The AI Reproducibility Index: Ranking Venues, Countries and Institutions by Reproducibility

Kevin L Coakley
San Diego Supercomputer Center
University of California San Diego
La Jolla, California, USA
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Trønderlag, Norway
kcoakley@sdsc.edu

Thijs Snelleman
Chair of AI Methodology
RWTH Aachen University
Aachen, North Rhine-Westphalia, Germany
snelleman@aim.rwth-aachen.de

Holger Hoos
Chair of AI Methodology
RWTH Aachen University
Aachen, North Rhine-Westphalia, Germany
Leiden Institute of Advanced Computer Science
Leiden University
Leiden, South Holland, Netherlands
hh@aim.rwth-aachen.de

Odd Erik Gundersen
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Trønderlag, Norway
odderik@ntnu.no

Abstract

The debate about a reproducibility crisis in science has intensified, with studies—including in AI—suggesting that many results lack reproducibility. However, prior work is often limited by small samples, narrow scopes, and short time spans, hindering longitudinal insight. We introduce the AI Reproducibility Index, a platform that automatically analyses papers from leading AI venues and ranks venues, countries, and institutions by indirectly estimated reproducibility, distributed on reproducibilityindex.ai. Extending our prior conference-only analysis, we add journals, country and institution rankings, author metadata extraction, and a continuously updated platform. Applying the quality-assured LLM-based method from our prior work, reproducibilityindex.ai presents the analysis of 74 972 open-access papers from major AI conferences and journals over 12 years. We compute two metrics, which we refer to as Reproducibility score and Documentation score. Both improve from 2014 to 2025. Conferences show slightly higher reproducibility, while journals score higher on documentation. European and Asian countries and institutions dominate the rankings. Our approach enables large-scale longitudinal tracking within the scope of all of AI research.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

Reproducibility, Artificial Intelligence, Automatisation

This work is licensed under a Creative Commons Attribution 4.0 International License.
ACM REP '26, Delft, Netherlands
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2778-8/26/07
https://doi.org/10.1145/3820002.3828592

ACM Reference Format:
Kevin L Coakley, Thijs Snelleman, Holger Hoos, and Odd Erik Gundersen. 2026. The AI Reproducibility Index: Ranking Venues, Countries and Institutions by Reproducibility. In *ACM Conference on Reproducibility and Replicability (ACM REP '26), July 20–22, 2026, Delft, Netherlands.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3820002.3828592

1 Introduction

The reproducibility crisis [1] has been assessed many times over the years, with various studies showing substantial issues with the reproducibility of Artificial Intelligence (AI) research [24, 25, 37]. However, these meta-studies are time-consuming, and it usually requires months or even years of time by highly trained individuals to collect the data. These limitations result in persistent issues when it comes to consistently tracking the reproducibility in AI research over time. Among other concerns, we note three main issues affecting reproducibility studies for AI research:

- The meta study has a relatively small sample size: 400 papers [25], 255 papers [37], and 30 papers [24].
- The meta study only covers a small (sub)field, due to the expertise of the independent investigators; paper selection in Raff [37] was driven by relevance to the JSAT library [36].
- The meta study only covers a small time-span or offers little statistical power for a year-on-year analysis: three discrete publication years (2012, 2014, and 2016) [25] and 2011-2016 [24].

To mitigate these limitations, we present reproducibilityindex.ai, an automatic meta-data extraction tool [8] providing AI-based analyses of reproducibility and documentation, allowing users to investigate trends of leading venues as well as different countries and institutions. We apply the methodology from Coakley et al. [8] and extend it to a repeatable automatic and scalable system. The extended methodology provides two key metrics for each study,

both building on prior work: the reproducibility score (Section 3.2), an AI-based estimate of the likelihood that the study is reproducible, calibrated using empirical reproduction rates from Gundersen et al. [24], and the documentation score (Section 3.1), an aggregate of seven automatically computed Boolean indicators derived from the reproducibility variables identified by Gundersen and Kjensmo [25].

The present work builds on our prior study [8], which analyzed 56 800 conference papers (52 328 empirical) across five conferences (AAAI, ICLR, ICML, IJCAI, NeurIPS) over 11 years (2014–2024) and validated the LLM-based system reused here. We extend that foundation in four respects: scope, the first large-scale conference-versus-journal comparison, country and institution rankings, an LLM-based author metadata extraction method, and a publicly accessible, continuously updated platform at reproducibilityindex.ai. Specifically, our contributions are:

- We extend the methodology of Coakley et al. [8] with an ETL system and reproducibilityindex.ai - the first continuously updated public platform for automated reproducibility assessment in AI - which computes paper-level Documentation and Reproducibility Scores and aggregates them across conferences, journals, countries, and institutions with 95% BCa bootstrap confidence intervals.
- We develop an LLM-based author metadata extraction method evaluated across 335 298 authorship attributions, achieving 98.99% overall accuracy across nine data fields.
- We produce the first country-level and institution-level rankings of reproducibility and documentation in AI research, covering 42 countries and 240 institutions.
- We expand the corpus of Coakley et al. [8] from five conferences to a multi-venue collection of 74 972 open-access papers spanning both conferences and journals (DMLR, JAIR, JMLR, and TMLR), conducting the first large-scale longitudinal comparison of reproducibility and documentation practices between conferences and journals (62 833 versus 6 269 empirical papers) over 12 years (2014–2025).

We find that both our reproducibility and documentation scores improved steadily from 2014 to 2025 across conferences and journals. Conferences show a slightly higher reproducibility score than journals (0.55 *vs* 0.53), though the effect size is negligible, while journals score higher on documentation (4.08 *vs* 3.77), with a small effect size; both gaps are narrowing over time. European and Asian countries and institutions dominate the top rankings on both indices, while the United States, the largest contributor with 28 163 empirical papers, ranks 32nd of 42 countries in reproducibility score and 36th in documentation score.

2 Related Work

The field of empirical meta-science is often plagued by time-intensive labour, and more often than not requires highly trained individuals to collect or evaluate data. For example, Gundersen et al. [24] conducted a large-scale reproducibility study, spanning several years and various venues within the field of artificial intelligence (AI). In this experiment, Gundersen et al. aimed to reproduce the thirty most highly cited studies over the past few years, limiting themselves to those with public data (yielding 22 candidates out of thirty) and a forty-hour working limit per study. They found that studies with public code were reproducible in 86% of the cases; those without code and only public data were reproducible in only 33% of the cases. Although a truly large-scale experiment, with an estimated upper limit of 880 working hours, using only highly trained individuals able to replicate state-of-the-art research, the sample size is still quite small at effectively 22 data samples. Raff [37] dealt with a substantially larger sample size with 255 studies spanning 4 decades, noting that it took approximately 20 minutes per paper to only denote the meta information of each study – an estimated workload of 85 working hours before even a single reproduction attempt was started. Although Raff [37] managed to do an experiment larger by an order of magnitude, achieving much higher coverage and statistical power, they still only considered a fraction of empirical work published in AI. For example, the Raff attempted to reproduce about 7.72 papers for each year covered by their sample, hence yielding relatively low statistical power for a year-on-year analysis; the experiments were performed by 'manually attempting to implement' the code of each study, making it costly to apply the experiment protocol to monitor reproducibility over time. Furthermore, after contacting the author of Raff [37], he stated that the experiment was conducted in a timespan of seven years, making it hardly a viable option to repeat on a frequent basis for monitoring purposes. Werner et al. [44] documented similar obstacles – missing versions, undocumented hyperparameters, inaccessible data – when reproducing and reimplementing a neuro-symbolic method across three datasets.

Yet, frequent monitoring and observing reproducibility is important for the community to be able to determine which policy changes or community efforts have strong positive impact on the reproducibility of science. For example, one (albeit debatable) metric often used to measure academic impact of a publication is citation rate. As Raff [38] points out, in the field of machine learning (ML), there is a moderate positive effect of the reproducibility of publications and their citation rate. If we could monitor the reproducibility of papers continuously and at scale, we might then also be able to automatically determine the increase or reduction of the citation rate between the reproducibility and documentation score, which may directly explain one incentive of authors to make their work reproducible.

In prior work, we addressed the scale limitations described above by applying an LLM-based system to 52,328 empirical papers from five top-tier conferences over 11 years (2014–2024), achieving F1 scores exceeding 90% for six of nine reproducibility variables [8]. That study quantified more than a factor two in the estimated reproducibility score, from 28% in 2014 to 64% in 2024; it found no statistical evidence that reproducibility checklists accelerated documentation improvements; and it identified a reversal beginning in 2018, when industry-affiliated papers began documenting reproducibility variables at higher rates than academic papers. The work presented here extends that system to journals, expands coverage to 74 972 papers over 12 years, introduces country and institution rankings, and makes results continuously accessible through reproducibilityindex.ai. The methodology, prompt optimization dataset, and LLM evaluation stages are inherited from Coakley et al. [8]; Sections 4.1–4.4 summarize those components for completeness. The Extract, Transform, and Load (ETL) stage, fractional-authorship

aggregation, author metadata extraction, multi-venue corpus, and public platform are new to this work.

3 Documentation and Reproducibility Scores

To quantify reproducibility at scale, we define two complementary measures: Documentation Score and Reproducibility Score. Documentation Score captures the extent to which a paper reports information that is relevant for independent reproduction, while Reproducibility Score estimates the probability that a study can in practice be reproduced from the artifacts made available by the authors. Together, these measures provide complementary information on reproducibility-relevant documentation and the estimated likelihood of being able to reproduce the research itself.

These two metrics build on the reproducibility types and degrees introduced by Gundersen [23]. Reproducibility types describe what documentation is used when reproducing an experiment: a textual description only (R1), code and description (R2), data and description (R3), or the complete experiment (R4). Reproducibility degrees describe the level at which results agree: an experiment is outcome reproducible if the reproduced outcome is identical (for example, same class), analysis reproducible if a different outcome yields the same conclusion under the same analysis, and interpretation reproducible if a different analysis still supports the same conclusion. The Documentation Score measures reporting practices along the R1–R4 axis, while the Reproducibility Score estimates the likelihood of reproduction along the outcome analysis dimension using observed reproduction rates under different artifact-sharing conditions [24].

3.1 Documentation Score

The Documentation Score $D_p \in [0, 7]$ is based on seven reproducibility variables: pseudocode, open code, open data, dataset splits, software dependencies, hardware specifications, and experiment details. These variables were derived from the broader framework introduced by Gundersen and Kjensmo [25] and narrowed to a subset that could be identified accurately and consistently at scale using an LLM [8]. The selection was motivated by two considerations: each variable captures information that is substantively important for reproducing machine learning research, and each can be detected with sufficient reliability in paper text to support large-scale automated analysis. For each paper, the LLM assigns a value of 1 if a feature is present and 0 otherwise. The Documentation Score is the equally weighted sum of these seven binary indicators.

We motivate this approach mainly based on the following three reasons. First, the features captured by our seven variables reflect core dimensions of experimental transparency. Access to code and data directly reduces barriers to reproduction, while information about dataset splits, software dependencies, hardware, and experimental details reduces ambiguity surrounding the acquisition of results. Pseudocode further helps clarify the intended method when implementation details are not fully recoverable from the text alone. Second, prior work has shown that these factors are closely related to whether AI research can be reproduced in practice; Raff [37] found that pseudocode, readability, and specification of hyperparameters were among the features most predictive of successful independent replication. Differences in software frameworks [27, 35], software versions [7, 12, 26, 42], ancillary software [7, 12, 26, 33, 34], processing units [7, 26, 28, 32], random seeds [3, 31, 33, 40, 46], and dataset splits [3, 4, 30] have each been shown to produce substantially different results when reproducing experiments. Third, a restricted set of variables is appropriate in a large-scale study; the purpose of the Documentation Score is not to exhaustively characterise all determinants of reproducibility, but to measure a reproducibility-relevant subset of reporting practices that can be observed consistently across the vast majority of published papers. Coakley et al. [8] confirmed F1 scores exceeding 90% for most variables on an independently labelled evaluation dataset of 160 papers. In this sense, the Documentation Score should be interpreted as a measure of documentation quality with respect to reproducibility, rather than a direct measure of successful reproduction.

3.2 Reproducibility Score

The Reproducibility Score $R_p \in [0, 1]$ complements the Documentation Score by estimating the extent to which a collection of papers is reproducible in practice. It is based on empirical results reported by Gundersen et al. [24], who found that papers sharing both open code and data could be reproduced in 86% of the cases, whereas papers sharing open data but not code could be reproduced 33% the cases. The authors did not attempt to reproduce papers without shared data and are assigned a score of 0. Based on these findings, we assign each paper a reproducibility score

$$R_p = \begin{cases} 0.86 & \text{if both open code and data are available,} \\ 0.33 & \text{if open data is available but code is not,} \\ 0 & \text{if open data is not available.} \end{cases} \quad (1)$$

The Reproducibility Score for a set of papers is the average of these values and therefore ranges from 0 to 1. Note that according to this definition, the range of R_p is effectively bounded by 0.86, since no publication can be awarded a score higher than 0.86 currently based on the empirical data.

This formulation is preferable to simpler additive alternatives, because it is empirically calibrated rather than normatively assigned. A scheme that awards one point for open code and one for open data would measure artifact sharing, but it would not estimate reproducibility. The present weighting reflects observed differences in reproduction outcomes under different sharing conditions. It also captures an important asymmetry between code and data. In many machine learning settings, the absence of open data is a more severe obstacle than the absence of code, because using a different dataset introduces its own sources of irreproducibility: dataset bias from collection methods [17, 21, 34, 39, 43], inconsistencies introduced by pre-processing [16] and annotation quality [2], and differences in how data is split across training, validation, and test sets [3, 4, 11, 30, 35]. At the same time, sharing data without code does not ensure reproducibility, since reproduction may still fail because of missing algorithmic and implementation details, such as hyperparameter choices, optimisation procedures [3, 4, 27, 40], initialisation [33, 46], random seeds [3, 31, 33, 40, 46], or undocumented engineering decisions. The Reproducibility Score therefore reflects both the central role of dataset availability and the additional contribution of code availability to successful reproduction.

We note that this score should be interpreted with appropriate caution. The underlying empirical estimates are based on a relatively small sample of reproduction attempts [24] and should therefore be understood as approximate probabilities rather than universal constants. Nevertheless, they provide a more defensible basis for large-scale estimation than arbitrary equal weighting, which, for the purposes of the AI Reproducibility Index, is an important advantage. The objective of the Index is to compare venues, countries, and institutions systematically with respect to the extent to which their research is documented and reproducible. The Documentation Score serves this objective by measuring the presence of key reproducibility-related reporting practices, while the Reproducibility Score translates the most consequential artifact-sharing practices into an empirically grounded estimate of reproducibility potential.

4 System Overview

To systematically evaluate reproducibility documentation across large corpora of AI conference and journal papers, we developed an end-to-end system consisting of six components: (1) Paper Preprocessing, (2) Prompt Optimisation, (3) LLM Evaluation, (4) Generalisation Evaluation, (5) Extract, Transform, and Load (ETL), and (6) a publicly accessible website. A high-level overview of our system is shown in Figure 1. Components 1 through 4 (Section 4.1–4.4) were first introduced and validated by Coakley et al. [8]; for the present work, the prompt was extended to extract author metadata (Section 5.1), which required re-executing both prompt optimization and generalisation evaluation. This work extends this framework by introducing the ETL stage (Section 4.5) and the reproducibility website (Section 4.6), which together transform the evaluation outputs into a continuously updated, publicly accessible resource. Section 4.1 through Section 4.4 briefly summarise these stages for completeness; full methodological details are reported in our earlier work [8].

4.1 Paper Pre-processing

We scraped PDFs from open-access conference proceedings and journal websites, and we converted each to plain text, retaining figure captions and table text but excluding figures and table formatting. This text served as input to the prompt optimisation and LLM Evaluation stages.

4.2 Prompt Optimisation

We optimised the LLM prompts on the dataset of Gundersen and Kjensmo [25] — 400 papers from AAAI (2014, 2016) and IJCAI (2013, 2016) manually evaluated for reproducibility variables. Using few-shot prompting [5], we iteratively refined candidate prompts against the ground truth until accuracy and F1 score stabilised; for each variable the LLM returned a binary classification and a supporting quote. Full prompt engineering details and per-variable metrics have been reported by Coakley et al. [8].

4.3 LLM Evaluation

For inference, the plain text of each paper was submitted with the optimised prompt. For each paper the model returned, in structured JSON, a binary answer and supporting quote for two meta variables — research type (empirical *vs* theoretical) and author affiliation — and the seven reproducibility variables presented in Section 3.1. We evaluated 17 models from Anthropic, Google, OpenAI, Alibaba Cloud, and Microsoft, selecting Google Gemini 2.5 Flash on four criteria: structured JSON support, processing time, cost, and accuracy. Selection criteria and per-variable F1 scores across five runs have been reported by Coakley et al. [8].

4.4 Generalisation Evaluation

To assess generalisation to the full corpus, we sampled 160 papers uniformly at random and labelled them manually as an evaluation dataset. Comparing LLM classifications against these labels yielded F1 scores exceeding 90% for most reproducibility variables, improving on the prompt optimisation dataset and confirming that the method generalises. Complete results are reported in Coakley et al. [8]. When the prompt or model is updated — for example, following improvements to classification accuracy or the addition of new extraction variables — Prompt Optimisation and Generalisation Evaluation are re-executed to verify classification accuracy before any updated prompt or model is deployed.

4.5 Extract, Transform, and Load

To aggregate the LLM evaluation results and prepare them for dissemination, we developed an Extract, Transform, and Load (ETL) stage. The stage receives the binary classifications and supporting text produced by the LLM Evaluation component and computes two metrics for each paper.

The first metric, the Documentation Score, quantifies the comprehensiveness of the documentation of a given publication as a sum over the seven reproducibility variables. The second metric, the Reproducibility Score, provides an estimated probability that the empirical results reported in the publication can be independently reproduced. Following the methodology of Coakley et al. [8], we applied the empirical reproducibility rates reported by Gundersen et al. [24] to derive the reproducibility estimate.

In addition to paper-level metrics, the ETL stage aggregates scores across three dimensions: Countries, reflecting the geographic affiliations of the contributing authors, Institutions, capturing institutional contributions of the contributing authors, and Proceedings and Volumes, enabling longitudinal analysis at the level of publication venues. For journals with multiple volumes per year, scores are consolidated into annual volumes to enable year-on-year comparison.

All structured outputs are loaded into a database that serves as the back-end for the public website. The stage is designed to be re-executable, enabling incremental updates as new conference proceedings are processed.

4.6 The AI Reproducibility Index Website

The website at https://reproducibilityindex.ai makes the outputs of the evaluation stage continuously available to the research community. It is designed to grow incrementally as new conference proceedings and journal volumes are published, following the model of continuously updated leaderboards and benchmarks in AI research [6, 18, 29, 41, 45], and establishes a persistent, updatable record of open science practices across the field.

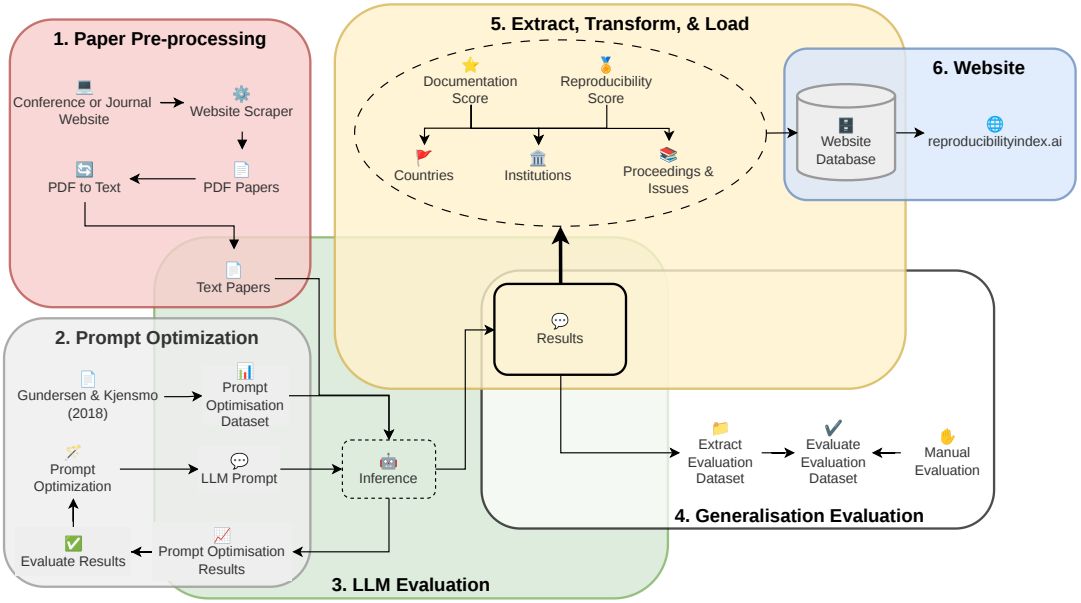

Figure 1: Overview of our system for automated reproducibility evaluation and dissemination. Our system consists of six components: (1) paper pre-processing, including PDF extraction, and conversion to plain text; (2) prompt optimisation, in which few-shot prompts are iteratively refined against a manually annotated dataset [25]; (3) LLM Evaluation, in which the optimised prompt and paper text are submitted to a large language model for inference across nine reproducibility variables; (4) generalisation evaluation, in which LLM-generated classifications are validated against a manually labelled evaluation dataset; (5) Extract, Transform, & Load (ETL), in which paper-level Documentation and Reproducibility Scores are computed and aggregated by country, institution, and publication venue before being loaded into a centralised database; and (6) website, in which the structured results are made publicly accessible at reproducibilityindex.ai. Components 1–4 were introduced and validated by Coakley et al. [8]; components 5 and 6 are original to this work.

Results are accessible at five levels of granularity: individual papers, publication venues, yearly volumes or proceedings, countries, and institutions. At the paper level, users can view the binary classification assigned to each reproducibility variable, the supporting quote extracted by the LLM, and the resulting Documentation Score. At the aggregate level, the platform surfaces reproducibility and documentation trends by venue, country, and institution, identifying where open science practices lead or lag.

The data page records each evaluation batch as a run, identified by a run date. Each paper is traceable to the run that produced its classifications, including the prompt and model used for inference. When the prompt or model is updated — for example, following improvements to classification accuracy — users can identify which papers were evaluated under which version and which editions were re-evaluated as a result. This traceability allows the community to understand how and when the underlying assessments were produced, an important property for a resource updated continuously over time.

5 Method

The system described in Section 4 produces paper-level Documentation and Reproducibility Scores for each empirical paper in the corpus, as defined in Section 3. To rank venues, countries, and institutions, we aggregate these scores across entities using a fractional authorship approach (Section 5.2), distributing credit for each paper proportionally among its authors' affiliated entities.

5.1 Author Metadata Extraction

To extend attribute paper-level scores to countries and institutions, the LLM prompt by Coakley et al. [8] was extended to extract author metadata alongside the reproducibility variables. Following the prompt optimisation procedure described in Section 4.2, the prompt was updated and optimised on the prompt optimisation dataset [25] to extract nine fields per author: number of authors, first name, last name, email address, institution, department, city, country, and first-author status. Extraction reliability was then evaluated on the evaluation dataset described in Section 4.4, comparing LLM-generated results against manually verified ground truth for all 703 authors across the 160 papers.

5.2 Entity-Level Attribution and Ranking

We distribute credit for each paper proportionally among its authors' affiliated entities, following the fractional counting approach [19, 20]. In the description below, we use the term entity to refer to either a country or an institution. We classify an institution as a

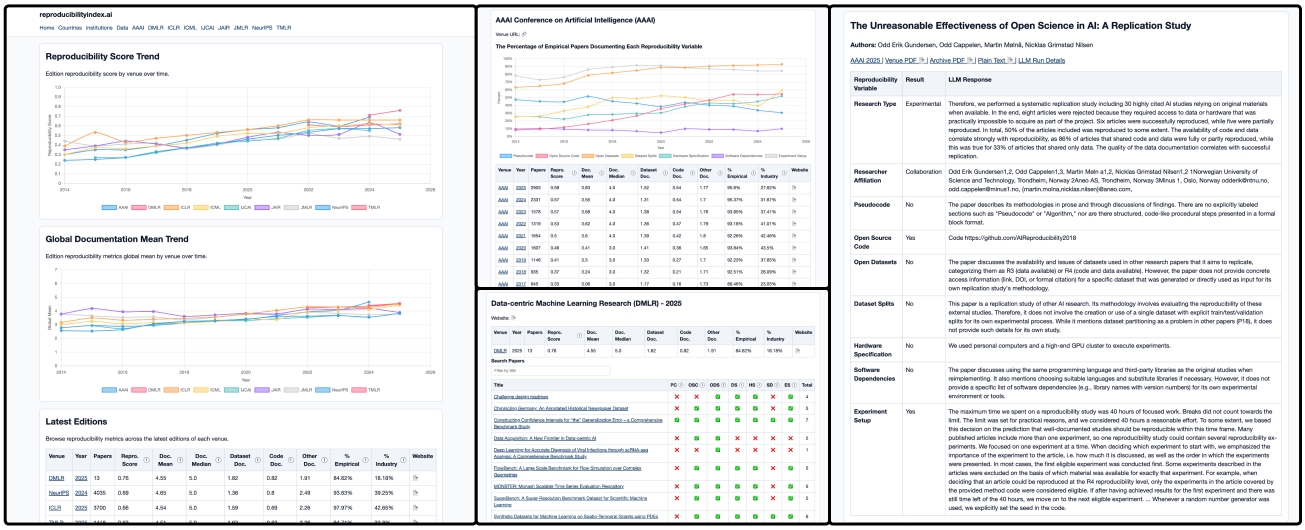

Figure 2: The AI Reproducibility Index website, showing four pages: the Home page (left), the Venue page (middle top), the Edition Detail page (middle bottom), and the Paper Detail page (right). The Home page displays Reproducibility Score and Documentation Score trends over time alongside a summary table of the most recent editions. The Edition Detail page lists all papers in a given edition with binary classifications for each reproducibility variable. The Venue page displays all proceedings for a venue with trends in reproducibility variables over time. The Paper Detail page presents binary classifications and supporting quotes extracted by the LLM for each reproducibility variable of an individual paper.

multinational corporation when it is a commercial company rather than a university or public research institute. Such corporations are treated as a single institution, aggregating contributions across all of their offices without distinguishing where each office resides. The full list of institutions classified as multinational corporations is available at reproducibilityindex.ai. If an author has more than one affiliation, only the first affiliation is used for attribution.

For a paper with N authors, each author's affiliated entity receives a share equal to $1/N$ of that paper's score. For example, a paper with four authors, two from entity A and one each from entities B and C, assigns half the paper's score to entity A and one quarter each to entities B and C. Authors with unknown affiliations are retained in the denominator to prevent artificial score inflation. The entity's overall score is the weighted mean of these fractional contributions across all papers to which it contributed. Only entities with at least 25 absolute contributing papers are included for countries and at least 100 for institutions. Entities are ranked by their mean weighted score.

Because entities with few contributing papers may exhibit extreme scores due to variance alone, we account for statistical uncertainty using a robust standard error that applies Bessel's correction based on the unweighted count of distinct papers per entity. We tested the score distributions for normality [13, 14] and rejected the null hypothesis of normality. We therefore report 95% confidence intervals using the bias-corrected and accelerated (BCa) bootstrap method [15] with 10 000 resamples, which does not assume normality of the underlying distribution.

5.3 Conference versus Journal Comparison

To compare reproducibility and documentation practices between conference and journal papers, we compute the unweighted mean and standard error for each venue type. Because each paper belongs to exactly one venue type, no fractional credit is needed; the computation reduces to a standard mean over all papers in each group.

We tested the score distributions for normality [13, 14] and rejected the null hypothesis of normality. We therefore report 95% BCa bootstrap [15] confidence intervals for each group with 10 000 resamples. To test whether the difference in means between conferences and journals is statistically significant, we apply Welch's t-test, which is robust to non-normality at the sample sizes in this study and accounts for unequal sample sizes and potentially unequal variances between the two groups. Because large sample sizes can produce statistically significant p-values even when the practical difference is negligible, we also report Cohen's d as a measure of effect size. Following standard conventions [9], $|d| < 0.2$ indicates a negligible effect, 0.2–0.5 a small effect, 0.5–0.8 a medium effect, and $|d| > 0.8$ a large effect.

6 Results

Our results are limited to empirical studies; theoretical papers were excluded from all analyses, since large parts of our methodology were designed for estimating the reproducibility of empirical research. Of the 74 972 papers in the corpus, 69 102 were classified as empirical (62 833 conference papers and 6 269 journal papers). All

reproducibility variables were evaluated by an LLM whose classifications achieved F1 scores exceeding 90% for most variables on an independently labelled evaluation dataset of 160 papers; full accuracy metrics are reported in Coakley et al. [8]. The proceedings for NeurIPS 2025 had not been published at the time of writing, hence NeurIPS is excluded from the 2025 results.

6.1 Author Metadata Extraction Accuracy

Following the verification procedure described in Section 5.1, the LLM correctly extracted 6 263 of 6 327 author-field pairs, yielding an overall accuracy of 98.99%. Four fields — author count, first name, last name, and first-author identification — achieved perfect accuracy across all 703 authors. Email address extraction exceeded 99%, while city and country each surpassed 98%. Institution and department were the only fields below 98%, at 96.7% and 97% respectively, with Wilson score 95% confidence intervals reported in Table 1. Qualitative review of the 64 extraction errors revealed five recurring causes: non-standard affiliation symbol conventions in the source document, ambiguity when an author holds multiple affiliations, use of informal laboratory names in place of formal departmental designations, disagreement over geographic granularity (e.g., Hong Kong listed as country versus as a city within China), and author contact information placed in footnotes rather than the standard affiliation block. Notably, these errors are attributable to structural heterogeneity in how author metadata is presented across publications rather than to model hallucination, suggesting that the LLM reliably extracts information as presented and that error rates reflect the inherent ambiguity of the source documents. Results are reported in Table 1 in the appendix.

6.2 Conferences versus Journals

Figure 3 shows the reproducibility and documentation score trends for conference and journal empirical papers over the full period. Both venue types improved on both scores from 2014 to 2025. Reproducibility scores rose from 0.28 to 0.61 for conferences and from 0.41 to 0.60 for journals; documentation scores rose from 2.78 to 4.25 for conferences and from 3.78 to 4.37 for journals. The strong upward trend in reproducibility begins two years later for journals (2018) than for conferences (2016).

On aggregate, conferences scored slightly higher on reproducibility (0.55 *vs* 0.53) and journals slightly higher on documentation (4.08 *vs* 3.77). Both differences are statistically significant ($p < 0.001$) but small to negligible in practice ($d = 0.05$ and $d = -0.25$, respectively).

The year-by-year results reveal a reversal in the reproducibility gap. Journals scored higher than conferences from 2014 to 2017 ($d = -0.48$ to -0.16); from 2018 onward, conferences overtook journals, though the effect sizes remained small ($d = 0.06$ to 0.22). The documentation gap followed a different pattern: journals scored consistently higher throughout, but the gap narrowed from a medium effect in 2014 ($d = -0.79$) to a negligible one in 2024 ($d = -0.01$).

These results indicate that the conference/journal distinction has little practical bearing on reproducibility or documentation quality. The year-over-year improvement within both venue types is substantially larger than the distance between them.

6.3 Countries

Figure 4 shows the rankings of the mean fractional reproducibility and documentation scores for all 42 countries with at least 25 contributing papers.

Luxembourg ranked first on both indices (reproducibility: 0.71; documentation: 4.65), albeit with only 45 contributing empirical papers and correspondingly wide confidence intervals. Among countries with more than 300 empirical papers included in our analysis, the United Arab Emirates (0.64, 360 papers), Denmark (0.63, 406 papers), and the Netherlands (0.62, 1 030 papers) ranked highest in reproducibility. Austria (4.13, 534 papers), the United Arab Emirates (4.12, 360 papers), and Belgium (4.05, 391 papers) ranked highest in documentation among this group. Overall, European and Asian countries dominated the top 10 on both indices. The largest contributors, by contrast, did not lead either ranking. Germany, the largest EU contributor with 3 422 contributing empirical papers, ranked eighth in reproducibility (0.61) and 13th in documentation (3.98). China, the second-largest contributor overall with 20 920 empirical papers, ranked 17th in reproducibility (0.57) and 16th in documentation (3.92). The United States produced the most empirical papers by a wide margin (28 163) but ranked in the bottom quartile in reproducibility (32nd; 0.51) and in documentation (36th; 3.62).

South Africa ranked last on both indices (reproducibility: 0.31; documentation: 2.67), albeit with only 37 contributing empirical papers. Among countries with at least 300 empirical papers, Israel ranked lowest in both reproducibility (0.49, 1 355 papers) and documentation (3.45, 1 355 papers).

6.4 Institutions

Figure 5 report the top ten and bottom ten institutions by mean fractional reproducibility and documentation score, respectively, among institutions with at least 100 contributing empirical papers.

Five institutions appeared in the top 10 on both indices: University of Freiburg (Germany), LMU Munich (Germany), Eindhoven University of Technology (Netherlands), Mohamed bin Zayed University of AI (United Arab Emirates), and Westlake University (China). European institutions dominate both lists, occupying 6 of the top 10 positions in reproducibility score and 7 of the top 10 in documentation score, with German institutions the most represented across the two lists. The Allen Institute for AI was the only US institution in either top-10 list (eighth in reproducibility score, 0.67). The institutions that led on one index but not the other illustrate an asymmetry between documentation and artifact sharing: Technical University of Wien ranked 1st in documentation score (4.58) but did not appear in the reproducibility score top 10, while Shanghai AI Lab ranked 3rd in reproducibility score (0.69) but did not appear in the documentation score top 10.

At the bottom of the distribution, Weizmann Institute of Science (Israel) and Polytechnic University of Milan (Italy) appeared in the bottom 10 on both indices. US institutions were the most represented in the bottom 10 in reproducibility score, holding five positions, while French institutions accounted for three of the bottom 10 in documentation score. The gap between the top- and bottom-ranked institutions was substantial: 0.42 in reproducibility score (0.72 for Eindhoven University of Technology versus 0.30 for

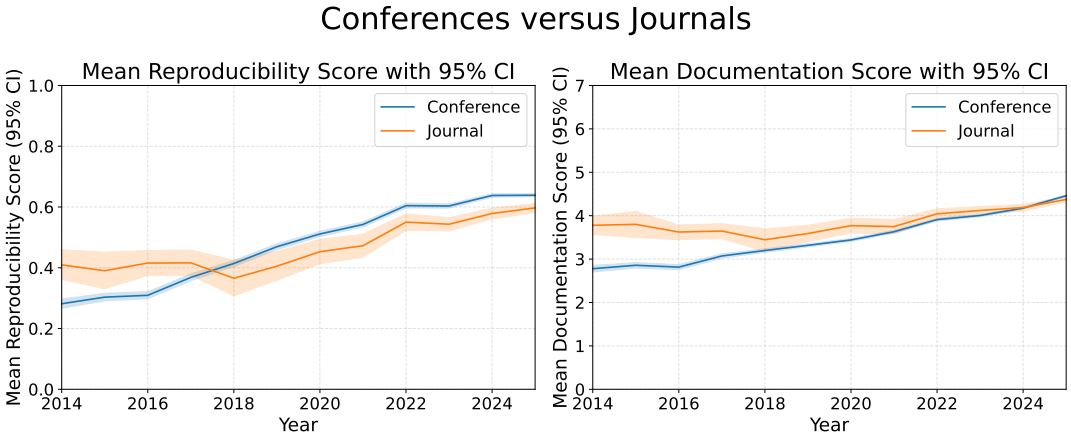

Figure 3: Mean reproducibility score (left) and mean documentation score (right) for conference and journal empirical papers from 2014 to 2025, with 95% confidence intervals. Both venue types improved steadily over the period. Journals scored higher than conferences on the reproducibility score from 2014 to 2017; conferences overtook journals from 2018 onward. The documentation score gap narrowed from a medium effect ($d = -0.79$) in 2014 to a negligible effect ($d = -0.10$) in 2025. For the full reproducibility and documentation score results, see Table 2 and Table 3 respectively in the appendix.

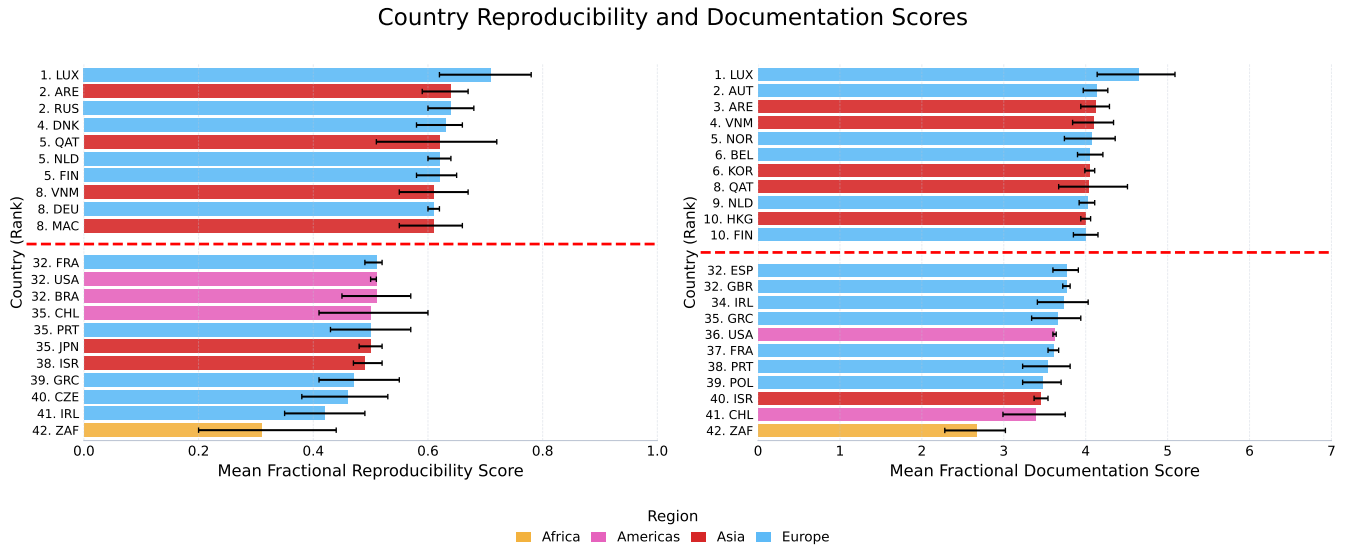

Figure 4: Mean fractional reproducibility score (left) and mean fractional documentation score (right) for the top and bottom 10 countries with at least 25 contributing empirical papers, ranked by mean score with 95% confidence intervals and coloured by region. Luxembourg ranked first on both indices. European and Asian countries dominated the top 10 on both indices, while the United States, the largest contributor with 28 163 empirical papers, ranked 32nd in reproducibility score and 36th in documentation score. Full results are reported in Table 4 (left) and Table 5 (right) in the appendix.

Polytechnic University of Milan) and 1.56 in documentation score (4.58 for Technical University of Wien versus 3.02 for Weizmann Institute of Science).

Figure 6 shows the scores for the 21 institutions with at least 1 000 contributing empirical papers. ETH Zurich ranked first in reproducibility among high-volume institutions (0.61, 1 061 papers) and fourth in documentation (3.96). Hong Kong University of Science and Technology ranked first in documentation (4.07, 1 010 papers).

The largest US universities ranked in the lower half of this subset. Of the 240 ranked institutions, Stanford University ranked 122nd overall in reproducibility (0.55, 2 033 papers; 49th percentile), Massachusetts Institute of Technology ranked 173rd (0.51, 2 158 papers;

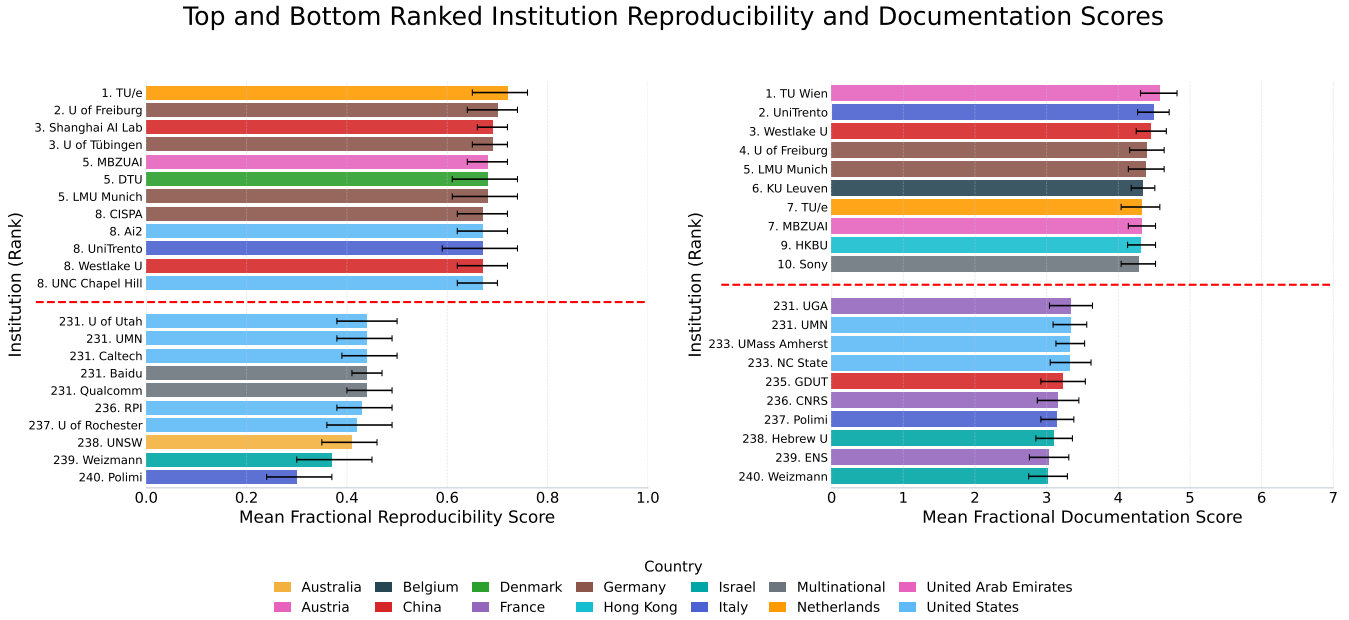

Figure 5: Mean fractional reproducibility score (left) and mean fractional documentation score (right) for the top 10 and bottom 10 institutions with at least 100 contributing empirical papers, ranked by mean score with 95% confidence intervals and coloured by country. Institution abbreviations correspond to the results listed in Table 6 (left) and Table 7 (right). European institutions occupied the majority of positions in both top-10 lists, with five institutions appearing in the top 10 on both indices.

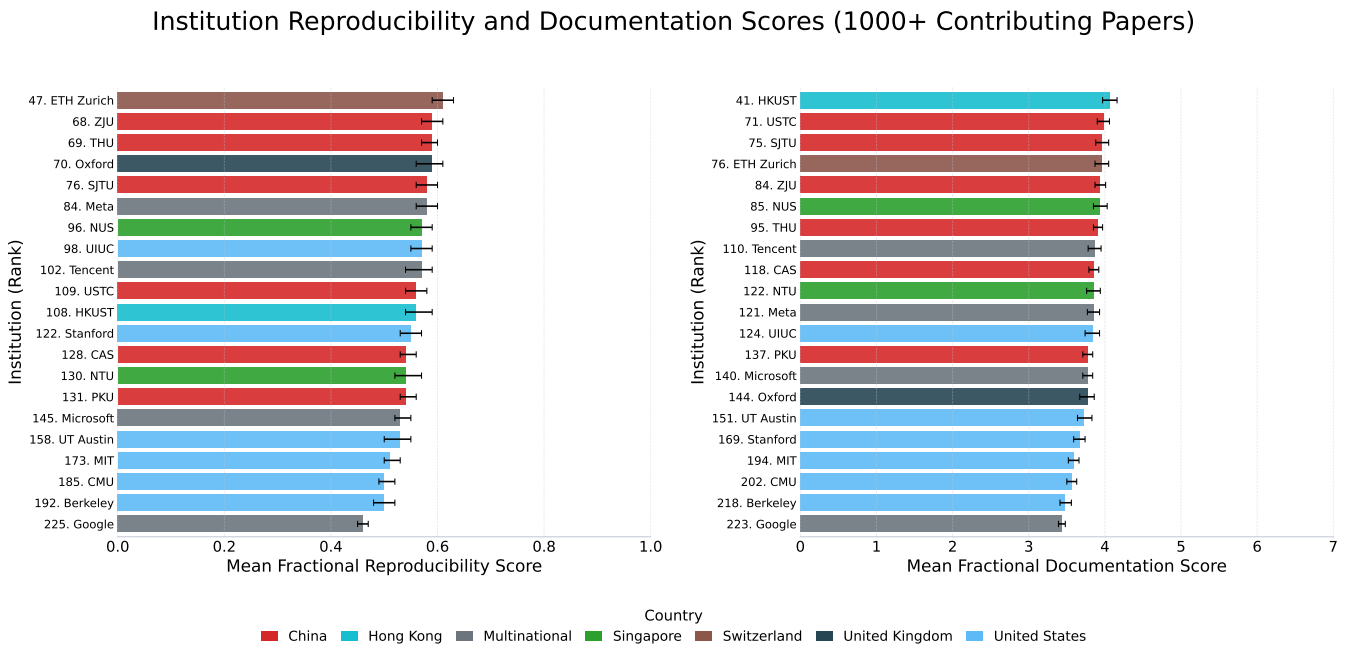

Figure 6: Mean fractional reproducibility score (left) and mean fractional documentation score (right) for all 21 institutions with at least 1 000 contributing empirical papers, ranked by mean score with 95% confidence intervals and colored by country. Institution abbreviations correspond to those in the full results of Table 8 (left) and Table 9 (right) in the appendix. ETH Zurich ranked first in reproducibility score (47th overall) and Hong Kong University of Science and Technology ranked first in documentation score (41st overall). The largest US universities and Google ranked in the lower half on both indices.

28th percentile), Carnegie Mellon University ranked 185th (0.50, 2 459 papers; 23rd percentile), and University of California Berkeley ranked 192nd (0.50, 1 778 papers; 20th percentile). The pattern was similar for documentation scores, where these four institutions ranked between 169th and 218th overall (9th to 30th percentile).

Among multinational corporations, Google ranked lowest in both reproducibility (0.46, 3 802 papers) and documentation (3.43, 3 802 papers), placing 225th and 223rd overall, the bottom 7% on both indices. Microsoft ranked 145th in reproducibility (0.53, 2 232 papers; 40th percentile) and 140th in documentation (3.77, 2 232 papers; 42nd percentile). Meta ranked 84th in reproducibility (0.58, 1 298 papers; 65th percentile) and 121st in documentation (3.85, 1 298 papers; 50th percentile).

7 Discussion

Prior studies of reproducibility in AI have been constrained by the labour required to evaluate individual papers. As discussed in Section 2, the analyses by neither Gundersen et al. [24] nor Raff [37] could be repeated at regular intervals or scaled to the tens of thousands of papers published each year. The AI Reproducibility Index addresses this gap by automating the extraction and scoring of reproducibility-relevant metadata with a quality-assured LLM system, producing a dataset of 74 972 papers that would be infeasible to construct manually.

The Reproducibility Score (Section 3.2) is calibrated against empirical reproduction rates reported by Gundersen et al. [24] at a specific point in time, treating these values as stable constants across the 12-year span we analyse and into the future. Community practices, tooling, and software ecosystems evolve, and the underlying rates may shift accordingly. New large-scale reproduction studies — potentially accelerated by AI agent-assisted reproduction — would provide updated calibration values and, where methodology differs substantially, may require revising the scoring formula.

Our system, Section 4, is designed to accommodate growth. New venues can be added by collecting papers and running them through the same pre-processing and LLM evaluation stages. Existing papers can be re-evaluated as LLMs improve or new reproducibility variables are introduced. The website at reproducibilityindex.ai provides interactive access to the full corpus at multiple levels of detail, from aggregate rankings down to individual paper assessments with the supporting quote obtained from the LLM for each variable. As our system is re-executable and the website is continuously updated, the Index can serve as infrastructure for evaluating the impact of policy changes – such as the introduction of reproducibility checklists, code submission mandates, or changes to page limits – on documentation and reproducibility practices over time.

The narrowing gap in Documentation Score between conferences and journals coincides with a more than twofold increase in average conference paper length, from 5 933 words in 2014 to 12 083 in 2025, while journal paper length remained stable between approximately 16 000 and 19 400 words (Table 10 in the appendix). Two structural changes contributed to this growth: ICML began including supplemental materials in its proceedings PDFs in 2022, and NeurIPS followed in 2023. Supplemental materials provide authors with space to report implementation details — such as hardware specifications, software dependencies, and experiment configurations — that would otherwise be omitted under strict page limits. The convergence in Documentation Score from a medium effect ($d = -0.79$) in 2014 to a negligible one ($d = -0.10$) in 2025 may therefore partially reflect that conference papers have grown closer to journal papers in the space available for reproducibility-relevant content rather than a fundamental shift in documentation practices.

European and Asian countries and institutions consistently outperformed their counterparts in the Americas on both indices, despite the United States producing the most empirical papers by a wide margin (28 163). This geographic pattern held at the institutional level as well: European institutions occupied 6 of the top 10 positions in reproducibility score and 7 of the top 10 in documentation score. The pattern is particularly pronounced among multinational corporations. Google ranked lowest on both indices (225th in reproducibility, 223rd in documentation, 3802 papers), and Microsoft ranked 145th and 140th respectively (2232 papers), mirroring the findings of Gundersen [22]. Intellectual property concerns are cited as a barrier to code and data sharing for researchers working in industry [10, 34], and this may partly explain the lower scores observed for multinational corporations relative to universities and research institutions.

Our system is domain-agnostic in principle. It requires open-access PDFs, an LLM prompt optimised for the target field's reproducibility variables, and a validation dataset of manually labelled papers. Adapting the methodology to other domains would require redefining the reproducibility variables and re-optimising the prompt, but our system could be reused without modification.

8 Limitations

Our work has a few limitations that affect the representation of the index. Reproducibility variables are evaluated using an automated LLM-based process, which introduces classification uncertainty; F1 scores and full accuracy metrics are reported in Coakley et al. [8]. Although these values certainly leave room for debate and improvement, we believe they are of sufficient quality to provide meaningful results. Furthermore, as shown in Figure 1, we provide room for continuous improvement in our methodology. While the LLM accurately extracted author affiliations as written, many authors use one of many variations of the official name of their institution – for example, University of California San Diego, UC San Diego, and UCSD all refer to the same institution. Furthermore, additional conflicts may occur when the institution uses both a native and an English name. Across 74 972 papers and 335 298 authorships, we performed manual post-processing to consolidate as many written variants of the same institution as possible. Resource and time constraints have limited the number of AI conferences and journals we could analyse, but this can be expanded upon in future work as this project develops. However, one hurdle that we cannot overcome currently is that the coverage is restricted to venues that provide open access to their papers. This limits our analysis and results to venues that have committed to open science.

9 Conclusions

In this work, we presented reproducibilityindex.ai, a platform that automatically analyses reproducibility-relevant documentation in

AI research papers and ranks venues, countries, and institutions by reproducibility. Building on the quality-assured LLM-based system validated in our prior conference-only study [8], we extended the corpus to 74 972 open-access papers from major AI conferences and journals spanning 12 years (2014–2025), computing for each paper a reproducibility score, which estimates the probability of successful reproduction, and a documentation score, an aggregate of seven key reproducibility variables. Both metrics improved steadily across conferences and journals alike (Section 6.2), and the conference/journal distinction has little practical bearing: the aggregate reproducibility gap is negligible ($d = 0.05$) and the documentation gap, while statistically significant, narrowed from a medium effect in 2014 ($d = -0.79$) to a negligible one in 2025 ($d = -0.10$). The year-on-year improvement within each venue type is substantially larger than the distance between them.

Beyond these aggregate trends, several patterns emerged from our results. First, documentation and reproducibility are only partially coupled. The two indices rank entities differently: Technical University of Wien led on documentation but fell outside the reproducibility top 10, while Shanghai AI Lab ranked third on reproducibility yet did not make the documentation top 10; at the venue level, journals scored higher on documentation while conferences edged ahead on reproducibility. Reporting reproducibility-relevant details and openly sharing code and data are therefore separable behaviours, not a single underlying tendency, and should be tracked as distinct targets.

Second, measured documentation may be shaped in part by venue policy, not by author practice alone. The narrowing conference–journal documentation gap coincides with a more than two-fold increase in mean conference paper length (from 5 933 to 12 083 words; Table 10) and with ICML and NeurIPS adding supplementary material to their proceedings PDFs in 2022 and 2023, while journal article length remained stable. We did not isolate how much of the increase reflects greater available space rather than changed reporting practice; both may contribute.

Third, research volume and reproducibility are dissociated. The United States produced by far the most empirical papers (28 163) yet ranked 32nd in reproducibility and 36th in documentation, and the largest single contributors — multinational corporations such as Google (3 802 papers, lowest among the 21 high-volume institutions on both indices) — clustered near the bottom. Intellectual-property and commercial constraints have been cited as barriers to sharing for such organisations [10, 34].

European and Asian countries and institutions dominate the top 10 on both indices. We do not read this as a cultural difference. The rankings are confounded by venue composition, field mix, institution size, the industry-to-academia ratio of each country's output, and differences in the open-science policy environment across funding bodies and regions. We report the geographic pattern as an observation that warrants further study.

The AI Reproducibility Index is not a one-off analysis. Because the underlying system is re-executable and the website at reproducibilityindex.ai is continuously updated, it provides the research community with persistent infrastructure for tracking how documentation and reproducibility practices respond to policy changes over time, subject to the caveats that our scores rely on automated LLM-based classification (Section 8) and that coverage is restricted to open-access venues. In principle, our system can be generalised to other computational sciences by redefining the reproducibility variables and re-optimising the prompt.

There are several directions for future work. Direct improvements to prompts or evaluation of new LLMs can increase classification accuracy, and enabling authors to manually correct their own assessments, for example by submitting correction requests for incorrect LLM classifications directly through the website, can yield larger and more reliable evaluation datasets. The platform could also verify whether shared code and data links remain live, as no such URL checking is currently performed. This infrastructure also enables a more targeted study of the geographic rankings: aligning the corpus with the introduction dates of specific open-science mandates would help isolate their effect from confounding factors such as venue composition and institutional mix. Our method can also be expanded to cover more empirical research venues, and, in principle, even subscription-based venues, as the AI Reproducibility Index publishes only metadata that does not conflict with restricted access policies. Finally, incorporating bibliometric data – such as citation rates – into the AI Reproducibility Index could clarify when and why reproducible work attracts more citations, as explored by Raff [38].

10 Supplementary Information

The code and data is available at:
https://doi.org/10.5281/zenodo.20145235. The website source is available on GitHub at
https://github.com/kevincoakley/reproducibilityindex-ai-web.

Acknowledgments

This work is partially supported through the US National Science Foundation award # 2226453, Cloud computing services provided by CloudBank: National Science Foundation award # 1925001. This work has been partly funded by the Alexander von Humboldt Foundation through an Alexander von Humboldt Professorship awarded to Holger Hoos. This work has been partially funded by the Norwegian Research Council through the grant agreement #329730 RICO - Robust Intelligent Control.

References

[1] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (May 2016), 452–454. doi:10.1038/533452a

[2] Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG: Overview and Results. In *Proceedings of the 14th International Conference on Natural Language Generation*. Association for Computational Linguistics, Aberdeen, Scotland, UK, 249–258. doi:10.18653/v1/2021.inlg-1.24

[3] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. 2021. Accounting for variance in machine learning benchmarks. In *Proceedings of the Fourth Conference on Machine Learning and Systems*, Vol. 3. 747–769.

[4] Xavier Bouthillier, César Laurent, and Pascal Vincent. 2019. Unreproducible research is reproducible. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. PMLR, 725–734.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 1877–1901.

[6] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I Jordan, Joseph E Gonzalez,

et al. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*. 8359–8388. doi:10.48550/arXiv.2403.04132

[7] Kevin Coakley, Christine R Kirkpatrick, and Odd Erik Gundersen. 2022. Examining the effect of implementation factors on deep learning reproducibility. In *Proceedings of the IEEE 18th International Conference on e-Science (e-Science)*. IEEE, 397–398. doi:10.1109/eScience55777.2022.00056

[8] Kevin L Coakley, Thijs Snelleman, Holger Hoos, and Odd Erik Gundersen. 2026. The Shift Toward Open and Reproducible AI Research. arXiv:2606.16974 [cs.AI] doi:10.48550/arXiv.2606.16974

[9] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge, 52 Vanderbilt Ave, Fl 11, New York, NY 10017, USA. doi:10.4324/9780203771587

[10] Christian Collberg and Todd A Proebsting. 2016. Repeatability in computer systems research. *Commun. ACM* 59, 3 (2016), 62–69. doi:10.1145/2812803

[11] Çağrı Çöltekin. 2020. Verification, Reproduction and Replication of NLP Experiments: a Case Study on Parsing Universal Dependencies. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*. Association for Computational Linguistics, Barcelona, Spain (Online), 46–56. https://aclanthology.org/2020.udw-1.6/

[12] Matt Crane. 2018. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics* 6 (2018), 241–252. doi:10.1162/tacl_a_00018

[13] Ralph B d'Agostino. 1971. An omnibus test of normality for moderate and large size samples. *Biometrika* 58, 2 (1971), 341–348. doi:10.1093/biomet/58.2.341

[14] Ralph B d'Agostino and Egon S Pearson. 1973. Tests for departure from normality. Empirical results for the distributions of b 2 and $\sqrt{b}$. *Biometrika* 60, 3 (1973), 613–622. doi:10.1093/biomet/60.3.613

[15] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC. doi:10.1201/9780429246593

[16] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark). Association for Computing Machinery, New York, NY, USA, 101–109. doi:10.1145/3298689.3347058

[17] Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. 2021. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine* 385, 3 (2021), 283–286. doi:10.1056/NEJMc2104626

[18] Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, runxin xu, Zhengyang Tang, Wang Benyou, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2025. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *Proceedings of the Thirteenth International Conference on Learning Representations*. 100540–100569.

[19] Marianne Gauffriau, Peder Larsen, Isabelle Maye, Anne Roulin-Perriard, and Markus von Ins. 2008. Comparisons of results of publication counting using different methods. *Scientometrics* 77, 1 (2008), 147–176. doi:10.1007/s11192-007-1934-2

[20] Marianne Gauffriau and Peder Olesen Larsen. 2005. Counting methods are decisive for rankings based on publication and citation studies. *Scientometrics* 64, 1 (2005), 85–93. doi:10.1007/s11192-005-0239-6

[21] Michael F Goodchild and Wenwen Li. 2021. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences* 118, 35 (2021), e2015759118. doi:10.1073/pnas.2015759118

[22] Odd Erik Gundersen. 2019. Standing on the Feet of Giants—Reproducibility in AI. *AI Magazine* 40, 4 (2019), 9–23. doi:10.1609/aimag.v40i4.5185

[23] Odd Erik Gundersen. 2021. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A* 379, 2197 (2021), 20200210. doi:10.1098/rsta.2020.0210

[24] Odd Erik Gundersen, Odd Cappelen, Martin Mølnå, and Nicklas Grimstad Nilsen. 2025. The Unreasonable Effectiveness of Open Science in AI: A Replication Study. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Vol. 39. 26211–26219. doi:10.1609/aaai.v39i25.34818

[25] Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the Art: Reproducibility in Artificial Intelligence. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, Vol. 1. 1644–1651. doi:10.1609/aaai.v32i1.11503

[26] Odd Erik Gundersen, Saeid Shamsaliei, and Richard Juul Isdahl. 2022. Do machine learning platforms provide out-of-the-box reproducibility? *Future Generation Computer Systems* 126 (2022), 34–47. doi:10.1016/j.future.2021.06.014

[27] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, Vol. 32. 3207–3214. doi:10.1609/aaai.v32i1.11694

[28] Song-You Hong, Myung-Seo Koo, Jihyeon Jang, Jung-Eun Esther Kim, Hoon Park, Min-Su Joh, Ji-Hoon Kang, and Tae-Jin Oh. 2013. An evaluation of the software system dependency of a global atmospheric model. *Monthly Weather Review* 141, 11 (2013), 4165–4172. doi:10.1175/MWR-D-12-00352.1

[29] Naman Jain, Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In *Proceedings of the Thirteenth International Conference on Learning Representations*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (Eds.). 58791–58831.

[30] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2018. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one* 13, 3 (2018), e0194889. doi:10.1371/journal.pone.0194889

[31] Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the State of the Art of Evaluation in Neural Language Models. In *Proceedings of the Sixth International Conference on Learning Representations*.

[32] Prabhat Nagarajan, Garrett Warnell, and Peter Stone. 2019. The Impact of Nondeterminism on Reproducibility in Deep Reinforcement Learning. In *2nd Reproducibility in Machine Learning Workshop at ICML 2018* (Stockholm, Sweden).

[33] Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. 2020. Problems and opportunities in training deep learning software systems: An analysis of variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. Association for Computing Machinery, New York, NY, USA, 771–783. doi:10.1145/3324884.3416545

[34] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research* 22, 1 (January 2021).

[35] Line Pouchard, Yuewei Lin, and Hubertus Van Dam. 2020. Replicating Machine Learning Experiments in Materials Science. In *Parallel Computing: Technology Trends*. Advances in Parallel Computing, Vol. 36. IOS Press, 743–755. doi:10.3233/APC200105

[36] Edward Raff. 2017. JSAT: Java statistical analysis tool, a library for machine learning. *Journal of Machine Learning Research* 18, 23 (2017), 1–5.

[37] Edward Raff. 2019. 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, 1–11.

[38] Edward Raff. 2023. Does the Market of Citations Reward Reproducible Work?. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability* (Santa Cruz, CA, USA) *(ACM REP '23)*. Association for Computing Machinery, New York, NY, USA, 89–96. doi:10.1145/3589806.3600041

[39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet?. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5389–5400. https://proceedings.mlr.press/v97/recht19a.html

[40] Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 22nd Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 338–348. doi:10.18653/v1/D17-1035

[41] Scale Labs. 2026. *Leaderboards - Scale Labs*. https://labs.scale.com/leaderboard

[42] Mostafa Shahriari, Rudolf Ramler, and Lukas Fischer. 2022. How do deep-learning framework versions affect the reproducibility of neural network models? *Machine Learning and Knowledge Extraction* 4, 4 (2022), 888–911. doi:10.3390/make4040045

[43] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO, USA). IEEE, 1521–1528. doi:10.1109/CVPR.2011.5995347

[44] Luisa Werner, Nabil Layaïda, Pierre Genevès, Jérôme Euzenat, and Damien Graux. 2024. Reproduce, replicate, reevaluate. the long but safe way to extend machine learning methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 15850–15858. doi:10.1609/aaai.v38i14.29515

[45] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. In *Proceedings of the Thirteenth International Conference on Learning Representations*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (Eds.). 91595–91631.

[46] Donglin Zhuang, Xingyao Zhang, Shuaiwen Song, and Sara Hooker. 2022. Randomness in Neural Network Training: Characterizing the Impact of Tooling. In *Proceedings of the Fourth Conference on Machine Learning and Systems*, D. Marculescu, Y. Chi, and C. Wu (Eds.), Vol. 4. 316–336.

A Appendix

A.1 Tables

Field	Total	Correct	Error	Accuracy	95% CI
# of Authors	703	703	0	100.0%	[99.5%, 100.0%]
First Name	703	703	0	100.0%	[99.5%, 100.0%]
Last Name	703	703	0	100.0%	[99.5%, 100.0%]
Email	703	701	2	99.7%	[99.0%, 99.9%]
Institution	703	680	23	96.7%	[95.1%, 97.8%]
Department	703	682	21	97.0%	[95.5%, 98.0%]
City	703	695	8	98.9%	[97.8%, 99.4%]
Country	703	693	10	98.6%	[97.4%, 99.2%]
First Author	703	703	0	100.0%	[99.5%, 100.0%]
Overall	**6327**	**6263**	**64**	**99.0%**	**[98.7%, 99.2%]**

Table 1: Accuracy of LLM-based author metadata extraction across 703 authors and 160 research papers [8], yielding an overall accuracy of 98.99%. Wilson score 95% confidence intervals are reported for all nine fields.

	Conference				Journal					
Year	Papers	Mean Score	Std. Error	95% CI	Papers	Mean Score	Std. Error	95% CI	P-value	Cohen's d
2014	1085	0.28	0.008	[0.27, 0.30]	149	0.41	0.025	[0.36, 0.46]	$p < 0.001$	-0.48
2015	1626	0.30	0.006	[0.29, 0.32]	115	0.39	0.031	[0.33, 0.45]	0.007	-0.35
2016	2007	0.31	0.006	[0.30, 0.32]	252	0.42	0.021	[0.37, 0.46]	$p < 0.001$	-0.38
2017	2409	0.37	0.006	[0.36, 0.38]	240	0.42	0.021	[0.38, 0.46]	0.03	-0.16
2018	3328	0.41	0.005	[0.40, 0.42]	119	0.37	0.03	[0.31, 0.42]	0.11	0.17
2019	4296	0.47	0.005	[0.46, 0.48]	202	0.41	0.024	[0.36, 0.45]	0.010	0.20
2020	5448	0.51	0.004	[0.50, 0.52]	279	0.45	0.021	[0.41, 0.49]	0.007	0.20
2021	6091	0.54	0.004	[0.53, 0.55]	314	0.47	0.019	[0.43, 0.51]	$p < 0.001$	0.22
2022	6471	0.60	0.004	[0.60, 0.61]	588	0.55	0.014	[0.52, 0.58]	$p < 0.001$	0.17
2023	8217	0.60	0.004	[0.60, 0.61]	980	0.54	0.011	[0.52, 0.56]	$p < 0.001$	0.17
2024	11350	0.64	0.003	[0.63, 0.64]	1347	0.58	0.009	[0.56, 0.60]	$p < 0.001$	0.18
2025	10505	0.61	0.003	[0.61, 0.62]	1684	0.60	0.008	[0.58, 0.61]	0.04	0.06
Overall	**62833**	**0.55**	**0.001**	**[0.54, 0.55]**	**6269**	**0.53**	**0.004**	**[0.53, 0.54]**	**0.003**	**0.05**

Table 2: Mean reproducibility score for conference and journal empirical papers by year, with standard errors and 95% confidence intervals, Welch's t-test p-values, and Cohen's d effect sizes. Conference scores rose from 0.28 in 2014 to 0.61 in 2025; journal scores rose from 0.41 to 0.60. The aggregate difference is statistically significant ($p = 0.003$) but negligible in magnitude ($d = 0.05$).

	Conference				Journal					
Year	Papers	Mean Score	Std. Error	95% CI	Papers	Mean Score	Std. Error	95% CI	P-value	Cohen's d
2014	1085	2.78	0.038	[2.70, 2.85]	149	3.78	0.108	[3.57, 3.99]	$p < 0.001$	-0.79
2015	1626	2.86	0.032	[2.79, 2.92]	115	3.80	0.153	[3.50, 4.10]	$p < 0.001$	-0.72
2016	2007	2.81	0.029	[2.75, 2.87]	252	3.62	0.087	[3.45, 3.79]	$p < 0.001$	-0.62
2017	2409	3.07	0.025	[3.02, 3.12]	240	3.65	0.089	[3.47, 3.82]	$p < 0.001$	-0.46
2018	3328	3.20	0.021	[3.16, 3.24]	119	3.45	0.131	[3.18, 3.70]	0.06	-0.20
2019	4296	3.32	0.018	[3.28, 3.35]	202	3.59	0.095	[3.40, 3.78]	0.005	-0.23
2020	5448	3.44	0.017	[3.41, 3.47]	279	3.77	0.085	[3.60, 3.94]	$p < 0.001$	-0.26
2021	6091	3.63	0.017	[3.60, 3.66]	314	3.75	0.085	[3.58, 3.91]	0.17	-0.09
2022	6471	3.91	0.016	[3.88, 3.94]	588	4.04	0.06	[3.92, 4.16]	0.03	-0.10
2023	8217	4.00	0.015	[3.98, 4.03]	980	4.12	0.045	[4.03, 4.21]	0.02	-0.08
2024	11350	4.17	0.012	[4.15, 4.20]	1347	4.19	0.038	[4.11, 4.26]	0.73	-0.01
2025	10505	4.25	0.012	[4.23, 4.28]	1684	4.37	0.033	[4.31, 4.44]	$p < 0.001$	-0.10
Overall	**62833**	**3.77**	**0.005**	**[3.76, 3.78]**	**6269**	**4.08**	**0.018**	**[4.04, 4.11]**	**p < 0.001**	**-0.25**

Table 3: Mean documentation score for conference and journal empirical papers by year, with standard errors and 95% confidence intervals, Welch's t-test p-values, and Cohen's d effect sizes. Conference scores rose from 2.78 in 2014 to 4.25 in 2025; journal scores rose from 3.78 to 4.37. The aggregate difference is small ($d = -0.25$) but statistically significant ($p < 0.001$)

Rank	Country	Mean Score	Std Error	95% CI	Contributing Papers	Rank	Country	Mean Score	Std Error	95% CI	Contributing Papers
1	LUX	0.71	0.04	[0.62, 0.78]	45	22	IRN	0.56	0.05	[0.47, 0.65]	67
2	ARE	0.64	0.02	[0.59, 0.67]	360	22	GBR	0.56	0.01	[0.55, 0.57]	5424
2	RUS	0.64	0.02	[0.60, 0.68]	307	22	CAN	0.56	0.01	[0.55, 0.57]	3573
4	DNK	0.63	0.02	[0.58, 0.66]	406	25	SGP	0.55	0.01	[0.54, 0.57]	2968
5	QAT	0.62	0.05	[0.51, 0.72]	35	25	ITA	0.55	0.01	[0.52, 0.57]	971
5	NLD	0.62	0.01	[0.60, 0.64]	1030	27	POL	0.54	0.03	[0.49, 0.60]	197
5	FIN	0.62	0.02	[0.58, 0.65]	383	27	ESP	0.54	0.02	[0.50, 0.59]	350
8	VNM	0.61	0.03	[0.55, 0.67]	111	27	AUS	0.54	0.01	[0.53, 0.56]	2570
8	DEU	0.61	0.01	[0.60, 0.62]	3422	27	IND	0.54	0.01	[0.52, 0.56]	1024
8	MAC	0.61	0.03	[0.55, 0.66]	133	27	SAU	0.54	0.02	[0.50, 0.58]	325
11	KOR	0.60	0.01	[0.59, 0.61]	2220	32	FRA	0.51	0.01	[0.49, 0.52]	2543
11	AUT	0.60	0.02	[0.57, 0.64]	534	32	USA	0.51	0.00	[0.50, 0.51]	28163
13	NOR	0.59	0.04	[0.50, 0.66]	127	32	BRA	0.51	0.03	[0.45, 0.57]	149
14	BEL	0.58	0.02	[0.54, 0.62]	391	35	CHL	0.50	0.05	[0.41, 0.60]	53
14	CHE	0.58	0.01	[0.57, 0.60]	2254	35	PRT	0.50	0.03	[0.43, 0.57]	118
14	NZL	0.58	0.04	[0.50, 0.64]	120	35	JPN	0.50	0.01	[0.48, 0.52]	1575
17	TWN	0.57	0.02	[0.54, 0.61]	343	38	ISR	0.49	0.01	[0.47, 0.52]	1355
17	SWE	0.57	0.02	[0.53, 0.61]	428	39	GRC	0.47	0.04	[0.41, 0.55]	96
17	CHN	0.57	0.00	[0.56, 0.57]	20920	40	CZE	0.46	0.04	[0.38, 0.53]	153
17	TUR	0.57	0.05	[0.46, 0.66]	73	41	IRL	0.42	0.04	[0.35, 0.49]	114
17	HKG	0.57	0.01	[0.55, 0.58]	2511	42	ZAF	0.31	0.06	[0.20, 0.44]	37

Table 4: Mean fractional reproducibility score by country for all 42 countries with at least 25 contributing empirical papers, ranked by mean score with standard errors and 95% confidence intervals. Luxembourg ranks first (0.71). The United States, the largest contributor with 28 163 empirical papers, ranks 32nd (0.51).

Rank	Country	Mean Score	Std Error	95% CI	Contributing Papers	Rank	Country	Mean Score	Std Error	95% CI	Contributing Papers
1	LUX	4.65	0.25	[4.14, 5.09]	45	20	SGP	3.88	0.03	[3.83, 3.94]	2968
2	AUT	4.13	0.07	[3.97, 4.27]	534	23	BRA	3.87	0.12	[3.64, 4.10]	149
3	ARE	4.12	0.09	[3.94, 4.29]	360	24	TWN	3.84	0.07	[3.69, 3.99]	343
4	VNM	4.09	0.13	[3.84, 4.34]	111	25	NZL	3.83	0.16	[3.52, 4.14]	120
5	NOR	4.07	0.16	[3.74, 4.36]	127	25	AUS	3.83	0.03	[3.77, 3.89]	2570
6	BEL	4.05	0.08	[3.90, 4.21]	391	27	CHE	3.82	0.03	[3.76, 3.89]	2254
6	KOR	4.05	0.03	[3.99, 4.11]	2220	28	CZE	3.81	0.14	[3.54, 4.09]	153
8	QAT	4.04	0.22	[3.67, 4.51]	35	29	CAN	3.80	0.03	[3.75, 3.85]	3573
9	NLD	4.02	0.05	[3.92, 4.11]	1030	29	IND	3.80	0.05	[3.70, 3.89]	1024
10	HKG	4.00	0.03	[3.94, 4.06]	2511	31	JPN	3.78	0.04	[3.71, 3.86]	1575
10	FIN	4.00	0.08	[3.85, 4.15]	383	32	ESP	3.76	0.08	[3.60, 3.91]	350
12	SAU	3.99	0.09	[3.82, 4.16]	325	32	GBR	3.76	0.02	[3.72, 3.81]	5424
13	DEU	3.98	0.03	[3.93, 4.03]	3422	34	IRL	3.73	0.16	[3.41, 4.03]	114
14	IRN	3.97	0.20	[3.54, 4.32]	67	35	GRC	3.66	0.15	[3.34, 3.94]	96
15	RUS	3.94	0.08	[3.78, 4.09]	307	36	USA	3.62	0.01	[3.60, 3.64]	28163
16	CHN	3.92	0.01	[3.90, 3.94]	20920	37	FRA	3.61	0.03	[3.54, 3.67]	2543
17	DNK	3.91	0.08	[3.76, 4.06]	406	38	PRT	3.53	0.15	[3.23, 3.81]	118
18	SWE	3.90	0.07	[3.76, 4.05]	428	39	POL	3.47	0.12	[3.23, 3.70]	197
19	MAC	3.89	0.13	[3.65, 4.13]	133	40	ISR	3.45	0.04	[3.37, 3.54]	1355
20	TUR	3.88	0.25	[3.37, 4.34]	73	41	CHL	3.39	0.19	[2.99, 3.75]	53
20	ITA	3.88	0.05	[3.78, 3.97]	971	42	ZAF	2.67	0.19	[2.28, 3.02]	37

Table 5: Mean fractional documentation score by country for all 42 countries with at least 25 contributing empirical papers, ranked by mean score with standard errors and 95% confidence intervals. Luxembourg ranks first (4.65). The United States ranks 36th (3.62) despite producing the most empirical papers.

Rank	Institution	Abbr	Country	Mean Score	Std. Error	95% CI	Contributing Papers
1	Eindhoven University of Technology	TU/e	Netherlands	0.72	0.03	[0.65, 0.76]	143
2	University of Freiburg	U of Freiburg	Germany	0.70	0.03	[0.64, 0.74]	147
3	Shanghai AI Lab	Shanghai AI Lab	China	0.69	0.02	[0.66, 0.72]	380
3	University of Tübingen	U of Tübingen	Germany	0.69	0.02	[0.65, 0.72]	273
5	Mohamed bin Zayed University of AI	MBZUAI	United Arab Emirates	0.68	0.02	[0.64, 0.72]	275
5	Technical University of Denmark	DTU	Denmark	0.68	0.03	[0.61, 0.74]	112
5	LMU Munich	LMU Munich	Germany	0.68	0.03	[0.61, 0.74]	102
8	CISPA Helmholtz Center for Information Security	CISPA	Germany	0.67	0.03	[0.62, 0.72]	151
8	Allen Institute for AI	Ai2	United States	0.67	0.03	[0.62, 0.72]	153
8	University of Trento	UniTrento	Italy	0.67	0.04	[0.59, 0.74]	113
8	Westlake University	Westlake U	China	0.67	0.02	[0.62, 0.72]	194
8	University of North Carolina Chapel Hill	UNC Chapel Hill	United States	0.67	0.02	[0.62, 0.70]	267
231	University of Utah	U of Utah	United States	0.44	0.03	[0.38, 0.50]	131
231	University of Minnesota	UMN	United States	0.44	0.03	[0.38, 0.49]	223
231	California Institute of Technology	Caltech	United States	0.44	0.03	[0.39, 0.50]	271
231	Baidu	Baidu	Multinational	0.44	0.02	[0.41, 0.47]	395
231	Qualcomm	Qualcomm	Multinational	0.44	0.02	[0.40, 0.49]	122
236	Rensselaer Polytechnic Institute	RPI	United States	0.43	0.03	[0.38, 0.49]	186
237	University of Rochester	U of Rochester	United States	0.42	0.03	[0.36, 0.49]	138
238	University of New South Wales	UNSW	Australia	0.41	0.03	[0.35, 0.46]	147
239	Weizmann Institute of Science	Weizmann	Israel	0.37	0.04	[0.30, 0.45]	115
240	Polytechnic University of Milan	Polimi	Italy	0.30	0.03	[0.24, 0.37]	130

Table 6: Mean fractional reproducibility score for the top 10 and bottom 10 institutions with at least 100 contributing empirical papers, ranked by mean score with standard errors and 95% confidence intervals. European institutions dominate the top 10, with German institutions the most represented; US institutions account for five of the bottom 10.

Rank	Institution	Abbr	Country	Mean Score	Std. Error	95% CI	Contributing Papers
1	Technical University of Wien	TU Wien	Austria	4.58	0.13	[4.31, 4.82]	146
2	University of Trento	UniTrento	Italy	4.49	0.11	[4.27, 4.71]	113
3	Westlake University	Westlake U	China	4.46	0.11	[4.25, 4.67]	194
4	University of Freiburg	U of Freiburg	Germany	4.40	0.12	[4.16, 4.64]	147
5	LMU Munich	LMU Munich	Germany	4.39	0.13	[4.14, 4.64]	102
6	KU Leuven	KU Leuven	Belgium	4.35	0.09	[4.18, 4.51]	200
7	Eindhoven University of Technology	TU/e	Netherlands	4.33	0.14	[4.04, 4.58]	143
7	Mohamed bin Zayed University of AI	MBZUAI	United Arab Emirates	4.33	0.10	[4.14, 4.52]	275
9	Hong Kong Baptist University	HKBU	Hong Kong	4.32	0.10	[4.13, 4.52]	293
10	Sony	Sony	Multinational	4.29	0.12	[4.04, 4.52]	140
231	Grenoble Alpes University	UGA	France	3.34	0.16	[3.04, 3.64]	137
231	University of Minnesota	UMN	United States	3.34	0.12	[3.09, 3.56]	223
233	University of Massachusetts Amherst	UMass Amherst	United States	3.33	0.10	[3.13, 3.53]	290
233	North Carolina State University	NC State	United States	3.33	0.14	[3.05, 3.62]	155
235	Guangdong University of Technology	GDUT	China	3.23	0.16	[2.92, 3.54]	103
236	French National Centre for Scientific Research	CNRS	France	3.16	0.15	[2.87, 3.45]	112
237	Polytechnic University of Milan	Polimi	Italy	3.14	0.12	[2.92, 3.38]	130
238	Hebrew University	Hebrew U	Israel	3.10	0.13	[2.85, 3.36]	158
239	Ecole Normale Supérieure	ENS	France	3.03	0.14	[2.76, 3.31]	102
240	Weizmann Institute of Science	Weizmann	Israel	3.02	0.14	[2.75, 3.29]	115

Table 7: Mean fractional documentation score for the top 10 and bottom 10 institutions with at least 100 contributing empirical papers, ranked by mean score with standard errors and 95% confidence intervals. European institutions again dominate the top 10; French and Israeli institutions account for five of the bottom 10.

Rank	Institution	Abbr	Country	Mean Score	Std. Error	95% CI	Contributing Papers
47	ETH Zurich	ETH Zurich	Switzerland	0.61	0.01	[0.59, 0.63]	1061
68	Zhejiang University	ZJU	China	0.59	0.01	[0.57, 0.61]	1501
69	Tsinghua University	THU	China	0.59	0.01	[0.57, 0.60]	2641
70	University of Oxford	Oxford	United Kingdom	0.59	0.01	[0.56, 0.61]	1264
76	Shanghai Jiaotong University	SJTU	China	0.58	0.01	[0.56, 0.60]	1412
84	Meta	Meta	Multinational	0.58	0.01	[0.56, 0.60]	1298
96	National University of Singapore	NUS	Singapore	0.57	0.01	[0.55, 0.59]	1293
98	University of Illinois Urbana-Champaign	UIUC	United States	0.57	0.01	[0.55, 0.59]	1122
102	Tencent	Tencent	Multinational	0.57	0.01	[0.54, 0.59]	1064
109	University of Science and Technology of China	USTC	China	0.56	0.01	[0.54, 0.58]	1187
108	Hong Kong University of Science and Technology	HKUST	Hong Kong	0.56	0.01	[0.54, 0.59]	1010
122	Stanford University	Stanford	United States	0.55	0.01	[0.53, 0.57]	2033
128	Chinese Academy of Sciences	CAS	China	0.54	0.01	[0.53, 0.56]	1772
130	Nanyang Technological University	NTU	Singapore	0.54	0.01	[0.52, 0.57]	1089
131	Peking University	PKU	China	0.54	0.01	[0.53, 0.56]	2020
145	Microsoft	Microsoft	Multinational	0.53	0.01	[0.52, 0.55]	2232
158	University of Texas Austin	UT Austin	United States	0.53	0.01	[0.50, 0.55]	1079
173	Massachusetts Institute of Technology	MIT	United States	0.51	0.01	[0.50, 0.53]	2158
185	Carnegie Mellon University	CMU	United States	0.50	0.01	[0.49, 0.52]	2459
192	University of California Berkeley	Berkeley	United States	0.50	0.01	[0.48, 0.52]	1778
225	Google	Google	Multinational	0.46	0.01	[0.45, 0.47]	3802

Table 8: Mean fractional reproducibility score for all 21 institutions with at least 1 000 contributing empirical papers, ranked by mean score with standard errors and 95% confidence interval. Overall ranks from the full institution ranking are shown. ETH Zurich ranks first in this subset (47th overall) at 0.61; Google ranks last (225th overall) at 0.46.

Rank	Institution	Abbr	Country	Mean Score	Std. Error	95% CI	Contributing Papers
41	Hong Kong University of Science and Technology	HKUST	Hong Kong	4.07	0.05	[3.97, 4.16]	1010
71	University of Science and Technology of China	USTC	China	3.98	0.04	[3.90, 4.06]	1187
75	Shanghai Jiaotong University	SJTU	China	3.96	0.04	[3.88, 4.05]	1412
76	ETH Zurich	ETH Zurich	Switzerland	3.96	0.05	[3.87, 4.05]	1061
84	Zhejiang University	ZJU	China	3.94	0.04	[3.87, 4.01]	1501
85	National University of Singapore	NUS	Singapore	3.94	0.05	[3.85, 4.03]	1293
95	Tsinghua University	THU	China	3.91	0.03	[3.85, 3.97]	2641
110	Tencent	Tencent	Multinational	3.87	0.04	[3.78, 3.95]	1064
118	Chinese Academy of Sciences	CAS	China	3.86	0.03	[3.79, 3.92]	1772
122	Nanyang Technological University	NTU	Singapore	3.85	0.05	[3.76, 3.94]	1089
121	Meta	Meta	Multinational	3.85	0.04	[3.77, 3.93]	1298
124	University of Illinois Urbana-Champaign	UIUC	United States	3.84	0.05	[3.74, 3.93]	1122
137	Peking University	PKU	China	3.78	0.03	[3.71, 3.84]	2020
140	Microsoft	Microsoft	Multinational	3.77	0.03	[3.71, 3.84]	2232
144	University of Oxford	Oxford	United Kingdom	3.77	0.05	[3.67, 3.86]	1264
151	University of Texas Austin	UT Austin	United States	3.73	0.05	[3.64, 3.83]	1079
169	Stanford University	Stanford	United States	3.67	0.04	[3.59, 3.74]	2033
194	Massachusetts Institute of Technology	MIT	United States	3.59	0.04	[3.52, 3.66]	2158
202	Carnegie Mellon University	CMU	United States	3.56	0.03	[3.50, 3.63]	2459
218	University of California Berkeley	Berkeley	United States	3.48	0.04	[3.41, 3.56]	1778
223	Google	Google	Multinational	3.43	0.02	[3.39, 3.48]	3802

Table 9: Mean fractional documentation score for all 21 institutions with at least 1 000 contributing empirical papers, ranked by mean score with standard errors and 95% confidence interval. Overall ranks from the full institution ranking are shown. Hong Kong University of Science and Technology ranks first in this subset (41st overall) at 4.07; Google ranks last (223rd overall) at 3.43.

	Conferences					Journals				**Average**	
Year	AAAI	ICLR	ICML	IJCAI	NeurIPS	DMLR	JAIR	JMLR	TMLR	**Conference**	**Journal**
2014	5637	5422	6828	-	5668	-	19090	14816	-	**5933**	**16361**
2015	5466	6931	6901	5987	5725	-	21951	14170	-	**5902**	**16419**
2016	5379	7429	6963	5631	5770	-	19572	15380	-	**5840**	**16285**
2017	5509	7885	7025	5895	6306	-	21557	16177	-	**6226**	**17388**
2018	6235	9055	7191	5921	6533	-	19937	16285	-	**6688**	**17932**
2019	6340	9826	7297	5904	6685	-	19338	16287	-	**6851**	**17170**
2020	6855	10471	7932	5894	7255	-	18721	17835	-	**7460**	**18056**
2021	7240	11561	8190	6080	7945	-	19986	19189	-	**8075**	**19399**
2022	7330	13125	13586	6001	8222	-	20240	19039	14217	**9459**	**17725**
2023	7275	13185	14208	7120	12522	-	20471	20006	14817	**11554**	**17193**
2024	7248	13882	14603	7013	16496	16237	18420	20657	15182	**13172**	**16955**
2025	7149	14836	14755	7383	-	14033	19256	21495	15632	**12083**	**16739**
Average	**6785**	**13047**	**11976**	**6315**	**10309**	**15521**	**19767**	**18402**	**15247**	**9746**	**17185**

Table 10: Average word count per empirical paper by venue and year. Journal papers are consistently longer than conference papers, averaging 17 185 words compared to 9 746 for conferences across the full period. Conference paper length increased from 5 933 words in 2014 to 12 083 in 2025, while journal paper length remained stable between 16 000 and 19 400 words.