

Tracking AI’s Scientific Anatomy: A Novel Framework for Analyzing the Use and Diffusion of AI in Science

Liangping Ding ^a, Cornelia Lawson ^a, and Philip Shapira ^{a,b,c}

liangping.ding@manchester.ac.uk

0000-0001-6832-4114

cornelia.lawson@manchester.ac.uk

0000-0002-1262-5142

pshapira@manchester.ac.uk

0000-0003-2488-5985

^a Manchester Institute of Innovation Research, Alliance Manchester Business School,
The University of Manchester, Manchester, UK.

^b The Jimmy and Rosalynn Carter School of Public Policy,
Georgia Institute of Technology, Atlanta, USA.

^c Turing Fellow, The Alan Turing Institute, London, UK

Abstract

Artificial intelligence (AI) promises to transform science by accelerating knowledge discovery, automating processes, and introducing new paradigms for research. However, there remains a limited understanding of how AI is being utilized in scientific research. In this paper, we develop a framework based on GPT-4 and SciBERT to identify AI’s role in scientific papers, differentiating between Foundational, Adaptation, Tool and Discussion modes of AI research. This allows us to capture AI’s diverse contributions, from theoretical advances to practical applications and critical analysis. We examine AI’s trajectory across these modes by analyzing time series, field-specific, and country trends. This approach expands on search-term based identification of AI contributions and offers insights into how AI is being deployed in science.

Keywords: Artificial Intelligence; AI in Science; GPT; SciBERT; OpenAlex

1. Introduction

Artificial intelligence (AI) is increasingly lauded as a transformative general-purpose technology, poised to reshape innovation and problem-solving across diverse fields (Crafts, 2021; Duede et al., 2024). The process of innovation has long been understood as a form of knowledge recombination, where novel ideas emerge from the creative reconfiguration of existing knowledge elements (Schumpeter, 1934 [1983]). AI promises to significantly extend this recombinatory potential by offering new methodological opportunities. Researchers and practitioners are increasingly expanding their toolkits to incorporate AI approaches such as machine learning, natural language processing, and generative models, which allow for the analysis of very large, complex, and diverse datasets (Von Krogh et al., 2023).

A growing body of literature has begun to consider the consequences and impacts of increased AI use in research. The potential for AI to accelerate scientific discovery is meaningful, with the OECD (2023) noting its capacity to automate routine tasks and introduce new paradigms for inquiry. However, relatively few studies to date draw on observable outcomes to infer how AI adoption is reflected in research activities. At least some scientists at the frontier of knowledge production may be cautious about integrating AI into their research workflows. The introduction of a new technology like AI might both consolidate prior knowledge within a field and destabilize its trajectory (Chen et al., 2021). This coexistence of consolidation and destabilization makes it essential to understand whether AI constitutes a technology with heterogeneous effects across fields and over time.

Importantly, AI's use in research is not confined to computer and data sciences but is being developed and applied across the natural sciences, engineering, medicine, agricultural sciences, the social sciences, and the humanities (Duede et al., 2024). Large-scale bibliometric studies have found that publications deemed to be AI-relevant have higher novelty and attract more citations than non-AI works (Bianchini et al., 2022) and that scientists who publish AI-relevant research are more productive, more highly cited, and progress faster to last-author positioning than their non-AI-using peers (Hao et al., 2025). These findings suggest that AI is correlated with productivity gains and career acceleration.

However, we argue that such studies face a critical methodological shortcoming: they often rely on keyword-based bibliometric methods that fail to distinguish between research that adopts or develops AI technologies and research that discusses AI without using it as a method or tool. In some cases, researchers looking for AI's impacts focus only on papers in the natural sciences. In fact, we argue that some of these publications may not involve the actual use of AI. For instance, Ding et al. (2025) find that more than 10 percent of AI-related publications contribute to scholarly discourse on its ethics, risks, and societal implications, but do not implement AI methods empirically. In short, the literature lacks a systematic framework to classify the distinct modes of AI use—such as whether AI is being developed, adapted, applied as a tool, or discussed critically. This gap hinders a precise assessment of AI's true scientific anatomy and its implications for research practices, productivity, and governance, which could lead to analytical inaccuracies. Mapping the extent of AI use in science, moreover, presents a foundational step toward addressing broader questions about its implications for research performance, reproducibility, the scientific labor force, and responsible research.

This paper proposes a systematic approach for assessing how AI use is reflected in scientific publications, by combining traditional bibliometric techniques with recent advances in language models. We develop a two-stage framework to track AI's scientific anatomy that identifies AI-relevant publications and classifies the role of AI in scientific research into distinct modes. Building on our prior work (Ding et al., 2025), which distinguished papers that apply AI techniques from those engaging primarily in conceptual or ethical discussions, we extend the framework to capture more nuanced modes of AI use – namely, foundational, adaptive, tool-based, and discussion modes – to better reflect the manifold roles of AI in science.

We offer our approach as a transparent, openly available, and accountable public tool. The code is accessible at <https://github.com/liangpingding/ai-in-the-lab>, supporting validation, reproducibility, knowledge exchange, and further advancement.

2. Data and Methodology

In this section, we provide a detailed explanation of our proposed two-stage framework and outline the details of model architecture and data collection.

2.1 Task formulation

To track the scientific anatomy of artificial intelligence, we develop a two-stage framework. The first stage identifies AI-relevant scientific publications, while the second classifies these publications into four distinct modes of AI use.

In this study, we define AI-related publications as whether a scientific paper is relevant to artificial intelligence technology. Prior studies mainly use a bibliometric keyword strategy (OECD, 2023). One of the drawbacks of the keyword-based method is that it introduces false positives due to polysemy and might miss false negatives because of an evolving terminology that makes it unfeasible to cover all the relevant keywords. For example, “neural network” may refer to an AI algorithm but could also refer to the interconnected neurons in neuroscience. To mitigate the problem of false positives introduced by bibliometric keyword search, we develop an automatic classifier based on large language model to filter out the non-AI relevant publications.

Following this identification, we distinguish four modes of AI use in scientific publications. The criteria underpinning this framework can be conceptualized as follows:

(1) **AI as Topic of Discussion.** Papers that primarily engage with conceptual, societal, or ethical aspects of AI—without implementing AI methods—are classified as *Discussion*. The release of generative AI in 2022 has sparked wide debate about its implications for research productivity, novelty, reproducibility, intellectual property rights, workforce dynamics, risk, bias, and ethics (Mukherjee & Chang, 2024). Such papers advance discourse on AI but do not include technical or applied AI implementation.

(2) **Advancing AI.** Papers that develop AI models are classified as *Foundational* where they contribute to the technical progress of AI itself (e.g., developing new architectures, algorithms, or optimization strategies).

(3) **Use of Existing AI Techniques.** Papers that use existing AI technologies without architectural changes are classified as *Tool*. They translate AI methods into other scientific fields with the aim of accelerating discovery and innovation. This aligns with observations that AI not only evolves as a general-purpose technology but also diffuses as an enabling tool across diverse disciplines (Cockburn et al., 2018).

(4) **Modification for Use in Domain Science.** By contrast, papers that refine or extend AI models are more than mere tools. They adapt AI techniques to specific research requirements, which we classify as *Adaptation*. This distinction echoes the difference between “technology-push” and “application-pull” innovation pathways (Nelson & Winter, 1985; Rosenberg, 1982) and reflects broader innovation dynamics where invention is largely driven by recombination and refinement (Strumsky & Lobo, 2015).

Based on these criteria, we define four modes of AI use in science. *Foundational* includes publications that propose new AI models, algorithms, training techniques, or theoretical advancements aimed at improving the core capabilities of AI. *Adaptation* refers to studies that modify or extend existing AI models for specific tasks or domains, without introducing fundamentally new techniques. *Tool* describes research in which AI is employed as an

application or instrument to solve domain-specific problems, without altering its underlying architecture. *Discussion* captures publications that engage with AI—through literature reviews, ethical and societal commentary, interviews, perception studies, or bibliometric analyses—without integrating AI methods into their research methodology.

Technically, we formulate the task of tracking AI’s scientific anatomy as a two-stage text classification task. In the first stage, given a sequence of tokens $X = [x_1, x_2, \dots, x_n]$ where n is the length of the sequence, the goal of model is to predict the corresponding label Y , s. t. $Y \in A$, where A is the label set for AI relevance consisting of two classes: AI-related and non-AI related. The annotated dataset with K samples can be regarded as a set of pairs of token sequence and label sequence: $\mathbf{D}_1 = \{(X^{(k)}, Y^{(k)})\}_{k=1}^K$, where $(X^{(k)}, Y^{(k)})$ is the k -th instance from dataset \mathbf{D}_1 . In the second stage, similarly, for the given sentence of tokens X , the goal of model is to predict the label Z , s. t. $Z \in B$, where B is the label set for AI modes consisting of four classes: Foundational, Adaptation, Tool and Discussion. The annotated dataset with M samples can be defined as $\mathbf{D}_2 = \{(X^{(k)}, Y^{(k)})\}_{k=1}^M$.

2.2 Model Architecture

To ensure robust model performance, we employ state-of-the-art large language models—GPT and SciBERT—as the backbone of a unified GPT–SciBERT pipeline, which is applied across both stages of our framework. The rationale for adopting this pipeline is considered below.

The Generative Pre-trained Transformer (GPT) (OpenAI, 2023) series represents a leading family of large-scale generative AI models built on the Transformer architecture. By leveraging self-attention mechanisms, GPT models capture long-range dependencies in text, producing coherent and contextually relevant outputs. Pretrained on massive corpora through unsupervised learning, GPT can be fine-tuned for downstream tasks or applied in zero-shot and few-shot settings, making it highly versatile for natural language generation, summarization, translation, and classification (De Kok, 2025).

SciBERT (Beltagy et al., 2019) is a domain-specific variant of BERT (Devlin et al., 2019), originally trained on general-purpose corpora such as Wikipedia and BookCorpus. In contrast, SciBERT was pretrained on over one million full-text scientific articles from the Semantic Scholar corpus, spanning a broad range of disciplines including computer science and biomedicine. This domain-oriented pretraining enables SciBERT to capture the specialized vocabulary, syntactic patterns, and conceptual relationships that characterize scientific writing. As our task involves the classification of scientific publications, fine-tuning SciBERT provides superior contextual understanding and semantic alignment between the model’s pretraining data and our target corpus.

The pipeline approach offers both practical and methodological advantages. It eliminates the need for large volumes of human-annotated training data, instead leveraging high-quality synthetic labels generated by GPT—a method shown to be effective for text classification (Bucher & Martini, 2024). Moreover, this design is cost- and resource-efficient, while also mitigating some of the limitations of generative models. By combining GPT’s generative strengths with SciBERT’s domain specialization, the pipeline reduces randomness and potential hallucinations in GPT outputs, thereby supporting greater scientific rigor and reproducibility.

2.3 Dataset construction

As is standard practice in supervised machine learning, the use of a training set, development set, and test set is essential to ensure scientific rigor and model reliability. In this study, we construct synthetic annotated training set using GPT-4o for both AI relevance classification stage and AI mode classification stage. To validate and benchmark model

performance, we also develop a human-annotated dataset that serves as a gold standard. This dataset is split into a development set—used for tuning model parameters—and a test set, which is reserved for evaluating the final performance of the classification models.

2.3.1 Stage I: AI relevance classification

The construction of the automatic classifier for AI relevance classification relies on annotated data comprising both AI and non-AI publications in the training set. To obtain publication data, we employ public and freely available database OpenAlex as our primary data source, leveraging its extensive repository of over 260 million scientific works, including journal articles, conference proceedings, preprints, and other scholarly outputs. Our dataset is drawn from the May 2025 snapshot of OpenAlex.

As the field of AI is evolving rapidly with new AI models and algorithms coming out very often, which makes it hard to construct an exhaustive list of AI relevant keywords. And currently, no consensus has been reached on the bibliometrics definition to quantify the boundary of AI. Therefore, we adopt a bibliometric definition of AI following prior studies (Ding et al., 2025; Liu et al., 2021; Van Noorden & Perkel, 2023). AI-related search terms from these sources are aggregated to construct a comprehensive keyword set for further step. We then query OpenAlex to identify AI-relevant publications by searching for these keywords in titles and abstracts. To implement this search, we design a Boolean query that filters SQL records based on exact and partial string matches. We employ the LIKE operator with the wildcard character % to capture variations in terms (e.g., “machine learning%” matches both “machine learning” and “machine learnings”). Both LIKE (case-sensitive) and ILIKE (case-insensitive) operators are used to account for differences in term capitalization. This approach yields 2,679,653 AI-related publications.

While OpenAlex is a comprehensive and openly available bibliographic database, it is not without issues such as missing records and duplicate entries (Zhang et al., 2024; Zheng et al., 2025). To address these limitations, we conduct a series of data preprocessing and cleaning steps. The key preprocessing steps are: (1) retain only English language publications; (2) drop records missing a DOI; (3) retain only article type; (4) exclude papers marked as retracted; (5) exclude comments, editorials, or similar non-research content; (6) drop records with missing titles or missing abstracts; (7) eliminate duplicates based on title, abstract, and DOI; and (8) retain publications published before 2025. Following this cleaning process, we obtain a refined dataset with 1,605,453 AI relevant publications.

To construct the training data, we draw stratified random samples of both AI and non-AI publications across 26 OpenAlex scientific fields based on OpenAlex topic feature, with a maximum threshold of 1,000 publications per field to ensure balanced disciplinary coverage. The dataset is sampled from publications containing both titles and abstracts to ensure sufficient textual content for reliable classification. Sampling rates differ by group to reflect underlying distributions: 1% for AI publications and 0.01% for non-AI publications. This process yields a total of 20,057 publications, consisting of 10,919 AI and 9,138 non-AI samples. The field distribution for the sampled dataset is shown in Appendix Table A1.

Using GPT-4o, we generate synthetic relevance labels based on the concatenation of title and abstract of the publications using zero-shot prompting, producing a high-quality training set. The prompt used for AI relevance classification is shown in Appendix A1. In the comparison between keyword-based labels and GPT-generated synthetic labels within the 20,057-sample dataset, 914 publications retrieved by AI keywords were classified as non-AI-relevant by GPT, while 197 non-AI publications identified through the keyword search were predicted as AI-relevant by GPT. To improve the precision of the resulting labels, we retain only those publications for which GPT and the keyword method produce the same label. This allowed us to construct a high-confidence training dataset without relying on costly human

annotation. This results in a final synthetic training set of 18,946 publications, comprising 10,005 AI and 8,941 non-AI publications.

To benchmark and validate model performance, we construct a smaller human-annotated dataset. From the pre-processed AI-relevant corpus, we select publications with abstracts between 100 and 500 words to ensure sufficient content for human judgment and comparability with model performance, given that the maximum sequence length for SciBERT is 512 tokens. A stratified sample of 582 publications across 26 scientific fields is drawn, and two independent annotators assess their AI relevance. Inter-rater reliability is substantial, with a 94% agreement rate and a Cohen’s kappa coefficient of 0.698. For the 36 inconsistent cases, a third annotator adjudicates the final label. The dataset is divided into development and test sets at a 1:2 ratio, and the distribution of the training, development, and test sets for AI relevance classification is reported in Table 1. Note that the training, development, and test sets are mutually exclusive.

<Table 1 about here>

2.3.2 Stage II: AI modes classification

The second stage of the framework focuses on classifying AI-relevant publications into four modes that reflect the nature of AI’s role in the research. To construct the training data for AI modes classification, we build on the training data from AI relevance classification stage and retain the 10,005 publications identified as AI-relevant. To generate high-quality synthetic label for the training data, we use GPT-4o to assign each publication to one of our four categories: *Foundational*, *Adaptation*, *Tool* and *Discussion*, and include a fifth category, *Unclear*, to allow the model to abstain in cases of ambiguity, thus mitigating potential hallucination effects common in generative models. The prompt used for AI mode classification is shown in Appendix A2. To address class imbalance and reduce noise from ambiguous cases, we remove all publications in the *Unclear* category and randomly downsample the *Tool* category to 2,000 publications. This results in a balanced and cleaner dataset better suited for model training. After these adjustments, the training dataset used for fine-tuning SciBERT in the AI mode classification step comprises 6,121 publications.

To validate the model performance for AI mode classification, we create a human annotated dataset building on the 530 AI relevant human annotated publications. Similarly, two human annotators conducted the first round of annotation, the consistent rate is 74% and reached a Cohen’s kappa coefficient of 0.635. For the 136 inconsistent cases, a third annotator made the final decision. The dataset is divided into development and test sets at a ratio of 1:2, and the distribution of the training, development, and test sets for AI mode classification is presented in Table 2.

<Table 2 about here>

2.4 Evaluation metrics

To evaluate model performance, we adopt standard machine learning evaluation metrics: precision (P), recall (R), and F1-score (F1). These metrics are defined as follows:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

where *TP* denotes true positives, *FP* false positives, and *FN* false negatives.

For the AI relevance classification task, we report precision, recall, and F1-score only for the positive class (AI-relevant publications). This emphasis reflects the primary objective of the task, which is to minimize false positives, and is further justified by the class imbalance in the human-annotated dataset, where AI-relevant publications represent the majority of samples.

For the AI mode classification task, we report macro-averaged precision, recall, and F1-score. Macro-averaging calculates the metric independently for each class and then averages the results, treating all classes equally regardless of their size. This approach is appropriate for multi-class settings where class distributions may be imbalanced, ensuring a balanced evaluation across all categories.

3. Experiments and Results

For both the AI relevance classification and AI mode classification stages, we conduct experiments using a GPT-SciBERT pipeline, in which SciBERT is fine-tuned on GPT-generated synthetic data, and model performance is evaluated against the human-annotated dataset. The trained AI relevance classification model is then applied to the keyword-retrieved corpus to filter out false positives, while the AI mode classification model is applied to the filtered AI-relevant publications after Stage I to classify them into four AI modes.

3.1 Experimental setting

In experiments using GPT to generate synthetic data, to minimize randomness and reduce the likelihood of hallucinations, the temperature parameter is set to 0, constraining the model to produce stable and reproducible predictions and SciBERT is then fine-tuned on those synthetic datasets. SciBERT is initialized with weights pre-trained on a large corpus of scientific text and optimized using a cross-entropy loss function. Input text consists of the concatenated title and abstract of each publication, tokenized according to the SciBERT vocabulary, with a maximum sequence length of 512 tokens. We use grid search to tune the key hyperparameters learning rate in range $\{1e-5, 2e-5, 5e-5, 1e-4\}$ and the number of training epochs in range $\{10, 20, 30, 50, 100\}$ for both stages. We find that setting learning rate as $1e-4$ with 20 epochs achieves the best F1 score on the development set for AI relevance classification model, and setting learning rate as $1e-5$ with 100 epochs achieves the best F1 score on the development set for AI mode classification model. These best-performing hyperparameters are applied in the final training. Specifically, the AI relevance classification model is trained with early stopping using the AdamW optimizer with an initial learning rate of $1e-4$, employing a linear learning rate decay schedule that gradually reduces the learning rate throughout training, and up to 20 epochs. The AI mode classification model is trained under similar conditions, with an initial learning rate of $1e-5$ and up to 100 epochs with early stopping. All experiments are implemented on a Nvidia L40s GPU and batch size for training is set to 48.

3.2 Model Performance

Table 3 presents the performance of keyword search, GPT, and SciBERT on the AI relevance classification task. The keyword-based approach achieves but relatively low precision (0.89 in development, 0.93 in test), reflecting its tendency to include false positives. In contrast, both GPT and SciBERT demonstrate higher precision while maintaining strong recall. GPT achieves the best overall performance with an F1-score of 0.98 on both the development and test sets, followed closely by SciBERT with F1-scores of 0.97. These results confirm that combining GPT-generated synthetic labels with SciBERT fine-tuning effectively reduces false positives while preserving high recall, thereby improving the quality of AI relevance classification. In addition, we also assess the potential false negative rate, that is, the proportion of AI-relevant publications that may have been missed by the keyword search. To do so, we randomly sample 0.1% of publications from each of the 26 OpenAlex fields, exclude records with missing titles

or abstracts, and obtain a set of 112,163 publications. Applying the AI relevance SciBERT model to this sample yields a false negative rate of 1% (1,059 out of 112,163), suggesting that our keyword approach captures the vast majority of AI publications while missing only a small fraction.

<Table 3 about here>

Table 4 reports the performance of GPT and SciBERT on the multi-class AI mode classification task. Overall, performance is lower than in the binary AI relevance classification, reflecting the increased difficulty of distinguishing between four modes of AI use. GPT outperforms SciBERT across all metrics, achieving macro F1-scores of 0.83 (development) and 0.83 (test), compared to SciBERT’s 0.78 (development) and 0.76 (test). The confusion matrix (Figure 1) further highlights the classification challenges. The *Tool* category is most accurately identified, with 185 correct predictions and F1-core of 0.88. However, *Discussion* is more frequently misclassified as *Tool*, and *Adaptation* papers occasionally overlap with *Tool* (32 misclassifications). The small sample size of *Foundational* papers limits their classification accuracy, which is largely influenced by performance on the other categories. This suggests that while the model effectively recognizes the more prevalent categories, greater ambiguity exists at the boundary between *Foundational* and *Adaptation*, as well as between *Discussion* and AI application categories.

<Table 4 about here>

<Figure 1 about here>

4. Findings

To track AI’s scientific anatomy, we applied our two-stage framework to AI relevant publications retrieved from AI keyword search to infer AI relevance and AI mode. To ensure the reliability of the predictions for AI mode classification model, we set the decision threshold for classification at a confidence score of 0.5 where predictions with a maximum class probability below this threshold are assigned to an *Unclear* category. This post-processing step ensures that low-confidence predictions are flagged for uncertainty rather than misclassified. After excluding non-AI publications predicted by SciBERT and those with “*Unclear*” AI mode classifications, 1,506,550 AI publications remained. Furthermore, we limit the dataset to publications from 2004–2024 to mitigate distortions resulting from IEEE’s digitization project between 2001 and 2003, which results in a total of 1,433,159 publications¹.

A wide range of analyses can be conducted using the framework we have developed. In this paper, we present illustrative results that offer insights from research management and policy perspectives, including temporal trends in AI publications by mode, field-level distributions of AI modes, and cross-country comparisons of AI modes. Zhang et al. (2024) noted frequent missing institution data in the OpenAlex database, which can hamper such comparison. We retain 1,075,782 publications with complete institutional information, including institution ID, name, and country code for analysis. This represents 75 per cent of classified AI publications published in 2004-2024.

Figure 2 presents the annual trends in the number and proportion of AI publications between 2004 and 2024. *Foundational* papers are a signal of the pace of science in developing underlying core AI models and techniques. The number of *Foundational* papers in AI published

¹ We restrict our analysis to this period as the number of AI publications before the early 2000s is small. There is a temporary upward blip in 2002–2003 that largely reflects IEEE’s digitization project (IEEE Timeline from 1984 to 2012, 2025). This made previously published works available on IEEE Xplore rather than representing genuine contemporary growth in AI research. We thus start our analysis from 2004.

worldwide was fairly stable through much of the 2000s to the mid-2010s, beginning to pick up at a modestly increasing rate from 2017 through to 2023. Due to indexing delays in bibliographic databases (Adeosun, 2024), the 2024 data remain incomplete, explaining the apparent decline in that year. There is also a noticeable exponential and sustained growth in *Adaptation* and *Tool* papers beginning around 2016. This is important in signaling ways in which scientists are incrementally improving AI models and apply AI as a tool for science in their labs. Finally, *Discussion* papers also increased from around 2017, with a marked uptick from 2022. This is significant from a societal perspective as it corresponds with the emergence of generative AI, increased concerns about AI ethics and risk, and increased attention to AI's managerial, educational, and governance implications.

The lower panel of Figure 2 reports the share of AI modes by year, i.e., the proportion of each mode relative to total AI publications in that year. Three insights stand out. First, the share of *Foundational* papers over all AI papers has declined over time, suggesting that relative contributions to advancing core AI methods are becoming proportionally smaller. Second, the share of *Discussion* papers has risen sharply since 2018, reflecting the growing role of conceptual and ethical debates. Third, *Tool* papers remain the dominant category, however, *Adaptation* papers exhibit a steady upward trajectory, and the gap between *Adaptation* and *Tool* papers is gradually narrowing, signaling the increasing modification of AI models to enhance their adaptability for domain-specific problem-solving. Together, these patterns illustrate a structural shift: AI research is moving from a phase dominated by core methodological advances to one where application and societal reflection play a growing role.

<Figure 2 about here>

Figure 3 presents the distribution of AI publications across scientific fields and the relative shares of different AI modes within each of 26 OpenAlex fields. As we can see, Computer Science overwhelmingly dominates the AI publication landscape, accounting for 38.4% of all AI-related works, followed by Engineering (24.6%) and Medicine (8.8%). Other fields such as Environmental Science, Neuroscience, and Biochemistry contribute smaller but still visible shares, while fields including the Chemistry, Arts and Humanities account for less than 1% each. Clear differences also emerge in *how* AI is used across fields. *Tool* papers are consistently the most common category across the sciences, underscoring the centrality of applying existing AI models as instruments for domain-specific research. *Adaptation* papers account for notable shares in technical fields such as Engineering, Neuroscience, and Physics and Astronomy, reflecting that researchers in these fields often possess both domain expertise and AI knowledge to modify existing models, or are more inclined to collaborate with AI specialists. *Foundational* papers are concentrated in Computer Science, reflecting its role as the primary locus of core AI model and algorithm development. By contrast, the *Discussion* mode is more prevalent in the Arts and Humanities, Social Sciences, Business, and related fields. In these fields, AI is often addressed conceptually, ethically, or theoretically, rather than being implemented methodologically. This pattern underscores the broader societal and epistemic engagement with AI beyond technical development and application. Collectively, these findings highlight the heterogeneous role of AI across disciplines: while technical sciences focus on developing and applying AI models, the social sciences and humanities tend to engage with AI as an object of analysis and debate.

<Figure 3 about here>

Figure 4 displays the distribution of publications across four AI modes within the 20 leading countries ranked by AI-relevant publications. In *Foundational* research, China (28.4%) and the

United States (19.8%) dominate, together accounting for nearly half of global output, with smaller but non-negligible contributions from India, the United Kingdom, Germany and Japan. In *Adaptation* research, China (36%) again holds a leading position, followed by the United States (12.5%), India (5.4%), and several European countries, reflecting a strong emphasis on integrating AI into disciplinary problem-solving. *Tool* papers show a more balanced global distribution. China (16.9%) and the United States (15.8%) contribute equally, while India (7.4%), the United Kingdom (4.3%), and Germany (3.6%) also make notable contributions. This pattern suggests that applying AI as a methodological instrument is a widely shared activity across countries, not limited to the dominant AI producers. By contrast, *Discussion* papers are disproportionately concentrated in the United States (18.8%) and China (11.1%), followed by the United Kingdom (7.6%) and India (6.4%). This distribution highlights how leading scientific nations are also shaping global debates around AI's societal, ethical, and governance implications. Collectively, these results suggest that while foundational and adaptation research are concentrated in a small number of countries, tool-based applications and conceptual discussions of AI are more globally dispersed, underscoring varying modes of engagement with AI across national research systems.

<Figure 4 about here>

Figure 5 presents radar charts for six representative disciplines for the top five AI-producing countries (China, United States, India, United Kingdom, and Germany). Unlike the global distribution shown in Figure 4, this figure highlights how each country strategically emphasizes different AI modes within specific disciplines. Several striking contrasts emerge. In Computer Science, China leads in *Adaptation* research (64.2%), while the United States contributes the highest share of *Foundational* work (17.6%) and the United Kingdom leads in *Discussion*-oriented papers (13.3%). India is especially prominent in *Tool* use (43%), reflecting a strong focus on applying existing AI models. In Engineering, China dominates *Adaptation* (58.76%), whereas India is the largest contributor of *Tool* papers (66.2%), and the United States has the largest share of *Foundational* contributions (5.4%). In more application-driven fields, *Tool* papers dominate across countries: Germany leads in Biochemistry (69.2%) and Medicine (70.4%), while India holds the highest *Tool* share in Materials Science (82.3%). *Adaptation*, however, remains most prominent for China in these fields (e.g., 48.3% in Biochemistry, 59.7% in Medicine). The Social Sciences present a different profile: the United Kingdom has a striking prominence of *Discussion* papers (62.5%), underscoring its strong role in ethical, conceptual, and societal debates around AI. By contrast, China continues to emphasize *Adaptation* (33.7%), and India dominates *Tool* usage (47.6%). Collectively, these results demonstrate that while China is consistently strong in *Adaptation*, the United States and the United Kingdom retain leadership in *Foundational* and *Discussion* contributions, and Germany and India excel in *Tool* applications across various fields. This reflects differentiated national strategies for engaging with AI across the sciences and social sciences.

<Figure 5 about here>

5. Discussion

AI is now widely discussed, yet systematic understanding remains limited on the impacts of AI in and for science as well as on the reciprocal role of science in shaping AI's development. The two-stage framework proposed in this study provides a novel perspective for uncovering the anatomy of AI's role in scientific research. We demonstrate the potential of free or low-cost generative AI and transformer models to analyze documentary sources at scale, offering a replicable approach for mapping the integration of AI into science. Going forward, similar AI-

driven classification techniques could be extended to finer-grained categorizations (e.g., by types of AI models and algorithms), regional and institutional analyses, and linkages with measures of scientific productivity, novelty, and disruption. Understanding the modes of AI use in science is critical for a wide range of stakeholders. For policymakers, it informs strategic investment in AI infrastructure and talent development. For research funders and managers, it enables benchmarking of institutional and disciplinary performance. For scientists, it reveals methodological trends and potential opportunities for collaboration across fields.

As science itself becomes increasingly reliant on AI, it is important that national resource commitments to scientific research are appropriately targeted. Strengthening foundational AI research is one critical priority—both to develop frontier AI capabilities and to train future cohorts of AI researchers and developers, thereby equipping universities with the expertise needed to attract industry and sustain global leadership. However, while frontier AI research often dominates policy discourse, adaptation and tool-based uses of AI are equally important. It is through these processes of modification and deployment that AI-enabled discoveries will be realized across diverse fields of science, generating broader economic and societal benefits.

Our results show that current scientific practice relies heavily on reusing existing AI models as tools, with relatively fewer efforts devoted to foundational advances. This raises an important question: has AI reached a saturation point where existing capabilities are sufficient to sustain innovation across science and society, or does the decline in foundational contributions risk undermining the ecosystem in the longer term? Addressing this question requires not only attention to frontier development but also recognition that all disciplines – not limited to the social sciences—should devote proportionate effort to reflecting on the economic, societal, and ethical implications of AI.

6. Conclusion

In this study, we present a two-stage framework to disentangle nuanced modes of AI use – namely, *foundational*, *adaptation*, *tool*, and *discussion* – to better capture the multifaceted role of AI in science. The framework is made available as a public tool that can be validated, refined and further extended using free or low-cost data sources and AI tools.

Several challenges were encountered in developing the proposed approach and framework. These included defining the scope of the AI domain in science (addressed by drawing on recent definitional sources), securing an up-to-date, open-access, and comprehensive publication database (addressed through the use of OpenAlex, including a locally installed snapshot), and cleaning and deduplicating the data (for which AI-enabled methods were deployed). Constraints on computational resources and concerns about the reproducibility of generative AI outputs were mitigated by adopting a two-stage strategy: generative AI was used to construct the training set, while SciBERT was employed for large-scale classification. Our results show that the GPT-SciBERT pipeline performs robustly for AI relevance classification, effectively filtering out false positives. For AI mode classification, performance is promising but highlights challenges in differentiating conceptually adjacent categories. Several limitations, however, should be acknowledged.

(1) AI relevance classification. First, the classification relies solely on titles and abstracts, which may not always provide sufficient information to accurately determine whether a publication is AI-related. Second, we exclude statistical methods such as regression and principal component analysis (PCA) from the AI category, although they are sometimes used in similar analytical contexts. Third, in disciplines such as cognitive science and neuroscience, terms like *neural network*, *reinforcement learning*, and *attention* are employed in theoretical or biological contexts rather than in connection with AI techniques. Although such works may contribute conceptual foundations to AI, they are not treated as AI-relevant in this study.

(2) **AI mode classification.** Because classification is based only on titles and abstracts, important methodological details may be missing, which can lead to misclassification. Moreover, some authors may exaggerate claims about novelty, making it difficult to assess whether genuine advances in AI algorithms have been achieved. Finally, due to the inherently interdisciplinary nature of AI, it is sometimes challenging to distinguish whether a publication's primary focus lies in advancing AI itself or in applying AI to address domain-specific problems.

(3) **Assessment on AI's scientific anatomy.** Additional limitations in assessing AI's scientific anatomy include restriction to English-language publications (to enable classification), potential biases induced by excluded publications (where data was missing for classification), and imperfections in AI-enabled classification.

Ethical and Legal Considerations

Our framework uses open-source public data and publicly available AI tools. All data used is in the public domain under OpenAlex's CC0 license (free to use, no rights reserved). No personal or sensitive information is collected or processed, and the work complies with UK General Data Protection Regulation (UK GDPR) and other applicable data protection laws. As a responsible research practice, we have been especially concerned to reduce errors in the data, to be aware of potential bias, and to be transparent about methods and analytical steps. The code used for analysis is accessible at <https://github.com/liangpingding/ai-in-the-lab>,

Acknowledgments

We acknowledge feedback received from the 2025 Global Tech Mining conference in Atlanta, US, and the AI in Government & Academia Summit 2025, Manchester Metropolitan University, Manchester, UK. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the authors' affiliated organizations.

Author contributions

Liangping Ding: Conceptualization; Methodology; Data curation; Formal Analysis; Writing.
Cornelia Lawson and Philip Shapira: Conceptualization; Methodology; Writing; Supervision; Funding acquisition

Competing interests

The authors declare no competing interests.

Funding information

This work was supported in part by the Project on Innovations in the Lab: Leveraging Transformations in Science, FHUMS Large Collaborative Grant, University of Manchester, UK

References

- Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: a pretrained language model for scientific text. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3615–3620). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1371>
- Bucher, M. J. J., & Martini, M. (2024). *Fine-Tuned “Small” LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification* (No. arXiv:2406.08660). arXiv. <https://doi.org/10.48550/arXiv.2406.08660>
- Chen, J., Shao, D., & Fan, S. (2021). Destabilization and consolidation: Conceptualizing, measuring, and validating the dual characteristics of technology. *Research Policy*, 50(1), 104115. <https://doi.org/10.1016/j.respol.2020.104115>
- Cockburn, I. M., Henderson, R., & Stern, S. (2018). The impact of artificial intelligence on innovation: An exploratory analysis. In: *The Economics of Artificial Intelligence: An agenda* (pp. 115–146). University of Chicago Press.
- Crafts, N. (2021). Artificial intelligence as a General-purpose Technology: An Historical Perspective. *Oxford Review of Economic Policy*, 37(3), 521–536. <https://doi.org/10.1093/oxrep/grab012>
- De Kok, T. (2025). ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research. *Management Science*, mnscl.2023.03253. <https://doi.org/10.1287/mnscl.2023.03253>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [Cs]*. <http://arxiv.org/abs/1810.04805>
- Ding, L., Lawson, C., & Shapira, P. (2025). Rise of Generative Artificial Intelligence in Science. *Scientometrics*. <https://doi.org/10.1007/s11192-025-05413-z>
- Duede, E., Dolan, W., Bauer, A., Foster, I., & Lakhani, K. (2024). *Oil & Water? Diffusion of AI Within and Across Scientific Fields* (No. arXiv:2405.15828). arXiv. <http://arxiv.org/abs/2405.15828>
- Hao, Q., Xu, F., Li, Y., & Evans, J. (2025). *AI Expands Scientists’ Impact but Contracts Science’s Focus* (No. arXiv:2412.07727). arXiv. <https://doi.org/10.48550/arXiv.2412.07727>
- IEEE Timeline from 1984 to 2012. (2025, October 4). ETHW. https://ethw.org/IEEE_Timeline_from_1984_to_2012
- Liu, N., Shapira, P., & Yue, X. (2021). Tracking developments in artificial intelligence research: Constructing and applying a new search strategy. *Scientometrics*, 126(4), 3153–3192. <https://doi.org/10.1007/s11192-021-03868-4>
- Mukherjee, A., & Chang, H. H. (2024). *AI Knowledge and Reasoning: Emulating Expert Creativity in Scientific Research* (No. arXiv:2404.04436). arXiv. <http://arxiv.org/abs/2404.04436>
- Nelson, R. R., & Winter, S. G. (1985). *An Evolutionary Theory of Economic Change*. Cambridge, MA: Harvard University Press.
- OECD. (2023). *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*. OECD. <https://doi.org/10.1787/a8d820bd-en>
- OpenAI. (2023). *GPT-4 Technical Report* (No. arXiv:2303.08774). arXiv. <http://arxiv.org/abs/2303.08774>
- Rosenberg, N. (1982). *Inside the Black Box: Technology and Economics*. Cambridge: Cambridge University Press.

- Schumpeter, J. A. (1934 [1983]). *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*. Translated by R. Opie. New Brunswick, NJ: Transaction Publishers.
- Strumsky, D., & Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44(8), 1445–1461. <https://doi.org/10.1016/j.respol.2015.05.008>
- Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature*, 621(7980), 672–675. <https://doi.org/10.1038/d41586-023-02980-0>
- Von Krogh, G., Roberson, Q., & Gruber, M. (2023). Recognizing and Utilizing Novel Research Opportunities with Artificial Intelligence. *Academy of Management Journal*, 66(2), 367–373. <https://doi.org/10.5465/amj.2023.4002>
- Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*. <https://doi.org/10.1007/s11192-023-04923-y>
- Zheng, M., Miao, L., Bu, Y., & Larivière, V. (2025). Understanding discrepancies in the coverage of OpenAlex: The case of China. *Journal of the Association for Information Science and Technology*, asi.70013. <https://doi.org/10.1002/asi.70013>

Figure. 1. Confusion matrix for AI mode classification model on test set

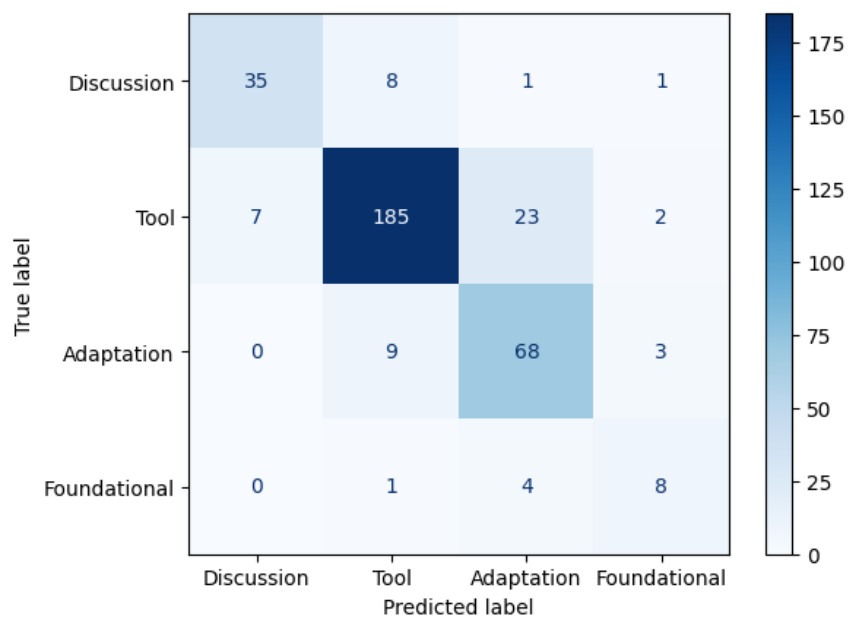
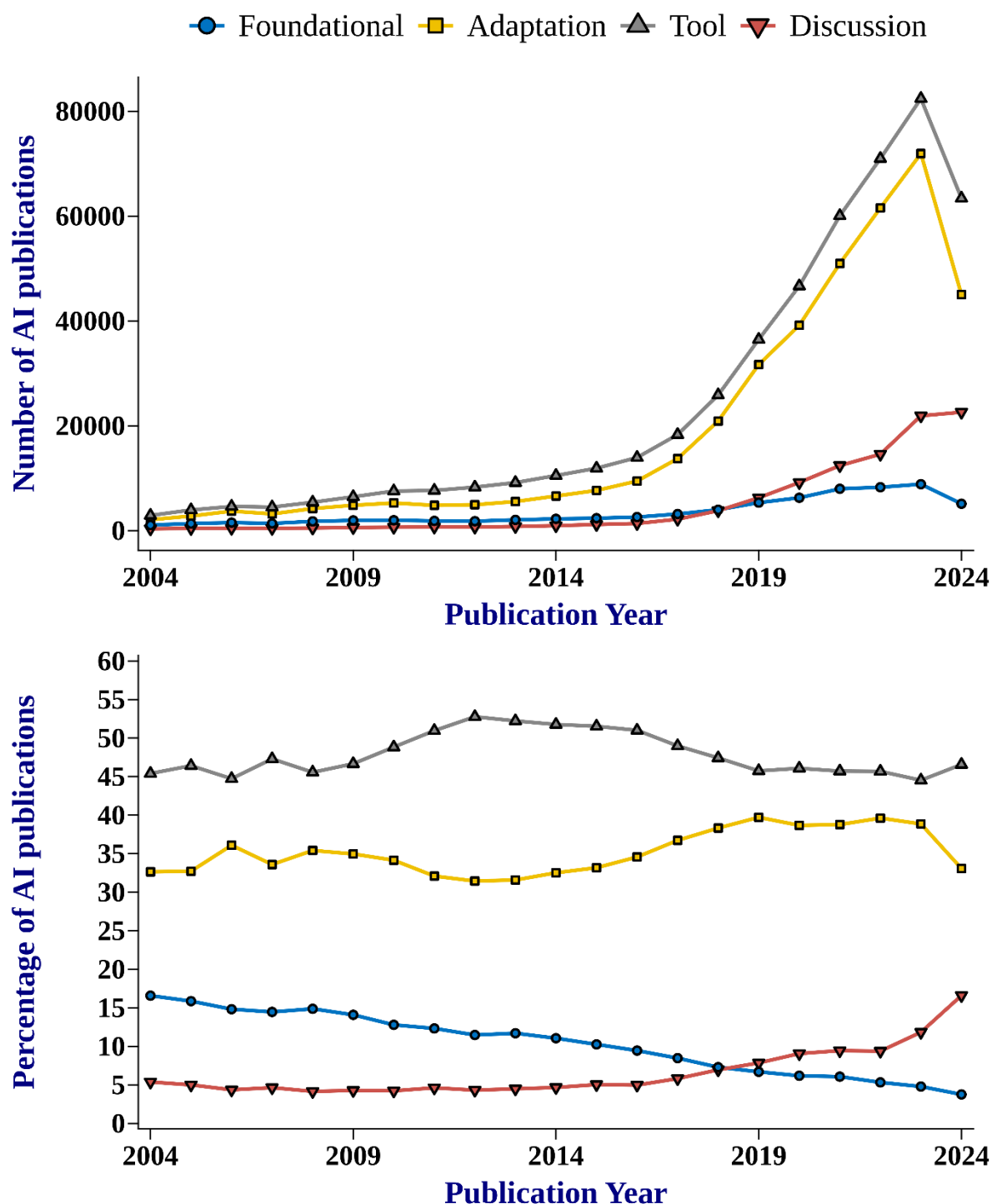
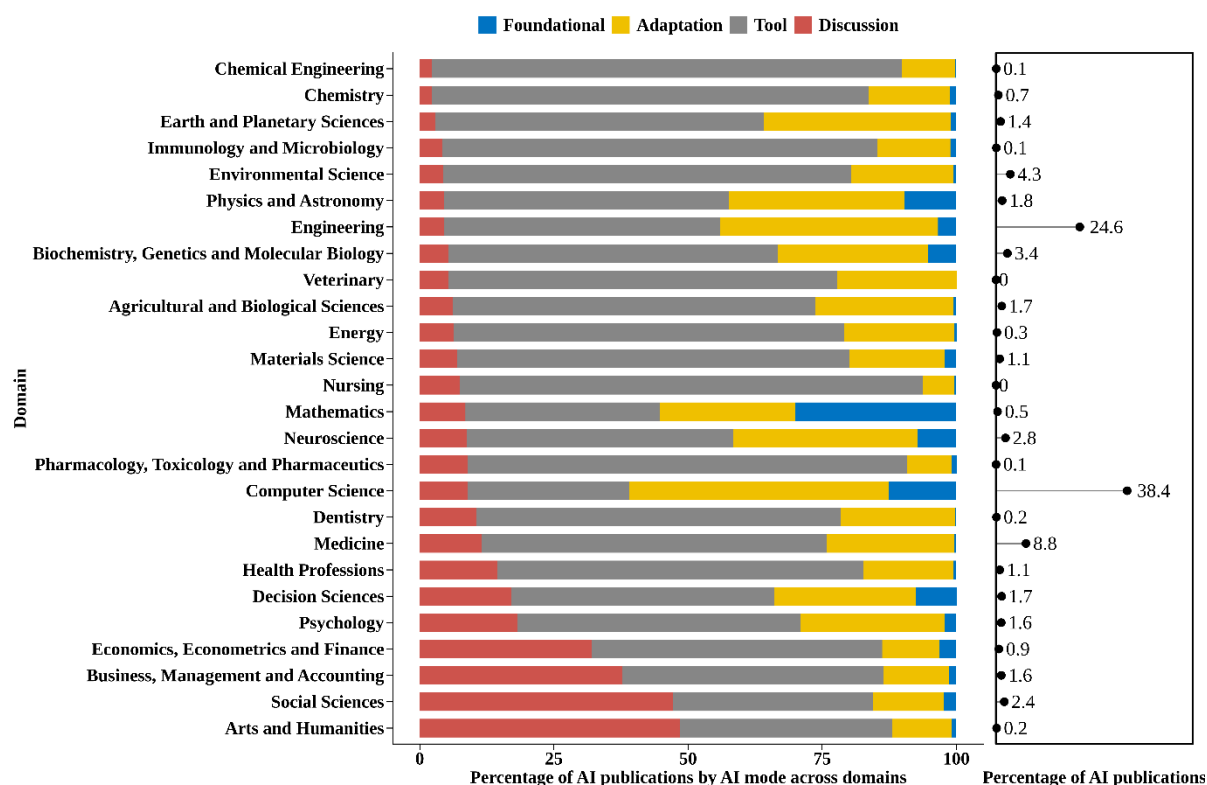


Figure. 2. Annual trends in AI-related publications by mode, 2004–2024



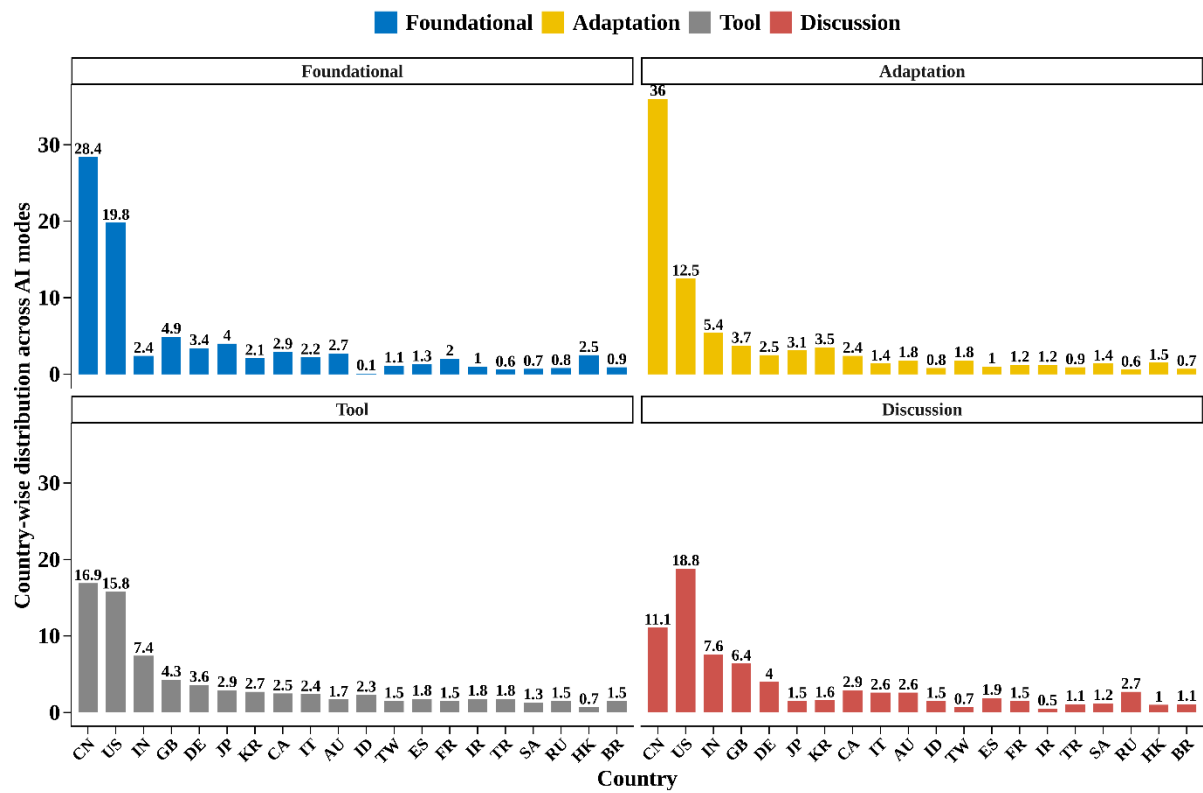
Note: Top: Number of AI publications by mode. Bottom: Percentage share of each mode within annual AI publications. Records are restricted to 2004–2024 to exclude distortions caused by IEEE’s digitization project in 2001–2003. The decline in 2024 reflects incomplete indexing of recent publications. Source: OpenAlex, May 2025 Snapshot. Number of AI publications, 2004–2024: 1,075,782.

Figure. 3. Field distributions of AI relevant publications by AI mode.



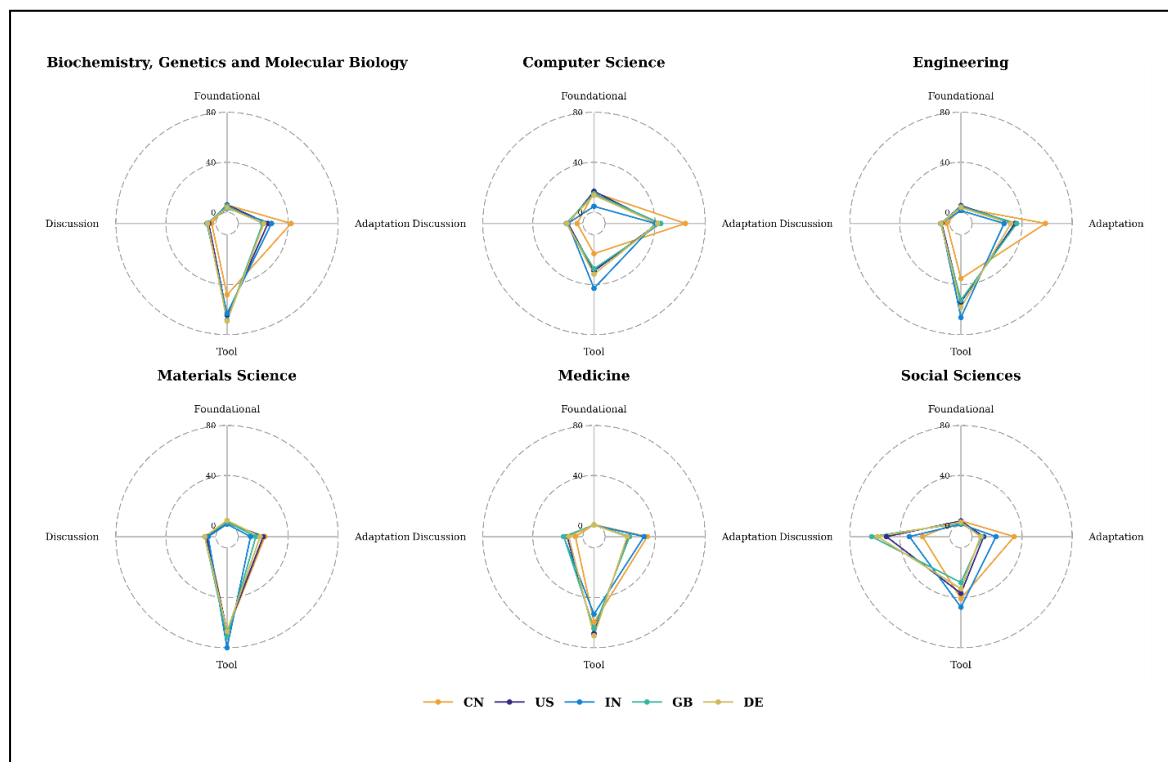
Note: The left panel illustrates the distribution of AI publications within each domain by AI mode, with each bar summing to 100% of the AI publications in that field. The right panel presents the overall distribution of AI publications across fields based on their primary field classification, with the shares likewise summing to 100%. Source: OpenAlex. May 2025 Snapshot. Number of AI publications, 2004-2024: 1,075,782.

Figure. 4. Scientific publications by AI mode, classified by foundational, adaptation, tool, and discussion modes for 20 leading countries.



Note: Within each panel, the denominator is the total number of publications in the focal mode, so the country-level shares sum to 100%. The countries are represented as ISO two-letter country code. Source: OpenAlex, May 2025 Snapshot. Authors' analysis. Number of AI publications, 2004-2024: 1,075,782.

Figure. 5. Radar charts of AI mode distribution across six disciplines for the top five AI-producing countries.



Note: The charts display the percentage of AI-relevant publications classified as Foundational, Adaptation, Tool, or Discussion in Biochemistry, Computer Science, Engineering, Materials Science, Medicine, and Social Sciences for the top five countries. For each country, the shares across the four AI modes sum to 100%. Source: OpenAlex, May 2025 Snapshot. Authors' analysis. Number of AI publications, 2004-2024: 1,075,782.

Table 1 Dataset statistics for AI relevance dataset

	Train	Dev	Test
AI	10,005	172	360
Non-AI	8,941	22	28
Total	18,946	194	388

Table 2 Dataset statistics for AI relevance dataset

	Train	Development	Test
Foundational	665	5	13
Tool	2,000	115	217
Adaptation	1,826	41	80
Discussion	1,630	16	45
Total	6,121	177	355

Table 3 AI relevance classification model performance

	Development			Test		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Keyword	0.89	1	0.94	0.93	1	0.96
GPT	0.98	0.99	0.98	0.98	0.98	0.98
SciBERT	0.96	0.98	0.97	0.96	0.98	0.97

Table 4 AI mode classification model performance

	Development			Test		
	Precision	Recall	F1-score	Precision	Recall	F1-score
GPT	0.83	0.87	0.83	0.84	0.82	0.83
SciBERT	0.76	0.80	0.78	0.76	0.77	0.76

Appendix A1: GPT prompt for AI relevance classification

```
""
You are a research assistant tasked with identifying whether a scientific paper is relevant to
Artificial intelligence (AI) technology, as indicated by AI-related terms in its title and abstract.
Output 1 if the paper is AI-relevant. Output 0 if the paper is not AI-relevant.
And return the results in the specified JSON format.
{
  "AI_relevance_GPT": "0 | 1",
}
""
```

Appendix A2: GPT prompt used for AI mode classification

""""

You are a research assistant tasked with analyzing the title and abstract of a scientific paper. Your goal is to classify the paper into one of the following categories:

- **Discussion**: The paper focuses on analyzing, critiquing, or reflecting on AI without implementing AI models in its research methodology. This includes literature reviews, ethical or societal commentary, perception studies, interviews, qualitative research, and bibliometric or meta-analyses.
- **Foundational**: The paper proposes new AI models, algorithms, training methods, or theoretical contributions that improve AI's core capabilities.
- **Tool**: The paper uses existing AI models (without changing their architecture) to solve domain-specific problems (e.g., in natural science, engineering, agriculture, medicine, education, law, social science, or the humanities).
- **Adaptation**: The paper modifies or adapts the model architecture of existing AI models to better suit specific tasks or domains, without proposing fundamentally new AI methods.
- **Unclear**: The abstract is too vague, lacks sufficient detail, or does not clearly fit any of the above categories.

Output Format

Return a JSON object in the following format:

```
{  
  "category": "<Discussion | Foundational | Tool | Adaptation | Unclear >",  
}
```

""""

Appendix A3: Field distribution of AI relevance sampled training dataset

Field	AI Sample size	Non-AI Sample size
Computer Science	1,000	686
Environmental Science	1,000	461
Engineering	1,000	1,000
Medicine	1,000	1,000
Biochemistry, Genetics and Molecular Biology	928	608
Social Sciences	821	1,000
Neuroscience	793	158
Decision Sciences	521	117
Business, Management and Accounting	520	338
Physics and Astronomy	504	418
Agricultural and Biological Sciences	440	533
Psychology	419	300
Earth and Planetary Sciences	376	182
Health Professions	332	275
Economics, Econometrics and Finance	296	324
Materials Science	269	399
Chemistry	191	283
Mathematics	188	156
Arts and Humanities	94	523
Energy	71	88
Immunology and Microbiology	43	104
Dentistry	43	45
Chemical Engineering	24	27
Pharmacology, Toxicology and Pharmaceuticals	18	40
Nursing	17	47
Veterinary	11	26
Total	10,919	9,138