Download Presentation

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

Kevin L Coakley[1,2], Thijs Snelleman[3], Holger Hoos[3,4], Odd Erik Gundersen[2]
[1]San Diego Supercomputer Center; [2]Norwegian University of Science and Technology; [3]RWTH Aachen University, [4]Leiden University

Tracking Reproducibility at Scale

- Reproducing a single AI study is slow and labor-intensive — often tens of hours per paper, even with code and data available[1]
- The compute to train the models is expensive and constrained due to the increased usage of generative AI
- AI research has reproducibility issues — but how do we track whether it's getting better or worse, and where?
- Policy changes (checklists, mandates, page limit changes) are introduced with no reliable way to measure their effect on reproducibility
- Answering these questions requires monitoring reproducibility-relevant practices continuously, across the full breadth of AI research

Prior Meta–Studies and Their Limitations

- Gundersen & Kjensmo[2], Raff[3], Gundersen et al.[1] — valuable but constrained by labor cost
- Small sample sizes (22–400 papers), narrow subfields/single-venue focus
- Motivates need for an automated, scalable approach

Documentation as a Proxy for Reproducibility

- Direct reproduction doesn't scale — an alternative is to assess whether a paper documents what's needed for reproduction, rather than attempting reproduction itself - not perfect, but with advances in LLMs it is possible to automate
- Gundersen[4] defines reproducibility degrees based on what documentation is shared:
  - R1 Description (text only)
  - R2 Code (text + code)
  - R3 Data (text + data)
  - R4 Experiment (text + code + data)

	Text	Code	Data
R1 Description	■		
R2 Code	■	■	
R3 Data	■		■
R4 Experiment	■	■	■

- More shared documentation types → easier for independent investigators to reproduce results; this documentation-availability spectrum is itself measurable at scale, even when full reproduction is not

Gundersen & Kjensmo (2018) — "State of the Art: Reproducibility in AI"

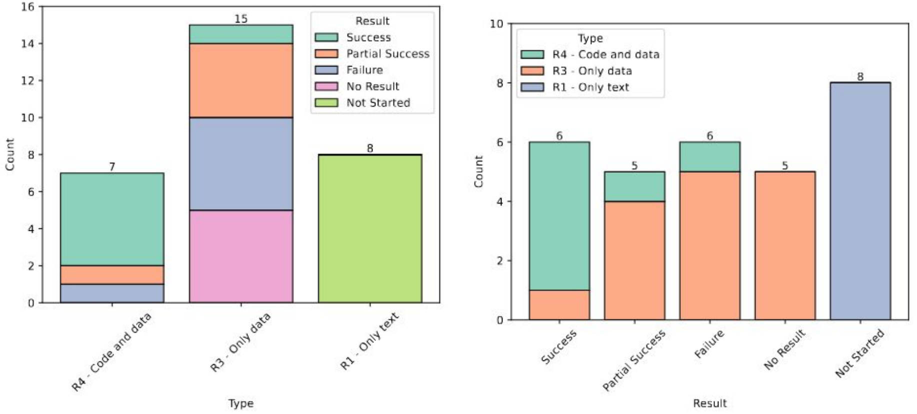
- Manually evaluated 400 papers (AAAI 2014/2016, IJCAI 2013/2016) across for the documentation of 20 reproducibility variables
- These 400 papers are the dataset that Coakley et al.[5] — and this work — build on

20 Reproducibility Variables

Problem	Prediction
Objective/Goal	Method source code
Research method	Hardware specifications
Research questions	Software dependencies
Pseudo code	Experiment setup
Training data	Experiment source code
Validation data	Research type
Test data	Research outcome
Results	Affiliation
Hypothesis	Contribution

Gundersen et al. (2025) — "The Unreasonable Effectiveness of Open Science in AI"

- 30 highly cited AI studies, systematic reproduction attempts (up to 40 hours/study)
- Empirical reproduction rates:
  - 86% (R4: code + data)
  - 33% (R3: data only)
  - No code (R1 & R2) was not attempted
- Establishes empirical grounding for translating documentation/artifact-sharing into an actual reproducibility likelihood

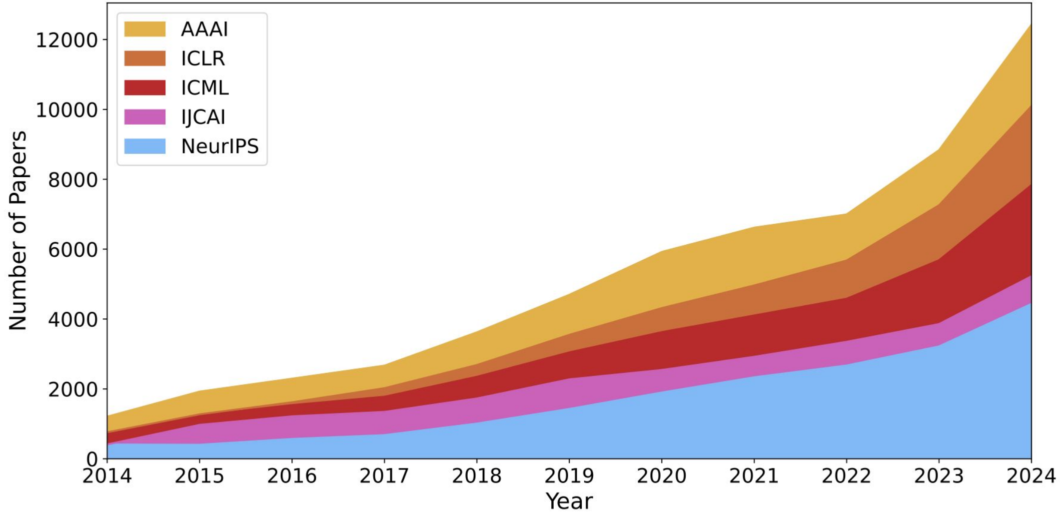


(a) Results per reproducibility type.

(b) Reproducibility types per result.

Coakley et al. (2026) — "The Shift Toward Open and Reproducible AI Research" (under review)

- Combined Gundersen & Kjensmo's[2] documentation variables with an LLM-based automated pipeline
- 56,800 papers evaluated
- 5 conferences, 2014–2024
- Used Gundersen et al.'s[1] empirical rates to estimate reproducibility from documentation



Growth of the number of papers accepted at the 5 major AI conferences (2014-2024)

The 9 Reproducibility Variables

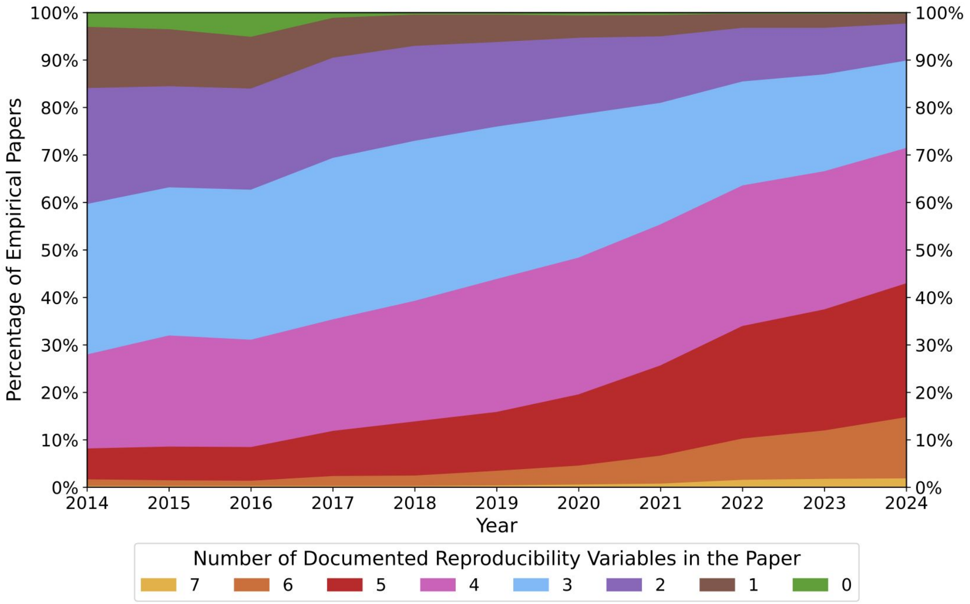
9 were selected for LLM-based automation (class balance, importance, F1 ≥ 75%)

Seven Reproducibility Variables

- Pseudo code
- Open Code
- Open Datasets
- Dataset Splits
- Hardware specifications
- Software dependencies
- Experiment setup

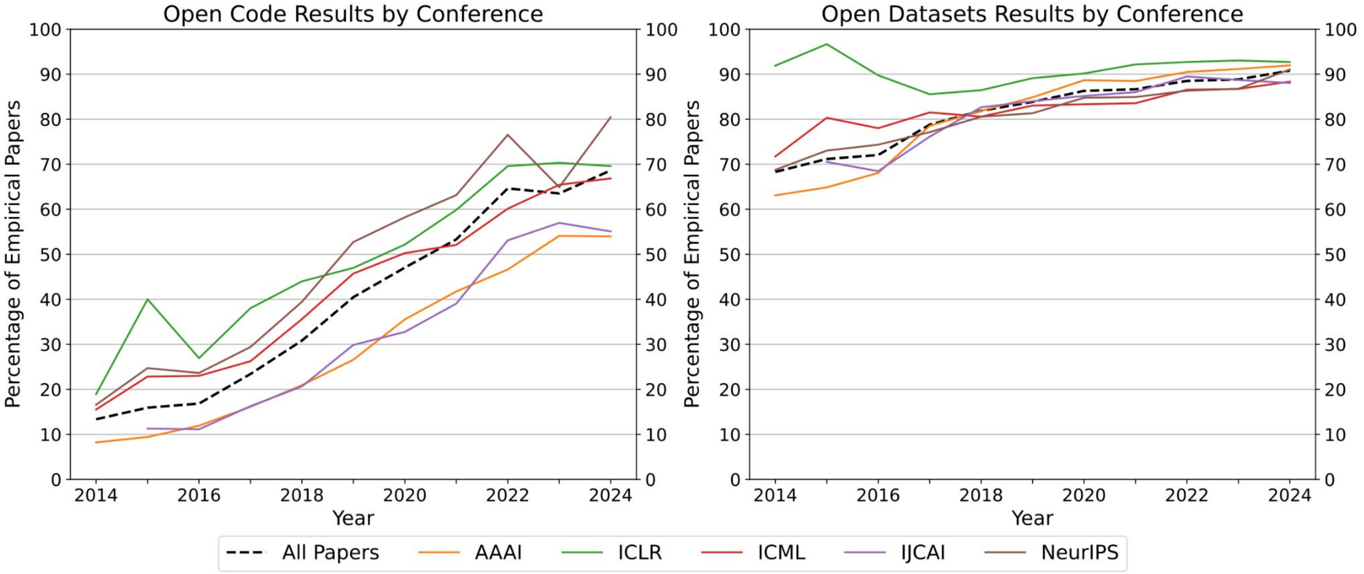
Two Meta Variables

- Research type: Experimental or theoretical
- Affiliation: Academia or industry



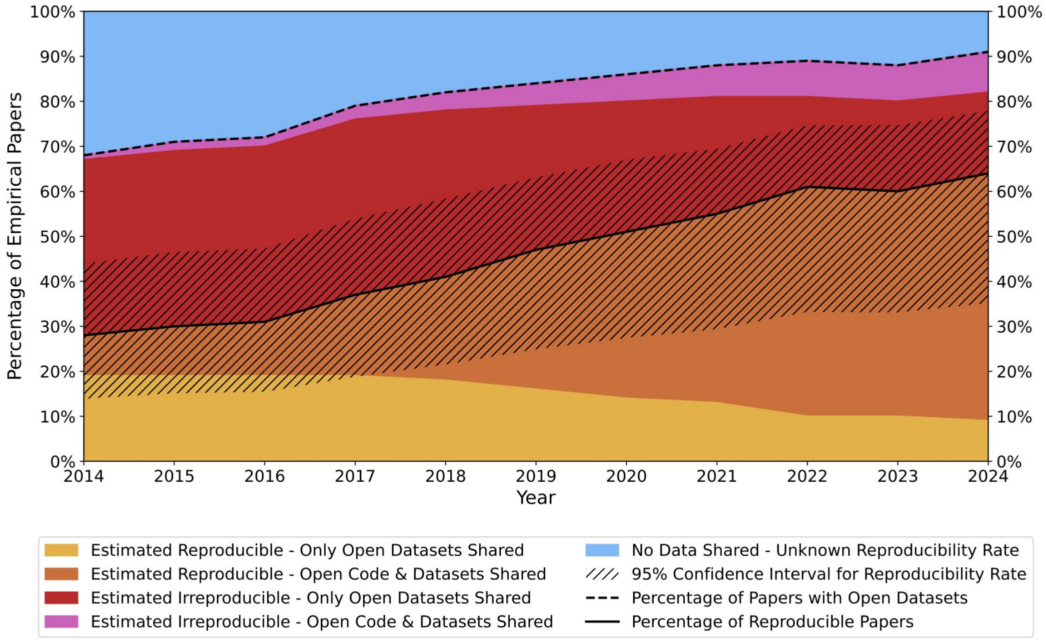
Each coloured band represents papers documenting a specific count of variables, from zero (top, dark green) to all seven (bottom, light orange). Documentation comprehensiveness improved substantially: the proportion of papers documenting none of these variables decreased from 3% to 0.3%, while papers 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.

AI Research is sharing more code and data



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.

Estimated AI Reproducibility increased from 28% to 64%

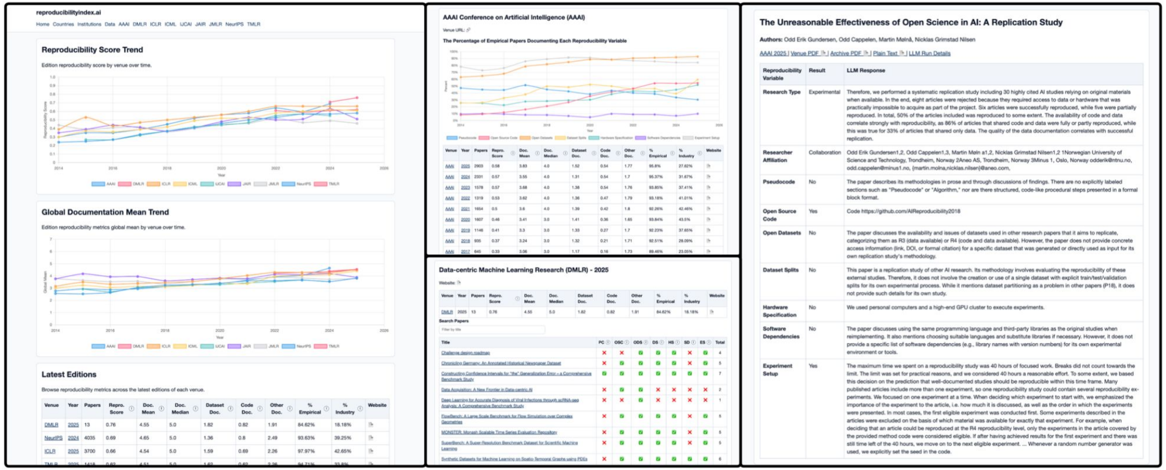


The estimated reproducibility of papers based on the availability of open code and open datasets, applying the empirical reproducibility rates reported by Gundersen et al.[1] The overall estimated reproducibility rate (dot-dash black line) more than doubled from 28% in 2014 to 64% in 2024. The dashed 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%.

Extending to reproducibilityindex.ai

Visit Website

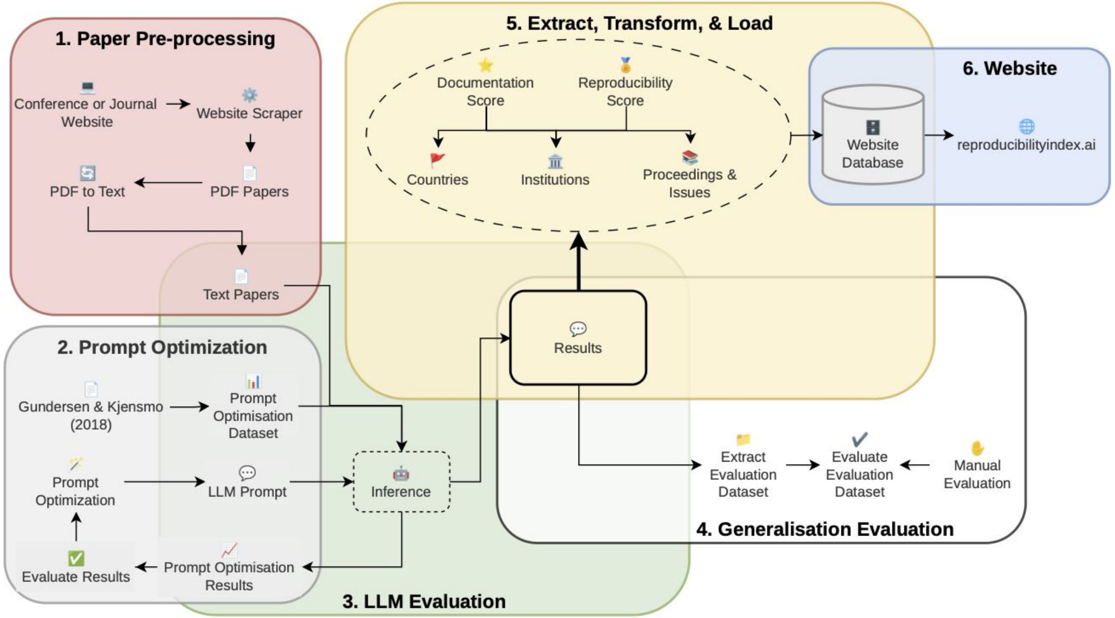
- From conferences-only to conferences + journals
- Expanded to 74,972 papers, 12 years (2014–2025)
- Added country and institution rankings
- Added LLM-based author metadata extraction
- Launched as a continuously updated public platform
- Five levels of granularity: paper, edition, venue, country, institution
- Continuously updated, traceable runs (prompt/model versioning)



The Two Metrics

- Documentation Score (0–7): seven Boolean reproducibility variables
  - Boolean because LLMs can accurately detect the presence of these reproducibility variables, but cannot necessarily access the quality
- Reproducibility Score (0–1): empirically calibrated using Gundersen et al.[1] rates (0.86 / 0.33 / 0)
  - The formula can be updated with new empirical evidence

System Architecture



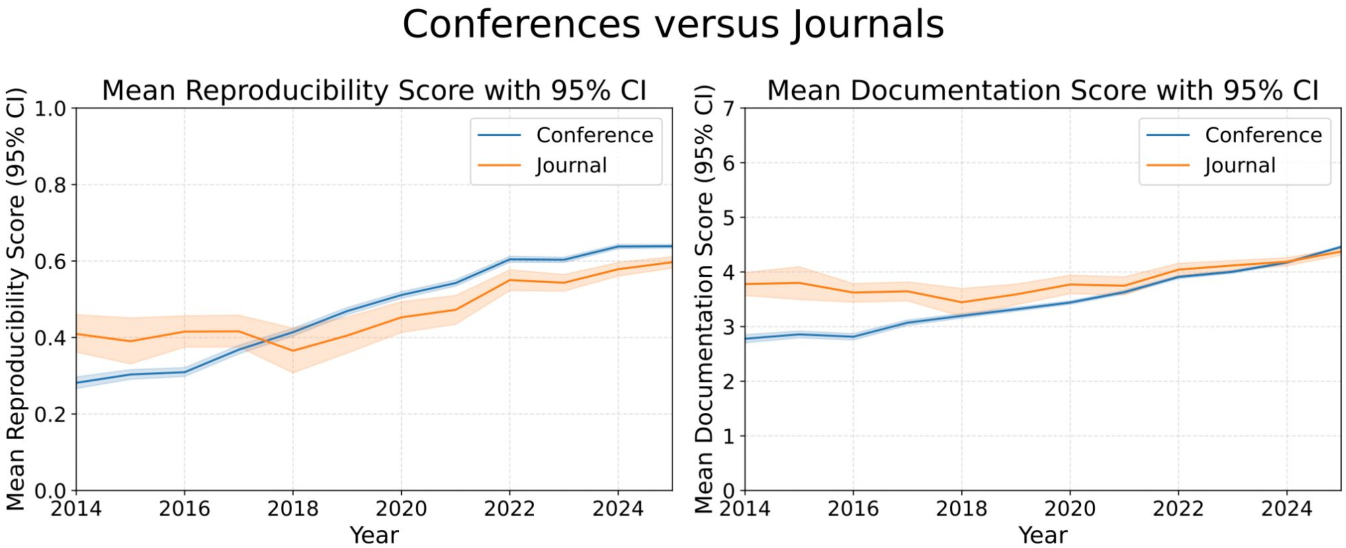
Six-component pipeline: Paper Pre-processing → Prompt Optimisation → LLM Evaluation → Generalisation Evaluation → ETL → Website. Components 1–4 inherited/validated from prior work[5]; 5–6 are new.

Author Metadata Extraction

- To analyze how well countries and institutions perform, we extended Coakley et al.[5] to capture author medata
  - We updated the LLM prompt and re-analyzed the papers using the Gundersen & Kjensmo[2] dataset
- Nine fields per author extracted via the LLM
- 98.99% overall accuracy across 335,298 authorships on the evaluation dataset

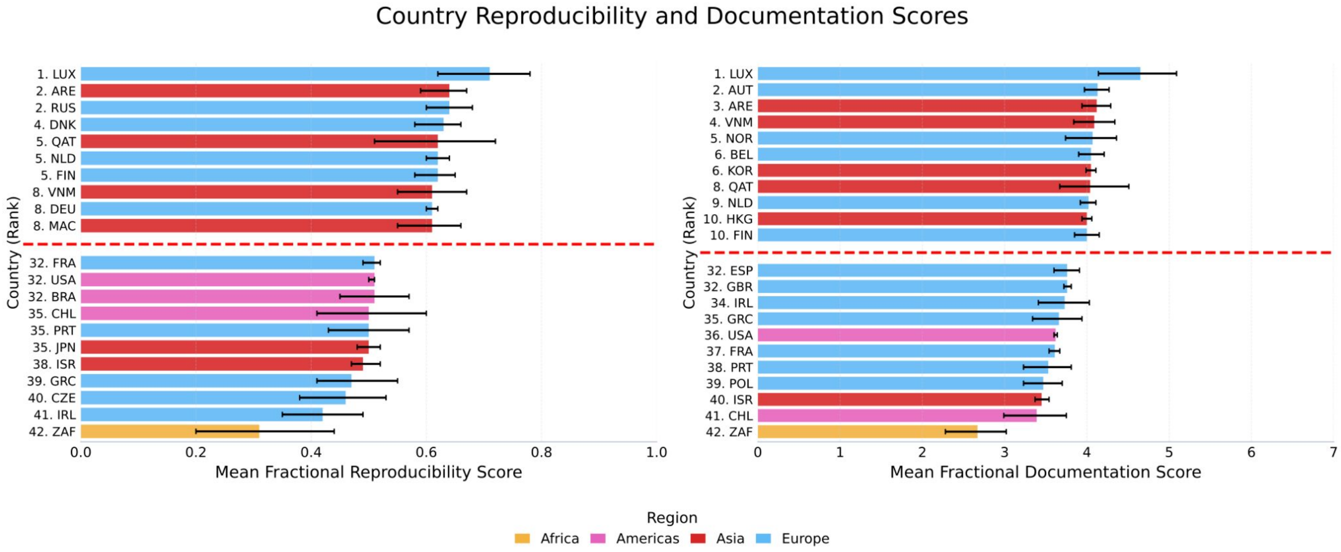
Field	Total	Correct	Error	Accuracy	95% CI
# of Authors	703	703	0	100.0%	[99.5%, 100.0%]
First Name	703	703	0	100.0%	[99.5%, 100.0%]
Last Name	703	703	0	100.0%	[99.5%, 100.0%]
Email	703	701	2	99.7%	[99.0%, 99.9%]
Institution	703	680	23	96.7%	[95.1%, 97.8%]
Department	703	682	21	97.0%	[95.5%, 98.0%]
City	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%]**

Results: Conferences vs. Journals



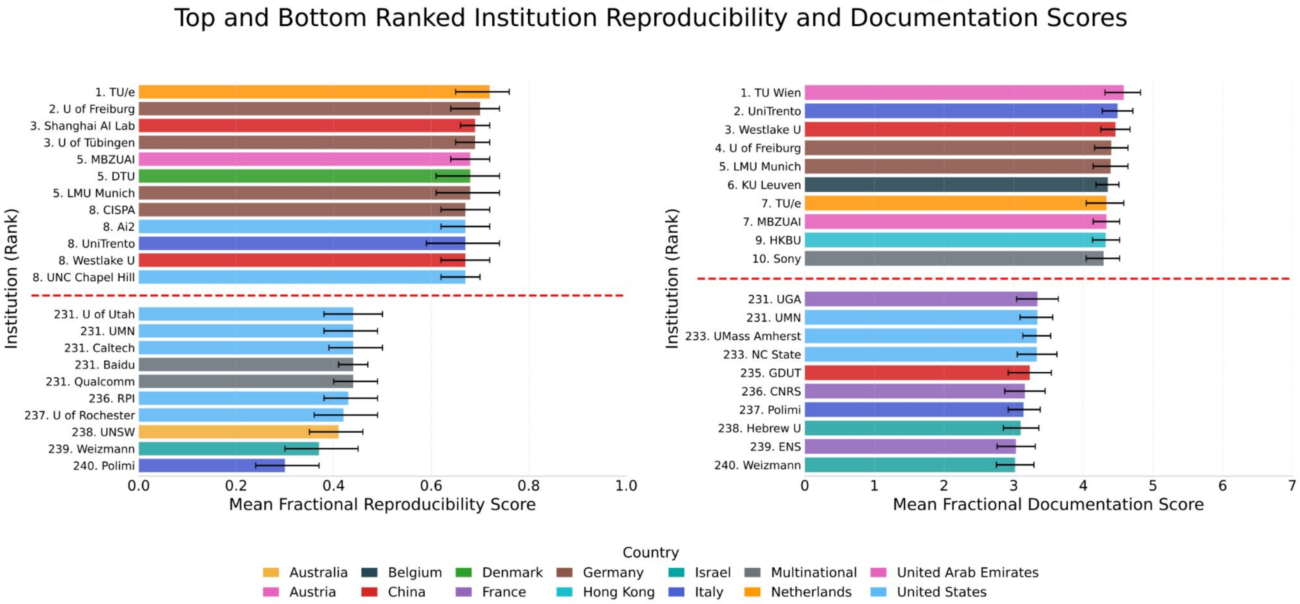
- Both improve steadily 2014–2025
- Conferences slightly higher reproducibility (0.55 vs 0.53, negligible effect)
- Journals higher documentation (4.08 vs 3.77, small effect)
- Gap narrowing

Results: Countries



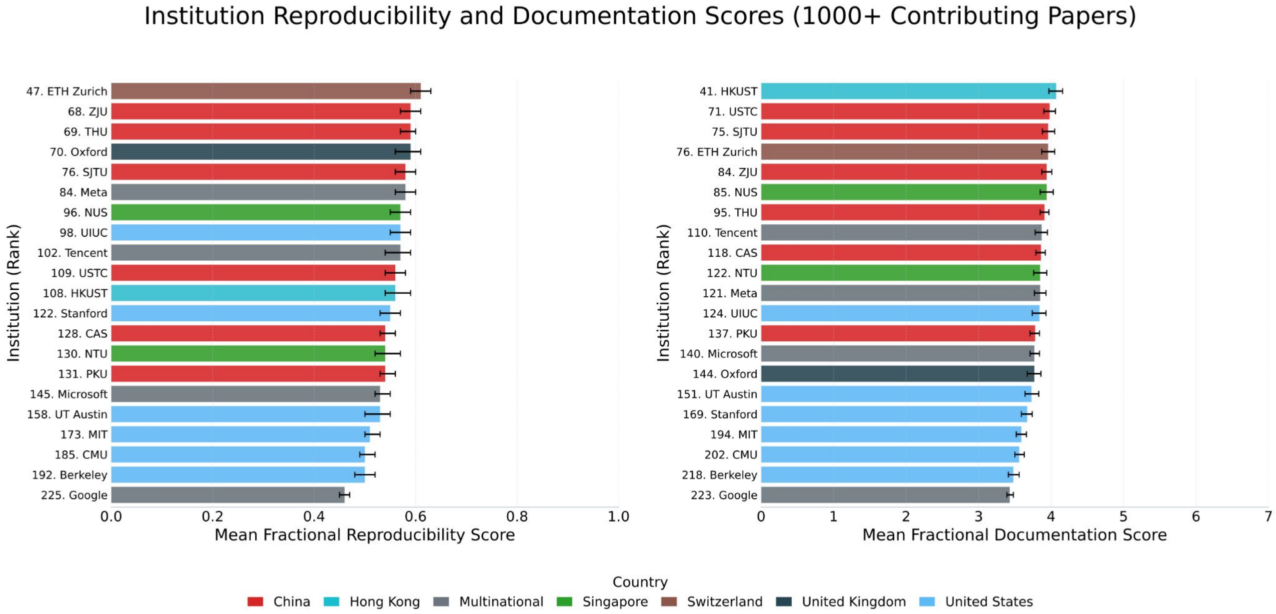
- Only countries with 25 or more contributing papers
- Luxembourg, United Arab Emmerates, Denmark, Netherlands lead
- Europe & Asia dominate top 10
- US: largest contributor (28,163 papers) but ranks 32nd/36th

Results: Institutions (100+ Contributing 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)

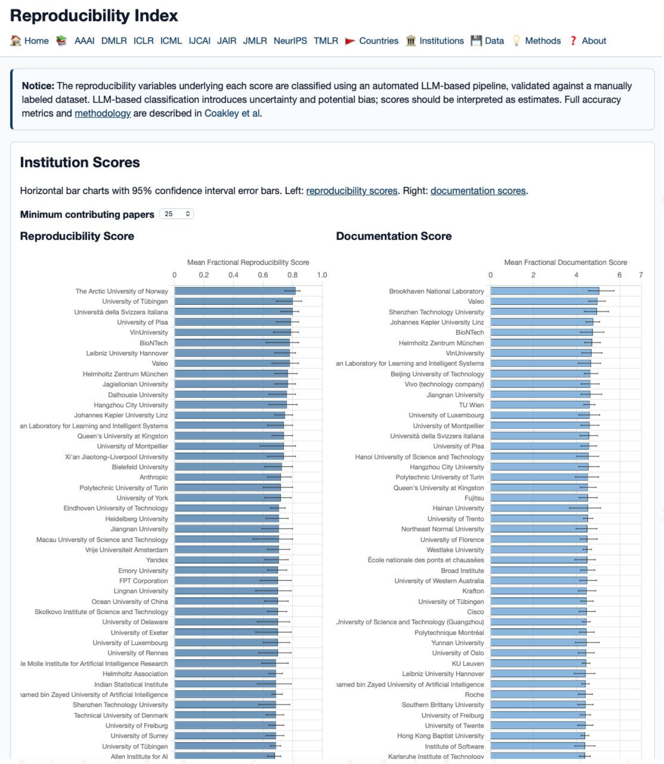
Results: Institutions (1000+ Contributing Papers)



- ETH Zurich ranked first in reproducibility and Hong Kong University of Science and Technology ranked first in documentation.
- Multinational corporations rank low — IP concerns as possible explanation
- Google - with the greatest number of contributing papers - ranked last of institutions with 1000+ CP

Since Submission: Website Updates

- Added NeurIPS 2025 (80,258 total papers now)
- Higher granularity institutional analysis – minimum lowered to 25 contributing papers



Code and Data

Code for evaluating papers and results:
https://doi.org/10.5281/zenodo.20145235

Code for Website:
https://github.com/kevincoakley/reproducibilityindex-ai-web

Contact information:
kcoakley@sdsc.edu

Download Presentation

NTNU SDSC RWTH AACHEN UNIVERSITY

References

1. 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
2. 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
3. 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.
4. 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
5. 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 (under review)