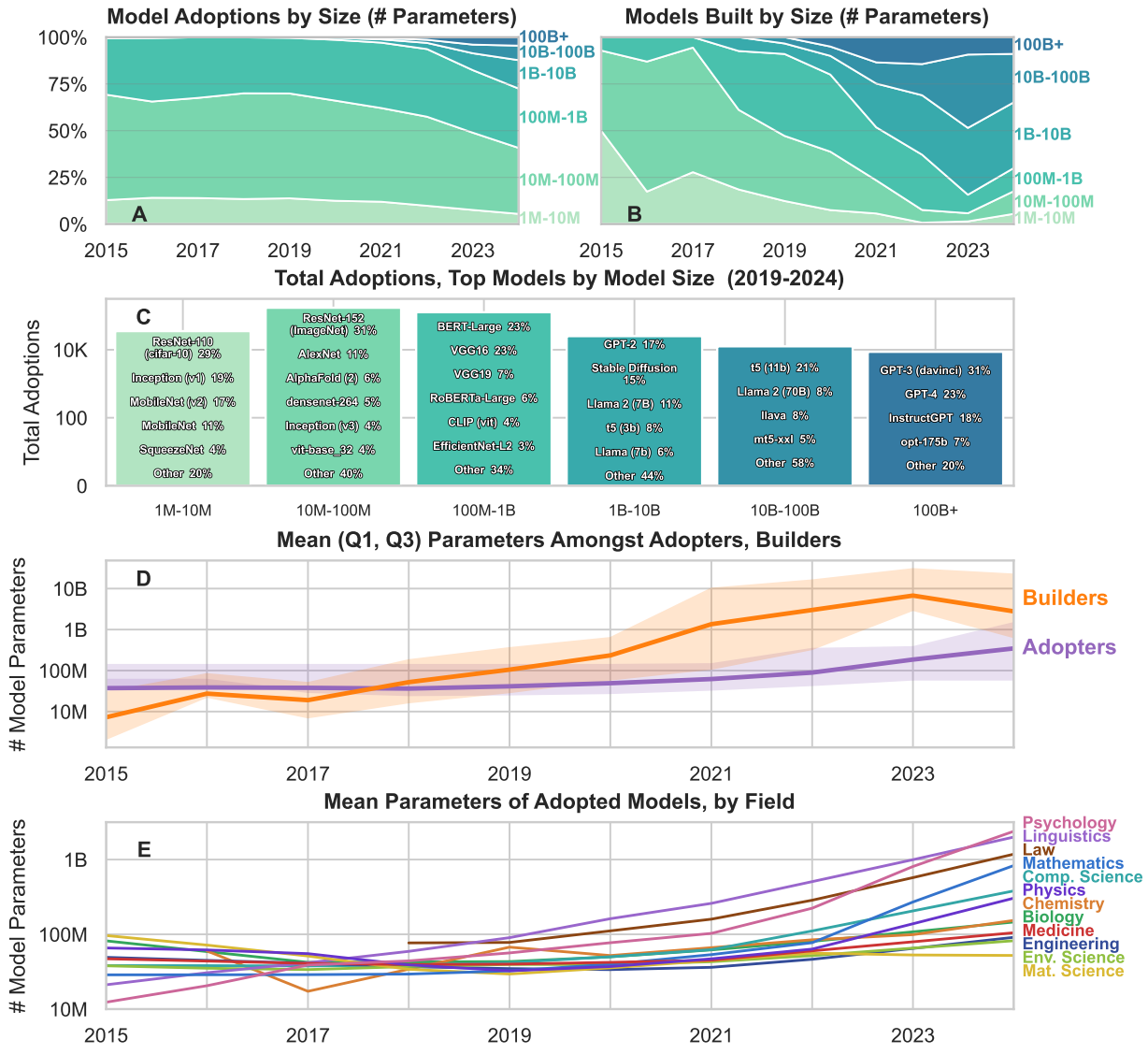


Figure 2: Foundation model adoption across scientific fields: (A) Share of adoptions by parameter count; (B) Share of models being built by parameter count; (C) Total adoption by model size (# parameters). Most-adopted models are listed in the bar; (D) Trends in mean model size for models built and adopted, with 25th-75th percentiles shaded; (E) Average model size adopted per year, by scientific field.

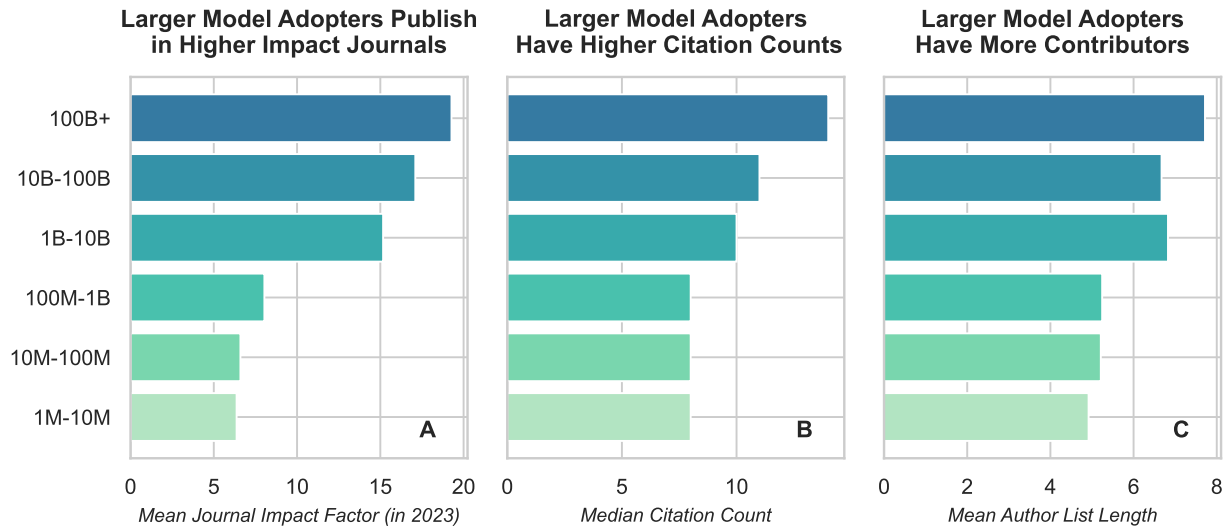


Figure 3: The scholarly impacts of foundation model adoptions: (A) average journal impact factor by adopted model size; (B) median citation counts by adopted model size; (C) average number of authors by adopted model size.

Large model adopters have greater impact

There is suggestive evidence that the use of large models may lead to stronger scientific outcomes. We observe this in two ways: the usage of larger models is correlated with publication in higher quality journals (Figure 3A) and with accruing greater numbers of citations (Figure 3B). These associations hold *after* controlling for publication year with regression fixed effects, indicating that they are not simply driven by recency bias. On average since 2020, a ten-fold increase in model parameters corresponds to an article being published in a journal with a 2023 impact factor that is 2.3 higher ($p < 0.1\%$, $r^2 = 0.053$). It also corresponds to accruing 1.3x more citations ($p < 0.1\%$, $r^2 = 0.145$).

We also consider the size of research teams using foundation models of different scale. It is plausible that the adoption of larger models demands greater resources, which could lead to larger research teams. By contrast, larger (more powerful) foundation models might allow authors to collaborate more with AI, replacing the roles previously done by human coauthors. This would imply smaller research teams for larger models.

In practice, we observe that larger models are used in papers with more contributors (Figure 3C)

($p < 0.1\%$, $r^2 = 0.022$ for all models; $p < 0.1\%$, $r^2 = 0.020$ for open models). There are, nevertheless, many small research groups that employ large-scale foundation models.

Vision models and open models play a critical role in research

Looking across scientific fields reveals enormous variation in the types of foundation models being used (Figure 4A). In many fields, the dominant models being used are vision models, likely because they allow the larger-scale analysis of image data (for example, as discussed by (*53*)). Despite the omnipresence of LLMs in the media, language models are only the dominant foundation models in three fields: Psychology, Law and Linguistics. Since so much of the data in these fields is textual, this suggests that language model adoption may also be driven by the ability to analyze larger corpora than was previously possible. In Biology and Chemistry, the use of Biology-specific foundation models is dominant.

While the importance of vision models is decreasing over time compared to other types, it still represents 65% of all usage overall, and nearly half of all usage in 2024 (Figure 4B). Language models are the next most-used foundation models, representing 23% of adoption, with their share growing to 26% of usage by 2024. Both of these percentages are part of the overall growth in adoption, with vision model usage growing at an annual rate of 7.7%, while language models grow at 35%. Use of other modalities has also grown: multimodal (9.0%, 955% growth rate), biology (6.0%, 309%), image generation (6.5%, 54%), and speech (2.2%, 43%).

This pattern of scientific adoption represents an entirely different mix of models than builders are creating. Of all the 1,942 foundation models being built, most are language models (52% overall), followed by vision (19%), biology (15%), multimodal (6.7%), and image generation (4.0%), video (3.1%), gaming (2.9%), speech (2.8%), and robotics models (2.0%). Figure 4C depicts these proportions over time.[4][5]

The adoption of foundation models in science is overwhelmingly of open-weight models, where anyone can download the models and apply or customize them. Across fields since 2019, 92% of adopted models are open. The ability to freely modify the model also leads to many more customizations of open models than closed ones (42% of adoptions for open models versus 23% for

[4]These categories overlap, as models may span multiple modalities.

[5]Additional modality-specific growth rates are provided in Figure S2.

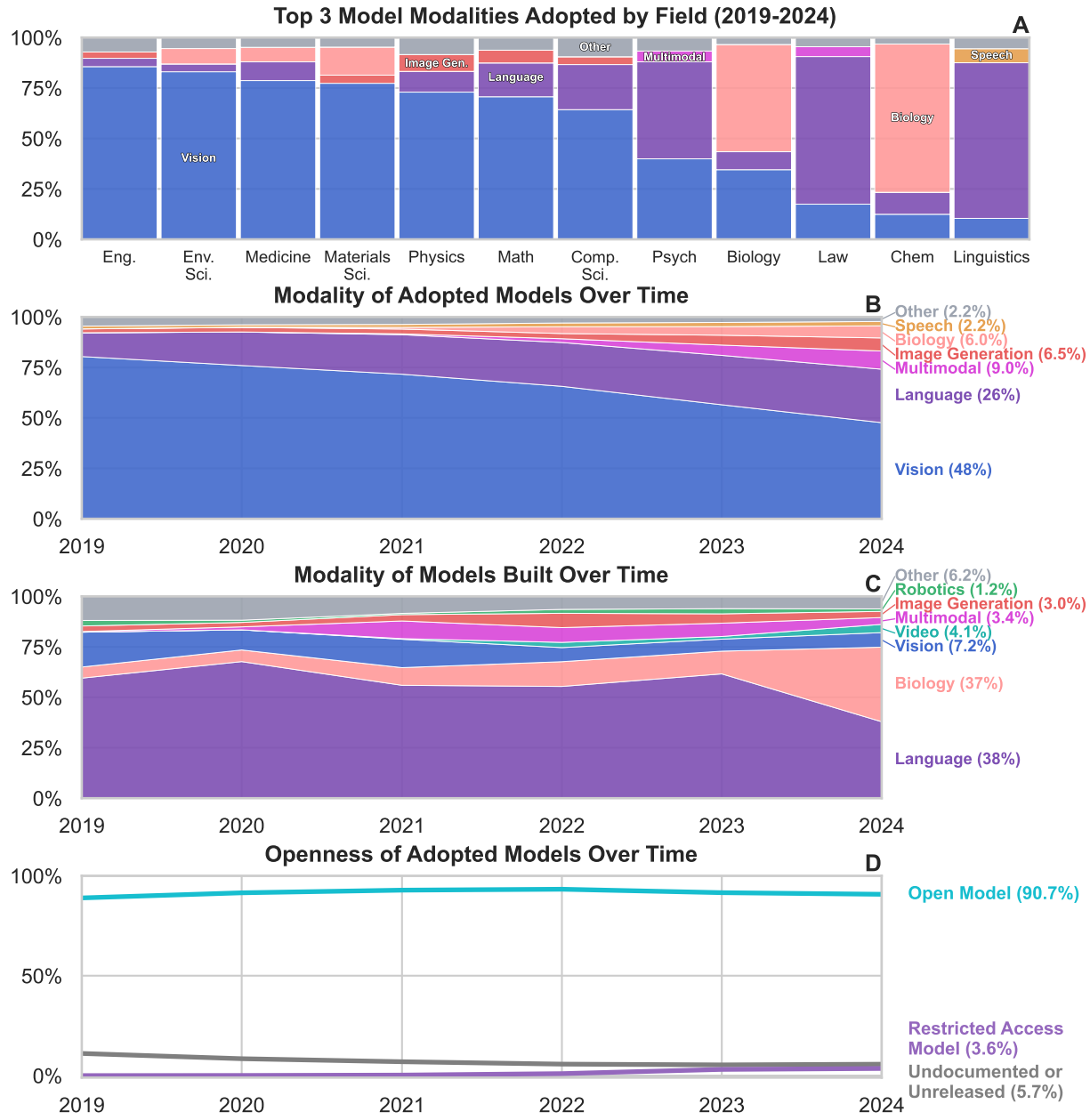


Figure 4: Share of foundation models by (A) adopted modality by field, (B) adopted modality over time, (C) built modality over time, and (D) openness of adopted models over time. In A-C, multiclass models are given equal weight per modality.

closed models). Despite closed-models limited prevalence, their adoption by scientists is growing more rapidly than open models. From 2020 through 2023, the usage of closed models grew 336% annually (vs 20% for open models). In that same period, customization grew by 298% (vs. 18% for open models). Much of this growth in restricted models is of extremely large (100B+ parameter) language models. This growing reliance on restricted-access models raises concerns about the long-term openness and accessibility of AI for scientific research, especially given the well-documented importance of open foundation models (*1, 27, 54*).

This pattern of adoption contrasts strongly with the mix of foundation models being built, where open-weight models comprise 52% of the total, while restricted-access models account for 7.9%. There are also 27% of models that are created but not released, and 13% of models for which we could not find documentation to determine openness.

Limitations

Our findings are limited both by the underlying data that we analyze and the limits of interpreting usage via the text around citations. For our foundation model data, we build on a set first gathered by Epoch AI (*55*). While a valuable record of foundation models, it likely omits some models as a consequence of fields evolving rapidly. For the identification of citations and adoption, we rely on the Semantic Scholar Academic Graph (*45*), which implies some limitations. Semantic Scholar primarily covers English-language publications, so our analysis largely represents English-language science. It is also updated only periodically and with a lag, hence at the time of writing the total publication count in 2024 is below that of 2023. For this reason, we largely report shares of publications and only report adoption CAGRs as of the end of 2023. Semantic Scholar's categorization of scientific fields is non-exclusive, meaning that individual publications can be present in multiple fields. For our analysis, such papers are counted in each field in which they are categorized.

Methodologically, our approach depends on the citation network, meaning that uses of foundation models outside formal publications and those not explicitly cited are not captured, a challenge previously noted for research software (*56*). Similarly, since we use the text around a citation to classify whether it is being cited/used/customized, if a model is cited once in a general context but later used without further citation, our analysis would not detect that usage. Because we analyze

the text of publications, we are only able to analyze open-access literature. We apply appropriate weighting schemes to adjust for this and other data gathering limitations (see Section 1.2.6).

Discussion

The adoption of AI in science is anticipated to be an enormous paradigm shift (*57*). In this article, we explore one of the key pathways for such a change: the adoption of foundation models. By 2024, we find that 0.9% of all scientific publications use foundation models without modification, while 0.4% customize them. To put these percentages in context, it is estimated that 36% of biomedical studies use microscopy (*58*), and 8–11% of computational science papers rely on supercomputing resources (*35*). This suggests that foundation models could still have significant room for further adoption if they are as important as these other tools. Furthermore, the near-exponential growth trajectory in the usage of foundation models signals that this shift is still very much ongoing.

Three methodological advances distinguish our approach and enable novel insights. First, we track individual foundation models rather than AI as a monolithic category, making it possible to analyze model characteristics including size, openness, modality, and release date. We find that older, small models are predominantly adopted, and that vision models still represent the most-used type. Overwhelmingly, scientists are currently using open models, but the use of restricted-use LLMs is growing rapidly. Second, we decompose engagement into distinct modes: citation without use, off-the-shelf usage, and active customization. This taxonomy reveals that a majority of papers discussing foundation models do not deploy them, revealing the gap between awareness and adoption. We also find remarkable differences in adoption, for example, in linguistics, a third of all papers are using foundation models. We also find clear evidence that, despite being easier to use, off-the-shelf implementations of foundation models are not yet sufficient for many scientists and thus a large minority are themselves customizing models to their particular needs. Third, our classification leverages LLMs for analysis of specific citation context rather than keyword matching or abstract classification, providing higher precision in distinguishing types of computational use from tangential references. Collectively, these advances answer fundamentally different questions: not just whether AI is being adopted, but which specific models are being used, how they are being deployed, and what characteristics make certain models more attractive for scientific applications.

While our research clearly indicates the importance of AI foundation models in science, it also highlights the yet-unrealized potential that remains. Such potential broadly falls into three categories: using more, using newer and using larger. *Using more* is clear, with only 1% of publications using foundation models, there is still enormous potential to use either general-purpose or field-specific tools more in science. *Using newer* reflects the fact that many scientists are using models 3-6 years old. Given the enormous progress in AI performance each year (*59*), this suggests there are significant untapped capabilities not yet being harnessed by scientists. *Using larger* captures that scientists are disproportionately using small models, despite clear evidence that large models yield better performance (*49, 50*) and our own suggestive evidence that using larger models achieves better journal and citation outcomes. We look forward to the scientific discoveries that will be unlocked as this unrealized potential turns into insights about the world.

References and Notes

1. R. Bommasani, *et al.*, The Foundation Model Transparency Index. *arXiv preprint arXiv:2310.12941* (2023), https://arxiv.org/abs/2310.12941.

2. R. Bommasani, *et al.*, On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* (2021).

3. H. Wang, *et al.*, Scientific Discovery in the Age of Artificial Intelligence. *Nature* **620** (7972), 47–60 (2023), doi:10.1038/s41586-023-06221-2, https://www.nature.com/articles/s41586-023-06221-2.

4. Z. Chen, L. Wei, G. Gao, Foundation models for bioinformatics. *Quantitative Biology* **12** (4), 339–344 (2024).

5. Y. Si, *et al.*, Foundation models in molecular biology. *Biophysics Reports* **10** (3), 135–151 (2024), doi:10.52601/bpr.2024.240006, http://www.biophysics-reports.org/en/article/doi/10.52601/bpr.2024.240006.

6. Y. Zimmermann, *et al.*, 32 examples of LLM applications in materials science and chemistry: towards automation, assistants, agents, and accelerated scientific discovery. *Machine Learning: Science and Technology* (2025).

7. K. G. Barman, *et al.*, Large physics models: towards a collaborative approach with large language models and foundation models. *The European Physical Journal C* **85** (9), 1066 (2025).

8. J. Birk, A. Hallin, G. Kasieczka, OmniJet-α: the first cross-task foundation model for particle physics. *Machine Learning: Science and Technology* **5** (3), 035031 (2024).

9. S. Subramanian, *et al.*, Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. *Advances in Neural Information Processing Systems* **36**, 71242–71262 (2023).

10. R. Yu, *et al.*, Physics-Guided Foundation Model for Scientific Discovery: An Application to Aquatic Science. *arXiv preprint arXiv:2502.06084* (2025).

11. J. Sivaloganathan, A. Trišović, N. Thompson, Mapping the Impact of Foundation Models on the UN Sustainable Development Goals, in *2025 IEEE International Conference on eScience (eScience)* (IEEE Computer Society) (2025), pp. 375–376, doi:10.1109/eScience65000.2025.00078, https://doi.ieeecomputersociety.org/10.1109/eScience65000.2025.00078.

12. J. Yin, S. Dash, F. Wang, M. Shankar, Forge: Pre-training open foundation models for science, in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2023), pp. 1–13.

13. K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit, Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **6** (2), 161–169 (2024).

14. A. Gray, ChatGPT" contamination": estimating the prevalence of LLMs in the scholarly literature. *arXiv preprint arXiv:2403.16887* (2024).

15. Y. Zhang, *et al.*, Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence* **1** (1), 14 (2025).

16. A. Narayan, I. Chami, L. Orr, S. Arora, C. Ré, Can foundation models wrangle your data? *arXiv preprint arXiv:2205.09911* (2022).

17. M. Binz, *et al.*, How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences* **122** (5), e2401227121 (2025).

18. Q. Hao, F. Xu, Y. Li, J. Evans, AI Expands Scientists' Impact but Contracts Science's Focus. *https://arxiv.org/abs/2412.07727* (2025).

19. E. Brynjolfsson, D. Rock, C. Syverson, *Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics*, Tech. rep., National Bureau of Economic Research (2017).

20. T. Besiroglu, N. Emery-Xu, N. Thompson, Economic impacts of AI-augmented R&D. *Research Policy* **53** (7), 105037 (2024).

21. P. Faraboschi, E. Giles, J. Hotard, K. Owczarek, A. Wheeler, Reducing the Barriers to Entry for Foundation Model Training. *arXiv preprint arXiv:2404.08811* (2024), version: 2, doi: 10.48550/arXiv.2404.08811.

22. N. D. Lawrence, J. Montgomery, Accelerating AI for science: open data science for science. *Royal Society Open Science* **11** (8), 231130, doi:10.1098/rsos.231130, `https://royalsocietypublishing.org/doi/10.1098/rsos.231130`.

23. R. Gruetzemacher, Leveraging Foundation Models for Scientific Research Productivity. *Global Journal of Computer Science and Technology: Neural & Artificial Intelligence* **23** (3) (2023).

24. J. Qiu, *et al.*, Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics* **27** (12), 6074–6087 (2023).

25. T. Besiroglu, *et al.*, The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?, doi:10.48550/arXiv.2401.02452, `http://arxiv.org/abs/2401.02452`.

26. H. Zenil, *et al.*, The future of fundamental science led by generative closed-loop artificial intelligence. *arXiv preprint arXiv:2307.07522* (2023).

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

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

29. Y. Hao, *et al.*, The Role of Computing Resources in Publishing Foundation Model Research. *arXiv preprint arXiv:2510.13621* (2025).

30. E. Duede, W. Dolan, A. Bauer, I. Foster, K. Lakhani, Oil & water? diffusion of ai within and across scientific fields. *arXiv preprint arXiv:2405.15828* (2024).

31. J. Gao, D. Wang, Quantifying the benefit of artificial intelligence for scientific research. *arXiv preprint arXiv:2304.10578* (2023).

32. S. Hajkowicz, C. Sanderson, S. Karimi, A. Bratanova, C. Naughtin, Artificial intelligence adoption in the physical sciences, natural sciences, life sciences, social sciences and the arts and humanities: A bibliometric analysis of research publications from 1960-2021. *Technology in Society* **74**, 102260 (2023).

33. M. R. Frank, D. Wang, M. Cebrian, I. Rahwan, The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence* **1** (2), 79–85, doi:10.1038/s42256-019-0024-5, https://www.nature.com/articles/s42256-019-0024-5.

34. L. Schmallenbach, T. W. Bärnighausen, M. J. Lerchenmueller, The global geography of artificial intelligence in life science research. *Nature Communications* **15** (1), 7527 (2024).

35. S. Bianchini, A. Geuna, F. Shermatov, AI and Supercomputing are Powering the Next Wave of Breakthrough Science - But at What Cost? (2025), https://arxiv.org/abs/2511.12686.

36. M. Chugunova, *et al.*, *Who Uses AI in Research, and for What? Large-scale Survey Evidence from Germany*, Research Paper 25-11, Max Planck Institute for Innovation and Competition (2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5259847.

37. R. Van Noorden, J. M. Perkel, AI and science: what 1,600 researchers think. *Nature* **621** (7980), 672–675 (2023), doi:10.1038/d41586-023-02980-0, https://doi.org/10.1038/d41586-023-02980-0.

38. Nature Methods, Preparing your material – Submission Guidelines (2025), https://www.nature.com/nmeth/submission-guidelines/preparing-your-submission.

39. A. Chatterji, *et al.*, *How People Use ChatGPT*, Working Paper 34255, National Bureau of Economic Research (2025), doi:10.3386/w34255, https://www.nber.org/papers/w34255.

40. A. Humlum, E. Vestergaard, *Large Language Models, Small Labor Market Effects*, Working Paper 33777, National Bureau of Economic Research (2025), doi:10.3386/w33777, https://www.nber.org/papers/w33777, revised October 2025.

41. S. Alfasly, *et al.*, When is a Foundation Model a Foundation Model (2023), https://arxiv.org/abs/2309.11510.

42. J. Schneider, Foundation models in brief: A historical, socio-technical focus (2022), https://arxiv.org/abs/2212.08967.

43. B. Benifei, I.-D. Tudorache, *Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, Tech. rep., Tech. rep., Committee on the Internal Market and Consumer Protection (2023).

44. A. Henninger, D. Kusnezov, Foundation Models at the Department of Homeland Security: Use Cases and Considerations. *The DHS Science and Technology Directorate (S&T), U.S. Department of Homeland Security* (2023).

45. A. D. Wade, The Semantic Scholar Academic Graph (S2AG), in *Companion Proceedings of the Web Conference 2022* (2022), p. 739.

46. J. Sevilla, *et al.*, Compute trends across three eras of machine learning, in *2022 International Joint Conference on Neural Networks (IJCNN)* (IEEE) (2022), pp. 1–8.

47. P. Villalobos, *et al.*, Machine learning model sizes and the parameter gap. *arXiv preprint arXiv:2207.02852* (2022).

48. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

49. J. S. Rosenfeld, A. Rosenfeld, Y. Belinkov, N. Shavit, A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673* (2019).

50. J. Kaplan, *et al.*, Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

51. J. Wei, *et al.*, Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

52. National Artificial Intelligence Research Resource Task Force, *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a*

National Artificial Intelligence Research Resource, Tech. rep., The White House, Office of Science and Technology Policy (2023), https://nsf-gov-resources.nsf.gov/2023-10/NAIRR-TF-Final-Report-2023.pdf, accessed 5 November 2025.

53. S. Zhang, D. N. Metaxas, On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis* **91**, 102996 (2024), doi:10.1016/j.media.2023.102996.

54. S. Kapoor, *et al.*, On the Societal Impact of Open Foundation Models, doi:10.48550/arXiv.2403.07918, http://arxiv.org/abs/2403.07918.

55. Epoch AI, Data on Notable AI Models (2024), https://epochai.org/data/notable-ai-models, accessed: 2024-09-05.

56. A. M. Smith, D. S. Katz, K. E. Niemeyer, Software citation principles. *PeerJ Computer Science* **2**, e86 (2016).

57. Y. Xu, Z. Zhang, Artificial intelligence for science—bridging data to wisdom. *The Innovation* **4** (6), 100525 (2023), doi:10.1016/j.xinn.2023.100525.

58. A. M. Reigoto, S. A. Andrade, M. C. Seixas, M. L. Costa, C. Mermelstein, A comparative study on the use of microscopy in pharmacology and cell biology research. *PloS one* **16** (1), e0245795 (2021).

59. Epoch AI, Machine Learning Trends, Epoch AI (2024), https://epoch.ai/trends, accessed November 18, 2025.

60. H. Touvron, *et al.*, Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

61. J. Achiam, *et al.*, GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

62. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012).

63. A. Vaswani, *et al.*, Attention Is All You Need, http://arxiv.org/abs/1706.03762.

64. R. Kinney, *et al.*, The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140* (2023).

65. K. MacMillan, S. Feldman, Announcing S2FOS: An Open Source Academic Field of Study Classifier, https://medium.com/p/9d2f641949e5 (2022), aI2 Blog.

66. K. Lo, L. L. Wang, M. Neumann, R. Kinney, D. Weld, S2ORC: The Semantic Scholar Open Research Corpus, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online) (2020), pp. 4969–4983, doi:10.18653/v1/2020.acl-main.447, https://www.aclweb.org/anthology/2020.acl-main.447.

67. K. Lo, L. L. Wang, M. Neumann, R. Kinney, D. S. Weld, S2ORC: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782* (2019).

68. L. Blecher, G. Cucurull, T. Scialom, R. Stojnic, Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418* (2023).

69. D. Jurgens, S. Kumar, R. Hoover, D. McFarland, D. Jurafsky, Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics* **6**, 391–406 (2018).

70. A. Athar, S. Teufel, Context-enhanced citation sentiment detection, in *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (2012), pp. 597–601.

71. A. Lauscher, *et al.*, MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *arXiv preprint arXiv:2107.00414* (2021).

72. D. Hu, *et al.*, VarMAE: Pre-training of Variational Masked Autoencoder for Domain-adaptive Language Understanding. *arXiv preprint arXiv:2211.00430* (2022).

73. A. Cohan, W. Ammar, M. Van Zuylen, F. Cady, Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608* (2019).

74. A. Lahiri, D. K. Sanyal, I. Mukherjee, CitePrompt: using prompts to identify citation intent in scientific papers, in *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (IEEE) (2023), pp. 51–55.

75. A. Fogelson, A. Trišović, N. Thompson, LLMs in Citation Intent Classification: Progress, Precision, and Reproducibility Challenges, in *Proceedings of the ACM Conference on Reproducibility and Replicability (ACM REP)* (Association for Computing Machinery) (2025), pp. 1–8.

76. Open AI, GPT-4o mini: advancing cost-efficient intelligence (2024), `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, accessed: 2024-09-05.

77. A. Dubey, *et al.*, The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

78. A. Grattafiori, *et al.*, The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

79. S. Colvin, Pydantic: Data validation and settings management using Python type annotations (2025), `https://github.com/pydantic/pydantic`.

80. A. Reuther, *et al.*, Interactive supercomputing on 40,000 cores for machine learning and data analysis, in *2018 IEEE High Performance extreme Computing Conference (HPEC)* (IEEE) (2018), p. 1–6.

Acknowledgments

The authors would like to express their sincere gratitude to Emanuele Del Sozzo for his generous support, insightful advice, and invaluable expertise in systems and hardware throughout the course of our experiments. The authors thank Zachary Brown for proofreading the paper and providing insightful comments. A.T. is grateful to Mercè Crosas for useful conversations during her stay at the Barcelona Supercomputing Center. We thank the undergraduate students who assisted with data collection for their contributions: Evan Zhang, Selinna Lin, Denis Siminiuc, Grace Yuan, Emma Li, Sri Saraf, Raul D Campos, Yibo Cheng, Alvin Banh, Dora M. Zhou, Ingrid Tomovski, Kristina Sakayeva and Maeve Zimmer.

Funding

This work is funded by Microsoft, Open Philanthropy/Good Ventures and the Alfred P. Sloan Foundation (G-2025-25164). We acknowledge support from OpenAI Research Credits & the UROP Program at MIT. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper.

Authors contributions

A.T. conceptualized the study and acquired funding, collected the initial dataset, generated the visualizations, performed preliminary analysis, and wrote the initial and subsequent drafts of the manuscript. A.F. developed the data pipeline software, collected and curated the dataset, performed formal data analysis, generated the visualizations and text statistics, and contributed to writing the manuscript. J.S. conducted data collection and integration, contributed to the preliminary analysis, and supported funding acquisition efforts. N.T. conceptualized the study and acquired funding, supervised the project and the methodological choices, provided guidance on data collection and visualization, and contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Data and materials availability

The dataset and executable notebooks will be publicly available on GitHub upon publication.

Supplementary materials

Materials and Methods

Supplementary Figures and Tables

Supplementary Materials for

The Rapid Growth of AI Foundation Model Usage in Science

Ana Trišović[*†], Alex Fogelson[†], Janakan Sivaloganathan, Neil Thompson[*]

MIT FutureTech, 32 Vassar St, Cambridge, MA 02139, USA.

[†]These authors share first authorship.

[*]Corresponding author: ana_tris@mit.edu, neil_t@mit.edu

This PDF file includes:

Materials and Methods

Supplementary Figures and Tables

Materials and Methods

1.1 Data Sources

Our analysis aggregates and extracts data from multiple sources. We begin by curating a custom dataset of foundation models and enhancing it with relevant metadata. We then retrieved papers which cite these foundation models, enabling us to extract key citation sentences and paper affiliations.

1.1.1 Foundation Models

The Epoch AI Index Dataset gathers a collection of foundational machine learning models and systems, many of which formally recognized as AI milestones (*55*). The model collection was guided by specific criteria, including the presence of a machine learning element, demonstration of experimental findings, and indication of a contribution to advance the field of AI (*46*). The majority of the models meet at least one criterion of significance, like receiving over 1000 citations, possessing historical relevance, contributing to a notable advancement in the field, being employed in an important setting, or fulfilling certain subjective criteria for newer models. The dataset is assembled from a variety of sources, such as highly cited papers from leading AI conferences, existing related datasets, literature reviews, Papers with Code[6], and recommendations from experts (*46*).

We obtained a snapshot of the Epoch AI Index dataset, covering 1,942 models released between 1962 and 2024, including notable foundation models like LLaMA (*60*), GPT-4 (*61*), and Claude [7]. Given the broad scope of the Epoch AI Index dataset, which includes foundation models, optimization algorithms, and model architectures, we applied a custom definition (see Section *Foundation Model Definition*) to identify and extract only those foundation models relevant to this study. This definition was deliberately designed to include historically significant releases such as AlexNet (*62*), while excluding purely architectural or theoretical contributions, such as the original Transformer (*63*), that were neither released nor verifiably used as pre-trained models.

Further, the original snapshot of the Epoch AI Index dataset lacked information on model size, model openness, and other relevant factors for many models. To improve its completeness,

[6]https://paperswithcode.com

[7]https://claude.ai

undergraduate students from MIT assisted in reviewing model entries, classifying the models, and adding missing information where possible. Model size was defined as the total number of trainable parameters. Model availability is classified based on whether the model can be downloaded, accessed via an API, or is entirely unavailable. We classify models as *open foundation models* (open weight models) if weights are publicly downloadable and modifiable, *restricted-access* if they are available only via API, and *unreleased* if they are not accessible for reuse. Our dataset also contains undocumented models for which openness could not be determined.

Many of these models were released alongside a paper publication, while others were introduced through blog posts or software repositories. For models documented with a paper URL, we located their Semantic Scholar Corpus IDs in the Semantic Scholar Academic Graph (S2AG) database to retrieve publication metadata. This metadata included details such as the paper's title, abstract, citation count, and a list of citing papers, facilitating the next step of our analysis, which was retrieving the citing papers.

1.1.2 Citing Papers

We leverage the S2AG: an extensive, open-access database that encompasses a vast collection of scientific publications and related academic information. It serves as a rich index of scientific literature, offering detailed metadata on hundreds of millions of academic articles and books. S2AG provides data about papers, authors, citations, and venues on over 200 million research papers across all fields of study (*45, 64*). It is intended as a free and open resource for the research community, offering both a downloadable dataset and access through APIs. When constructing our dataset, we utilize the citation graph, paper metadata (i.e., title, publication year, field of study), pre-extracted paper plain-text, and open-access download links. We retrieve and work with the S2ORC corpus from March 12th, 2025.

Semantic Scholar employs a machine learning classification model to assign fields of study to academic papers, based on paper's title and abstract. Each paper can be assigned up to three fields of study, reflecting its primary areas of focus. At the time of writing, the field of study classification system is limited to English-language papers. The model performs most accurately when both the title and abstract are available, although it can still operate using the title alone (*65*). We determine the field of study for a publication, as well as the total publication counts per field per year, using

Semantic Scholar metadata, allowing interdisciplinary papers to be included in all respective fields.

Semantic Scholar, along with the S2AG dataset, provides plain-text formatted version of millions of open-source papers through the S2ORC dataset (*66, 67*) enhanced with citation annotations and download links to the open-access PDF files. We aim to maximally expand our coverage of the academic literature: we do this by capturing both the pre-extracted plain text from S2ORC, while also downloading and processing open-access PDFs through our own custom pipeline. Where available for download, we parse the PDF files using Meta's Nougat model (*68*) which has the distinct advantage over other models we tested (e.g. Marker[8]) of being trained on specifically academic papers and standardizing bibliographies. Our text conversion is followed by extraction of in-text citations and affiliations sections using regular expressions. When available, our analysis defaults to the S2ORC plain-text, which includes annotations for these features. The text for in-text citations and affiliations are processed later in our methodology to create a structured dataset.

We find 19% of papers in our final dataset are detected in both pipelines, 62% is detected by just S2ORC plain-text, and 19% detected by our custom built pipeline. Our final sample covers 268,694 unique papers citing foundation models through paper citations. Across a randomly selected set of 50 citation sentences per pipeline, the false positive rate for both methods was 0%, giving high confidence in the fidelity of our detection mechanism.

Our initial analysis using the S2ORC API to retrieve references indicated that 62% of open-access papers lacked this annotated data, despite the presence of complete metadata. This suggested a significant issue with the completeness of the S2ORC plain text papers, leading us to develop our own custom pipeline. After developing and processing our own extracted in-text citations, it was discovered that these missing references resulted from an API bug, where even modest batch sizes (200 and above) sporadically returned missing results, while smaller batch sizes did not. The bug has since been reported to the Semantic Scholar team.

1.2 Methods

To understand how each foundation model is referenced within the citing papers and the scientific literature more generally, we apply state of the art large language models to classify the

[8]https://github.com/VikParuchuri/marker

citing sentences, structure affiliations metadata, and categorize institutions. Moreover, to adjust for both closed-access publications and open-access publications which are not captured through our pipeline, we employ sample up-weighting when presenting our results.

1.2.1 Field Selection

When selecting fields to present in our tables and graphics, we apply a uniform criteria to ensure the reliability of our results. In particular, we only include fields if their minimum sample weight in 2024 is less than 15; that is, we exclude results where each 2024 sample within a field requires 15× or more upsampling (see section 1.2.6 below for our weighting scheme). Fields with higher sample weights yield results which are less robust to individual sample variation. Those fields are Agricultural And Food Sciences, Art, Business, Economics, Education, Geography, Geology, History, Philosophy, Political Science, and Sociology. For completeness, Figure S2 shows adoption over time in these fields.

1.2.2 Citation Classification

Citation context analysis is an important task in natural language processing that examines the ways and reasons scholars reference each other's work. This task is framed as a multi-class classification problem (*69*), where each class signifies a distinct citation intent. Early research focused on single-sentence citation contexts, but it was later recognized that a broader context window is essential for a more accurate understanding (*70, 71*). Recent methods employ a citation labeling scheme that categorizes citations into classes such as *background*, *motivation*, *similarities*, *differences*, *uses*, and *extends* (*71–74*).

Following the mixed success of existing methods like `MultiCite` (*71*) for our problem, we turned to language models and focused on three classification categories inline with our research question: *background*, *uses*, and *extends*. This classification allows us to distinguishing how researchers reference models, whether citing foundational work, applying models, or customizing them for specific needs. We experimented extensively with various classification methods (*75*), including existing small language models like MultiCite, position-based classification (e.g. all introduction citations are "background"), open-source language models, large proprietary language models, chain-of-thought and question-answer sequences, and training small feed-forward networks

on the output of the aforementioned methods. In addition, we conducted citation-intent classification using full-paper context with OpenAI's GPT-4 models, chosen for their cost-efficiency and context window capabilities at the time of the experiment. Among all approaches, we found that a three-sentence context classified using zero-shot prompting of OpenAI's GPT-4.1-mini (*76*) delivered the best results. The resulting confusion matrix (Figure S1), was produced by labeling 100 randomly selected sentences for each classification category and weighting the columns according to our dataset's empirical distribution. We used the following prompt:

System Instruction

> You are an expert in many areas of the scientific literature, with a specialty in how machine learning models are used.

Prompt Template

> You will pretend to be the author of some sentences from an academic paper which reference {model_descriptor}. Your goal is to determine the extent to which you used the cited model in your paper. The specific citation which references the foundation model is highlighted using HTML style ⟨cite⟩ brackets.
>
> Choose from the following list to determine the extent to which you adopted the model:
>
> (1) I merely referenced the cited model as relevant background, for its methodology, or its dataset.
> (2) I use the cited model or a part of the model itself, but I don't alter the weights.
> (3) I make updates to the cited model's weights, through additional gradient based training such as fine-tuning.
>
> Please respond in JSON format only, with key "answer" and value either 1, 2, or 3.
>
> The sentences are as follows:
>
> "{multisentence}"

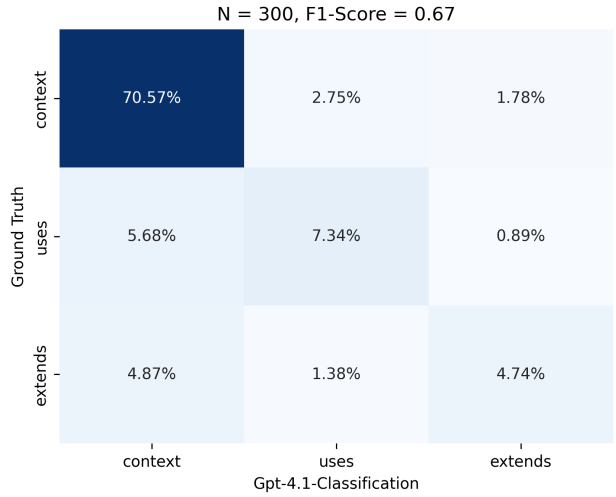


Figure S1: gpt-4.1-mini classifier confusion matrix, column-weighted by observed distribution.

1.2.3 Model Version Disambiguation

For many foundation model papers, multiple models are introduced within a single paper. When citing these papers, the citation graph alone leaves ambiguity as to which specific model is being referenced. To mitigate this, we group our foundation model list by their Semantic Scholar Corpus ID, noting when a single ID has multiple associated model keys. For those such citations, we disambiguate the individual model keys for each citation using Llama3.1:8B. The language model is prompted using (a) the name of the class of models (e.g. "Llama"), (b) the list of specific model keys (e.g. "llama_70B", "llama_7B"), and (c) the full three-sentence context of the citation. The model is allowed to use the key "UNCLEAR" if the context is still ambiguous. In our final analysis, unclear sentences split the associated weight per sample to each model key according to the distribution of unambiguous sentences (or evenly, in the rare case of all samples being ambiguous).

1.2.4 Extracting Paper Affiliations

To obtain accurate information on adopters' affiliations and countries, we extracted the first section of each paper listing the authors and their institutions. We employ Meta's Llama 3.1 8B open-source language model (*77, 78*) to (a) return the content of these sections in a normalized, structured JSON format and (b) to infer the country of origin for each institution, as well classifying the organization as either academic or industry. We validate the structure of these outputs using Pydantic (*79*).

1.2.5 Hardware and Computational Resources

Across our pipeline, we have various uses for GPU acceleration: conversions of PDFs to Markdown using Nougat (described in Section 1.1.2), extraction of structured paper affiliations using LLAMA 3.1 (Section 1.2.3), disambiguation of model versions within citing sentences (Section 1.2.2), and initial experiments with various citation classification techniques. For these computations, we used NVIDIA V100 Tensor Core GPUs and NVIDIA A100 Tensor Core GPUs distributed across 32 GPU Nodes on Lincoln Laboratory's Supercloud (*80*).

1.2.6 Paper-Level Classification and False Positive Corrections

When we analyze citation behavior, we classify individual citing sentences as reflecting one of three types of model engagement: CONTEXT for mere background citations, USES for deploying models unchanged, and EXTENDS for model customization. When aggregating sentences within paper, we assign the paper a label if at least one of its citation sentences matches the category: a paper is labeled EXTENDS if any sentence is classified as such; USES if no extends are found but at least one USES is present; and CONTEXT otherwise. Adoption is defined as either extends or uses.

A challenge arises as papers with more citation sentences are more likely to be (mis)classified into one of these categories, as each additional sentence increases the chance of a *false positive*. In contrast, estimating false positives at the sentence level is straightforward: we can take a random sample and manually check the accuracy. At the paper level, however, aggregation introduces bias that must be modeled.

To correct for this, we use a Bayesian approach. Let $FP(p)$ denote the event that paper p is a false positive. Let $\hat{e}$ and $\hat{u}$ be the observed counts of extends and uses sentences, respectively, and let $\mathbb{P}_n$ denote probabilities conditioned on a paper containing n citing sentences. Then by Bayes' theorem:

$$\underbrace{\mathbb{P}_n(FP(p))}_{\textbf{desired}} \cdot \underbrace{\mathbb{P}_n(\hat{e}=m,\hat{u}=k \mid FP(p))}_{\text{modeled via Multinomial}} = \underbrace{\mathbb{P}_n(FP(p) \mid \hat{e}=m,\hat{u}=k)}_{\text{sampled empirically}} \cdot \underbrace{\mathbb{P}_n(\hat{e}=m,\hat{u}=k)}_{\text{observed directly}}.$$

By estimating $\mathbb{P}_n(FP(p))$ for different kinds of paper-level false positives, we can down-weight the total paper counts when grouping by classifications. We consider adopts and extends false

positives as two separate cases (from which uses false positives can be derived), and outline below how the conditional terms above are approximated to derive our desired false positive rate, $P_n(FP(p))$, for each false positive type. Finally, we explain when and how these adjustments are incorporated into our results.

Binomial Model Given Paper False Positives We first aim to approximate the distribution of sentences labeled as uses/extends within a paper of size n, conditioned on this paper being a false positive. The parameters of the multinomial distribution are derived from hand labeled sentence classifications samples, using values which we argue result in upper bounds on the false positive rates. By using upper-bound false positive rates, we systematically err on the side of conservatism in our aggregated counts—that is, our corrections are designed to underestimate rather than overestimate the true prevalence. For transparency, we also include uncorrected counts in Figure S2.

Adoption false positives are papers which contain no true uses or extends sentences, yet have at least one (possibly more) false positives for either class. We denote the statement "paper p is an adoption false positive" as $FP_a(p)$. We use the *context to uses* and *context to extends* false positive rates from our sentence level confusion matrix ($r_{c\to u}$ = 2.8%, $r_{c\to e}$ = 1.8%) as parameters in a 3-class multinomial distribution:

$$\mathbb{P}_n(\hat{u} = k, \hat{e} = m | FP_a(p)) = \frac{Multinom(n, r_{c\to u}, r_{c\to e})(k, m)}{1 - Multinom(n, r_{c\to u}, r_{c\to e})(0, 0)}$$

Since our estimates only involve cases where $k + m = 1$, we can (and do) use an equivalent binomial with rate $(r_{c\to u} + r_{c\to e})$ to estimate $\mathbb{P}_n(\hat{u} + \hat{e} = 1 | FP_a(p))$.

Extends false positives (FP_e) are papers with no true extends sentences and at least one false positive extends label. We denote the statement "paper p is an extends false positive" as $FP_e(p)$, Note that these papers may be true usage papers, and we therefor must consider both *context-to-extends* and *uses-to-extends* false positive rates. Unlike *adopts* false positives, our confusion matrix does not provide enough information to estimate the Bernoulli probability of the binomial, since the true distribution contains not only true *context* sentences, but true *uses* sentences as well, the

distribution of which is unknown. Therefore, transition probabilities are insufficient to glean a precise estimate.

The true parameter is necessarily between $r_{c\to e}$ and $r_{u\to e}$. In order to *upper bound* our false positive rate, we want to *lower bound* this binomial term. As we discuss below, our estimation uses $m = 1$ (or $\mathbb{P}_n(\hat{e} = 1|FP_e(p))$), and therefor *upper bounding* the binomial rate will minimize this value (which can be thought of as minimizing the likelihood of getting only one false positive). We thereby use a probability of $r_{c\to e} + r_{u\to e}$ for our binomial distribution, normalized since $\hat{e} \neq 0$ when conditioning on a false positive:

$$\mathbb{P}_n(\hat{e} = m|FP_e(p)) = \frac{Binom(n, r_{c\to e} + r_{u\to e})(m)}{Binom(n, r_{c\to e} + r_{u\to e})(0)}$$

Sampling for Conditional False Positive Rate To estimate the term $\mathbb{P}_n(FP(p) \mid \hat{e} = m, \hat{u} = k)$, we select papers with specific (m, k) configurations and manually label their sentences to obtain an empirical estimate of the false positive probability. In particular, we leverage the most sample efficient m, k characteristics, using $(m, k) \in \{(0, 1), (1, 0)\}$ for adoption false positives, and $m = 1$ for extends, where k varies randomly. This allows us to determine the false positive rate with only one label per paper, and therefor gives us a more robust sample size per unit of labeling time.

More precisely, for $n \in \{1, 3, 5, 9, 13\}$, we label 20 sentences for each of the three characteristics: $(m, k) = (1, 0)$, $(m, k) = (0, 1)$, and $m = 1$. For extends false positives, we observe directly from our sample $\mathbb{P}_n(FP(p) \mid \hat{e} = 1)$, whereas for adoption false positives, we directly observe values $\mathbb{P}_n(FP(p) \mid \hat{e} = 1, \hat{u} = 0)$ and $\mathbb{P}_n(FP(p) \mid \hat{e} = 0, \hat{u} = 1)$, weighting them by their frequencies in the dataset to obtain the more general $\mathbb{P}_n(FP(p) \mid \hat{e} + \hat{u} = 1)$.

The resulting probabilities, each conditioned on some number of sentences n, are linearly extrapolated point-by-point to fill in missing n values. At this point, we have an empirical estimation of $\mathbb{P}_n(FP(p))$, for both adopts and extends false positives. We now explain how these are incorporated into our results.

Adjustments to Aggregated Counts When reporting *total counts* of papers which use, extend, or generally adopt models, we adjust our total counts based on the estimated false positive rates. Consider some subset of papers, for example, Computer Science papers from 2020. If we call $f(n)$

the proportion of papers in this subset which have n sentences, then we can compute the overall false positive rate as follows:

$$\mathbb{P}(FP(p)) = \sum f(i) \cdot \mathbb{P}_i(FP(p))$$

Thus given a subset of paper containing all classifications with total sample weight W, and total classification weights $W = W_c + W_u + W_e$, we make the following adjustments:

$$\begin{aligned}
W_c &\to W_c + W \cdot \mathbb{P}(FP_a(p)) \\
W_u &\to W_u + W \cdot \mathbb{P}(FP_e(p)) - W \cdot \mathbb{P}(FP_a(p)) \\
W_e &\to W_e - W \cdot \mathbb{P}(FP_e(p))
\end{aligned}$$

Similarly, when aggregating extends and uses into one adoption category, our transformation is equivalent to summing W_u and W_e in the transformations above.

In Figure S2 we display Figures 1 before and after making false positive adjustments.gi

Using Unadjusted Counts for Growth Rates For data subsets with particularly low rate of adoption, usage, or customization, on rare occasions we find our expected false positive correction is greater than the total number of observations. We avoid reporting absolute figures in these cases, but report qualified growth rates (Table 1).

False Negatives Our corrections are limited to false positives, and do not include false negatives. To see why, consider two potential effects: a false negative rate of 5% and false positive rate of 5%. The share of publications that customize foundation models is roughly 0.41%. If there is a false negative rate, then this number should actually be 0.43%, which does not meaningfully change our answers. By contrast, because there are 4.4× as many citations as customizations, a 5% false positive rate means that 5% of this larger pool gets added to customizations which would make the underlying rate appear to be 0.63%, or 54% larger than it should be.

More specifically, the average number of sentences in adoption papers is 9.0, with 3.0 of those sentences exhibiting adoption on average. This introduces an asymmetry in our favor, in which all adoption sentences must be false negatives in order for a paper to be a false negative, whereas with

false positives, only one such false positive sentence is needed. Also, when aggregating averages within USES or EXTENDS classification, false negatives do not affect our outcomes provided they are not systematic for our desired metric (e.g. model size, model openness), which is untrue of false positives. Finally, in our manually labeled sample, false negatives were generally more ambiguous to the labeler than false positives.

1.2.7 Weighting Schema

Our primary dataset consists of open-access academic papers. To account for the broader landscape of scientific literature, we leverage the full citation graph from Semantic Scholar and apply a weighting schema that allows us to adjust our estimates to the full population of publications.

Fix a year T and a foundation model M. There are two ways of determining the number of unique papers citing M which are published in year T: (1) through the S2AG dataset, which includes the widest swath of academic papers, which we denote Y_T^M; and (2) observationally through our sample, which is limited to open-access publications extracted through either of our extraction pipelines, which we denote X_T^M. Thus for each sample (that is, each sentence from a paper published in year T citing model M), we could assign a weight given by:

$$W_T^M = \frac{Y_T^M}{X_T^M}$$

However, a small fraction of papers are missing publication years. For each set of foundation model citers, we adjust our weights such that the distribution of years among papers without year metadata matches the distribution among papers where we do have year metadata. Let $T = 0$ index a year which is missing, and let p_T be the proportion of papers citing M with known years from year T (that is for $T > 0$, $p_T = X_T^M / \sum_{\tau \neq 0} X_\tau^M$). Then the property we desire from our weights is that for all $t \neq 0$:

$$W_t^M \left[X_t^M + p_t X_0^M \right] = Y_t^M$$

Intuitively, this means that the weight from our known sample, together with a proportional share of the unknown sample, should sum to the correct total number of papers. To satisfy this constraint, we augment the naive weight and also define a weight for papers with missing years:

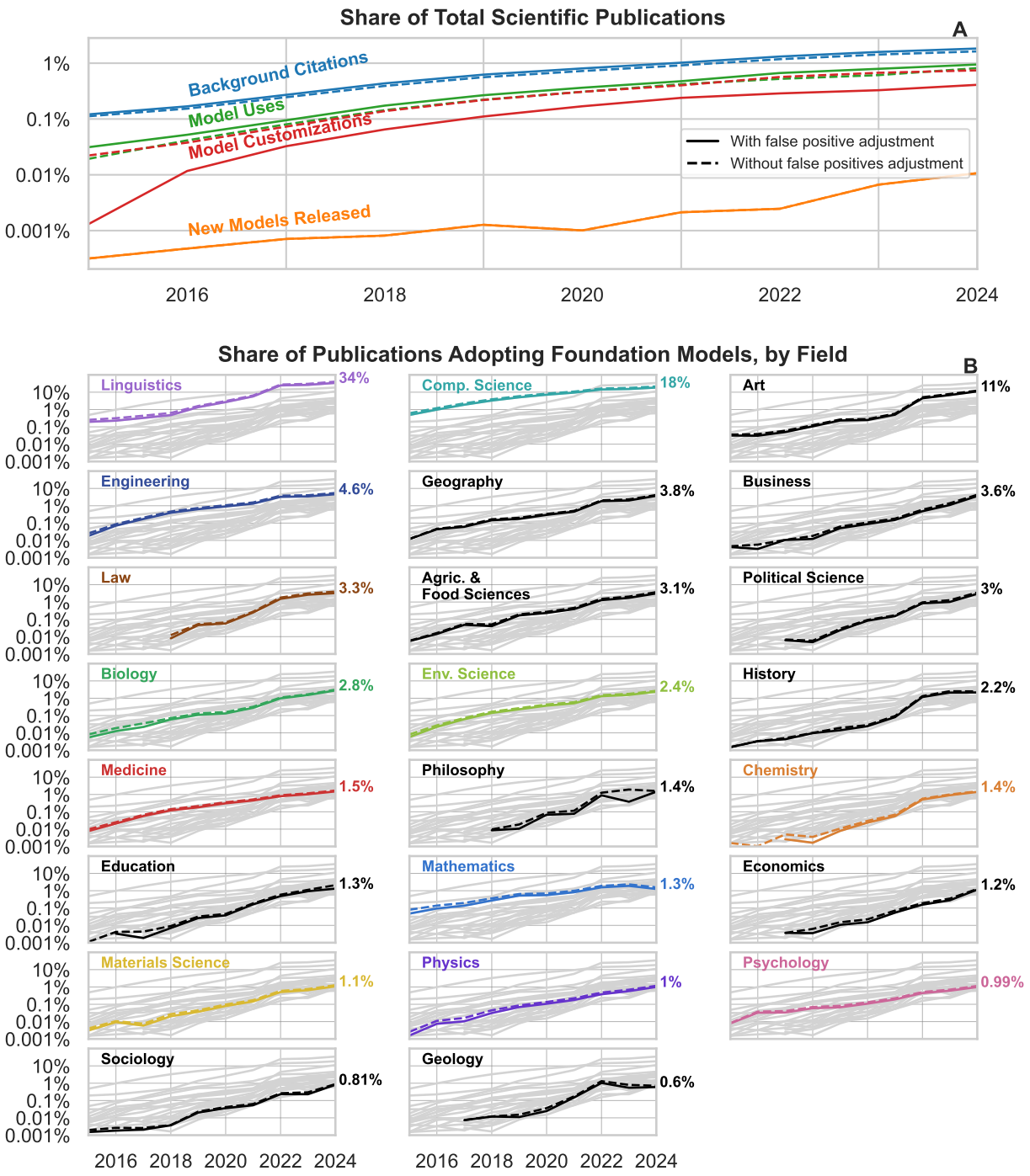


Figure S2: Figure 1 with false positive adjusted and unadjusted proportions. Includes fields which did not meet our selection criterion, but are included for completeness. Solid lines are adjusted, dashed lines are unadjusted.

$$W_t^M = \begin{cases} \dfrac{Y_t^M}{X_t^M + p_t X_0^M} & t \neq 0 \\ \\ \displaystyle\sum_{\tau \neq 0} p_\tau W_\tau^M & t = 0 \end{cases}$$

Finally, we note that the estimation of Y_t^M implicitly relies on an additional assumption, since the S2AG may contain missing years as well. To address this, we apply a smoothing procedure to account for missing years in Y_t^M as well, assuming they follow the same distribution, thus $Y_t^M = Z_t^M + q_t Z_0^M$, where Z is the true distribution of across years, Z_0^M is the number of M citers missing a publication year, and q_t is defined similarly to p_t.

Empirically, we find that weights within a paper are generally quite similar: the standard deviation of weights within a paper's citing sentences is, on average, less than 2. Among different foundation models cited by the same paper, we don't expect the availability of papers citing those models to vary substantially, thus this is consistent with expectations. When aggregating to the paper level – that is, aggregating information from multiple classified sentences to one paper – we always use the maximum weight among all the paper's sentences for that sample. Perhaps counterintuitively, this is a *lower bound* on the true number of papers represented by this sample. Consider $\mathcal{M}$ to be the set of models cited by some paper p. For a particular sentence citing model $m \in \mathcal{M}$, the weight tell us how many unseen papers also citing m are represented by the paper p. However, if the paper cites two (or more) distinct models $m_1, m_2 \in \mathcal{M}$, then p represents the unseen papers citing m_1 (denoted $\mathcal{P}_{m_1}$) and the unseen papers citing m_2 (denoted $\mathcal{P}_{m_2}$). By the union bound, we clearly have $\max(|\mathcal{P}_{m_1}|, |\mathcal{P}_{m_2}|) \leq |\mathcal{P}_{m_1} \cup \mathcal{P}_{m_2}| \leq |\mathcal{P}_{m_1}| + |\mathcal{P}_{m_2}|$.

Supplementary Tables

Field of Study	Model Parameters Avg (Q1, Q3)	Top 3 Foundation Models	Model Age in Years Avg (Q1, Q3)	Total Unique Models Adopted	Open-Weight	Customizers (% of Adopters)
Biology	130M (22M,758M)	AlphaFold (2) (45%) ResNet-152 (ImageNet) (6%) alphafold-multimer (5%)	4.8 (3.0, 6.0)	154	96%	27%
Chemistry	127M (31M,519M)	AlphaFold (2) (68%) alphafold-multimer (4%) esm2-15b (2%)	4.0 (3.0, 4.0)	65	98%	14%
Computer Science	275M (18M,4B)	ResNet-152 (ImageNet) (10%) BERT-Large (5%) VGG16 (3%)	5.1 (3.0, 8.0)	589	91%	43%
Engineering	79M (8M,746M)	ResNet-152 (ImageNet) (13%) VGG16 (4%) yolo (4%)	6.5 (5.0, 8.0)	277	94%	45%
Environmental Science	74M (14M,394M)	AlphaFold (2) (12%) ResNet-152 (ImageNet) (12%) faster_r-cnn (5%)	6.3 (4.0, 8.0)	185	98%	42%
Law	1B (70M,15B)	BERT-Large (23%) RoBERTa-Large (15%) GPT-3 (davinci) (6%)	4.0 (2.0, 5.0)	51	86%	48%
Linguistics	2B (111M,23B)	BERT-Large (15%) RoBERTa-Large (8%) xlm-roberta (5%)	3.6 (2.0, 5.0)	200	86%	42%
Materials Science	63M (6M,620M)	AlphaFold (2) (16%) ResNet-152 (ImageNet) (13%) AlexNet (3%)	6.7 (4.0, 8.0)	89	94%	47%
Mathematics	256M (13M,5B)	ResNet-152 (ImageNet) (14%) VGG16 (5%) diffusion models (lsun bedroom) (5%)	5.1 (2.0, 8.0)	142	88%	44%
Medicine	93M (12M,719M)	ResNet-152 (ImageNet) (14%) AlphaFold (2) (10%) VGG16 (5%)	6.3 (4.0, 8.0)	253	95%	50%
Physics	194M (16M,2B)	ResNet-152 (ImageNet) (8%) image-to-image_cgan (5%) Yolo (v3) (5%)	6.1 (3.0, 8.0)	104	86%	51%
Psychology	1B (58M,24B)	BERT-Large (12%) RoBERTa-Large (6%) GPT-3 (davinci) (6%)	5.2 (2.0, 7.0)	85	80%	38%

Table S1: Model adoption data aggregated by field. Asterisk indicates false-positive unadjusted data (see false positives in Section 1.2.5). This table restricts the domain to more recent papers than Table 1.

Model Category	CAGR (2021–2024)
Biology Models	137%
Video Models	227%
Robotics Models	100%
Speech Models	65%
Image Generation Models	53%
Multimodal Models	50%
Vision Models	46%
Language Models	36%

Table S2: Annual growth rates of different model categories from 2021 to 2024.