Data Brief

Identifying Emerging Technologies in Research

Authors
Catherine Aiken
James Dunham
Jennifer Melot
Zachary Arnold

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

December 2024

Executive Summary

The challenge of identifying emerging areas and technologies in research is not new, but more data, new methods, and more computational power allow for novel approaches. We build on existing research to develop two new solutions for identifying research relevant to emerging technology areas, specifically artificial intelligence (AI), cybersecurity, as well as chip design and fabrication. First, we trained and deployed machine learning models to predict publication relevance to select emerging technology topics. Second, we assigned publications into research fields within those topics, based on a hierarchical research field taxonomy. We deployed our solutions for emerging technology topic classification and research field scoring over a large corpus of scientific literature. Our evaluated solutions reliably identify research relevant to important emerging technology areas, enabling analysis and monitoring of developments and applications in these areas. We share our publication-level predictions and top fields of study in open datasets. Through interactive tools, we enable researchers to make use of them for analysis.

Background

Classifying scientific literature is an important part of researching innovation, technological development, and scientific progress. Yet, it is a task that presents many challenges. The scientific literature is vast and quickly growing, field and discipline distinctions are blurry, and terminology is evolving. These challenges are especially salient when researching emerging technologies. For example, when the Center for Security and Emerging Technology (CSET) was founded in 2019, our research required overcoming an unsolved problem: How do we find research that is relevant to the development and application of AI?

Researchers have leveraged different methods to identify topic-relevant research within the broader scientific literature (Gläser et al. 2017). A common approach is a keyword query, searching for publications that use a select set of words or phrases, often curated through expert input and sometimes supplemented with citation analysis or dynamic query expansion (Arora et al. 2013, Chou 2022, Huang et al. 2015, Mogoutov and Kahane 2007). Though practical, keyword queries are time-intensive to develop, evaluate and maintain, and run the risk of going stale.

Another approach, drawing on network analysis, is clustering scientific literature into concentrations of research based on citation linkages (Boyack and Klavans 2020, Klavans and Boyack 2011, Small et al. 2014, Waltman and van Eck 2012). At CSET, we maintain a set of citation-based research clusters (ETO 2023, Rahkovsky et al. 2021, Toney 2021). By connecting literature via citations, the clustering approach helps expand the search and identify relevant research that might not use exact terms. However, citation-based clusters do not neatly correspond to specific topics and have variable performance as literature search tools (Bascur et al. 2023).

Other approaches draw on advances in natural language processing (NLP). For example, assigning publications field relevance labels based on the proximity between embeddings of their publication text and text representing a given field of research, taken from Wikipedia articles and the academic sources they cite (Shen et al. 2018, Toney and Dunham 2022, Gelles and Dunham 2024). Another NLP approach involves fine-tuning transformer-based models (e.g., SciBERT, SPECTER) on dynamic, community-based subject tags (Dunham et al. 2020, Schoeberl et al. 2023) or programmatic labeling with expert-informed labeling functions (Ratner et al. 2020, Zhang et al. 2022). More recent research incorporates generative AI models into these solutions, using prompt engineering to enable large language model (LLM) data labeling and annotation for improved and more comprehensive training data (Tan 2024, Toney-Wails et al. 2024).

Problem

Beyond classifying research into subjects or topics, it is especially difficult to determine research relevance to emerging technologies. Widely-applicable classification criteria and general field taxonomies typically group research according to traditional academic disciplines (e.g., biology, psychology), making it difficult to surface research relevant to domains like AI, which span research areas, evolve quickly, and involve concepts that are poorly defined or lack consensus (Dunham et al. 2020, Krafft et al. 2019). This means many “off-the-shelf” classification solutions are not suited for analysis of research relevant to AI and other emerging technologies.

Meanwhile, building in-house, project-specific solutions is not feasible for many researchers. The classification approaches outlined above are resource-intensive. They require access to subject-matter experts, data science and engineering teams, and troves of data and computational resources. This leaves many researchers to rely on suboptimal but less resource-intensive solutions. Even with new open resources like OpenAlex (Priem et al. 2022), Semantic Scholar (Lo et al. 2019), and SciSciNet (Lin et al. 2023), the task of identifying research relevant to emerging technologies takes time and a well-resourced team.

Solutions

To address this problem, we developed two solutions for identifying research relevant to emerging technology topics and incorporated them into open resources. First, we classified research publications into select emerging technology topics by training machine learning models to predict publication relevance to the topics. This solution provides publications relevant to three emerging technology topics—cybersecurity, LLM development, and chip design and fabrication—expanding CSET's existing set of topic classifications for AI, computer vision, NLP, robotics, and AI safety (Dunham et al. 2020, Schoeberl et al. 2023, ETO 2023). Second, we categorized research publications according to their fields of study by computing publication field scores. This solution resulted in publication field scores for more than 1,100 fields of study, with a focus on fields within AI, cybersecurity, biotechnology, and chip design and fabrication.

Both solutions were deployed over CSET's merged academic corpus, containing over 260 million publications compiled from six scholarly literature databases: Clarivate's Web of Science, OpenAlex, Semantic Scholar, the Lens, arXiv, and Papers With Code.[*] We deduplicated publications following the method outlined in our public code repository. We extracted six document identifiers (DOI, citations, normalized abstract, normalized author names, normalized title, and publication year) for each document. Where certain sets of identifiers between documents are equal, we assigned those

[*] Certain data included herein are derived from Clarivate Web of Science. Copyright Clarivate 2024. All rights reserved.

documents the same merged ID.* The remaining articles were included in the final corpus as unique documents. We selected merged document metadata using heuristics like metadata source quality and frequency of appearance. Data pipelines needed to maintain these datasets were written using Apache Airflow and Apache Beam.

Emerging Technology Topic Classification

For the first solution, we developed machine learning models to predict relevance to three emerging technology topics for English-language publications in our corpus: cybersecurity, LLM development, and chip design and fabrication.† We selected these topics because they are emerging in the sense that they are evolving rapidly, driving innovation, and motivating policy debate, but do not fit neatly within traditional subjects or disciplines. Priority areas were selected in consultation with subject-matter experts at CSET and academic researchers studying scientific innovation and technological development.

To identify cybersecurity research, we trained a model on arXiv data, following the method used for our AI, computer vision, NLP, and robotics research classifiers (Dunham et al. 2020, ETO 2023). Articles in arXiv include subject tags that are initially provided by arXiv authors and revised by arXiv editors as appropriate. Leveraging those subject tags, we trained SPECTER (Cohan et al. 2020), a transformer language model pre-trained on scientific text, to predict cybersecurity relevance for all English-language publications in our corpus.

To identify LLM and chip research, we took a slightly different approach. For these topics, we applied a series of prompts to a generative LLM, specifically Google's Gemini 1.5 Flash. In the first prompt, we instructed the LLM to write a one-sentence summary of the work described in a publication's title and abstract, to include the

* We consider articles that match on their normalized title, normalized abstract, citations, or DOIs, plus either one other identifier in that set, publication year, or normalized author last names, to be the same article. Titles, DOIs, or abstracts that occur more than 10 times in the corpus are excluded from the set of data that can be used to match. We also merge articles based on vendor-provided cross-dataset links. For more details, see this description.

† We also updated our existing AI classifier (Schoeberl et al. 2023), originally developed in 2019, and experimented with using GPT-3.5-Turbo and GPT-4 for data annotation (Toney-Wails et al. 2024).

motivation and then the problem or research task(s) addressed and the methods applied. Then, in a second prompt, we instructed the model to classify each publication, based on the summary output from the first prompt, as relevant to the development of LLMs, chip design and fabrication, or neither.

This zero-shot approach offered substantial efficiency gains. For each model, we manually labeled a small set of papers for use in prompt development. We then drew and labeled a larger random sample for evaluation purposes (see Table 1), but overall annotated many fewer papers than would have been necessary under a supervised approach.

Running our emerging technology topic classification models over our corpus identified 507,828 cybersecurity relevant publications, 58,764 LLM development publications, and 1,198,381 chip design and fabrication publications published since 2010, as displayed in Table 1.*

Table 1. Emerging Technology Topic Classification Evaluation and Results

Topic	Number of Publications	Precision	Recall	F1
Cybersecurity	507,828	0.8	0.75	0.77
LLM development	58,764	0.88	1.0	0.93
Chip design and fabrication	1,198,381	0.86	0.73	0.79

Source: CSET merged academic corpus.

Fields of Study

We also expanded our scientific publication field scoring beyond Microsoft Academic Graph's field of study taxonomy (Shen et al., 2018). This taxonomy contains a hierarchy of scientific concepts (fields of study), ranging from high-level L0 fields like computer science and biology, to more granular L1-L3 fields. L1 includes broader

* For exploratory analysis of cybersecurity relevant publications, see Emerging Technology Observatory blog posts, "Key trends in global cybersecurity research: growth, leaders, dark horses" and "Hot topics in cybersecurity research: insights from the Map of Science" at https://eto.tech/blog/.

subfields such as AI and immunology, while L3 includes narrower subfields like cryptography and differential privacy.

Our previous research updated and expanded MAG's field of study (Toney and Dunham 2022), assigning field scores for the 19 L0 fields to all English-language publications in our corpus. That involved representing field descriptions and publication abstracts and titles in embedding form. We used Wikipedia pages and their academic references to create field text embeddings using a FastText (Bojanowski et al. 2017) model pre-trained on a corpus of scientific literature. Then we computed the cosine similarities between field text embeddings and publication text embeddings to measure the relevance of each field to each publication.

We then expanded our solution to include 1,089 L1-L3 fields (Gelles and Dunham 2024). We did not expand to all L1-L3 fields in the original MAG taxonomy. Instead, we focused on 284 L1 fields and 805 L2 and L3 fields, identified in consultation with subject-matter experts as relevant to emerging technologies of interest. The selected L2 and L3 fields fall under the following L1 fields: artificial intelligence, computer security, semiconductors, genetics, virology, immunology, neuroscience, biotechnology, and bioinformatics.

As before, we took the Wikipedia text and text of the page's citations as a representation of the chosen field. In cases where a field did not have a specific Wikipedia page, we identified sections of related Wikipedia pages to substitute. We used the extracted text to compute our embeddings for each chosen field and use cosine similarity to calculate a similarity score between fields and publications in our corpus. For our 207,231,266 publications, we calculated 230,026,705,260 initial field scores.

Each publication received a field score for each of the 1,108 fields, indicating relevance to each field. We assigned "top fields" to each publication using the three highest scoring fields at each level of the taxonomy (L0-L3).[*] We first identify the top L0 and L1

[*] We consider "top fields" as useful for describing a publication or for assessing the distribution of publications across fields, but note that they are not directly comparable across different fields. For example, a publication assigned a top L0 biology and a publication assigned a top L0 computer science will not necessarily have the same relevance to that L0 field.

fields for a publication, and then within each of those fields, we identify the top L2 and L3 fields. For example, to assign a publication's "top" L2 and L3 fields as cryptography and differential privacy, one of its "top" L0 fields must be computer science and a "top" L1 field must be computer security.* Figure 1 displays the count of publications in our corpus by highest scoring L0 field.

Figure 1. Publication Counts by Top L0 Field of Study

Top Field	Number of Publications
Biology	38,612,800
Medicine	26,121,950
Psychology	17,610,897
Materials science	14,960,555
Chemistry	14,796,493
Computer science	12,753,764
Environmental science	10,761,756
History	10,625,528
Economics	9,871,512
Business	8,461,657
Geology	7,138,889
Sociology	5,792,415
Engineering	5,466,194
Mathematics	5,040,931
Physics	4,711,357
Art	4,674,436
Political science	4,657,120
Philosophy	3,952,267
Geography	1,220,745

Source: CSET merged academic corpus.

* This requirement applies only to the scores for the included fields. Because we did not include all L2 and L3 fields (from the original MAG taxonomy), and rather focused on 805 L2 and L3 fields that fall under a subset of L1 fields, publications did not get a field score for all possible L2 or L3 fields.

We evaluated the field scores in several ways. For one, we expected pairwise cosine similarities for related fields to be closer together. We find that fields like computer science and engineering or business and economics have relatively high cosine similarities, while fields that are less related like biology and political science have low cosine similarities. We also inspect the relative position of fields in the embedding space using t-Distributed Stochastic Neighbor Embedding (t-SNE) to locate the 250-dimensional field embeddings in a 2-D plane. We see intuitive groupings like artificial intelligence and human computer interaction, computer security and computer networks, and computer architecture and operating systems near each other in the t-SNE plot.

Implementation

We deployed our solutions for emerging technology classification and research field scoring over our merged academic corpus. This enables novel analysis at CSET, but also provides the foundation for open data and analytic resources. We provide an open dataset of our emerging technology topic classifications for all publications in OpenAlex. We also provide an open dataset of country research output for our AI-related topics in our Country AI Activity Metrics.* We provide our code and data in a GitHub repository.

We also enable exploration of this data through our interactive online tools. Our Research Almanac displays data on research output and producers in each emerging technology topic. Our Map of Science displays clusters of research which can be filtered according to the emerging technology topics. Our Private-sector AI-Related Activity Tracker (PARAT) includes data on company publishing activity in our AI-related topics, while our Country Activity Tracker (CAT) displays country research output for our AI-related topics.

* These datasets are available on CSET's Emerging Technology Observatory at https://eto.tech/datasets/ and on Zenodo at https://zenodo.org/records/13836025 and https://zenodo.org/records/14183470.

Acknowledgements

This work was supported in part by the Alfred P. Sloan Foundation under Grant No. G-2023-22358. For feedback on earlier drafts, the authors thank Kevin Boyack and Autumn Toney. For editorial assistance, we thank Jason Ly and Shelton Fitch.

Authors

Catherine Aiken is the director of data science and research at CSET.

James Dunham is an NLP engineer at CSET.

Jennifer Melot is the technical lead of the Emerging Technology Observatory at CSET.

Zachary Arnold is the analytic lead of the Emerging Technology Observatory at CSET.

Document Identifier: doi: 10.51593/20240050

References

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré, “Snorkel: Rapid Training Data Creation with Weak Supervision,” *The VLDB Journal* 29, no. 2 (May 2020): 709–30, https://doi.org/10.1007/s00778-019-00552-1.

Andrei Mogoutov and Bernard Kahane, “Data Search Strategy for Science and Technology Emergence: A Scalable and Evolutionary Query for Nanotechnology Tracking,” *Research Policy* 36, no. 6 (July 2007): 893–903, https://doi.org/10.1016/j.respol.2007.02.005.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld, “SPECTER: Document-level Representation Learning using Citation-informed Transformers,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (July 2022): 2270–2282, https://aclanthology.org/2020.acl-main.207/.

Autumn Toney and James Dunham, “Multi-Label Classification of Scientific Research Documents Across Domains and Languages.” *Proceedings of the Third Workshop on Scholarly Document Processing* (October 2022): 105-114, https://aclanthology.org/2022.sdp-1.12.

Autumn Toney, "Creating a Map of Science and Measuring the Role of AI in it," Center for Security and Emerging Technology, June 2021, https://cset.georgetown.edu/publication/creating-a-map-of-science-and-measuring-the-role-of-ai-in-it/.

Caroline Schuerger, Steph Batalis, Katherine Quinn, Ronnie Kinoshita, Owen Daniels, and Anna Puglisi, “Understanding the Global Gain-of-Function Research Landscape,” Center for Security and Emerging Technology, August 2023, https://cset.georgetown.edu/publication/understanding-the-global-gain-of-function-research-landscape/.

Christian Schoeberl, Autumn Toney, and James Dunham, "Identifying AI Research," Center for Security and Emerging Technology, July 2023, https://cset.georgetown.edu/publication/identifying-ai-research/.

Daniel Chou, "Counting AI Research," Center for Security and Emerging Technology, July 2022, https://cset.georgetown.edu/publication/counting-ai-research/.

"Documentation: Research Cluster Dataset," Emerging Technology Observatory, November 22, 2023, https://eto.tech/dataset-docs/mac-clusters/#overview.

Dunham, James, Jennifer Melot, and Dewey Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," arXiv preprint, arXiv:2002.07143v2 (2020), http://arxiv.org/abs/2002.07143.

Henry Small, Kevin W. Boyack, and Richard Klavans, "Identifying Emerging Topics in Science and Technology," *Research Policy* 43, no. 8 (October 2014): 1450–67, https://doi.org/10.1016/j.respol.2014.02.005.

"How we define 'AI safety research' in our tools," Emerging Technology Observatory, June 20, 2023, https://eto.tech/blog/how-we-define-ai-safety-tools/.

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,* Volume 6 (April 2021), https://doi.org/10.3389/frma.2021.630124.

Jochen Gläser, Wolfgang Glänzel, and Andrea Scharnhorst, "Same Data—Different Results? Towards a Comparative Approach to the Identification of Thematic Structures in Science," *Scientometrics* 111, no. 2 (May 2017): 981–98, https://doi.org/10.1007/s11192-017-2296-z.

Juan Pablo Bascur, Suzan Verberne, News Jan van Eck, and Ludo Waltman, "Academic Information Retrieval Using Citation Clusters: In-Depth Evaluation Based on

Systematic Reviews," *Scientometrics* 128 (March 2023): 2895-2921, https://doi.org/10.1007/s11192-023-04681-x.

Kevin W. Boyack and Richard Klavans, "A Comparison of Large-Scale Science Models Based on Textual, Direct Citation and Hybrid Relatedness," *Quantitative Science Studies* 1, no. 4 (December 2020): 1570–1585, https://doi.org/10.1162/qss_a_00085.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld, "S2ORC: The Semantic Scholar Open Research Corpus," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (July 2020): 4969–83, https://doi.org/10.18653/v1/2020.acl-main.447.

Ludo Waltman and Nees Jan van Eck, "A New Methodology for Constructing a Publication-Level Classification System of Science," *Journal of the American Society for Information Science and Technology* 63, no. 12 (November 2012): 2378–2392, https://doi.org/10.1002/asi.22748.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics* 5 (June 2017): 135–146, https://aclanthology.org/Q17-1010.pdf.

P.M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo, "Defining AI in Policy Versus Practice," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (February 2020): 72–78, https://dl.acm.org/doi/10.1145/3375627.3375835.

Priem, Jason, Heather Piwowar, and Richard Orr, "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts," arXiv preprint, arXiv:2205.01833 (2022), https://doi.org/10.48550/arXiv.2205.01833.

Rebecca Gelles and James Dunham, "Multi-Label Field Classification for Scientific Documents using Expert and Crowd-sourced Knowledge," *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia* (2024): 14-20, https://aclanthology.org/2024.wikinlp-1.7/.

Richard Klavans and Kevin W. Boyack, “Using Global Mapping to Create More Accurate Document-Level Maps of Research Fields,” *Journal of the American Society for Information Science and Technology* 62, no. 1 (January 2011): 1–18, https://onlinelibrary.wiley.com/doi/full/10.1002/asi.21444.

Sanjay K. Arora, Alan L. Porter, Jan Youtie, and Philip Shapira. “Capturing New Developments in an Emerging Technology: An Updated Search Strategy for Identifying Nanotechnology Research Outputs,” *Scientometrics* 95, no. 1 (April 2013): 351–70, https://doi.org/10.1007/s11192-012-0903-6.

Tan, Zhen, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu, “Large Language Models for Data Annotation: A Survey,” arXiv preprint, arXiv:2402.13446 (2024), https://arxiv.org/abs/2402.13446.

Toney-Wails, Autumn, Christian Schoeberl, and James Dunham, “AI on AI: Exploring the Utility of GPT as an Expert Annotator of AI Publications,” arXiv preprint, arXiv:2403.09097 (2024), https://arxiv.org/abs/2403.09097.

Ying Huang, Jannik Schuehle, Alan L. Porter, and Jan Youtie, “A Systematic Method to Create Search Strategies for Emerging Technologies Based on the Web of Science: Illustrated for ‘Big Data,’” *Scientometrics* 105, no. 3 (December 2015): 2005–22. https://doi.org/10.1007/s11192-015-1638-y.

Zhang, Jieyu, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner, “A Survey on Programmatic Weak Supervision,” arXiv preprint, arXiv:2202.05433 (2022), https://doi.org/10.48550/arXiv.2202.05433.

Zhihong Shen, Hao Ma, and Kuansan Wang, “A Web-Scale System for Scientific Knowledge Exploration,” *Proceedings of Association for Computational Linguistics 2018, System Demonstrations,* July 2018, https://doi.org/10.18653/v1/P18-4015.

Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang, “SciSciNet: A Large-Scale Open Data Lake for the Science of Science Research” *Scientific Data* 10, no. 1, June 2023, 315, https://doi.org/10.1038/s41597-023-02198-9.