The uneven embedding of AI in scientific workflows: Technological generations and the restructuring of scientific work

Abstract:
The rapid diffusion of artificial intelligence (AI) is transforming scientific research, yet little is known about how different generations of AI technologies are embedded across scientific workflows. This study develops a large language model (LLM)-based framework to systematically identify and classify AI use in scientific knowledge production. We distinguish four workflow stages capturing AI's roles in research design and hypothesis generation (S1), data collection and curation (S2), experiment and simulation control (S3), and analysis and interpretation (S4). Using abstracts of AI-assisted articles published between 2021 and 2024, we apply structured prompting and rule-based extraction to identify workflow stages, AI model generations, data-processing tasks, and analytical functions performed by AI systems. Preliminary findings reveal substantial heterogeneity in the embedding of AI across scientific workflows. S4 remains the dominant stage, accounting for over 80% of identified cases, suggesting that AI is still primarily used for downstream analytical and interpretive tasks. By contrast, S3 exhibits relatively higher shares in the physical sciences and technology-related fields. The framework further captures differences between technological generations of AI, ranging from classical machine learning and deep learning models to transformer-based and generative AI systems, enabling analysis of how distinct AI capabilities correspond to different forms of scientific work. This study contributes a scalable computational framework for mapping AI-enabled scientific workflows and provides new insights into the uneven restructuring of scientific labor and knowledge production under the diffusion of advanced AI technologies.

Keywords: AI use, workflows, LLM-based approach, knowledge production, AI generations

1. Introduction

Artificial intelligence (AI) is increasingly integrated into scientific research processes, assisting scientists to generate ideas and hypotheses, collect and curate datasets, and interpret research findings, etc (Messeri & Crockett, 2023; de Kok, 2025; Qu & Wang, 2025; He & Bu, 2026). This allows improving efficiency of scientific discovery compared to traditional methods, even catalyzing a paradigm shift in the age of AI (Noy & Zhang, 2023; Wang et al., 2023; Zheng, et al., 2025; Hao et al., 2026).

Despite considerable research in this area, most existing studies tend to treat AI adoption as a binary or aggregate phenomenon, focusing on whether and to what extent AI is adopted in scientific research. Yet scientific knowledge production is inherently multi-stage, involving distinct activities such as problem formulation, data collection, experimentation, and analysis. Distinguishing AI integration across research stages is critical, as these stages involve different levels of control over knowledge production. As a result, stage-specific patterns of AI use can reveal inequalities that remain obscured in aggregate measures, offering a more nuanced understanding of how AI reshapes scientific work.

Existing studies have made important progress in examining the integration of AI into research process and stages, often through systematic reviews and development of conceptual frameworks. For example, Wang et al. (2023) review major advance over the past decade and summarize how AI has been incorporated into scientific discovery to augment and accelerate

research processes. Similarly, Agrawal et al. (2026) conceptualize AI as an augmentative tool in knowledge production, demonstrating that its impact varies across domains and workflow stages. By proposing a foundational three-level taxonomy, Tool, Analyst, and Scientist, Zheng et al. (2025) theorize the evolving roles of AI across the research lifecycle. Complementing these frameworks.

While this body of research offers valuable conceptual architecture for understanding AI in scientific discovery, it tends to emphasize the evolution of AI techniques or high-level categorizations of AI applications. Consequently, the current literature may overlook the empirically positioning of AI within scientific workflows. In particular, the literature lacks a systematic and scalable approach to identify where and how AI is used across the research process. Addressing this gap requires moving beyond broad conceptualizations toward a more fine-grained, workflow-oriented analysis of AI's functional roles in knowledge production.

To systematically identify AI use across key research stages, this study develops a novel three-stage approach by combining bibliometric methods with large language models. Building upon recent work (Ni et al., 2026), which identifies AI-assisted papers from AI-related research, we further distinguish key research stages for each AI-assisted publication. In our framework, employing LLM-based approach, we classify main research workflows through the abstract that involving AI as tool into four categories, including design/hypothesis (S1), data collection/curation (S2), experiment/simulation (S3), analysis/interpretation (S4). This categorization follows the canonical decomposition of scientific research into distinct and sequential stages. Importantly, each stage is associated with different forms of knowledge production and capability requirements, thereby enabling a systematic examination of stage-specific patterns of AI use. The framework further captures differences between technological generations of AI, ranging from classical machine learning and deep learning models to transformer-based and generative AI systems, enabling analysis of how distinct AI capabilities correspond to different forms of scientific work. By mapping AI use across workflows at scale, this study provides new insights into the evolving organization of scientific work.

This study makes three main contributions. First, it advances the conceptualization of AI use in science from a workflow-oriented perspective, enabling a more fine-grained understanding of AI's roles across the research process. Second, it contributes methodologically by developing an LLM-based framework for the scalable and systematic identification of AI use in scientific publications. Third, it provides new empirical evidence on the uneven embedding of AI across scientific work, revealing a persistent concentration in downstream analytical and infrastructural tasks alongside substantial disciplinary variation in workflow integration and AI capabilities.

2. Data and methods

2.1 Data source

To identify and classify AI use across research workflow stages, this study develops a three-stage approach.

In the first stage, we identify AI-related articles. Specifically, using an expanded AI and Gen-AI keyword search strategy, 554,543 AI-related articles from the Web of Science (SCIE and SSCI) published between 2021 and 2024 were retrieved, a period capturing the rapid diffusion of Gen-AI technologies in scientific research while ensuring data comparability.

The second stage further identifies AI-assisted articles, in which we retain papers using AI techniques to augment research and exclude those AI fundamental studies. We classified journals according to Journal Citation Reports subject categories. Articles published in the 1,619 journals belonging to the 14 "Computer Science" categories were coded as CS_J and excluded from further analysis, as these journals predominantly publish methodological and technical advances in AI rather than applying AI as a research tool in domain-specific knowledge production. While all others were coded as NCS_J and remained for further processing, which reflecting AI applications across diverse domains, resulting in an intermediate dataset of 359,769 articles. Next, we identify AI-assisted research by processing each abstract using LLM-based approach with a zero-shot prompt that extracts methodological information and determines whether AI is used as a research tool (e.g., for idea generation, data processing, modeling, or analysis). Papers that merely review AI-related developments are excluded. At last, we obtained 283,068 unique publication records with explicitly stated AI-assisted research.

2.2 Research framework

AI use in scientific workflows

The third stage, following the identification of AI-assisted papers, employs an LLM-based approach to analyze abstracts and capture how AI is used within the research process. Rather than relying solely on paper keywords, this approach enables interpretation of AI's functional role. Each AI-assisted paper is classified according to the primary role AI plays in the research workflow building on the canonical decomposition of scientific research (Wang et al., 2023; Zheng et al., 2025). Each workflow stage is modeled as an independent binary classification task; stages S1–S4 may co-occur within the same paper, whereas S5 is mutually exclusive with S1–S4. It worth noting that category S5 is introduced to account for cases in which papers are identified as AI-related in the previous step but do not substantively use AI within the research workflow, thereby reducing the impact of false positives arising from keyword-based identification.

- S1 – Design and hypothesis generation: AI used to formulate research questions, hypotheses, or conceptual frameworks.
- S2 – Data collection and curation: AI used to generate, collect, label, or preprocess data.
- S3 – Experiment or simulation control: AI used to guide experimental procedures, simulations, or automated systems.
- S4 – Analysis and interpretation: AI used for data analysis, modeling, pattern recognition, or interpretive tasks.
- S5 – Not AI-assisted or other: Papers that do not substantively use AI as a research tool.

AI generations in scientific workflows

(to be refined)

The flowchart in Fig. 1 shows the procedure of selection of AI-assisted articles and classification of AI use in scientific workflow, AI generations and functions.

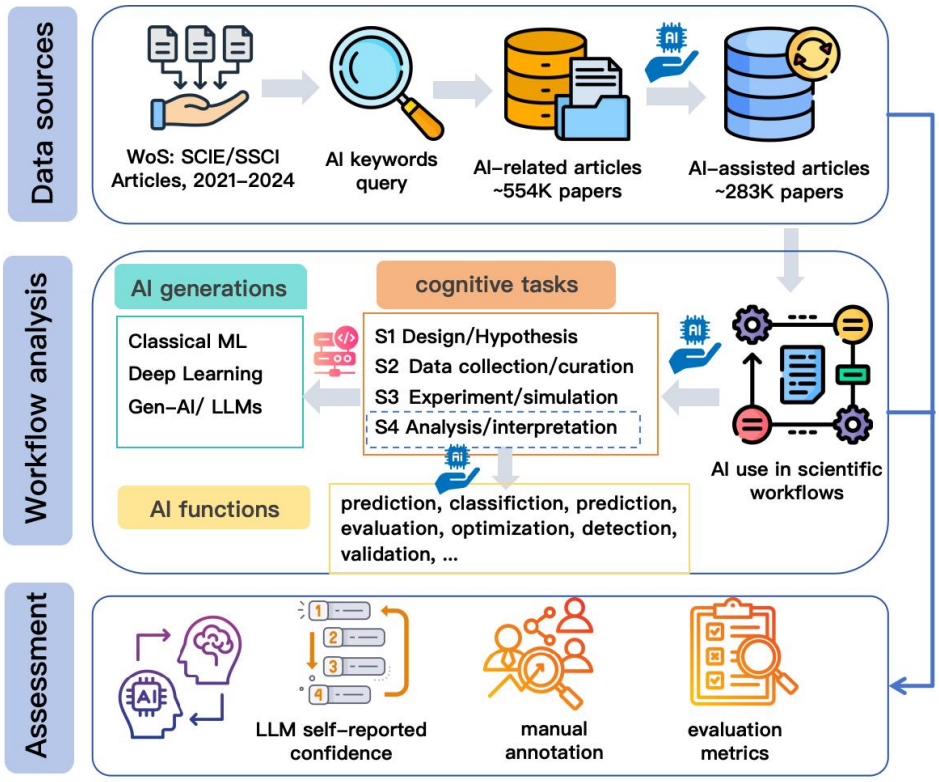


Figure 1. Procedure for scientific workflow analysis.

2.3 Workflow classification model and validation

LLM-based workflow classification

We developed a structured prompt to enable large language models (LLMs) to perform the classification task. The model was deployed through Azure OpenAI's GPT-5o API and operated in a zero-shot setting, without task-specific fine-tuning. The prompt was refined through five rounds of iteration, in which we systematically adjusted task instructions, category definitions, and output formats. The prompt instructs the LLM to assign binary indicators (0/1) to each of the five stages (S1-S5), indicating the presence or absence of AI use at each stage. In addition, the model is required to provide textual evidence and a brief rationale supporting its classification for each stage, thereby enhancing transparency and enabling downstream validation. To further assess prediction reliability, the LLM is also asked to report a self-assessed confidence score for each classification instance (Appendix 1).

Evaluation dataset construction

To evaluate the classification performance, we constructed a gold-standard dataset through manual annotation of abstracts. The annotation process followed a two-round procedure designed to standardize coding criteria and improve inter-coder reliability.

In the first round, we randomly selected 50 abstracts, which were independently coded into categories S1–S5 by two researchers. The coders then compared their results and discussed discrepancies, with a third researcher consulted to resolve ambiguous cases. Based on these discussions, the annotation scheme was refined, and a consensus set of labels was established. In the second round, an additional 50 abstracts were randomly sampled and independently coded by the same two researchers using the refined guidelines. Inter-coder agreement improved to approximately 80% from 50% in the first round, indicating increased consistency in applying the classification criteria. Remaining discrepancies were resolved through

discussion until full consensus was reached. Representative examples of manually annotated cases for each workflow stage are provided in Appendix 2.

Model evaluation: metrics and results

Table 1 reports the model's classification performance across stages using standard evaluation metrics. Overall, the model achieves strong performance, with particular high presision, recall and F1 for S1 and S4, indicating reliable identification of these stages. Performance is more moderate for S2 and S5, with balanced precision and recall. While S3 shows comparatively lower recall and F1 scores, reflecting challenges in identifying less frequent stages. Notably, accuracy remains high across all stages, partly reflecting class imbalance, as indicated by the varying base rates.

Table 3. Model performance on the test dataset

Stage	Base rate (manual=1)	Precision	Recall	F1	Accuracy
S1	0.02	1.00	1.00	1.00	1.00
S2	0.15	0.63	0.67	0.65	0.89
S3	0.03	0.50	0.33	0.40	0.97
S4	0.61	0.94	0.84	0.89	0.87
S5	0.27	0.71	0.93	0.81	0.88

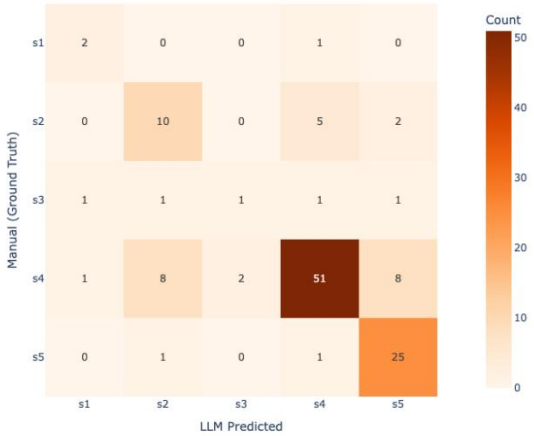


Figure 2. Confusion matrix for AI use in scientific workflows

2.4 AI generations recognition model and validation
(to be refined)

3. Preliminary results

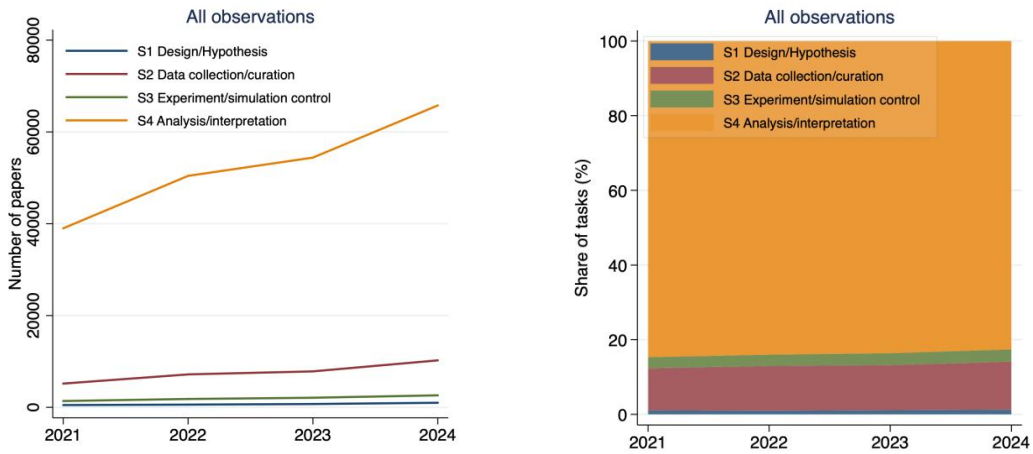
3.1 Overall distribution

In this analysis, we exclude category S5 (accounting for 22.4% of the sample), as it does not represent AI use within specific research workflow stages, yielding 219,558 observations. Figure 3 illustrates both the absolute and relative distribution of AI use across core research workflow stages (S1–S4) within AI-assisted publications. Fig 3(a) shows that the number of publications involving AI use has increased steadily across all workflow stages from 2021 to 2024. The most pronounced growth is observed in the analysis and interpretation stage (S4), which consistently dominates in absolute terms. Fig 3(b) presents the corresponding proportional distribution. AI use is overwhelmingly concentrated in the analysis and interpretation stage (S4), which accounts for approximately 85% of all cases throughout the

observation period. Data collection and curation (S2) constitutes over 10%, whereas the other stages collectively represent only a minor fraction.

Despite the substantial increase in the absolute number of publications, the relative distribution across workflow stages remains remarkably stable over time. This stability in relative shares indicates a persistent functional specialization of AI within the research workflow. Moreover, this pattern suggests that, while AI adoption in scientific research is expanding rapidly, its functional role is consistently concentrated in downstream analytical tasks rather than upstream stages such as research design or experimental control.

Further, we restrict the sample to observations with LLM self-reported confidence ≥ 0.8 (N = 161,592; 73.6% of all observations). Results based on this higher-confidence subsample remain consistent with the main findings, demonstrating the robustness of our results (Appendix 3).



(A) (B)

Figure 3. Overall distribution of AI-assisted tasks

3.2 Distribution by disciplines

Across disciplines, AI use is predominantly concentrated in the analysis and interpretation

Across disciplines (Figure 4), AI use is consistently concentrated in the analysis and interpretation stage (S4), highlighting a common functional pattern in how AI is integrated into scientific workflows. In the Arts and Social Sciences (ASS), approximately 80% of AI-assisted activities occur in S4, while in the Life Sciences and Biomedicine (LSB) this share rises to around 85%. Similar patterns are observed in the Physical Sciences (PhS) and Technology fields, where roughly 80% of AI use is devoted to analytical and interpretative tasks.

Despite this overall convergence, notable disciplinary differences emerge in the distribution of AI use across other stages. In particular, the share of AI use in experimentation and simulation (S3) is more pronounced in PhS and Technology compared to ASS and LSB. This likely reflects the greater reliance on computational modeling, simulation-based experimentation, and system control in these fields, where AI can be directly embedded into experimental workflows.

By contrast, upstream stages, which refer to earlier phases of the research process such as design and hypothesis formulation (S1) and data collection and curation (S2), remain relatively limited across all disciplines. Taken together, these findings suggest that while AI adoption is widespread, its functional role remains systematically skewed toward downstream

analytical processes, which correspond to later stages involving analysis and interpretation (S4), with discipline-specific extensions into experimental contexts.

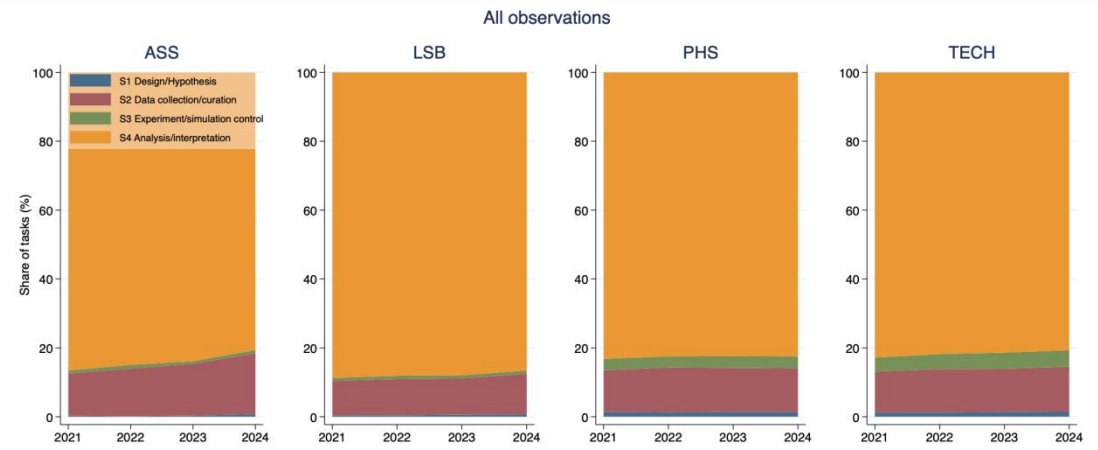


Figure 4. AI-assisted task composition across disciplines

3.3 AI generations and workflows
(to be refined)

4. Discussion and conclusions

Understanding how AI is embedded in scientific workflows is critical to assessing its transformative impact on knowledge production. This study provides a systematic mapping of AI use across scientific workflows, by integrating a three-stage framework with LLM-based methods. The findings consistently show that AI use is heavily concentrated in downstream analytical stages, with relatively limited involvement in upstream activities such as research design and data collection. This pattern holds across disciplines, despite some field-specific variations. One possible explanation lies in the current technological ability of AI systems, which are particularly effective in tasks involving navigating large and complex information spaces. Our findings align with emerging evidence that the impact of AI on scientific discovery is highly uneven across domains and contexts. In particular, AI appears to be most effective in contexts characterized by complex and combinatorial problem spaces, where it can support the exploration of large knowledge domains (Bianchini et al., 2026). This may help explain why AI use in our analysis is concentrated in downstream analytical stages, which rely heavily on processing and interpreting complex information.

These findings have important implications for the governance of AI use in scientific research. To capture the nuances of how AI is used within the full research process, there is a need to develop standardized disclosure frameworks for AI use in scientific publications. In particular, journals and publishers could establish formal reporting requirements that prompt authors to explicitly document the use of AI tools at different stages of the research workflow. One practical approach would be to incorporate structured AI-use statements into existing sections such as the acknowledgements or methods, specifying the role, scope, and extent of AI assistance. Such measures would enhance transparency, improve the transparency and reproducibility of research, and support the responsible integration of AI into scientific practice. More broadly, the development of shared norms and reporting standards would enable more systematic monitoring of AI's role in knowledge production, thereby informing evidence-based policy discussions on the governance of emerging research technologies.

Several limitations should be acknowledged in this study. First, the analysis relies on abstracts, which may not fully capture the extent and nuances of how AI is used within the full research process. Second, although the LLM-based classification framework is validated against manually annotated data, it remains subject to potential biases related to prompt design and model behavior. While the use of confidence thresholds and robustness checks helps mitigate these concerns, classification errors cannot be entirely eliminated. Future research can build on this framework in several directions. Extending the analysis to full-text data would allow for a more comprehensive identification of AI use across research workflows.

Selected references

Bianchini, S., Di Girolamo, V., Ravet, J., & Arranz, D. (2026). AI in science: When and where it makes a difference. Research Policy, 55(6), 105478.

Hao, Q., Xu, F., Li, Y., & Evans, J. (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 1-7.

Ni, R., Qu, Y. & Wang, J. (2026). Bridging or Widening the Gap? The Role of AI in Shaping Global Research Performance. Journal of Chinese Political Sciences. https://doi.org/10.1007/s11366-026-09931-2

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science,* 381(6654), 187-192.

Stein, C., & Hill, R. (2026). How Artificial Intelligence Shapes Science: Evidence from AlphaFold, NBER Working Paper 35143, https://doi.org/10.3386/w35143.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., ... & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. Nature, 620(7972), 47-60. https://www.nature.com/articles/s41586-023-06221-2

Zheng, T., Deng, Z., Tsang, H. T., Wang, W., Bai, J., Wang, Z., & Song, Y. (2025). From automation to autonomy: A survey on large language models in scientific discovery. arXiv preprint arXiv:2505.13259.