Data Brief

Identifying AI Research

Author
Christian Schoeberl
Autumn Toney
James Dunham

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

July 2023

Table of Contents

Introduction

Analyzing scholarly literature provides insight into scientific research activity, including international collaboration rates,[1] countries' research ecosystems,[2] and research funding portfolios.[3] At CSET, our analytic questions required overcoming a challenge: how do we find research that is relevant to artificial intelligence and machine learning (AI/ML)? How we define and identify relevant literature matters for the conclusions we draw and the policy recommendations they inform.[4] Meanwhile, the methods available for identification are driven by analytic requirements that can vary from project to project, resulting in varied approaches.

Policy analysis at CSET and elsewhere has leveraged different methods for finding AI/ML-relevant research within the broader scientific literature.[5] Here we provide an overview of different methods used and evaluate the impact of choosing one over another. We assess four methods:

- keyword search,
- field of study classification,
- arXiv-based classification, and;
- research clustering.

We look for variation in the publications each method identifies as AI/ML-relevant.

We show that the choice of method matters for identifying AI/ML-relevant publications. Our results suggest using a supervised approach for English-language analysis; a model fine-tuned on expert labels from arXiv performs much better than alternatives in a series of evaluations. When extending classification to include Chinese-language publications, we recommend applying the arXiv classifier to English-language text, and then using keyword search over Chinese-language text.[6]

Scholarly Literature Data

We use three datasets, each providing a different corpus from which to compare our methods for identifying AI/ML-relevant publications. The first is arXiv, an open preprint repository. The second is AI/ML conference publications, which includes peer-reviewed publications accepted at top AI/ML conferences since 2010. Third is CSET's full merged corpus of scholarly literature, which combines (and deduplicates) publications from Digital Science Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers with Code.[7] The arXiv and AI/ML conference publications represent topical subsets of CSET's merged corpus.

arXiv

Our first source of scientific publications is the open preprint repository arXiv. Authors who contribute publications to arXiv—primarily in computer science, math, and physics—provide one or more subfield tags to describe their work. Volunteer subject editors review and amend these tags for accuracy. In this way, arXiv provides a set of expert-labeled publications. Within computer science, there are 40 subfield labels for areas such as artificial intelligence, machine learning, natural language processing, computer vision, and robotics.[8]

When classifying publications for analytic relevance, we want a method that performs well on hard examples. In other words, we want to distinguish relevant from irrelevant publications in fields that might overlap with AI/ML. The arXiv corpus allows us to evaluate methods in their ability to make this distinction in science and mathematics.

arXiv has some limitations for current purposes. Its coverage emphasizes a particular subset of research. Publications in AI/ML-relevant categories began appearing on the platform only around 2010. Lastly, many papers on arXiv are pre-prints, which may differ from peer-reviewed papers.[9] Publishing research on arXiv is an emerging norm in the AI/ML community;[10] thus, this corpus constitutes a large sample of AI/ML research, with fine-grained expert labels describing its contents.[11]

AI/ML Conference Publications

As a complement to the arXiv corpus, we identified peer-reviewed publications accepted at top AI/ML conferences since 2010.[12] This set of papers represents AI/ML research recognized by reviewers to advance the field. Additionally, analysis that relies on the successful identification of AI/ML-relevant publications is likely to consider those appearing at top conferences as important. We therefore use the AI/ML conference corpus to evaluate the different methods in terms of a fundamental goal: whether they can correctly identify papers at the forefront of AI/ML research as relevant to AI/ML.

CSET Merged Corpus

We extend our analysis to a broader variety of publications using CSET's merged corpus of scholarly literature, which captures journal articles, conference proceedings, dissertations, thesis papers, books, and other scientific documents spanning from the 1700s until the present day.[13] For this analysis, we restrict the corpus to publications that cite or have been cited by at least one other publication and were published in 2010 and later.

CSET's merged corpus includes no fine-grained expert labels indicating AI/ML relevance, but it provides far greater coverage of scientific literature across topics and languages. This allows us to evaluate the different identification methods on a broader set of literature. It additionally delivers metadata from multiple sources about authors, organizational affiliations, and citations. Analysis at CSET typically relies upon this information.

Summary

CSET's merged corpus contains all publications from arXiv and the AI/ML conferences, and many of the conference papers also appear on arXiv. But the three datasets differ substantially, as seen in Table 1.

Table 1. Corpus Statistics

	arXiv	AI/ML Conferences	CSET Merged Corpus[*]
Publication Count	1,451,888	88,305	72,584,078
Publication Language[†]			
English	99%	96%	85%
Chinese	2%	1%	21%
Author Affiliation Available[‡]	72%	98%	77%
United States	25%	45%	13%
China	9%	25%	39%
EU-27	30%	20%	14%
Country Unknown	35%	13%	23%

Source: arXiv and CSET's merged corpus.

CSET's merged corpus is larger than the arXiv and AI/ML conference corpora by orders of magnitude. Fifteen percent of its publications appear only in a language other than English, compared to only 1–4 percent of the other two datasets. Coverage across scientific domains and languages is a defining feature of CSET's merged corpus.

As a predominantly English-language platform, arXiv contains more work whose authors have an organizational affiliation in the United States or European Union. Although we observe a similar language distribution in the AI/ML conference corpus, about one quarter of its papers report a Chinese institutional affiliation.

[*] CSET's merged corpus restricted to publications that cite or have been cited by at least one other publication and were published in 2010 and later.

[†] As identified by CLD2. Titles and abstracts can appear in more than one language; we count a publication toward each observed. Corpus language percentages sum to more than 100%.

[‡] Percentages indicate publications with 1+ affiliation reported. See methodological details.

Methods for Identifying AI/ML Research

We assess four possible methods for identifying AI/ML-relevant publications in each corpus introduced above. To understand the analytic implications of using one method over another, we then examine each set of results.

Keyword Search

We implement a keyword search over publication titles and abstracts using 35 Chinese terms and 104 English terms that CSET developed in 2019 and 2020 by manual curation (see Appendix C for keywords).[14] Previous CSET research used this approach to count AI publications by contributing countries and explore global AI research output.[15] In one sense, this is a straightforward solution: select terms related to the topic of interest—here, AI—and then find publications that use those terms. Low barrier to implementation is a primary appeal of keyword search.

Yet developing, evaluating, and maintaining performant queries is a time-intensive undertaking. The terms most associated with AI/ML research in 2022 will be different from relevant terms in 2012. Prior analysis has demonstrated that the effect of such drift is significant.[16] An additional limitation is that titles and abstracts of technical work often include only quite obscure terms. Use of more general, frequently observed terms (e.g., algorithm) in a query may return more publications that are not AI/ML specific and miss correct publications that use topic-distinct terms (e.g., convolutional neural network).

Fields of Study

In the second approach, we evaluate fields of study, a system for categorizing scientific publications that originated in Microsoft Academic Graph (MAG).[17] Computer science is one of 19 top-level categories in the field taxonomy; within it are 34 subfields, including artificial intelligence and machine learning. These categories offer a method for identifying AI/ML-relevant research.

In brief, publications receive scores that indicate their association with each field of study. Scoring is based on the proximity between embeddings for the publication text and text that represents each field, with field representations drawing on Wikipedia articles and the academic sources they cite. For current purposes, we consider a publication AI/ML-relevant if any of its three highest-scoring subfields is artificial intelligence or machine learning.[18]

Like keyword search, fields of study require language-specific implementation. In CSET's merged corpus, we assign fields of study to English-language publications and then impute them for others, using the field scores of citation-graph neighbors. (This yields coverage of one in five non-English publications.)[19] Fields of study are best suited to topical analysis of English-language scientific literature.

In analysis of AI/ML research, fields of study offer a unique ability to surface publications in application areas. For example, finding research that uses AI/ML techniques in biotech research and development could be achieved by selecting AI/ML publications that are categorized in the field of biology or its subfields.

arXiv Classifier

The fine-grained categories in the arXiv corpus offer training data for a supervised solution to identifying AI/ML-relevant publications. Under this approach, arXiv's expert labels represent a dynamic, implicit definition of AI/ML relevance that we extend to publications beyond arXiv by fine-tuning a SPECTER-based transformer model.[20]

Among the 40 categories of computer science research on arXiv, we consider six AI/ML-relevant in training: artificial intelligence, computation and language, computer vision, machine learning, multiagent systems, and robotics. arXiv papers without any of these labels provided negative examples.[21]

Table 2 reports the performance of an arXiv-trained classifier on a hold-out test set of papers.[22] We observe 89 percent precision and 87 percent recall, for an F1 score of 88 percent. This improves upon an earlier model developed in 2019 using SciBERT.[23] Since then, the number of papers with AI/ML-relevant labels on arXiv has grown by 130 percent, allowing better performance with a more efficient model.

Table 2. Test Metrics for arXiv-Trained Classifiers

	Precision	Recall	F1
Earlier model (2019)	83%	85%	84%
Current model	89%	87%	88%

Source: Author analysis. For details of training and evaluation, see our replication materials.

An arXiv-trained classifier can be maintained with relatively little effort, particularly in contrast with keywords. But like fields of study, its application is limited to English-language publications.

Map of Science Research Clusters

Our last approach to identifying AI/ML-relevant publications leverages ETO's Map of Science, which includes CSET research clusters that have at least 50 publications and five publications in the past five years. ETO's Map of Science contains more than 120,000 research clusters generated from publication citation links.[24] The CSET research clusters are derived from CSET's merged corpus, where each research cluster can be analyzed using aggregated metadata from its member publications.

We observe our 2019 arXiv classifier predictions for English-language publications in each cluster and keyword search results for Chinese-language publications; if a majority of these is predicted AI/ML-relevant based on classifier predictions or keyword search, we consider the cluster's publications AI/ML-relevant.[25]

The main appeal of this approach is that it extends coverage to languages other than English or Chinese via citation networks. But cluster publications' relatively dense citation ties will not always reflect a common AI/ML relevance. Additionally, AI/ML-classified publications that are in clusters not labeled as AI/ML relevant will not be considered.

Results

CSET Merged Corpus

We begin by examining the results from each method when applied to our dataset of 72 million publications from CSET's merged corpus. We cannot calculate performance metrics like precision or recall here, because ground-truth labels are unavailable.* Instead, we assess how summary statistics vary among the publications each method surfaces.

The fields of study and arXiv classifier methods yield the most publications identified as AI/ML-relevant (2.7 million), as seen in Table 3. We surface 2.5 million publications using keyword search. The Map of Science solution identifies the fewest (1.7 million).

While the percentage of publications with one or more U.S. institutional affiliations is 13–14 percent across methods, the percentage with a Chinese affiliation varies, between 35 percent when using the Map of Science and 44 percent with keyword search. Judging by language, the Map of Science identifies more non-English and non-Chinese publications as AI/ML-relevant. This highlights that analysis of AI research output that examines the institutional affiliations of authors will be sensitive to the methodology used to identify AI research.[26]

* See the following sections for those evaluations in our other datasets.

Table 3. AI/ML-Relevant Publications in CSET's Merged Corpus, by Identification Method

	Approach 1: Keyword Search	Approach 2: Fields of Study	Approach 3: arXiv Classifier	Approach 4: Map of Science*
AI/ML Publication Count	2,489,773	2,719,355	2,711,210	1,733,379
AI/ML Publication Languages†				
English	96%	95%	100%	95%
Chinese	26%	21%	19%	16%
AI/ML Author Affiliation Available‡	78%	80%	83%	79%
United States	13%	13%	14%	13%
China	44%	41%	38%	35%
EU-27	14%	14%	16%	15%
Country Unknown	18%	20%	18%	22%

Source: arXiv and CSET's merged corpus.

The variation we observe in descriptive statistics across these results points to disagreement between methods. Five million publications are identified as AI/ML-relevant by at least one approach, but only 2.6 million are selected by two or more methods. All four approaches agree on about 595,000 publications. Our keyword search resulted in the largest share of unique publications tagged as AI/ML-relevant, with 28 percent not matching relevant publications from any other method.

Table 4 shows the proportion of results that are common across methods. For instance, 55 percent of keyword search results are also identified using the arXiv

* Map of Science research clusters with a majority of their papers predicted AI/ML-relevant.

† Titles and abstracts can appear in more than one language, so we count a publication toward however many we observe. Corpus language percentages sum to more than 100%

‡ Percentages indicate publications with at least one author affiliation reported.

classifier, while 50 percent of arXiv classifier results can be found using keywords. The largest overlap is 70 percent of publications identified using Map of Science research clusters which were also identified using our arXiv classifier.

Table 4. Overlap Between Identification Methods in CSET's Merged Corpus

Keyword Search Overlap		Field of Study Overlap	
Keyword Search — No Overlap	*28%*	Keyword Search	45%
Fields of Study	49%	*Fields of Study — No Overlap*	*22%*
arXiv Classifier	55%	arXiv Classifier	56%
Map of Science	35%	Map of Science	41%
arXiv Classifier Overlap		**Map of Science Overlap**	
Keyword Search	50%	Keyword Search	50%
Fields of Study	56%	Fields of Study	64%
arXiv Classifier — No Overlap	*23%*	arXiv Classifier	70%
Map of Science	45%	*Map of Science — No Overlap*	*10%*

Source: CSET analysis.

AI/ML Conference Publications

We turn to papers from top AI/ML conferences for more insight into the disagreement between methods. Ideally, any solution for identifying AI/ML-relevant publications would return all of the publications in this corpus.

Table 5 shows that Fields of Study correctly identifies only 44 percent of these conference papers as AI/ML-relevant. The keyword search and Map of Science methods achieve 50 and 76 percent recall, respectively. The arXiv classifier, trained on data that includes many of the papers in the AI/ML conference corpus, shows the highest performance, with 81 percent of AI/ML conference papers predicted to be AI/ML relevant.

Table 5: Recall on AI/ML Conference Publications

Method	Recall
Fields of Study	44%
Keyword Search	50%
Map of Science	76%
arXiv Classifier	81%

Source: CSET analysis.

These are cautionary results. Keyword search may seem like a low-cost solution for document retrieval, but the best terms developed at CSET for identifying AI/ML-relevant publications fail to return half of the papers from top AI/ML conferences. The approach based on fields of study yields only 44 percent of the papers.

The arXiv classifier performs best, identifying 81 percent of the conference corpus as AI/ML-relevant. To provide a better understanding of the remaining 18 percent of AI conference publications that the arXiv classifier does not surface, Table 6 displays five publications from various conferences and publication years.

Table 6: Sample of Top AI Conference Publications Classified as Not AI

Paper Title	Conference Published	Year
Differentially Private Data Release for Data Mining	KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining	2011
Traveling the Silk Road: a Measurement Analysis of a Large Anonymous Online Marketplace	WWW '13: Proceedings of the 22nd international conference on World Wide Web	2013
Recovering from Selection Bias in Causal and Statistical Inference	Twenty-Eighth AAAI Conference on Artificial Intelligence	2014
Guarantees for Greedy Maximization of Non-submodular Functions with Applications	ICML'17: Proceedings of the 34th International Conference on Machine Learning	2017
Optimal Algorithms for Non-Smooth Distributed Optimization in Networks	NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems	2018

Source: CSET's merged corpus

arXiv

We evaluate methods against the expert labels associated with each arXiv preprint with an eye to how well they distinguish between relevant and non-relevant examples across computer science, math, and physics.

In this corpus, keyword search performs worst. It returns just 62 percent of publications with AI/ML expert labels, while 79 percent of the publications it returns as relevant are labeled by experts to be AI/ML research. By contrast, the corresponding metrics for the arXiv classifier are 89 percent precision and 87 percent recall, after training on these same labels.[27]

Table 7: Evaluation Against Expert Labels in the arXiv Corpus

Method	Precision	Recall	F1
Keyword Search	79%	62%	69%
Fields of Study	80%	64%	71%
Map of Science	89%	72%	80%
arXiv Classifier	89%	87%	88%

Source: CSET analysis.

The arXiv corpus is not representative of all AI/ML-relevant publications, even those in English, but our results are instructive considering the platform's popularity among researchers and practitioners.

Conclusion

The divergence in results from these four methods when applied to CSET's merged corpus indicates that analytic results will often be sensitive to the choice of method for identifying AI/ML-relevant publications.

Our evaluations recommend the arXiv classifier for identifying AI/ML-relevant publications in English, due to its performance and support for updates from new expert labels over time. By comparison, other methods exhibit lower recall on analytically important AI/ML conference publications, while making far more errors in the STEM preprints available on arXiv.

This analysis has not directly evaluated the performance of Chinese-language keyword search on Chinese-language publications, but our English-language keyword search results suggest careful manual review of results is required. In cross-language analysis, we recommend applying the arXiv classifier to English-language text and keyword search to Chinese-language text, favoring performance over methodological consistency.

Authors

Christian Schoeberl is a Data Research Analyst at CSET.

Autumn Toney is a Data Research Analyst at CSET.

James Dunham is an NLP Engineer at CSET.

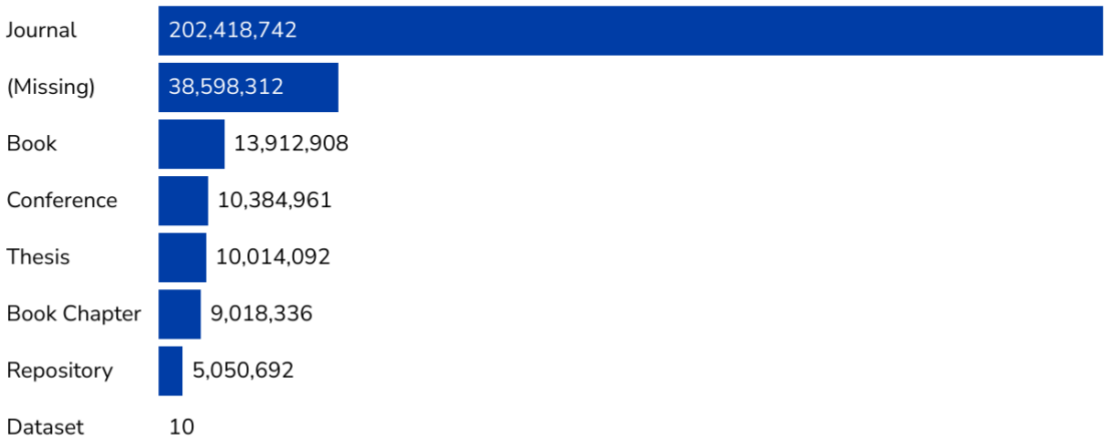
Acknowledgments

For feedback on earlier drafts, we thank Kyle Lo, Nestor Maslej, and Catherine Aiken.

Document Identifier: doi: 10.51593/20220030

Appendix A: Publication Types in CSET’s Merged Corpus

Figure 1. Distribution of Publication Types in CSET’s Merged Corpus



Source: CSET’s merged corpus.

Appendix B: Top AI/ML Conferences

Table 8. AI/ML Conferences
AAAI Conference on Artificial Intelligence
International Joint Conference on Artificial Intelligence
IEEE Conference on Computer Vision and Pattern Recognition
European Conference on Computer Vision
International Conference on Computer Vision
International Conference on Machine Learning
International Conference on Knowledge Discovery and Data Mining
Neural Information Processing Systems
Annual Meeting of the Association for Computational Linguistics
North American Chapter of the Association for Computational Linguistics
Conference on Empirical Methods in Natural Language Processing
International Conference on Research and Development in Information Retrieval
International Conference on World Wide Web

Appendix C: English and Chinese Keywords for Keyword Search Term Methodology

Table 9. Keyword Search Terms	
active learning	object recognition
adaptive learning	one shot learning
anomaly detection	pattern matching
artificial intelligence	pattern recognition
artificial neural network	random forest
associative learning	recommend* system*
autonomous navigation	recurrent network
autonomous system*	recurrent neural network
autonomous vehicle*	reinforcement learning
average link clustering	restricted Boltzmann machine
back propagation	scene* classification
Backpropagation	scene* understanding
binary classification	self driving car*
bioNLP	semi supervised learning
boltzmann machine	sentiment classification
character recognition	single link clustering
classification algorithm	spatial learning
classification label*	speech processing
clustering method*	speech recognition
complete link clustering	speech synthesis
computer aided diagnosis	statistical learning
computer vision	strong artificial intelligence
convolutional neural network	supervised learning
deep learning	support vector machine

ensemble learning	text mining
evolutionary algorithm	text processing
fac* expression recognition	transfer learning
fac* identification	translation system
fac* recognition	unsupervised learning
feature extraction	video classification
feature learning	video processing
feature matching	weak artificial intelligence
feature selection	zero shot learning
feature vector	人工智能
feedforward network	知识表示
feedforward neural network	信息抽取
fuzzy clustering	模式识别
generative adversarial network	计算机视觉
gradient algorithm	人脸识别
graph matching	面部识别
graphical model	面像识别
handwriting recognition	面容识别
hierarchical clustering	深度学习
hierarchical model	深层学习
human robot	一次性学习
image annotation	强化学习
image classification	监督学习
image matching	零次学习
image processing	玻尔兹曼机
image registration	生成式对抗网络
image representation	图模型
image retrieval	机器学习

incremental clustering	神经网络
information extraction	随机森林
information fusion	循环网络
information retrieval	支持向量机
k nearest neighbor	语音识别
knowledge based system*	机器翻译
knowledge discovery	自然语言处理
knowledge representation	自然语言理解
language identification	迁移学习
machine learning	人机交互
machine perception	机器视觉
machine translation	人工神经网络
multi class classification	卷积神经网络
multi label classification	循环神经网络
multi task learning	递归神经网络
natural language generation	受限玻尔兹曼机
natural language processing	
natural language understanding	
neural network	

Endnotes

[1] Autumn Toney and Melissa Flagg, “Research Impact, Research Output, and the Role of International Collaboration” (Center for Security and Emerging Technology, November 2021). https://cset.georgetown.edu/publication/research-impact-research-output-and-the-role-of-international-collaboration/.

[2] Husanjot Chahal, Sara Abdulla, Jonathan Murdick, and Ilya Rahkovsky, “Mapping India’s AI Potential” (Center for Security and Emerging Technology, March 2021). https://cset.georgetown.edu/publication/mapping-indias-ai-potential/; see also CSET Emerging Technology Observatory’s Country Activity Tracker at https://cat.eto.tech/.

[3] Ilya Rahkovsky, Autumn Toney, Kevin W. Boyack, Richard Klavans, and Dewey A. Murdick. “AI research funding portfolios and extreme growth.” Frontiers in Research Metrics and Analytics 6 (2021): 630124.

[4] For prior discussion, see Dewey Murdick, James Dunham, and Jennifer Melot, “AI Definitions Affect Policymaking” (Center for Security and Emerging Technology, June 2020). https://cset.georgetown.edu/publication/ai-definitions-affect-policymaking/.

[5] For example, see Daniel Chou, “Counting AI Research Exploring AI Research Output in English- and Chinese-Language Sources” (Center for Security and Emerging Technology, July 2022). https://cset.georgetown.edu/publication/counting-ai-research/.

[6] Replication materials can be found on GitHub https://github.com/georgetown-cset/identifying-ai-research/.

[7] Data sourced from Dimensions, an inter-linked research information system provided by Digital Science (http://www.dimensions.ai). All China National Knowledge Infrastructure content is furnished for use in the United States by East View Information Services, Minneapolis, MN, USA.

[8] See https://arxiv.org/category_taxonomy for the taxonomy of fields and subfields on arXiv.

[9] Some papers on arXiv are peer-reviewed. For example, conference proceedings are often republished on arXiv. But arXiv requires instead that submissions be “of interest, relevance, and value to [specific scientific] disciplines” as judged by moderators who are “volunteer subject matter experts with terminal degrees in their field.” For more, see https://info.arxiv.org/help/moderation/index.html.

[10] “arXiv submission rate statistics: Data for 1991 through 2019, updated 1 January 2020” https://web.archive.org/web/20201206003554/https://arxiv.org/help/stats/2019_by_area/index.

[11] Clement, Colin B., Matthew Bierbaum, Kevin P. O’Keeffe and Alexander A. Alemi, “On the Use of ArXiv as a Dataset,” arXiv abs/1905.00075 (2019).

[12] We selected these papers by searching CSET's merged corpus for the 13 top AI/ML conferences according to CSRankings metrics. See the Appendix for conference names and https://csrankings.org/ for more on CSRankings, "a metrics-based ranking of top computer science institutions around the world."

[13] See the appendix for a distribution of document type in CSET's merged corpus. This is a broader set of document types than we see in the conference paper or arXiv dataset.

[14] These terms can be found in the Appendix. We developed the English- and Chinese-language queries independently from sets of seed terms (e.g., "machine learning" and "artificial intelligence"). We iteratively searched for publications with matching titles and abstracts; identified co-occurring terms; evaluated the contexts in which they appeared; and kept those tending to appear in relevant publications. This yielded more English terms than Chinese terms. We explored two lower-cost solutions for maintaining AI/ML keyword queries, but the terms developed in 2019–2020 performed better in a series of evaluations. The replication materials describe these experiments.

[15] Daniel Chou, "Counting AI Research: Exploring AI Research Output in English- and Chinese-Language Sources," (Centre for Security and Emerging Technology, July 2022) https://cset.georgetown.edu/publication/counting-ai-research/.

[16] James Dunham, Jennifer Melot, and Dewey Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," https://arxiv.org/abs/2002.07143 (2020).

[17] When MAG reached end-of-life in 2021, we reimplemented fields of study for use in CSET's merged corpus. See Autumn Toney and James Dunham, "Multi-label Classification of Scientific Research Documents Across Domains and Languages," ACL Anthology, 2022, https://aclanthology.org/2022.sdp-1.12/. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 105–114, Gyeongju, Republic of Korea, Association for Computational Linguistics.

[18] That is, we rely exclusively on the ranking of subfield (level-one) scores for each publication. Artificial intelligence and machine learning are subfields under computer science.

[19] Specifically, we take the average of field scores available among publications linked by in- or out-citation. Eight percent of publications with field scores are imputed in this way.

[20] Arman Cohan et al., "SPECTER: Document-level Representation Learning using Citation-informed Transformers" (2020), https://arxiv.org/abs/2004.07180.

[21] According to arXiv's documentation, its Artificial Intelligence category, "Covers all areas of AI except Vision, Robotics, Machine Learning, Multiagent Systems, and Computation and Language (Natural Language Processing), which have separate subject areas." We considered papers in any of these six categories examples of AI/ML-relevant publications. Other categories such as Neural and Evolutionary Computing seemed partially overlapping. We found that in practice, papers in such categories often appeared in one of our six AI/ML categories too. Further details on model development can be found in our replication materials.

[22] The test set contains 234,353 publications, with 37,136 positive ("is AI") publications.

[23] Iz Beltagy, Kyle Lo, and Arman Cohan, "SciBERT: A Pretrained Language Model for Scientific Text" (2019), https://arxiv.org/abs/1903.10676.

[24] See Emerging Technology Observatory, Map of Science, https://sciencemap.cset.tech/.

[25] For a discussion of thresholds for cluster relevance, see Autumn Toney, "Locating AI Research in the Map of Science" (Center for Security and Emerging Technology, July 2021). https://cset.georgetown.edu/publication/locating-ai-research-in-the-map-of-science/.

[26] For analysis on best practices, see https://cset.georgetown.edu/article/studying-tech-competition-through-research-output-some-cset-best-practices/.

[27] For the arXiv classifier results, we report the same precision, recall, and F1 values from Table 2, which were computed on our validation set for unbiased results.