# The Shift Toward Open and Reproducible AI Research

Kevin L Coakley[1,2*], Thijs Snelleman[3], Holger Hoos[3,4], Odd Erik Gundersen[1*]

[1*]Department of Computer Science, Norwegian University of Science and Technology, Sem Sælands Vei 7-9, Trondheim, 7031, Trønderlag, Norway.
[2]San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Dr, La Jolla, 92093, California, USA.
[3]Chair of AI Methodology, RWTH Aachen University, Theaterstrasse 35-39, Aachen, 52062, North Rhine-Westphalia, Germany.
[4]Leiden Institute of Advanced Computer Science, Leiden University, Einsteinweg 55, Leiden, 2333 CC, South Holland, The Netherlands.

*Corresponding author(s). E-mail(s): kcoakley@sdsc.edu; odderik@ntnu.no;
Contributing authors: snelleman@aim.rwth-aachen.de; hh@aim.rwth-aachen.de;

**Abstract**

The reproducibility crisis has directed the AI research community toward improving documentation practices. Several studies have identified methodological issues, and in response, the most impactful venues in the field have introduced reproducibility checklists. We seek to understand whether documentation practices have changed over time by assessing all published papers at five leading AI conferences over the past decade. Seven reproducibility variables were identified, quality-assured and used to analyse 56 800 publications. Our analysis reveals that in the period 2014 to 2024, documentation practices have improved; papers sharing both code and data increased nearly sixfold, from 11% to 64% Building on empirical reproducibility rates from a prior study, we estimate — inferred from documentation practices, not direct testing — that reproducibility increased from 28% in 2014 to 64% in 2024. Improvements in documentation practices predate the introduction of reproducibility checklists, suggesting these changes reflect a broader movement toward open science rather than a direct response to formal requirements.

**Keywords:** Reproducibility, Artificial Intelligence, Machine Learning

Many scientific results cannot be reliably reproduced [1], and therefore conclusions cannot be trusted [2]. This phenomenon is called the reproducibility crisis [3] and has been described in a number of studies covering a broad range of disciplines, including psychology [4–6], economics [7], medicine [8–10], neuroscience [11], and genetics [12]. Reproducibility concerns also gained prominence in artificial intelligence (AI) [13, 14], where rapid growth in the field and increasing social importance have intensified the urgency to mitigate the problem.

One way to mitigate the reproducibility crisis is to provide detailed descriptions of the research alongside open code and data, and therefore a cultural shift toward open science [15] has emerged, advocating for transparent and accessible knowledge. As a large portion of empirical research in computer science and AI is conducted on computers and experiments are described in code, sharing detailed descriptions of experiments should, in principle, be straightforward [16]. Achieving this relies critically on high-quality documentation of experimental methods and artifacts, such as code, data, and workflows [17–20]. This has again led communities and venues to demand stronger requirements from authors, reflected in the introduction of reproducibility checklists for example at NeurIPS [21] and JAIR [22]. However, the effect of these procedural and communal changes remains difficult to measure, especially at scale, due to the manual labour required to reproduce studies and the tenfold increase in published studies in the past decade, see Figure 1.

It is widely acknowledged that empirical AI research can fail in subtle ways. Nondeterminism due to variability between execution (or computing) environments can cause results to vary so much that it affects reproducibility [23–27]. Poor data treatment, including poorly documented datasets, dataset splits, and data leakage also pose challenges to reproducibility [28–30]. Inadequate hyperparameter optimisation can lead to irreproducible results [23, 31–33], and the same holds for parallelization [34], compiler settings, and operating system [35]. Even when code is provided, software dependency issues, errors, and post-publication updates can prevent empirical AI research from being reproduced [36, 37]. Gundersen et al. [38] categorizes different sources of irreproducibility, but few papers investigate to what degree empirical AI research results can be trusted.

Several attempts have been made to quantify the reproducibility of empirical research in computer science and AI. Following Gundersen [39], we use reproducibility to refer to the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators. No distinction is made between reproducibility and replication, consistent with Gundersen [39], Schmidt [40], Nosek and Lakens [41], Goodman et al. [42], and Gundersen and Kjensmo [13]. Collberg and Proebsting [43] were only able to execute the code of 48.3% of the 601 computer science papers that shared code as part of the publication, not even attempting to evaluate whether the results were reproduced. Raff [44] was able to reproduce 63.4% of 255 papers by reimplementing algorithms from scratch,

while Gundersen et al. [45] reproduced 33% of papers that only provided data and 86% of papers that shared both code and data.

These studies, while insightful, reveal intrinsic limitations of manual reproduction. Efforts are limited by the time required for manual reproduction, which introduces selection biases when choosing which papers to replicate and reduces sample size. Accurately measuring reproducibility trends in a field in which the leading five conferences publish more than 12 000 papers annually cannot be done manually.

Another approach avoids reimplementing experiments entirely and instead focuses on the quality of documentation and the inclusion of open artifacts as a practical means of evaluating reproducibility in the spirit of open science. Gundersen and Kjensmo [13] manually evaluated 400 randomly selected papers for characteristics associated with reproducibility. Random selection reduced selection bias and documentation analysis allowed for increasing the sample size to 400, yet their results still represent only a fraction of conference papers published between 2013 and 2016. The reliance on manual effort, whether for full reproduction or documentation analysis, renders a large-scale, longitudinal analysis of the impact of open science on empirical AI research infeasible.

To address this methodological gap and, to the best of our knowledge, provide the first comprehensive longitudinal analysis of open science in AI research, in this study, we automate the evaluation of reproducibility documentation at scale leveraging large language models (LLMs). We systematically assess documentation practices and the use of open artifacts, such as code and datasets, of 56 800 papers from five leading AI conferences — AAAI (Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence), ICLR (International Conference on Learning Representations), ICML (International Conference on Machine Learning), IJCAI (International Joint Conference on Artificial Intelligence), and NeurIPS (Conference on Neural Information Processing Systems) — spanning from 2014 to 2024. We confirmed the reliability of our approach through prompt optimisation against a manually annotated dataset; furthermore, we manually evaluating 160 randomly sampled papers from the resulting dataset. This automated method circumvents the prohibitive costs and inherent selection biases of manual reproduction, as well as the costs of manual documentation analysis that have constrained previous work, enabling an, previously infeasible, exhaustive analysis. With this scalable framework, our study tracks trends in artifact sharing, quantifies the impact of community practices, such as requirements for reproducibility checklists, and characterizes the evolution of the estimated reproducibility rate of AI research.

Our analysis reveals a quantifiable cultural shift toward open science through improved quality of documentation and increased artifact sharing within the AI research community. We estimate that the reproducibility rate of empirical research — inferred from documentation practices and the empirical rates reported by Gundersen et al. [45], not direct testing — has more than doubled, increasing from 28% in 2014 to 64% in 2024, a trend supported by a fivefold increase in papers sharing both code and relying on open datasets. Our results show that improvements in documentation practices were underway before reproducibility checklists were introduced. Our study advances the literature on AI reproducibility by demonstrating quantifiable trends

**Annual publication counts across five leading AI conferences from 2014 to 2024.**

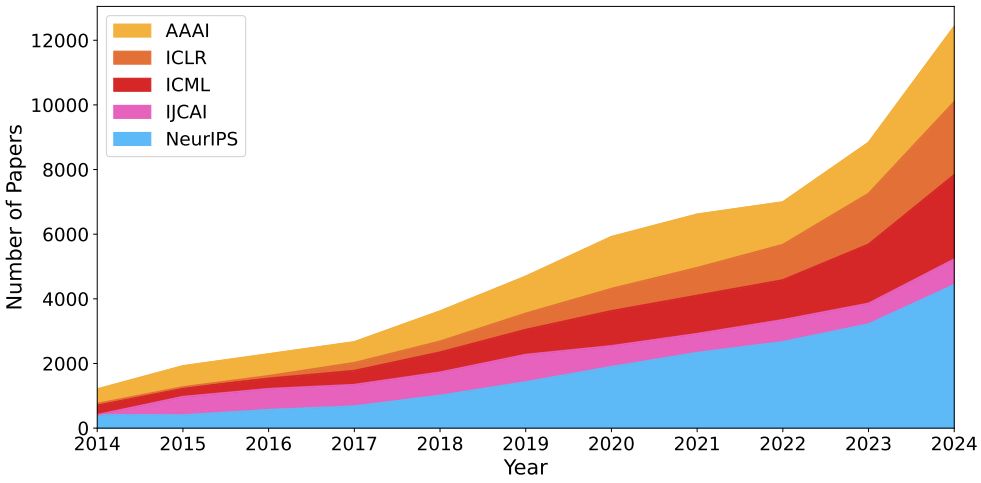


**Fig. 1** The number of papers published at five leading AI conferences (AAAI, ICLR, ICML, IJCAI, and NeurIPS) from 2014 to 2024. Publications increased tenfold over this period, from 1 206 papers in 2014 to 12 026 in 2024, reflecting the rapid expansion of AI research. IJCAI did not hold a conference in 2014. A tabular representation of this data can be found in Supplementary Table 10.

in documentation practices, validated against manually annotated datasets, while establishing an LLM-based methodology for conducting meta-science at scale.

# 1 Results

We analysed 56 800 papers from five top-tier AI and ML conferences, AAAI, ICLR, ICML, IJCAI, and NeurIPS, published between 2014 and 2024. The number of publications at these conferences has grown substantially over this period, from 1 206 papers in 2014 to 12 026 in 2024, a tenfold increase. 52 328 papers (92%) were identified as empirical research and thus form the basis of our analysis; we excluded theoretical papers, as the reproducibility of theoretical work warrants an analysis of an entirely different nature.

## 1.1 Reproducibility Variables

To quantify the progress of open science and documentation practices in empirical AI research, we evaluated each paper in our dataset for a set of variables. These variables have been proposed in other studies [13, 44, 46] and cover the previously introduced reproducibility checklists (see Supplementary Table 1). We use the *reproducibility variables* as seven Boolean indicators that represent whether a given paper contains the following information: (1) availability of open source code; (2) use of one or more open datasets; (3) specification of dataset splits; (4) pseudocode describing the algorithm(s) used; descriptions of (5) the hardware used to conduct the experiment; (6) the software

**Distribution of empirical papers by number of documented reproducibility variables from 2014 to 2024.**

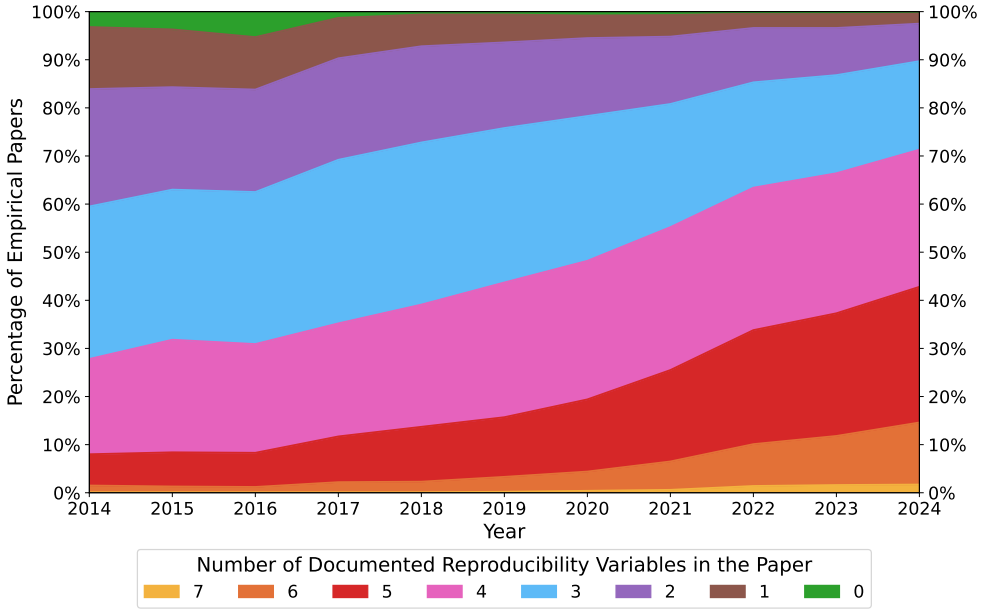


**Fig. 2** The distribution of empirical papers grouped by how many of the seven reproducibility variables (pseudocode, open code, open datasets, dataset splits, hardware specification, software dependencies, and experiment setup) they document, shown for each year from 2014 to 2024. Each coloured band represents papers documenting a specific count of variables, from zero (top, dark green) to all seven (bottom, light orange). Documentation comprehensiveness improved substantially: the proportion of papers documenting none of these variables decreased from 3% to 0.3%, while papers documenting all seven increased from 0.1% to 1.7%. Most notably, papers documenting at least five of the seven variables increased more than fivefold, from 8% in 2014 to 43% in 2024, demonstrating a sustained shift toward more thorough documentation practices.

dependencies; and (7) the setup of the experiment. In addition, we distinguish whether the study is considered empirical or theoretical, and capture whether at least one author has an industry (rather than academic) affiliation to assess how institutional context influences documentation practices.

## 1.2 Improved Documentation and Open Science Practices

Research documentation has improved from 2014 to 2024; we observed a substantial positive trend for reproducibility variables. In 2014, only 8% of the empirical papers provided documentation for five or more reproducibility variables (Figure 2), while by 2024 this figure had increased to 43%, representing more than a fivefold increase over the decade. The fraction of papers providing documentation for none of these variables vanished by 2019; at the same time, papers providing documentation for all variables emerged.

**Open code and open dataset documentation rates across five AI conferences from 2014 to 2024.**

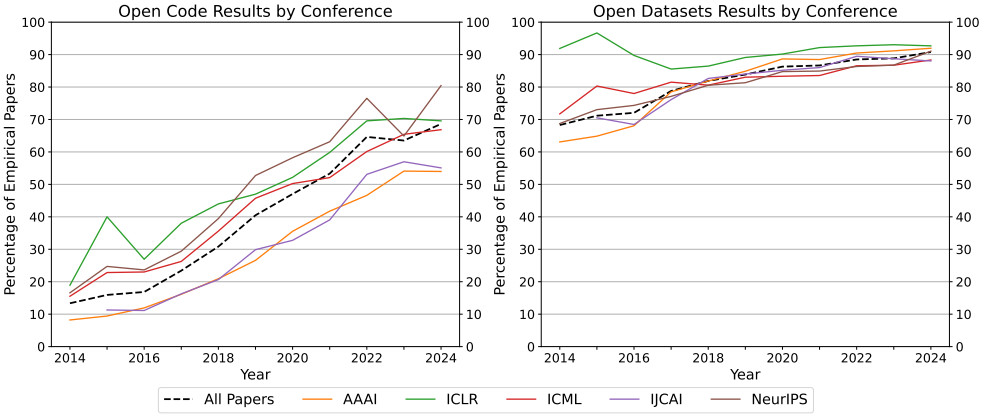


**Fig. 3** The percentage of empirical papers documenting the availability of open code (left) and the use of open datasets (right) for each of the five conferences from 2014 to 2024. IJCAI did not hold a conference in 2014. Both metrics show substantial increases across all venues. Open code availability rose from 13% in 2014 to 69% in 2024 across all conferences (dashed black line), while open dataset usage increased from 68% to 91% over the same period.

The seven reproducibility variables were selected because each captures a distinct artifact or methodological description that reduces ambiguity for an independent team attempting to reproduce an experiment. Because empirical AI research encompasses a wide variety of methodologies, from reinforcement learning in simulated environments to benchmarking on static corpora, none of the seven is strictly required for every paper, and documenting all seven does not guarantee reproducibility. However, the absence of each variable introduces a specific barrier: without pseudocode or open code, algorithmic ambiguity increases [44, 47]; without open datasets or dataset splits, an independent team must invest substantial time and compute identifying data and splits that yield comparable results [45, 48, 49]; without hardware and software specifications, environmental differences can shift results enough to alter conclusions, particularly when performance margins are small [35, 50, 51]; and without experiment setup details, hyperparameter search requires additional time and compute [44, 52, 53]. Each documented variable therefore removes one such barrier. At the population level, papers that document more variables leave fewer barriers in place, increasing the likelihood of successful reproduction. The year-over-year increase in the proportion of papers documenting five or more variables reflects a measurable shift in the expected reproducibility of the published literature, not a guarantee for any individual paper.

Our analysis shows that the AI community has broadly adopted open science practices. Open science promotes transparency, accessibility and reproducibility in research, and sharing of code and data is a cornerstone for achieving this [16]. Our results show an increasing trend towards open source code and open datasets, as seen in Figure 3; the use of open datasets increased by 23 percentage points from 68% to a of 91%, whereas open source code increased five-fold from 13% to 69%.

**Rates of empirical papers documenting both open code and open datasets versus neither from 2014 to 2024.**

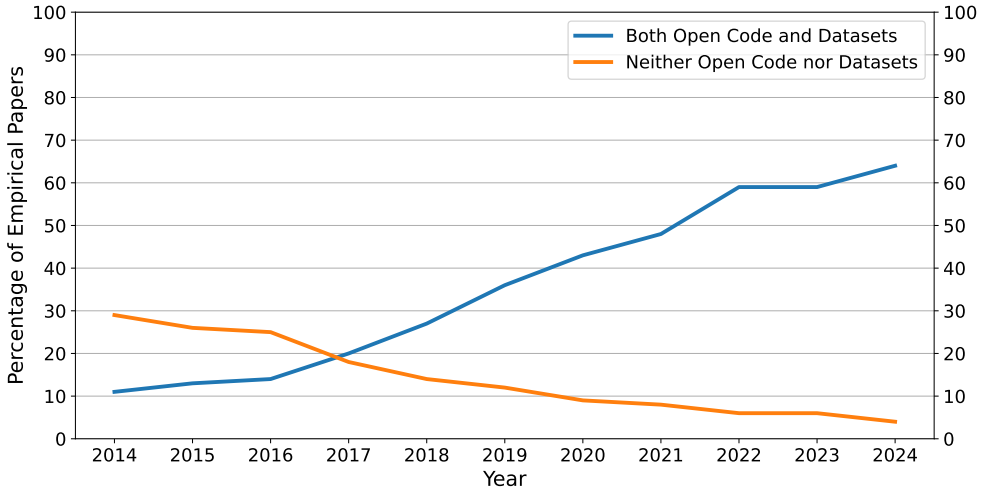


**Fig. 4** The percentage of all empirical papers across the five conferences that document both the availability of open code and the use of open datasets (blue) versus papers that document neither (orange), from 2014 to 2024. Papers sharing both code and data increased nearly sixfold, from 11% in 2014 to 64% in 2024, while papers sharing neither decreased sevenfold, from 29% to 4% over the same period. This shift represents a fundamental change in research transparency, as the combination of open code and open datasets is most strongly associated with reproducible findings [45]. The crossing point occurred around 2017, after which sharing both became more common than sharing neither.

Sharing code and data correlates with successful replication [45]. Over the time period we analysed, the number of papers that provide open source code and open data, the combination most strongly associated with reproducible findings, increased fivefold, while the proportion that shares neither code nor data, which requires the most effort to reproduce, decreased by a factor of seven, as illustrated in Figure 4. The substantial growth in papers sharing both code and data is associated with a higher probability of reproducible findings.

To estimate what proportion of publications can be reproduced, we performed an analysis based on the empirical reproducibility rates reported by Gundersen et al. [45]. Specifically, we applied two fixed rates from that study: 33% for papers sharing only open datasets, and 86% for papers sharing both open code and open datasets. For each year, we computed a weighted sum of these rates using the observed proportions of papers in each category as weights, providing the estimated reproducibility rate for each year (Figure 5). This procedure assumes that the reproducibility rates from Gundersen et al. [45] hold as constants across all five conferences and the full 2014–2024 period, and that papers using private datasets can be excluded from the estimate; the resulting values should therefore be interpreted as approximations rather than precise measurements.

To quantify the uncertainty in these estimates, we computed Wilson score 95% confidence intervals for the two empirical reproducibility rates reported by Gundersen

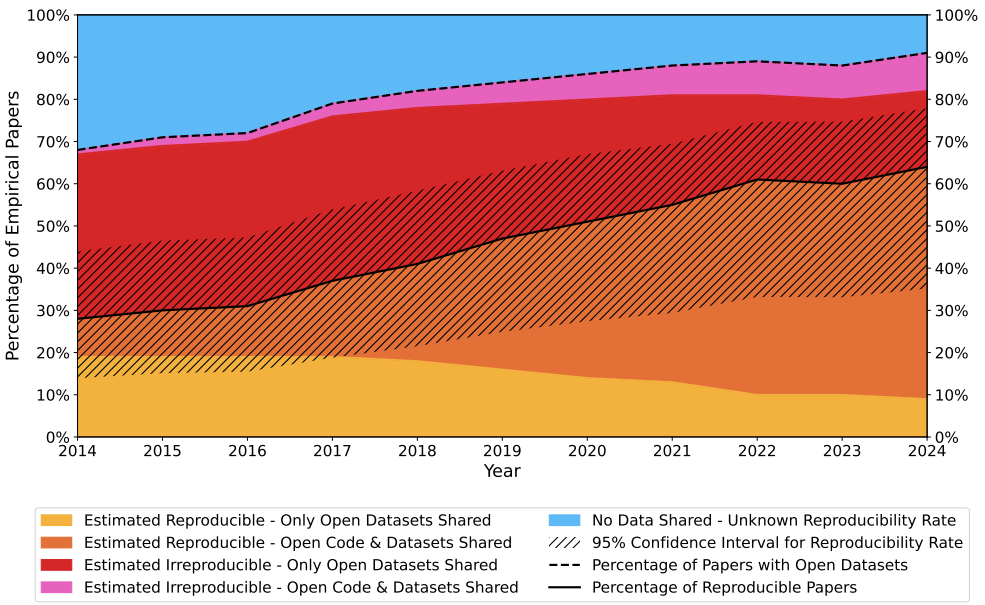


**Fig. 5** Estimated reproducibility rate of empirical AI research from 2014 to 2024. The estimated reproducibility of papers based on the availability of open code and open datasets, applying the empirical reproducibility rates reported by Gundersen et al. [45]: 33% for papers using open datasets only and 86% for papers sharing both code and open datasets. The overall estimated reproducibility rate (dot-dash black line) more than doubled from 28% in 2014 to 64% in 2024. The dashed black line shows the percentage of papers using open datasets, which directly determines whether reproducibility can be estimated, increased from 68% to 91% of all empirical papers, while papers with unknown reproducibility (blue, no open datasets) decreased from 32% to 9%. The hatched band represents the 95% confidence interval for the estimated reproducibility rate, derived from a Monte Carlo simulation using Beta distributions parametrised from the replication counts of Gundersen et al. [45].

This substantial improvement demonstrates the tangible impact of increased code and data sharing on the reproducibility of AI research.

et al. [45]: 86% (6/7) for papers sharing both open code and open datasets and 33% (5/15) for papers sharing only open datasets. The lower and upper bounds were then applied to the per-year reproducibility rate estimate (Figure 5).

Over the 11 years covered by our study, our results show a substantial increase in the estimated reproducibility rate, inferred from the documentation of the reproducibility variables, not direct testing, more than doubled from 28% in 2014 to 64% in 2024 (Figure 5). The proportion of empirical papers using private datasets, for which reproducibility could not be estimated, decreased by 72%, from 32% of the research analysed in 2014 to 9% in 2024. Although the overall number of published empirical papers increased tenfold during that time, the number of papers estimated to be reproducible increased nearly 24-fold, from approximately 305 in 2014 to approximately 7265 in 2024. These results illustrate how increased documentation, especially in the

form of open code and datasets, and the broader adoption of open science practices have tangibly improved the rigour of AI research, even as the field has expanded rapidly.

## 1.3 The Effect of Introducing Reproducibility Checklists

**Reproducibility variable trend slopes (pp/year) before and after checklist introduction, by conference.**

| Reproducibility Variable | All Papers | | AAAI | | ICLR | |
|---|---|---|---|---|---|---|
| | Before 2014-2018 | After 2019-2024 | Before 2014-2020 | After 2021-2024 | Before 2014-2021 | After 2022-2024 |
| Pseudocode | -0.7 | 0.5 | -1.12 | -3.1 | 3.52 | -1.44 |
| Open Code | 4.24 | 5.76 | 4.47 | 4.42 | 4.92 | 0 |
| Open Datasets | 3.48 | 1.26 | 4.66 | 1.11 | -0.38 | -0.01 |
| Dataset Splits | 2.19 | 0.41 | 5.09 | -3.15 | 0.01 | -3.27 |
| Hardware Specification | 0.51 | 7.58 | 1.05 | 1.92 | 1.62 | 2.17 |
| Software Dependencies | -0.38 | 1.1 | -0.78 | -0.86 | -0.07 | 0.68 |
| Experiment Setup | 1.19 | 1.26 | 3.17 | -1.36 | 0.29 | 0.07 |
| Reproducibility Variable | ICML | | IJCAI | | NeurIPS | |
| | Before 2015-2022 | After 2023-2024 | Before 2014-2020 | After 2021-2024 | Before 2014-2018 | After 2019-2024 |
| Pseudocode | -0.31 | -4.44 | -2.13 | -0.35 | -0.94 | 0.59 |
| Open Code | 5.67 | 1.42 | 4.8 | 5.2 | 5.04 | 4.91 |
| Open Datasets | 1.35 | 1.62 | 3.61 | 0.53 | 2.77 | 1.6 |
| Dataset Splits | -0.15 | -0.61 | 3.91 | -2.24 | 0.28 | 2.1 |
| Hardware Specification | 1.56 | 7.35 | -0.34 | 1.23 | 0.81 | 11.74 |
| Software Dependencies | 0.37 | -0.25 | -1.3 | 0.73 | 0 | 1.64 |
| Experiment Setup | -0.02 | -0.78 | 1.98 | -0.69 | -0.8 | 3.06 |

**Table 1** The slope of the trend line for each reproducibility variables before and after reproducibility checklists were introduced by the conference organizers. For all papers, we used the year 2019, when NeurIPS introduced the first reproducibility checklist of the five leading AI conferences.

Reproducibility checklists have been introduced to improve and standardize documentation practices in the field, as checklists help establish a higher standard of baseline performance [54] yet little information on the effect of these checklists is available. To determine whether reproducibility checklists have had a positive effect on actual or predicted reproducibility of AI research, we analysed the correlation between the introduction of reproducibility checklists and the reproducibility variables. Four of the five leading AI conferences have introduced reproducibility checklists (Supplementary Table 11). NeurIPS was first in 2019, followed by AAAI and IJCAI in 2021, and ICML in 2023. ICLR has introduced a guideline in 2022 to include an optional reproducibility statement. Authors may self-select into conferences based on their requirements, including the presence of reproducibility checklists, a factor we are unable to control for. We fitted a trend line to each reproducibility variable. The slope quantifies the average rate of change in percentage points per year (pp/year). We expect to see an effect of

**Reproducibility variable documentation rates with pre- and post-checklist trend lines, across all five conferences from 2014 to 2024.**

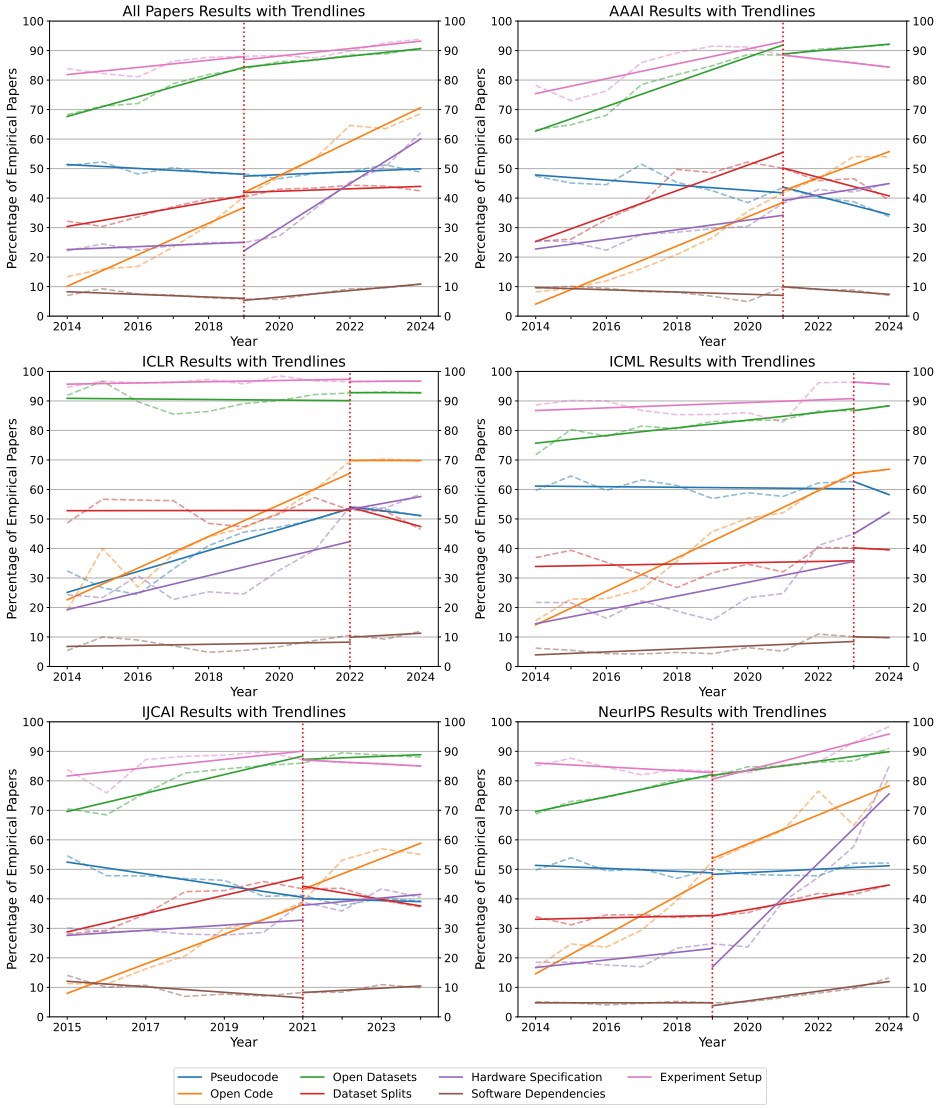


**Fig. 6** The percentage of empirical papers documenting seven reproducibility variables across all five conferences and for each conference individually from 2014 to 2024. Dashed lines show the observed annual percentages for each reproducibility variable; solid lines show the trend lines fitted before and after reproducibility checklist introduction, with slopes quantifying the rate of change in percentage points per year (pp/year). For aggregated results across all conferences, 2019 serves as the reference year, corresponding to NeurIPS's introduction of the first reproducibility checklist. IJCAI did not hold a conference in 2014. The trends demonstrate that improvements in documentation practices were largely established before formal checklist requirements, with the notable exception of hardware specification, which showed increased slopes after checklist introduction, particularly at NeurIPS.

a reproducibility checklist – if there is any – would appear as a change in the slope of the trend line after it is introduced. To evaluate the impact of the reproducibility checklist, we test the hypothesis that the pp/year increased after checklist introduction, comparing pre- and post-checklist slopes using a one-sided binomial test.

Our analysis shows that reproducibility checklists did not systematically improve the documentation rates of the reproducibility variables; the trends towards better documentation practices were present before formal procedural mandates were introduced.

Figure 6 illustrates these trends graphically for the aggregate dataset of all papers across the conferences and for each conference individually, while Table 1 provides the slope (pp/year) values for each reproducibility variable before and after the reproducibility checklist was introduced.

Across the seven reproducibility variables and five conferences (n = 35), 15 exhibited a higher post-checklist pp/year (Supplementary Table 3). A one-sided binomial test ($H_0 : p = 0.5$) indicates this is consistent with chance variation ($p = 0.84$, Cohen's $h = -0.16$), providing no statistical evidence that the adoption of the reproducibility checklist increased the rate of improvement in documentation. The exception is *Hardware specification* (purple line in Figure 6), which experiences an increase in all conferences after the introduction of reproducibility checklist. Although the increase is moderate for most conferences, the steep increase in Figure 6 is mainly driven by NeurIPS.

## 1.4 Comparison of Academia and Industry Documentation Practices

Observations suggest that papers associated with industry-affiliated authors share less code and data than papers solely by academic authors [43, 55]. Industry-affiliated authors may have incentives to withhold code or data to protect intellectual property and to thus maintain a competitive advantage [21, 43], or may be required to do so by their employer. To examine whether there are systematic differences in documentation practices, we evaluated which affiliation type documented each reproducibility variable at a higher rate. We applied one-sided binomial tests to assess whether the proportion of academic or industry papers with a higher percentage exceeded the 0.5 expected by chance if both groups were equally likely. Across the seven reproducibility variables, the five conferences, and 11 years (n = 378), academic papers documented more variables 177 times, while industry-affiliated papers had a higher percentage 200 times, with one instance being a statistical tie, see Supplementary Table 4. This result according to the binomial is consistent with chance variation and does not provide statistical evidence that there are systematic differences between academia and industry.

Although it may appear that these findings are in disagreement with prior research, our analysis — in which industry-affiliated papers documented more variables overall, is broadly consistent with Gundersen [55], whose analysis found academic papers to document reproducibility variables at higher rates, and indicates a shift in the trend of open code and open dataset sharing by industry-affiliated authors beginning in 2018. When restricting our analysis to the set of conferences common to both Gundersen [55] and our study and aggregating the results following their methodology, we observe comparable patterns: academic papers document six of the seven reproducibility

**Industry affiliation rates and documentation practices relative to academia across five AI conferences from 2014 to 2024.**

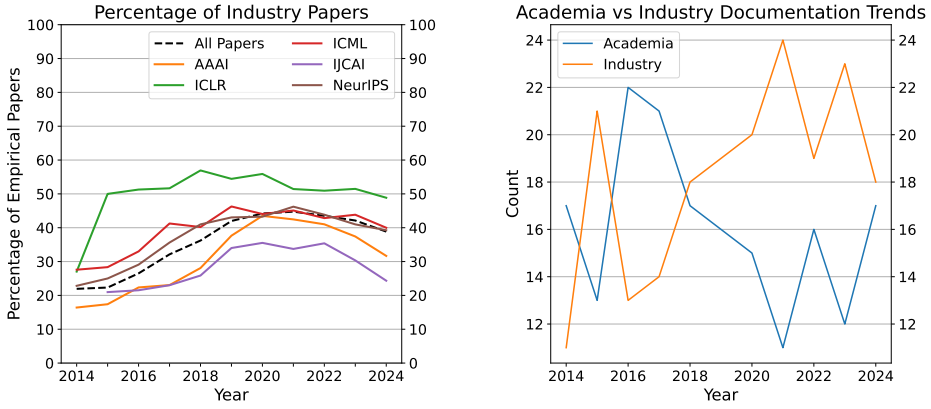


**Fig. 7** **Left**: The percentage of empirical papers at the five AI conferences that have at least one or more authors with industry affiliation. While the overall number of industry-affiliated authors has increased from 2014 to 2024, there has been a notable decline since the peak between 2018 and 2021. IJCAI did not hold a conference in 2014. **Right**: For each year, we counted — across all 35 variable-conference combinations (seven reproducibility variables × five conferences) — how many combinations academic papers (blue) documented at a higher rate than industry-affiliated papers, and how many industry-affiliated papers (orange) documented at a higher rate. The pattern reveals a notable shift: academic papers consistently demonstrated superior documentation practices in early years (2014-2017), but industry-affiliated papers began documenting variables more comprehensively starting in 2018, a trend that has persisted through 2024. This reversal coincides with the decline in overall industry participation.

variables at higher rates, with dataset splits, a variable not examined by Gundersen [55], showing a higher rate of documentation in industry papers, see Supplementary Table 5. Extending the analysis to the entire decade revealed a notable shift beginning in 2018, when industry-affiliated papers began to document reproducibility variables at higher rates than academic papers and continued to do so in subsequent years (Figure 7 right). This improvement in documentation quality among industry-affiliated papers coincides with a decline in their overall publication share (Figure 7 left), suggesting that while fewer industry-affiliated papers are now being published, those that are exhibit more comprehensive documentation of reproducibility variables, a reversal of the trend observed earlier in the decade. Gundersen [55], based on analyses from AAAI and IJCAI conferences in 2014 and 2016, and Collberg and Proebsting [43], published in 2016, both reported that industry-affiliated papers shared code less frequently than academic ones, a pattern consistent with our results, but began to change in 2018, when industry-affiliated papers started to exhibit more comprehensive documentation practices.

A possible explanation for the relative improvement in documentation quality among industry-affiliated papers is that the number of industry papers has declined since the 2018–2021 peak (Figure 7 left). What remains as industry submissions may be essentially academic research conducted by researchers with academic backgrounds

following similar standards. Consequently, when industry-affiliated researchers do choose to publish at conferences, they may do so with the intention of including code and data; the materials may no longer offer a competitive edge, or to align with strategic or reputational goals.

### 1.5 The Prevalence of Theoretical Research

The trend in the last decade reveals a diminishing proportion of purely theoretical work, driven by the rapid expansion of empirical AI research rather than a contraction of theoretical work in absolute numbers (Supplementary Figure 1). Across the five leading AI conferences, the proportion of theoretical papers decreased from 10% in 2014 to 6% in 2024. This decline has been steady after a peak of 15% in 2015. As the field has shifted toward empirical research, open science practices have become increasingly important. Unlike theoretical work, where the logical and mathematical rigour of proofs is self-contained within papers, empirical research depends on access to external artifacts and explicit methodological descriptions for independent verification. With empirical papers now constituting 94% of the published works at the five leading AI conferences in 2024, open science practices, such as comprehensive documentation, open source code, and open data, are now even more important to ensure the rigour, reliability, and trustworthiness of AI research.

An analysis of NeurIPS 2019 found that 9% of the submissions were designated by the authors as mainly theoretical [21]. Our LLM-based methodology, which analyses published papers rather than author-reported metadata, identified 10.3% of accepted papers as theoretical. The close agreement between these estimates, despite their derivation from different data sources and methodologies, provides evidence for the validity of our approach.

## 2 Discussion

We analysed 11 years of AI research from five top-tier AI conferences to assess how documentation trends have changed and influenced the reproducibility of empirical AI studies. Across 56 800 papers, documentation improved substantially over time: the share of empirical papers that provided documentation for five or more of the reproducibility variables increased from 8% in 2014 to 43% in 2024. This improvement included increased sharing of source code and datasets, which raised the estimated reproducibility rate of papers from 28% to 64%, under the assumptions described in Section 1.2. Statistical analysis revealed no evidence that the introduction of reproducibility checklists accelerated these changes. The upward trend in documentation started several years before the introduction of reproducibility checklists, indicating that broader shifts toward open science and improved documentation practices either preceded formal reproducibility requirements at these conferences or that formal requirements emerged as a response to the same underlying trends driving improved documentation practices.

Given this progress, a critical question emerges: can this trajectory be sustained? The slowing growth in our estimated reproducibility rate in recent years suggests the community may be approaching practical limits. One limit is the use of closed datasets,

which accounted for 9% of empirical papers in 2024. Achieving 100% open data is neither realistic nor desirable, as legitimate privacy and confidentiality concerns renders in some cases openly sharing data unrealistic.

A more tractable frontier for improvement lies with the 27% of papers that use open datasets but do not share code. Gundersen et al. [45] finds one third of these papers to be reproducible; closing this gap represents the most direct path to increasing the estimated reproducibility rate as the mentioned concerns of sharing of data rarely applies to code. Intellectual property is cited as a concern for researchers working in industry [21, 43]. As AI research has become increasingly profitable, industry may have become more reluctant to publish work that is closest to actual applications to maintain a competitive advantage. However, if the implementation is described in sufficient detail to enable reproduction, code sharing reveals no additional proprietary information while reducing barriers to building on existing work, though this efficiency gain may diminish authors' competitive advantage in pursuing follow-up research.

Withholding code reduces transparency and hampers reproducibility, as code often contains crucial implementation details that may not be documented. Although there may be good reasons for withholding both code and data, the default should be an expectation to share. Those who cannot share should be asked to state why.

Overcoming these challenges to push beyond the current plateau will require a more fundamental evolution in how we author and publish AI research. The principles behind initiatives such as the "Geoscience Paper of the Future", articulated by Gil et al. [56], would strengthen AI reproducibility by ensuring that data and software are reusable, properly licensed, and persistently identifiable, while computational workflows are explicitly documented to capture provenance and support transparent verification and reuse. This is critical not just for human researchers, but for the next generation of AI research. As argued by Gil [57], published papers are often too unstructured for automated analysis and reproduction. Tools that analyse papers, code, and data – such as Reproscreener [58], which automatically assesses computational reproducibility – and Reproducibility Copilots – such as OpenPub [59], which uses AI to facilitate replication by generating Jupyter Notebooks with code and actionable recommendations – demonstrate the use of AI to work around the challenges in extracting reproducibility information from traditional paper formats. These tools must work around incomplete or unstructured documentation and would benefit from an approach to publication that prioritizes documentation and open artifacts. By making publications machine-readable, we facilitate the development of AI systems that can independently reproduce, extend, and ultimately generate new scientific discoveries on their own. Researchers have already developed tools to help scientists generate novel hypotheses and research proposals, such as Google AI co-scientist [60]; although not without limitations [61], it has already shown to be valuable [62, 63].

Our study has several limitations. First, our LLM-based analysis relied exclusively on the plain text extracted from papers, excluding information potentially contained in figures, tables, supplementary materials, or external repositories not mentioned in the text. Second, our assessment of reproducibility variables is binary; it does not capture the quality or completeness of the shared artifacts. For example, a paper sharing partial code is treated the same as one with a complete, well-documented

repository. Third, LLMs can exhibit generalisation bias [64], and while our prompts were designed to mitigate this, a subtle influence on the model cannot be ruled out. Finally, our estimation of reproducibility is based on the availability of code and data and is highly uncertain. The numbers on which we base the analysis come from a small study that evaluated whether the authors were able to draw the same conclusions from the same data. The study did not evaluate whether the claims can be generalised to new data. Still, these are the best source data that we have found to base a field at scale estimation of reproducibility from.

Future work should focus on addressing these limitations, and on expanding the scope of inquiry. Methodologically, future tools could move beyond binary classifications and plain text to assess documentation quality, parse code repositories, and interpret figures. Additional descriptive analyses, breakdowns by institution type, subfield, and geographic region, and a before/after analysis centered on the public release of AI coding assistants would complement the longitudinal trends we report. Expanding the analysis to include more conferences and journals would enable a more rigorous examination of researcher self-selection into venues with or without reproducibility checklists, and could support an empirical estimate of the probability of reproducibility when all seven variables are documented.

Furthermore, comparing trends in AI to other scientific disciplines would provide valuable context, revealing whether this embrace of open science is unique to AI or part of a larger movement.

# 3 Methods

## 3.1 Reproducibility Variable Selection

The 20 variables identified by Gundersen and Kjensmo [13] provide the basis for our selection of the reproducibility variables used to assess the progress of open science in AI research. Gundersen and Kjensmo [13] provided both the conceptual framework and, critically, a prompt optimization dataset of 400 manually annotated papers. Building on this existing dataset allowed us to begin prompt engineering without first investing substantial time and resources in manual annotation, and the breadth of the variable set covers the major dimensions of empirical AI research documentation — method, data, experiment, and metadata — making it a well-motivated starting point. These include aspects of the method documentation (problem, objective/goal, research method, research questions, pseudocode), data documentation (training data, validation data, test data, results), experiment documentation (hypothesis, prediction, method source code, hardware specifications, software dependencies, experiment setup, experiment source code), and miscellaneous information about the research (research type, research outcome, affiliation, and contribution). The necessity of documenting these variables is rooted in the principle that sufficient detail must be provided for an independent team to reproduce the work. The variables are primarily Boolean, with one binary variable, research type (empirical/theoretical), and one ternary variable, author affiliation (academia/industry/collaboration).

A subsequent replication study by Gundersen et al. [45] reinforced the critical importance of sharing code and data. This finding prompted us to refine the original

data-related variables. Instead of checking for separate training, validation, and test data sharing, we replaced them with two new reproducibility variables, reducing the effective variable set from 20 to 19:

- Open Datasets: This is true if the paper uses a well-known public dataset or shares the dataset via a URL, DOI, or formal citation.
- Dataset Splits: This is true if the paper specifies how data is split, whether through exact percentages, absolute counts, predefined splits, stratified methods, or cross-validation.

To select a robust subset of variables for our automated analysis, we applied three criteria. First, we required class balance, excluding variables where one class had too few instances for reliable LLM evaluation; the first criterion eliminated seven variables, Result Outcomes, Research Method, Research Question, Hypothesis, Prediction, Open Experiment Code, and Results, from consideration. Supplementary Table 6 shows the class balance for each reproducibility variable. Second, we prioritized variables deemed most important for reproducibility based on the findings of Gundersen et al. [45], which emphasised that providing code and data are paramount. Accordingly, we classified pseudocode, open datasets, dataset splits, results, open code, hardware specification, software dependencies, experiment setup, and open experiment code as important. We also included affiliation and research type as important variables, as their results may influence whether code and data are shared. Third, we required the LLM to achieve an $F_1$ score of 75% or greater for the reproducibility variable during our prompt engineering phase on the prompt optimisation dataset; this threshold was selected to ensure reliable classification while retaining sufficient variables for a comprehensive analysis. Supplementary Table 2 shows the $F_1$ scores for the reproducibility variables on the prompt optimisation dataset. The three remaining excluded variables, Problem Description, Goal/Objective, and Contribution, passed the class balance criterion but did not achieve F1 scores of 75% or greater on the prompt optimization dataset. These variables are not always stated explicitly using consistent terminology in papers and require a higher level of contextual understanding than we were able to achieve reliably through prompting on the LLMs we tested. The excluded variables are not necessarily less important in principle; rather, their reliable automated extraction was not achievable within our current methodology, and we chose not to include any variable for which we lacked confidence in the results.

We made modifications to the original dataset from Gundersen and Kjensmo [13] to better suit the capabilities of our LLM-based evaluation. The open code variable was reevaluated; the original study verified if the code was still accessible at the provided link, a task current LLMs cannot perform. Our revised criterion only checks if the paper mentions a link to the code. We also reevaluated software dependencies to ensure the criterion was consistently applied, requiring the mention of both a software package and its version number. Justifications, with quotes from the papers, for the changes to open datasets, dataset splits, open code and software dependencies are provided in supplementary material. For the affiliation variable, we noticed that the original dataset contained only 11 papers in the industry category (i.e., with all authors from industry). With such a small sample size, we would not have been able to reliably

assess the ability of our method to detect this type of papers. We thus used a binary classification instead: papers written by authors solely from academia *vs* ones with at least one author affiliated with industry.

## 3.2 Automated Evaluation of Reproducibility Variables

To develop and validate the LLM prompts used for our automated evaluation method, we used the dataset created by Gundersen and Kjensmo [13] containing 400 papers from AAAI (2014 and 2016) and IJCAI (2013 and 2016) and manually evaluated for their 20 reproducible variables, which we refer to as the prompt optimisation dataset. We downloaded PDF versions of the 400 papers from the AAAI and IJCAI proceedings websites and extracted just the text of the papers, excluding figures but including figure captions and the text of the tables, without the table formatting. Then we created a script to automate the process of submitting the paper text and the reproducibility variable evaluation prompt to the APIs for Anthropic, Google, OpenAI and open weight models for inference and save the results in a structured format.

We undertook an iterative prompt engineering process to optimise LLM accuracy on each of the 20 reproducibility variables with the prompt optimisation dataset. We did not modify the weights of the model or use other techniques, such as Retrieval-Augmented Generation (RAG). Instead, we utilised a few-shot prompting technique. A few-shot prompt provides a few natural language demonstrations of the task at inference time, but no weight updates are performed [65]. As the papers analysed in this study were almost certainly used to train the LLMs we evaluated, the degree to which responses reflect in-context reasoning versus recall remains an open question; the method nonetheless achieved F1 scores exceeding 90.0% for six of nine variables on the Evaluation Dataset (Supplementary Table 9). For each variable, we drafted a few-shot prompt and used it to evaluate a subset of the 400 papers. For each reproducibility variable we asked the LLM to return two values: (1) the Boolean, binary, or ternary result of the engineered prompt and (2) a string result that includes a quote of the text from the paper that supports the first result, otherwise explain why the information in the paper is insufficient. The classifications obtained from the LLM were then compared against the ground truth of the prompt optimisation dataset using a confusion matrix. We closely examined the false positives and false negatives, reviewing the quotes from the LLMs string output with the source text of the papers to identify patterns of error. This analysis informed subsequent revisions of the few-shot prompt and the cycle was repeated until the performance of the LLM stabilised at a high level of accuracy and $F_1$ score. Examples demonstrating the model's handling of negated statements, such as explicit declarations that a dataset is not publicly available, are provided in Supplementary Section S6.

The final prompt submitted to the model for each paper consisted of a single API query containing a brief task description followed by the full plain text of the paper and the specific few-shot prompts for each of the 20 variables. The prompts for each variable asked the LLM to determine if the paper met the criterion, provided a list of positive and negative examples to guide its decision, requested a direct quote from the paper to support its answer, and concluded with a final yes/no question. The model was instructed to return its output in a structured JSON format, containing its binary

answer and the supporting text from the paper. The full prompt, excluding the text of the individual papers due to size limitations, is available in supplementary material.

### 3.3 Model Selection and Prompt Engineering

We evaluated 17 models from Anthropic, Google, OpenAI, Alibaba Cloud, and Microsoft, of which three were open weight models (Supplementary Table 8). We selected the Google Gemini 2.5 Flash model for the final large-scale evaluation based on four key criteria:

*LLM API Features:* The task required reliable structured output. The Google, OpenAI, and Ollama (used with the open weight models) APIs support generating JSON output based on a predefined schema, ensuring consistency. The Anthropic API lacked this native capability and was therefore eliminated.
*Time:* Due to limited access to GPU resources, we excluded the three open weights models, since we could not process 58 600 papers within a reasonable timeframe.
*Cost:* Reserving funds for reasoning and output tokens, which are more expensive and difficult to predict the length of, we selected models that would fit our budget for input tokens for all 56 800 papers, based on an estimated 13 000 input tokens per paper (Estimated costs are in Supplementary Table 8).
*Accuracy:* In preliminary tests, Gemini 2.5 Flash demonstrated superior accuracy on our task compared to other models within our budget constraints, a result corroborated by the public LLM leaderboard LMArena (https://lmarena.ai/) [**?** ].

The LLM-generated results for each of the 400 papers were systematically evaluated against the ground truth from the prompt optimisation dataset over five runs. We calculated the mean accuracy and mean $F_1$ score for each reproducibility variable to quantify model performance and selected affiliation, research type, pseudocode, open datasets, dataset splits, open code, hardware specification, software dependencies, and experiment setup as the final variables for our large-scale analysis. The results for each run can be found in Supplementary Table 7.

Google Gemini 2.5 Flash demonstrated high performance and stability across five evaluation runs. For affiliation and pseudocode, mean $F_1$ scores greater than 90% was achieved. For research type and hardware specification, open code, open datasets, and dataset splits, we found $F_1$ scores greater than 80%. Software dependencies and experiment setup were also predicted reliably, with $F_1$ scores of 78.3% and 78.8%, respectively. The variance across runs differed substantially for the different variables; the accuracy range varied from 0.5% (hardware specification) to 4.8% (dataset splits), and the $F_1$ score range varied from 1.0% (hardware specification) to 11.9% (software dependencies) – see Supplementary Table 7 for all means and ranges. Furthermore, the consistency of classifications for individual papers was high. For research type, affiliation, pseudocode, open code, hardware specification, and software dependencies, the evaluation was identical across all five runs over 95% of the time. Open dataset, dataset splits, and experiment setup also showed high consistency, with consistent evaluations 87%, 84%, and 74% of the time, respectively, see Figure 8.

**LLM classification consistency across five runs for each reproducibility variable.**

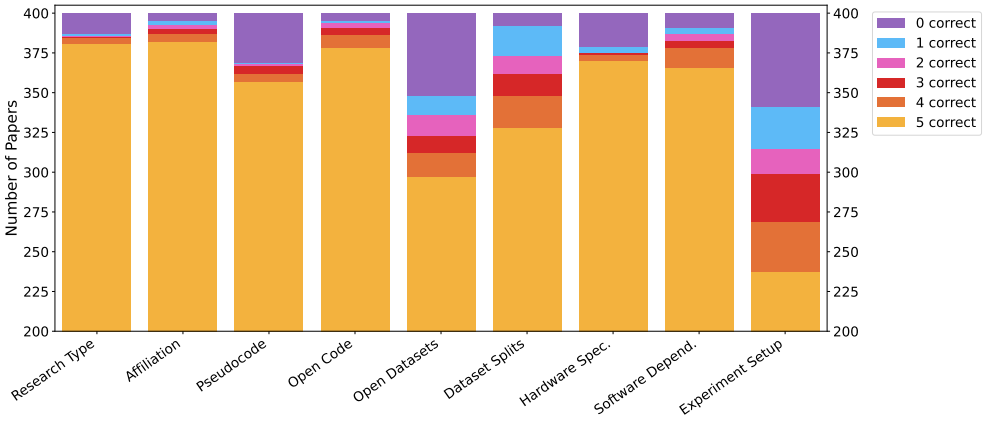


**Fig. 8** The consistency of classification from the LLM over five runs for each of the reproducibility variables. Light orange represents the number of papers the LLM consistently correctly classified reproducibility variable and the purple represents the number of papers the LLM consistently incorrectly classified reproducibility variable. The dark orange, red, pink and blue, represents the number of papers the LLM did not consistently make the same classification over multiple runs.

## 3.4 Large-Scale Conference Paper Analysis

For our large-scale analysis, we selected five top-tier AI conferences: AAAI, ICLR, ICML, IJCAI, and NeurIPS. We collected all published papers over an 11-year period to analyse trends over time. After excluding papers from affiliated workshops and papers from joint proceedings, which often have different reviewing criteria, our final dataset comprised 56 800 main conference papers. We did not include journals in our analysis. Although these journals, such as JAIR and JMLR, play major roles in AI research, we excluded these due to time constraints. This set also included 300 papers from the original analysis by Gundersen and Kjensmo [13] (AAAI 2014, AAAI 2016, and IJCAI 2016). A detailed description of the paper selection and collection process is available in Supplementary Section S3.

The workflow for processing the 56 800 papers mirrored the one used for the initial 400-paper evaluation. Each PDF was converted to text, and the same engineered prompt was used with Google Gemini 2.5 Flash for all of the reproducibility variables. Only the nine selected reproducibility variables from Section 1.1 were analysed. The results from the 56 800 paper LLM analysis containing the binary answer and the supporting text from the paper are provided as supplementary material.

## 3.5 Does the Method Generalise to the Population?

To validate how well our approach generalised from the prompt optimisation dataset to the full set of papers, we sampled 160 papers uniformly at random from full set of 56 800 papers and labelled them manually to create an evaluation dataset. The results are presented in Supplementary Table 9. The year and conference distributions of the

evaluation dataset closely mirror those of the full corpus (Supplementary Figure 2 and Supplementary Figure 3), supporting the generalisability of the results reported below. Surprisingly, the $F_1$ scores show a substantial improvement for all classes, achieving greater than 90.0% with the exception of three variables; data set splits (81.48%), software dependencies (81.25%) and experiment setup (86.21%). This increase in performance can be attributed to various factors. First, the prompt optimisation dataset is a biased subset of the population, representing only AAAI and IJCAI in 2013, 2014 and 2016. Secondly, the LLM was not actually fitted to the training data, but rather the prompts were optimised, which prevents possible overfitting through backpropagation. Only for one variable a decrease in performance was measured between the prompt optimisation and evaluation datasets: for dataset splits performance dropped from 81.48% to 79.56%, a decrease of 1.92 percentage points. We found that the majority of the papers in the evaluation set that had mislabelled data splits were caused by inflexibility in the prompt; in 60.0% of the errors, the procedure produced a false negative due to the authors choosing not to use a validation set and therefore not providing the full data split containing training, validation and test sets. Overall, the results show us that our method generalises well to the full dataset of papers.

**Supplementary information.** The LLM prompt, code, results, and analysis can be found online: https://doi.org/10.5281/zenodo.20785801 [66].

**Acknowledgements.** Cloud computing services provided by CloudBank: National Science Foundation [# 1925001].

# Declarations

## 3.6 Author contributions

K.L.C. conceived and designed the experiments, performed the experiments, analysed the data, contributed materials/analysis tools, and wrote the paper. T.S. analysed the data, contributed materials/analysis tools, and wrote the paper. H.H. wrote the paper. O.E.G. conceived and designed the experiments, analysed the data, contributed materials/analysis tools, and wrote the paper.

## 3.7 Competing interest

The authors declare no competing interests.

## 3.8 Data availability

The datasets generated during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.20785801 [66]. We do not include the PDFs for of all 56 800 papers analysed for this study due to copyright concerns. The Zenodo repository includes the code to download and preprocess all 56 800 papers.

## 3.9 Code availability

The code used for the experiment, include preprocessing and analysis, is available in the Zenodo repository, https://doi.org/10.5281/zenodo.20785801 [66].

## 3.10 Funding declaration

K.L.C discloses support for the research of this work from National Science Foundation [# 2226453]. O.E.G. discloses support for the research of this work from RICO (Robust Intelligent Control, Research Council of Norway) [# 329730]. T.S. and H.H. discloses support for the research of this work from through an Alexander von Humboldt Professorship held by Holger Hoos from the Alexander von Humboldt Foundation.

## References


[1] Ioannidis, J.P.: Why most published research findings are false. PLoS medicine **2**(8), 124 (2005) https://doi.org/10.1371/journal.pmed.0020124

[2] McNutt, M.: Reproducibility. American Association for the Advancement of Science (2014). https://doi.org/10.1126/science.1250475

[3] Baker, M.: 1,500 scientists lift the lid on reproducibility. Nature Publishing Group UK London (2016). https://doi.org/10.1038/533452a

[4] Pashler, H., Wagenmakers, E.-J.: Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? Perspectives on psychological science **7**(6), 528–530 (2012) https://doi.org/10.1177/1745691612465253

[5] Open Science Collaboration: Estimating the reproducibility of psychological science. Science **349**(6251), 4716 (2015) https://doi.org/10.1126/science.aac4716

[6] Klein, R.A., Ratliff, K.A., Vianello, M., Adams Jr, R.B., Bahník, Š., Bernstein, M.J., Bocian, K., Brandt, M.J., Brooks, B., Brumbaugh, C.C., *et al.*: Investigating variation in replicability. Social psychology (2014) https://doi.org/10.1027/1864-9335/a000178

[7] Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., *et al.*: Evaluating replicability of laboratory experiments in economics. Science **351**(6280), 1433–1436 (2016) https://doi.org/10.1126/science.aaf09

[8] Prinz, F., Schlange, T., Asadullah, K.: Believe it or not: how much can we rely on published data on potential drug targets? Nature reviews Drug discovery **10**(9), 712–712 (2011) https://doi.org/10.1038/nrd3439-c1

[9] Collins, F.S., Tabak, L.A.: Policy: NIH plans to enhance reproducibility. Nature **505**(7485), 612–613 (2014) https://doi.org/10.1038/505612a

[10] Begley, C.G., Ellis, L.M.: Raise standards for preclinical cancer research. Nature **483**(7391), 531–533 (2012) https://doi.org/10.1038/483531a

[11] Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R.: Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews neuroscience **14**(5), 365–376 (2013) https://doi.org/10.1038/nrn3475

[12] Hewitt, J.K.: Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. Behavior genetics **42**(1), 1–2 (2012) https://doi.org/10.1007/s10519-011-9504-z

[13] Gundersen, O.E., Kjensmo, S.: State of the art: Reproducibility in artificial

intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.11503

[14] Hutson, M.: Artificial intelligence faces reproducibility crisis. Science **359**(6377), 725–726 (2018) https://doi.org/10.1126/science.359.6377.725

[15] Vicente-Saez, R., Martinez-Fuentes, C.: Open science now: A systematic literature review for an integrated definition. Journal of Business Research **88**, 428–436 (2018) https://doi.org/10.1016/j.jbusres.2017.12.043

[16] Bischl, B., Casalicchio, G., Das, T., Feurer, M., Fischer, S., Gijsbers, P., Mukherjee, S., Müller, A.C., Németh, L., Oala, L., *et al.*: OpenML: Insights from 10 years and more than a thousand papers. Patterns (2025) https://doi.org/10.1016/j.patter.2025.101317

[17] Foster, E.D., Deardorff, A.: Open science framework (OSF). Journal of the Medical Library Association: JMLA **105**(2), 203 (2017) https://doi.org/10.5195/jmla.2017.88

[18] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L.B., Bourne, P.E., *et al.*: The FAIR guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016) https://doi.org/10.1038/sdata.2016.18

[19] White, M., Haddad, I., Osborne, C., Liu, X.-Y.Y., Abdelmonsef, A., Varghese, S., Hors, A.L.: The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence. arXiv preprint arXiv:2403.13784 (2024) https://doi.org/10.48550/arXiv.2403.13784

[20] Wilkinson, S.R., Aloqalaa, M., Belhajjame, K., Crusoe, M.R., Paula Kinoshita, B., Gadelha, L., Garijo, D., Gustafsson, O.J.R., Juty, N., Kanwal, S., *et al.*: Applying the FAIR principles to computational workflows. Scientific Data **12**(1), 328 (2025) https://doi.org/10.1038/s41597-025-04451-9

[21] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Larochelle, H.: Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). Journal of machine learning research **22**(164), 1–20 (2021)

[22] Gundersen, O.E., Helmert, M., Hoos, H.: Improving reproducibility in AI research: Four mechanisms adopted by JAIR. Journal of Artificial Intelligence Research **81**, 1019–1041 (2024) https://doi.org/10.1613/jair.1.16905

[23] Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are gans created equal? a large-scale study. Advances in neural information processing systems **31** (2018)

[24] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D.: Deep reinforcement learning that matters. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.11694

[25] Ferrari Dacrema, M., Cremonesi, P., Jannach, D.: Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp. 101–109 (2019). https://doi.org/10.1145/3298689.3347058

[26] Belz, A.: A metrological perspective on reproducibility in NLP. Computational Linguistics **48**(4), 1125–1135 (2022) https://doi.org/10.1162/coli_a_00448

[27] Gundersen, O.E., Shamsaliei, S., Kjærnli, H.S., Langseth, H.: On reporting robust and trustworthy conclusions from model comparison studies involving neural networks and randomness. In: Proceedings of the 2023 ACM Conference on Reproducibility and Replicability, pp. 37–61 (2023). https://doi.org/10.1145/3589806.3600044

[28] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021) https://doi.org/10.1145/3458723

[29] Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine **5**(1), 48 (2022) https://doi.org/10.1038/s41746-022-00592-y

[30] Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns **4**(9) (2023) https://doi.org/10.1016/j.patter.2023.100804

[31] Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Directors Shraddha Thakkar 35 Kusko Rebecca 36 Sansone Susanna-Assunta 37 Tong Weida 35 Wolfinger Russ D. 38 Mason Christopher E. 39 Jones Wendell 40 Dopazo Joaquin 41 Furlanello Cesare 42, M.A.Q.C.M.S.B., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., *et al.*: Transparency and reproducibility in artificial intelligence. Nature **586**(7829), 14–16 (2020) https://doi.org/10.1038/s41586-020-2766-y

[32] Belz, A., Agarwal, S., Shimorina, A., Reiter, E.: A systematic review of reproducibility research in natural language processing. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 381–393 (2021). https://doi.org/10.18653/v1/2021.eacl-main.29

[33] Bouthillier, X., Laurent, C., Vincent, P.: Unreproducible research is reproducible. In: International Conference on Machine Learning, pp. 725–734 (2019). PMLR

[34] Hunold, S., Carpen-Amarie, A.: Reproducible MPI benchmarking is still not as

easy as you think. IEEE Transactions on Parallel and Distributed Systems **27**(12), 3617–3630 (2016) https://doi.org/10.1109/TPDS.2016.2539167

[35] Hong, S.-Y., Koo, M.-S., Jang, J., Esther Kim, J.-E., Park, H., Joh, M.-S., Kang, J.-H., Oh, T.-J.: An evaluation of the software system dependency of a global atmospheric model. Monthly Weather Review **141**(11), 4165–4172 (2013) https://doi.org/10.1175/MWR-D-12-00352.1

[36] Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P., Taufer, M.: Enhancing reproducibility for computational methods. Science **354**(6317), 1240–1241 (2016) https://doi.org/10.1126/science.aah6168

[37] Ajayi, K., Choudhury, M.H., Rajtmajer, S.M., Wu, J.: A study on reproducibility and replicability of table structure recognition methods. In: International Conference on Document Analysis and Recognition, pp. 3–19 (2023). https://doi.org/10.1007/978-3-031-41679-8_1 . Springer

[38] Gundersen, O.E., Coakley, K., Kirkpatrick, C., Gil, Y.: Sources of irreproducibility in machine learning: A review. arXiv preprint arXiv:2204.07610 (2022) https://doi.org/10.48550/arXiv.2204.07610

[39] Gundersen, O.E.: The fundamental principles of reproducibility. Philosophical Transactions of the Royal Society A **379**(2197), 20200210 (2021) https://doi.org/10.1098/rsta.2020.0210

[40] Schmidt, S.: Shall we really do it again? the powerful concept of replication is neglected in the social sciences. Review of general psychology **13**(2), 90–100 (2009) https://doi.org/10.1037/a0015108

[41] Nosek, B.A., Lakens, D.: A method to increase the credibility of published results. Social Psychology **45**(3), 137–141 (2014) https://doi.org/10.1027/1864-9335/a000192

[42] Goodman, S.N., Fanelli, D., Ioannidis, J.P.: What does research reproducibility mean? Science translational medicine **8**(341), 341–1234112 (2016) https://doi.org/10.1126/scitranslmed.aaf5027

[43] Collberg, C., Proebsting, T.A.: Repeatability in computer systems research. Communications of the ACM **59**(3), 62–69 (2016) https://doi.org/10.1145/2812803

[44] Raff, E.: A step toward quantifying independently reproducible machine learning research. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, vol. 32. Curran Associates Inc., Red Hook, NY, USA (2019)

[45] Gundersen, O.E., Cappelen, O., Mølnå, M., Nilsen, N.G.: The unreasonable effectiveness of open science in AI: A replication study. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 26211–26219 (2025). https://doi.org/10.1609/aaai.v39i25.34818

[46] Magnusson, I., Smith, N.A., Dodge, J.: Reproducibility in NLP: What have we learned from the checklist? In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 12789–12811 (2023). https://doi.org/10.18653/v1/2023.findings-acl.809

[47] Gundersen, O.E., Gil, Y., Aha, D.W.: On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. AI magazine **39**(3), 56–68 (2018) https://doi.org/10.1609/aimag.v39i3.2816

[48] Makridakis, S., Spiliotis, E., Assimakopoulos, V.: Statistical and machine learning forecasting methods: Concerns and ways forward. PloS one **13**(3), 0194889 (2018) https://doi.org/10.1371/journal.pone.0194889

[49] Pouchard, L., Lin, Y., Van Dam, H.: Replicating machine learning experiments in materials science. In: Parallel Computing: Technology Trends, pp. 743–755. IOS Press, Amsterdam (2020). https://doi.org/10.3233/APC200105

[50] Coakley, K., Kirkpatrick, C.R., Gundersen, O.E.: Examining the effect of implementation factors on deep learning reproducibility. In: Proceedings of the IEEE 18th International Conference on e-Science (e-Science), pp. 397–398 (2022). https://doi.org/10.1109/eScience55777.2022.00056 . IEEE

[51] Zhuang, D., Zhang, X., Song, S., Hooker, S.: Randomness in neural network training: Characterizing the impact of tooling. In: Marculescu, D., Chi, Y., Wu, C. (eds.) Proceedings of the Fourth Conference on Machine Learning and Systems, vol. 4, pp. 316–336 (2022)

[52] Cooper, A.F., Lu, Y., Forde, J., De Sa, C.M.: Hyperparameter optimization is deceiving us, and how to stop it. Advances in Neural Information Processing Systems **34**, 3081–3095 (2021)

[53] Reimers, N., Gurevych, I.: Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 338–348 (2017). https://doi.org/10.18653/v1/D17-1035

[54] Gawande, A.: The Checklist Manifesto: How to Get Things Right. Metropolitan Books, New York City, New York (2010)

[55] Gundersen, O.E.: Standing on the feet of giants—reproducibility in AI. Ai Magazine **40**(4), 9–23 (2019) https://doi.org/10.1609/aimag.v40i4.5185

[56] Gil, Y., David, C.H., Demir, I., Essawy, B.T., Fulweiler, R.W., Goodall, J.L., Karlstrom, L., Lee, H., Mills, H.J., Oh, J.-H., *et al.*: Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance. Earth and Space Science **3**(10), 388–415 (2016) https://doi.org/10.1002/2015EA000136

[57] Gil, Y.: Will AI write scientific papers in the future? AI Magazine **42**(4), 3–15 (2022) https://doi.org/10.1609/aaai.12027

[58] Bhaskar, A., Stodden, V.: Reproscreener: Leveraging LLMs for assessing computational reproducibility of machine learning pipelines. In: Proceedings of the 2nd ACM Conference on Reproducibility and Replicability, pp. 101–109 (2024). https://doi.org/10.1145/3641525.3663629

[59] Bibal, A., Minton, S.N., Khider, D., Gil, Y.: AI copilots for reproducibility in science: A case study. arXiv preprint arXiv:2506.20130 (2025) https://doi.org/10.48550/arXiv.2506.20130

[60] Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., *et al.*: Towards an AI co-scientist. arXiv preprint arXiv:2502.18864 (2025) https://doi.org/10.48550/arXiv.2502.18864

[61] Cheetham, A.K., Seshadri, R.: Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery. Chemistry of Materials **36**(8), 3490–3495 (2024) https://doi.org/10.1021/acs.chemmater.4c00643

[62] Guan, Y., Cui, L., Inchai, J., Fang, Z., Law, J., Brito, A.A.G., Pawlosky, A., Gottweis, J., Daryin, A., Myaskovsky, A., *et al.*: AI-assisted drug re-purposing for human liver fibrosis. Advanced Science **12**(44), 08751 (2025) https://doi.org/10.1002/advs.202508751

[63] He, L., Patkowski, J.B., Wang, J., Miguel-Romero, L., Aylett, C.H.S., Fillol-Salom, A., Costa, T.R.D., Penadés, J.R.: Chimeric infective particles expand species boundaries in phage-inducible chromosomal island mobilization. Cell **188**(23), 6636–665317 (2025) https://doi.org/10.1016/j.cell.2025.08.019

[64] Peters, U., Chin-Yee, B.: Generalization bias in large language model summarization of scientific research. Royal Society Open Science **12**(4), 241776 (2025) https://doi.org/10.1098/rsos.241776

[65] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[66] Coakley, K.L., Snelleman, T., Hoos, H., Gundersen, O.E.: GitHub: Kevincoakley/ai-research-moves-towards. https://doi.org/10.5281/zenodo.20785801