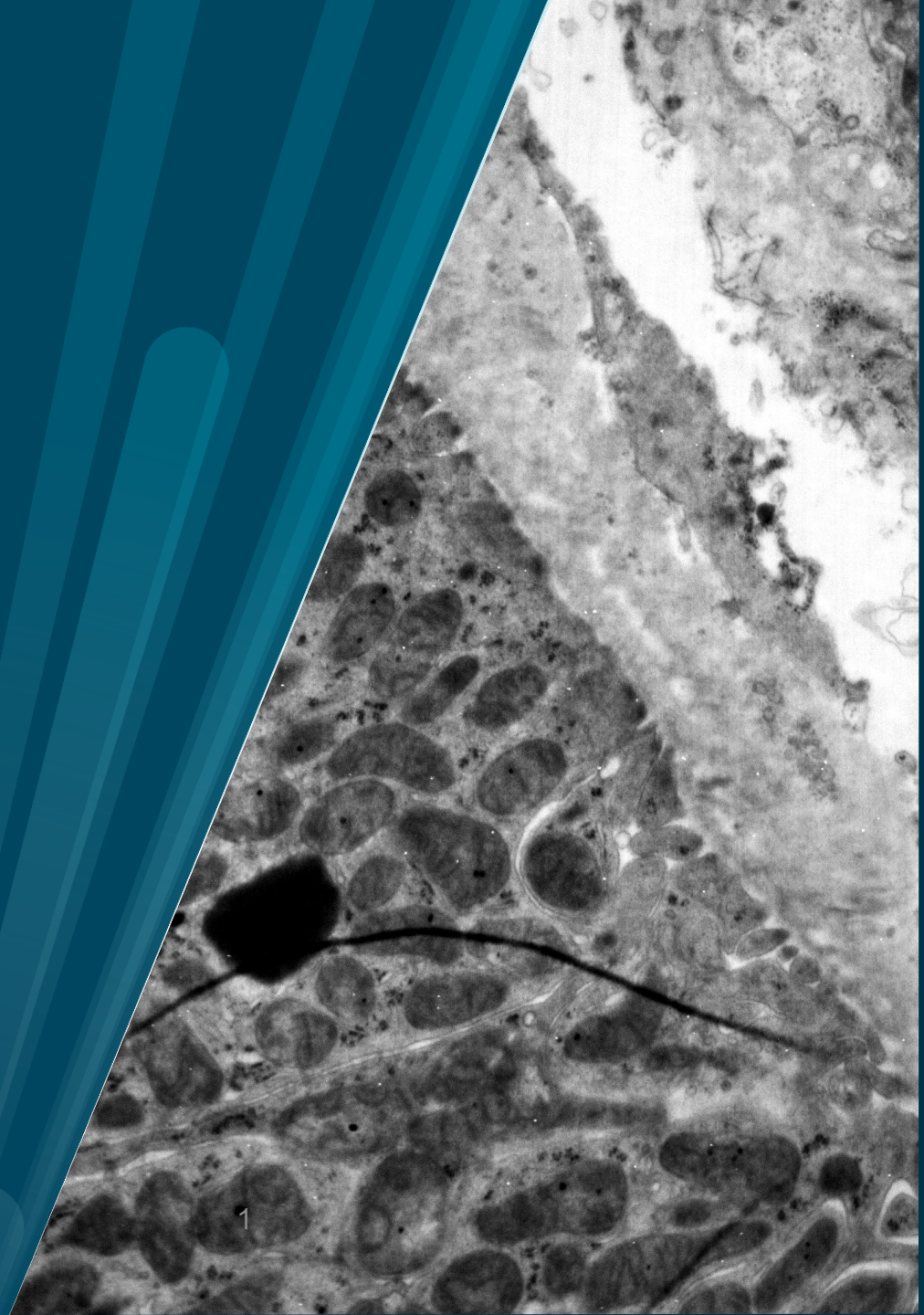
UiT The Arctic University of Norway

Code, Chaos, and Collaboration

The Quest for Reproducible AI Research

Elisabeth Wetzer, Associate Professor in Machine Learning
Machine Learning Group
elisabeth.wetzer@uit.no



Machine Learning Group

- machine-learning.uit.no
- visual-intelligence.no
- integreat.no

What is Reproducibility?

Reproducibility

External researchers can validate an experiment by following its documentation—either using the original data/code or reimplementing it —to confirm the results and implementation correctness beyond mere repeatability.

Replicability

Independent researchers confirm the same hypothesis with intentionally varied implementations or using fundamentally different experimental designs—showing the findings hold across altered conditions.

Repeatability

Repeatability is the ability of the original researchers to rerun an experiment with the same setup, procedures, and analysis and obtain the same results and conclusions.

The Reproducibility Crisis

Is the widespread failure to reliably replicate published machine learning results.

Erodes trust and credibility

If results can’t be reliably reproduced, researchers, funders, and the public lose confidence in published findings and scientific claims

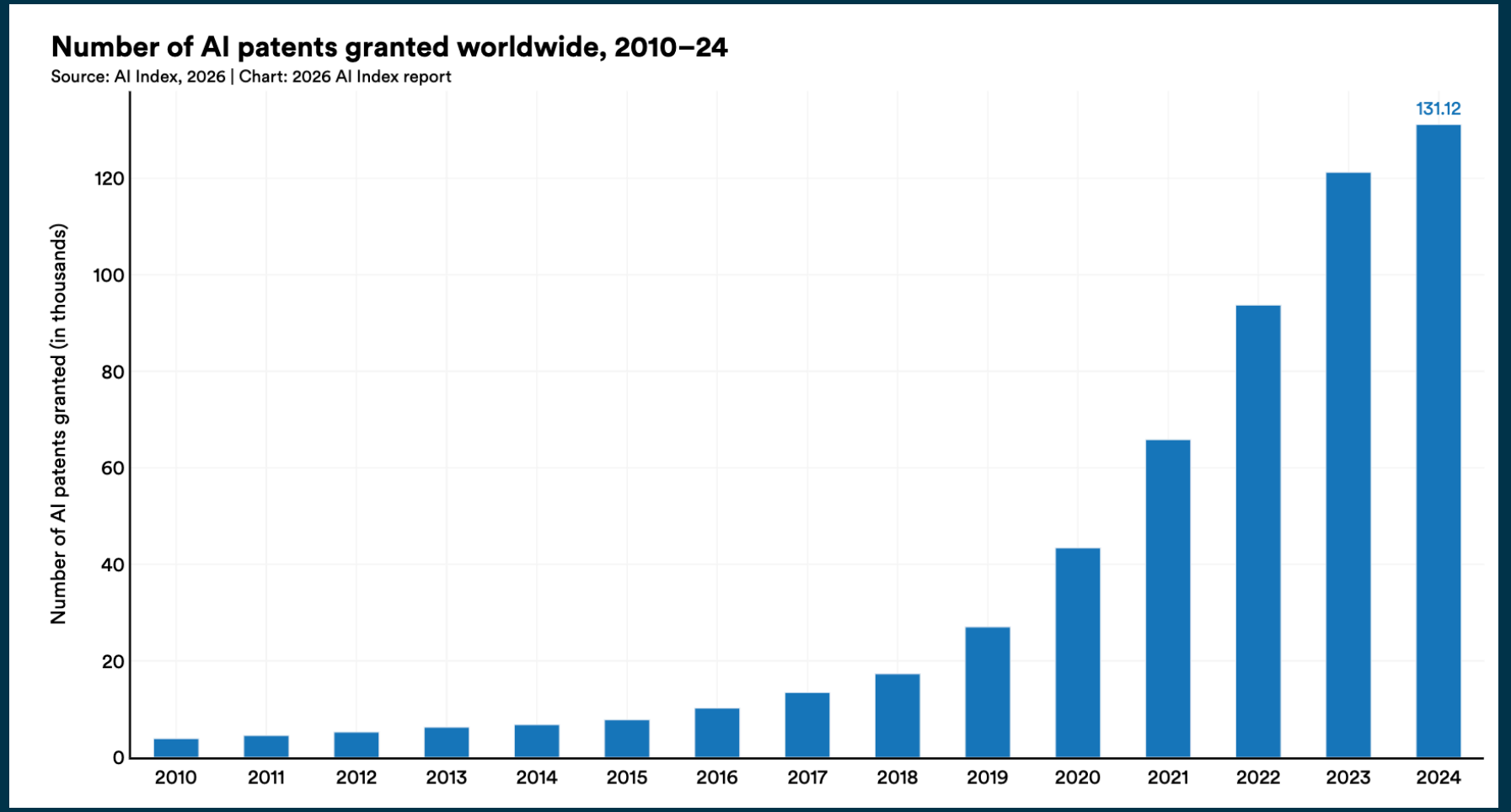
Wastes time and resources and slows progress

Irreproducible work forces others to chase false leads, repeat experiments, or build on unstable foundations

Misguides priorities and harms real-world impact

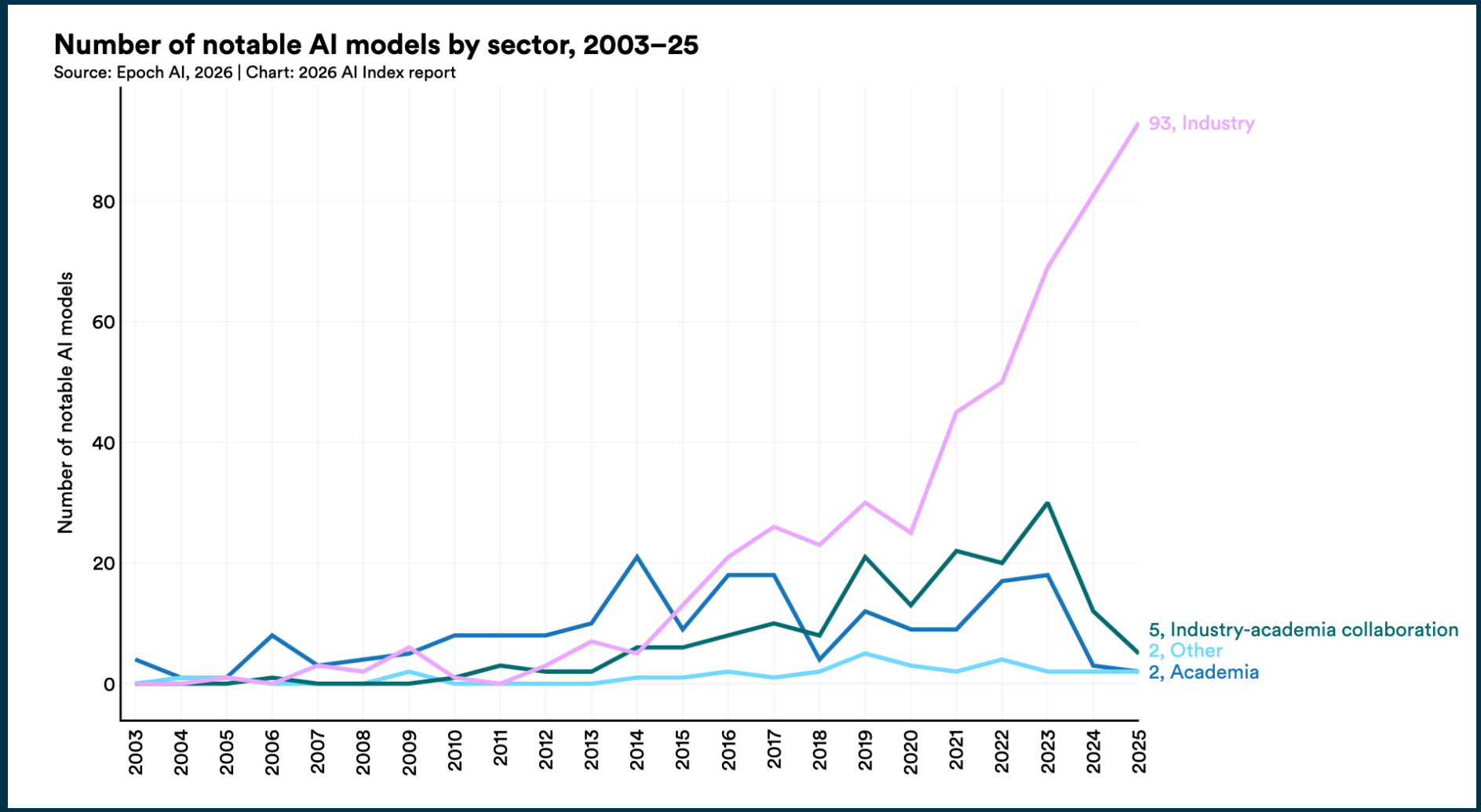
Incentives that reward flashy but fragile results skew research agendas

Machine Learning Explosion



AI Index Report 2026, Human-Centered AI, Stanford University

Machine Learning Explosion



AI Index Report 2026, Human-Centered AI, Stanford University

Machine Learning Explosion

Submission Trend

+136.5%

Last 5 events (2021-2025): 9,122 → 21,575 submissions (+12,453).

NEURAL INFORMATION PROCESSING SYSTEMS

Yearly Statistics

YEAR	LOCATION	ACCEPTED	SUBMITTED	ACCEPTANCE RATE	SECOND TRACK	NOTE
2025	San Diego, California, USA	5,290	21,575	24.5%		
2024	Vancouver, Canada	4,043	15,671	25.8%		
2023	New Orleans, USA	3,218	12,343	26.1%		
2022	New Orleans, USA	2,672	10,411	25.7%		
2021	Online, Online	2,334	9,122	25.6%		
2020	Online, Online	1,898	9,467	20.0%		
2019	Vancouver, Canada	1,428	6,743	21.2%		
2018	Montreal, Canada	1,011	4,856	20.8%		
2017	Long Beach, California, USA	678	3,240	20.9%		
2016	Barcelona, Spain	568	2,406	23.6%		

Machine Learning Explosion

Submission Trend

+136.5%

Last 5 events (2021-2025): 9,122 → 21,575 submissions (+12,453).

NEURAL INFORMATION PROCESSING SYSTEMS

Yearly Statistics

YEAR	LOCATION	ACCEPTED	SUBMITTED	ACCEPTANCE RATE	SECOND TRACK	NOTE
2025	San Diego, California, USA	5,290	21,575	24.5%		
2024	Vancouver, Canada	4,043	15,671	25.8%		
2023	New Orleans, USA	3,218	12,343	26.1%		
2022	New Orleans, USA	2,672	10,411	25.7%		
2021	Online, Online	2,334	9,122	25.6%		
2020	Online, Online	1,898	9,467	20.0%		
2019	Vancouver, Canada	1,428	6,743	21.2%		
2018	Montreal, Canada	1,011	4,856	20.8%		
2017	Long Beach, California, USA	678	3,240	20.9%		
2016	Barcelona, Spain	568	2,406	23.6%		

In 2026, the number of submissions likely hit over 40 000

Machine Learning Explosion

Submission Trend

+136.5%

Last 5 events (2021-2025): 9,122 → 21,575 submissions (+12,453).

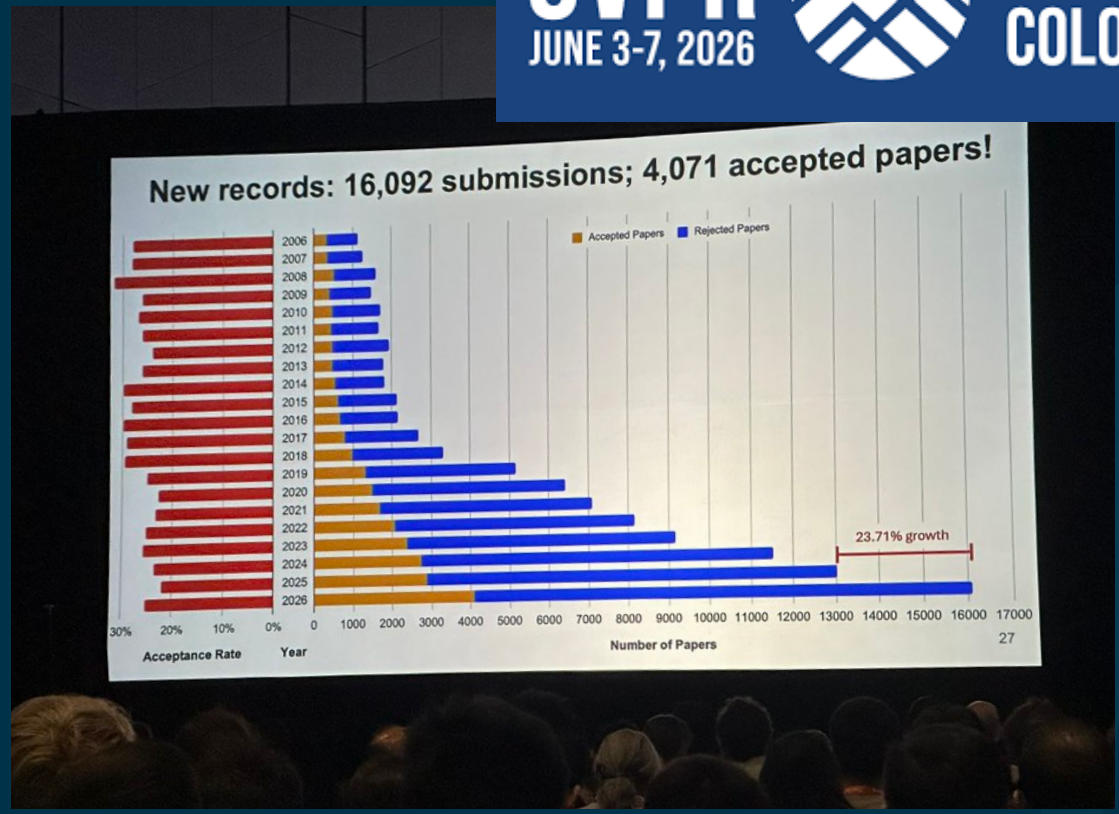
NEURAL INFORMATION PROCESSING SYSTEMS

Yearly Statistics

YEAR	LOCATION	ACCEPTED	SUBMITTED	ACCEPTANCE RATE	SECOND TRACK	NOTE
2025	San Diego, California, USA	5,290	21,575	24.5%		
2024	Vancouver, Canada	4,043	15,671	25.8%		
2023	New Orleans, USA	3,218	12,343	26.1%		
2022	New Orleans, USA	2,672	10,411	25.7%		
2021	Online, Online	2,334	9,122	25.6%		
2020	Online, Online	1,898	9,467	20.0%		
2019	Vancouver, Canada	1,428	6,743	21.2%		
2018	Montreal, Canada	1,011	4,856	20.8%		
2017	Long Beach, California, USA	678	3,240	20.9%		
2016	Barcelona, Spain	568	2,406	23.6%		

https://openaccept.org/c/ai/neurips/

CVPR JUNE 3-7, 2026 DENVER COLORADO



digg.com/tech/13bjrm9s

Machine Learning Explosion

- It is the same trend across all the big ML conferences:

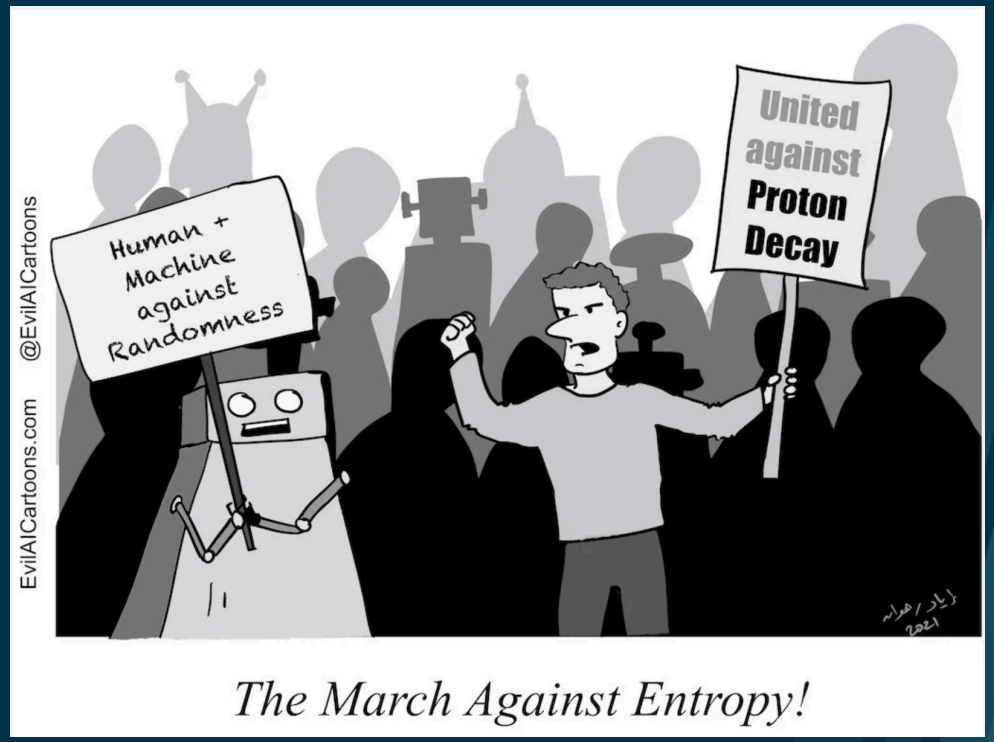
  ICLR, ICML, NeurIPS, AAAI, ECCV, ICCV, CVPR,…

- The number of available reviewers does not scale with this trend

Machine Learning Explosion

Result

- Lots of noise in reviewing
- We need quality insurance beyond the review process



Machine Learning Explosion

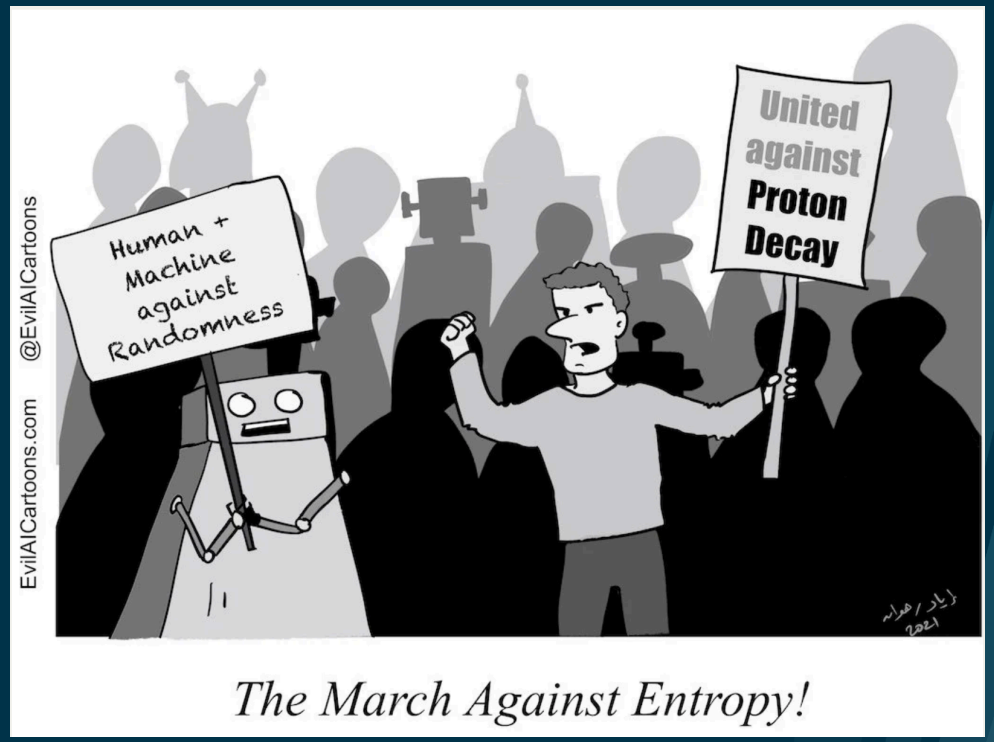
Result

- Lots of noise in reviewing
- We need quality insurance beyond the review process

NeurIPS Consistency Experiment:

23% of submissions received different decisions

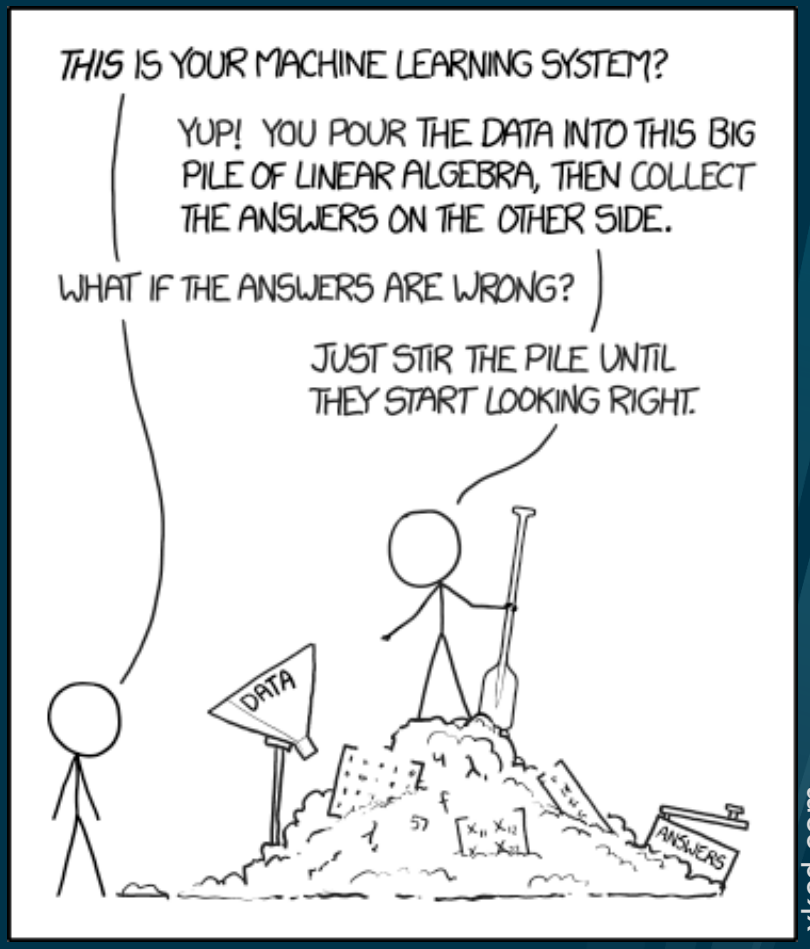
50.6% of accepted papers would have been rejected



Machine Learning Explosion

- We have a problem far beyond reproducibility

- Nonetheless, if scientific fields like medicine, psychology or other experimental fields can operate under uncontrollable factors like patient diversity and massive experimental costs, we should be able to perform sound research



What makes for a good ML paper?

What makes for a good ML paper?

New method addressing knowledge gap

What makes for a good ML paper?

New method addressing knowledge gap

Apply method to a handful of benchmark datasets

What makes for a good ML paper?

New method addressing knowledge gap

Apply method to a handful of benchmark datasets

Compare to 3-4 competing methods

What makes for a good ML paper?

New method addressing knowledge gap

Apply method to a handful of benchmark datasets

Compare to 3-4 competing methods

What makes for a good ML paper?

New method addressing knowledge gap

Apply method to a handful of benchmark datasets

Compare to 3-4 competing methods

Implementation Challenge

- Alternative I:
  No code available; method is described in paper

Implementation Challenge

- Alternative I:
  No code available; method is described in paper

  - Page limits imposed by conferences and journals
  - Default settings often not reported, but may change over time
  - A lot of additional time used for implementation, leaves no resources for fine-tuning

Implementation Challenge

- Alternative I:
  No code available; method is described in paper

- Alternative II:
  Code repo is available; but takes significant work to make it work

Implementation Challenge

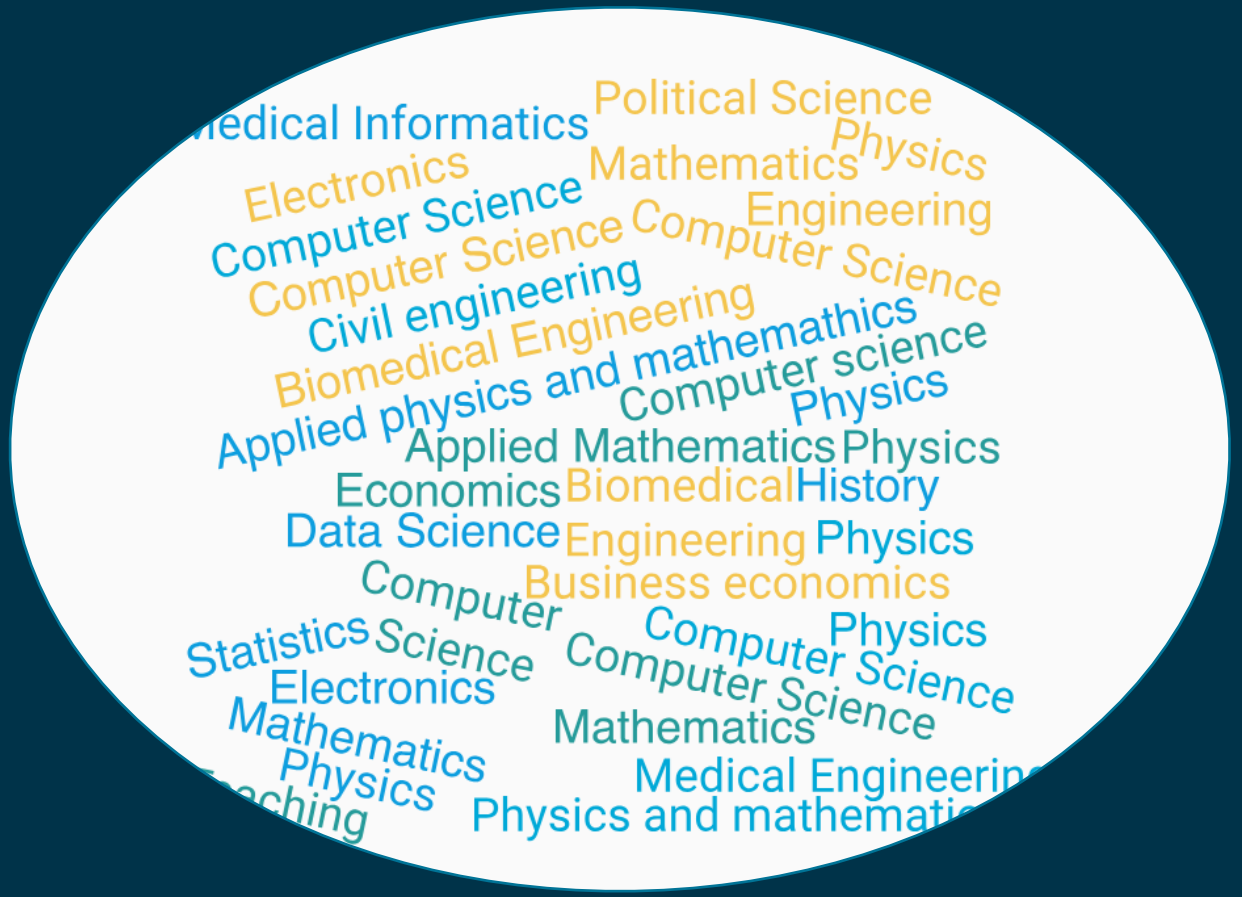
- Alternative I:
  No code available; method is described in paper

- Alternative II:
  Code repo is available; but takes significant work to make it work

- Alternative III:
  Code repo is available and actually usable ❤️

Why is that?

ML as a diverse field



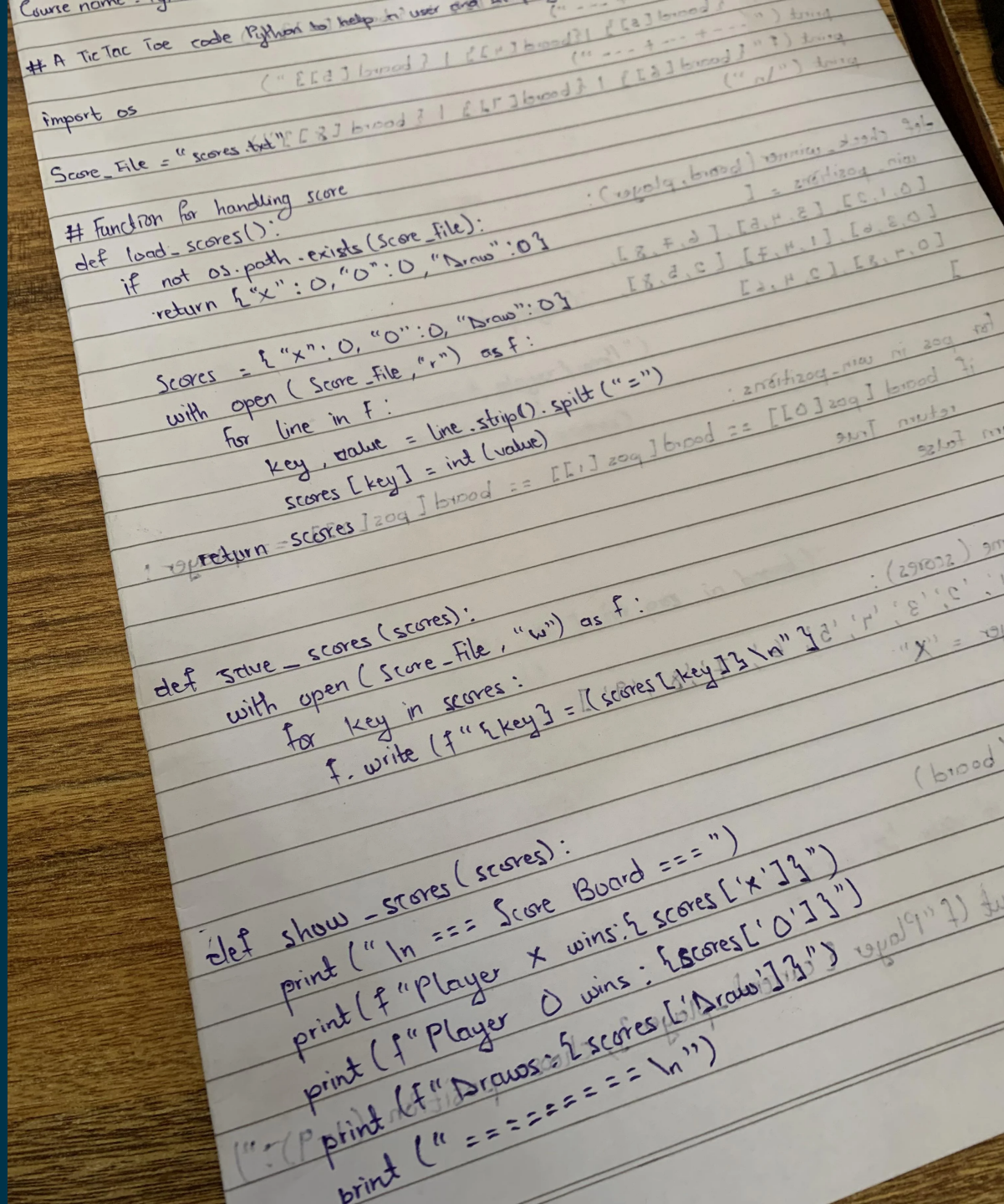
ML as a diverse field

- Many come from adjacent fields such as signal processing, mathematics or physics
- About 30% had more than two codings courses in their life
- Most reported being self-taught programmers or having had one course in programming, but no training in best coding practices

ML as a diverse field

- Some seniors who ended up in ML may not be passionate programmers
- Underestimate the time and effort that goes into code clean-up
- Don’t value research time spent on software publishing

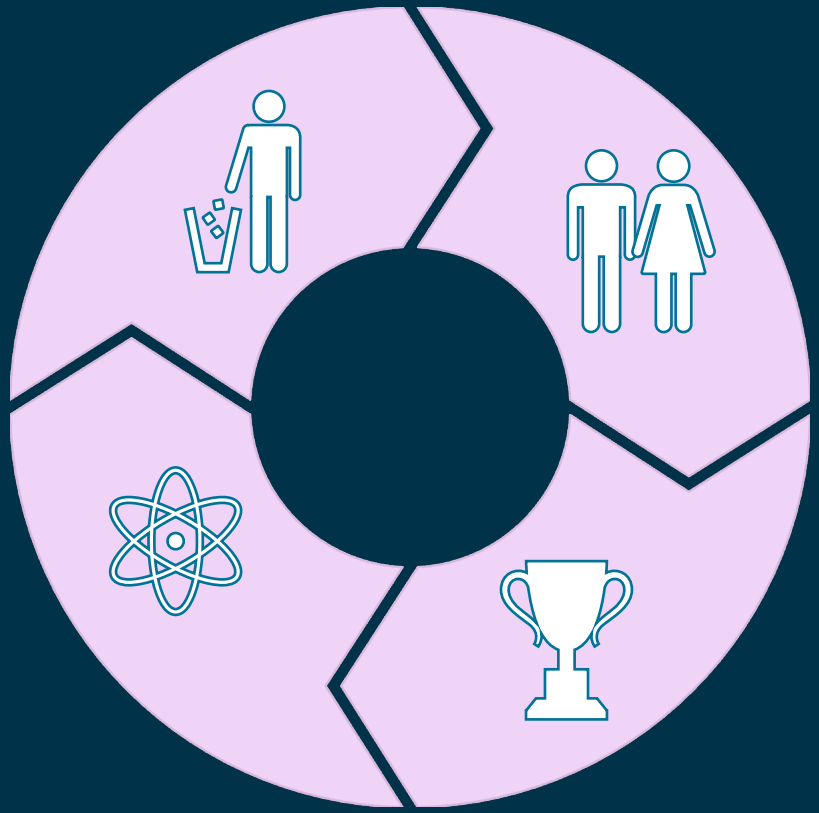
reddit: Bluebill_365



Reasons you should care about your research code

Efficiency

We have a responsibility to not waste resources



Collaborations

Good code can spark community contributions and collaborations

Scientific Integrity

It should be in your interest to have your findings challenged if your arguments or experiments were flawed

Impact

You want to be the competing method everyone picks because they do not have to spend weeks getting to run

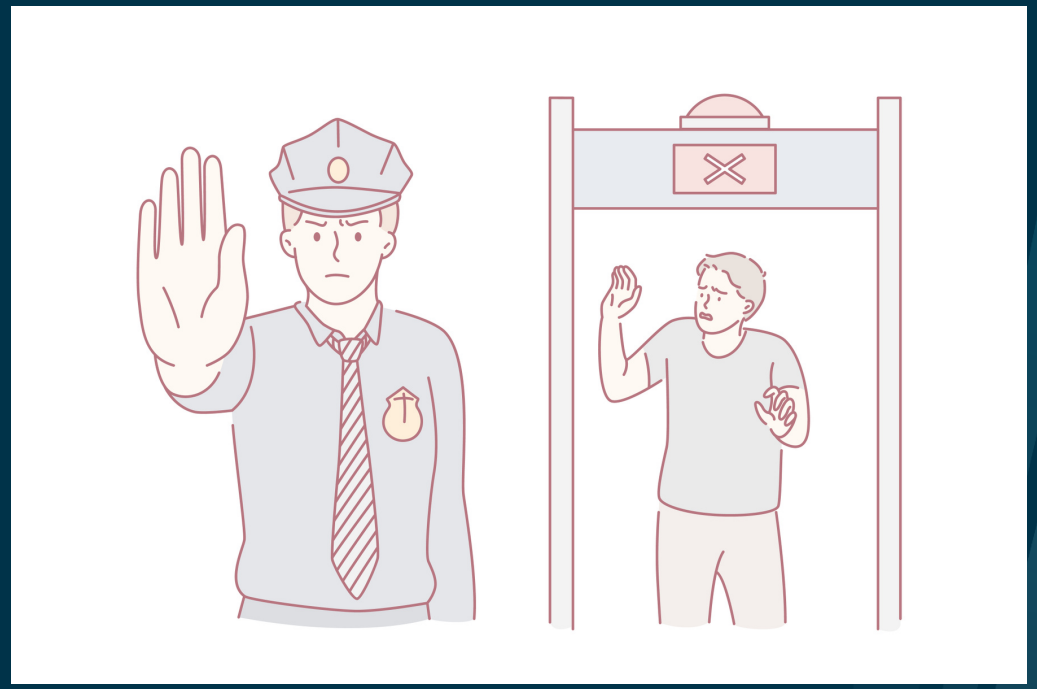
The purpose is not to keep tabs on you

README

All the libraries and the python version installed are listed below. All of them are not necessary, the core ones are e2cnn and pytorch.

Package Version

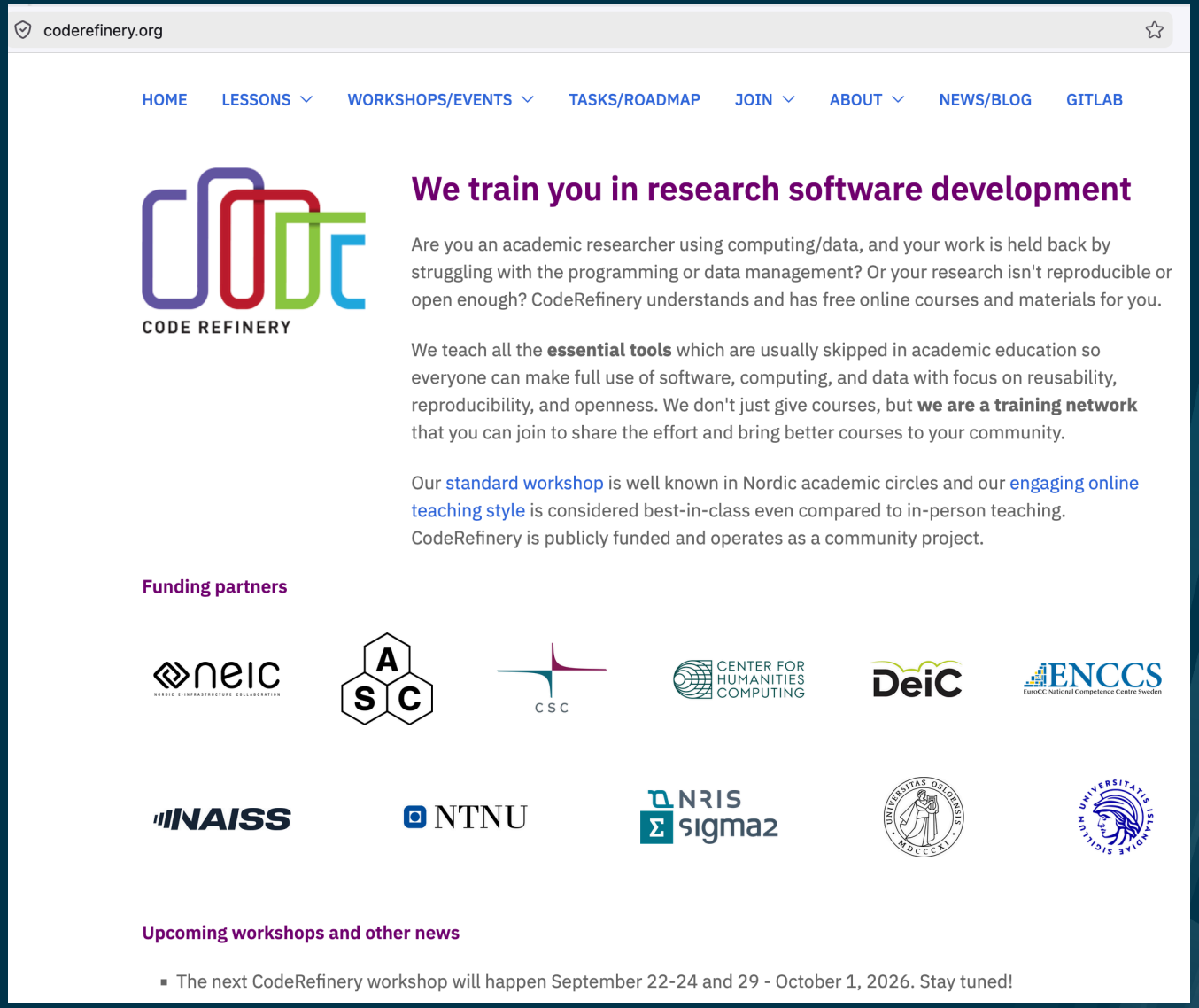
alabaster 0.7.12
albumentations 0.5.2
appdirs 1.4.4
astroid 2.3.3
atomicwrites 1.3.0
attrs 19.3.0
autopep8 1.5
Babel 2.8.0
backcall 0.1.0
bleach 3.1.0
cachey 0.2.1
certifi 2019.11.28 cffi 1.14.0
chardet 3.0.4
cloudpickle 1.3.0
cryptography 2.9
cycler 0.10.0
dask 2021.2.0
decorator 4.4.1
defusedxml 0.6.0
diff-match-patch 20181111
docstring-parser 0.7.3
docutils 0.16
e2cnn 0.1
entrypoints 0.3
flake8 3.7.9
freetype-py 2.1.0
HeapDict 1.0.1



So how do we get better in this?

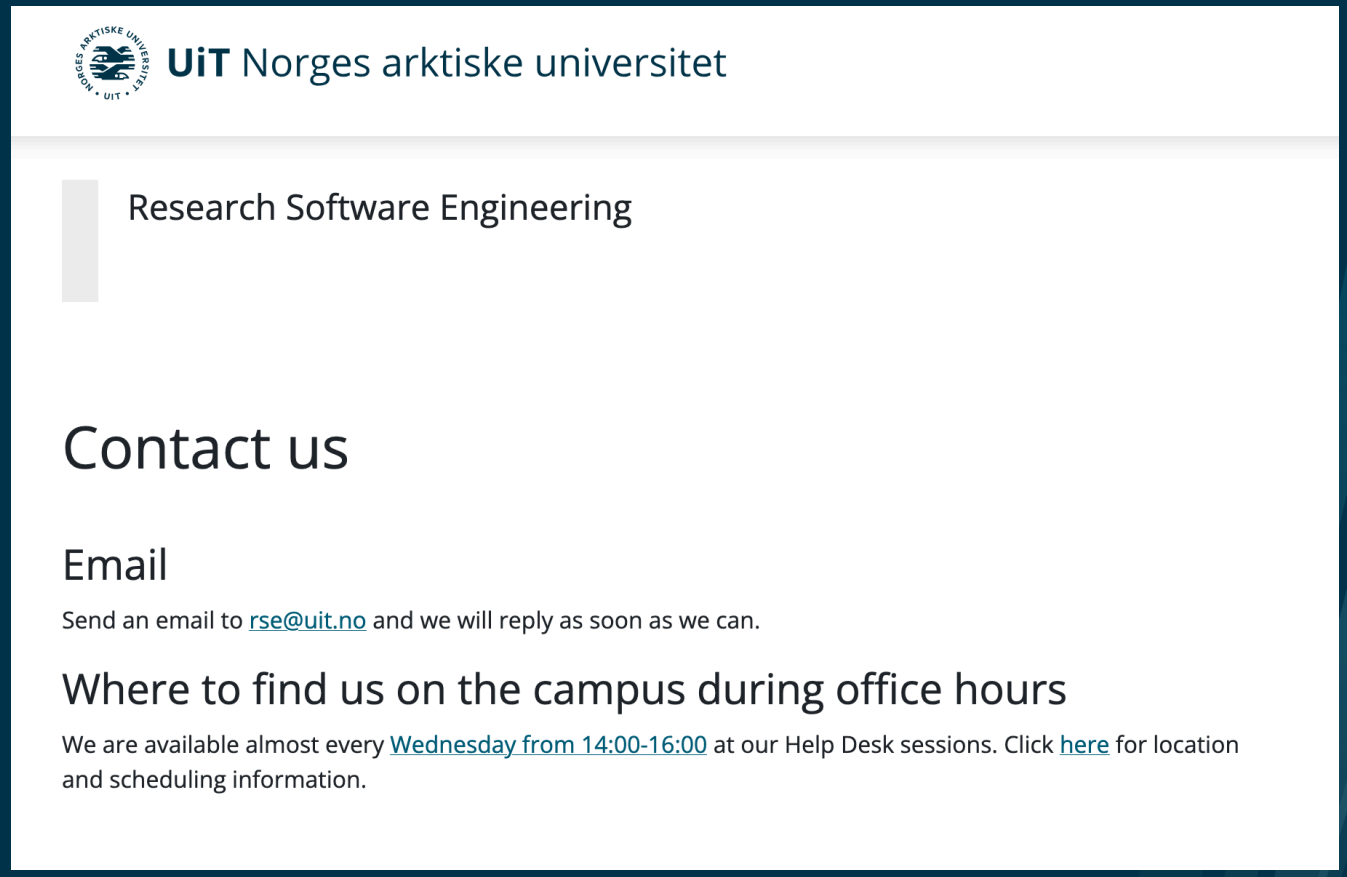
Increase RSE literacy

- Code Refinery offers plenty of learning material, courses and workshops



Increase RSE literacy

- RSE Group at UiT offers drop-in help desk every week



Increase RSE literacy

- Offer specialized training for PhD students

Increase RSE literacy

- Increase Reward and Appreciation in the Field

Who are our repos for?

- If we all use Claude & co to implement the competing method for our setting

… should we consider our paper and code well enough documented and reproducible if Claude can replicate our results?

Who are our repos for?

- If we all use Claude & co to implement the competing method for our setting

… should we consider our paper and code well enough documented and reproducible if Claude can replicate our results?

… should this be a service journals provide to test min. requirement?

Trends

- Agentic Reproducibility

Preprint. Under review.

Read the Paper, Write the Code: Agentic Reproduction of Social-Science Results

Benjamin Kohler
ETH Zurich

David Zollikofer
ETH Zurich

Johanna Einsiedler
University of Basel

Alexander Hoyle*
ETH Zurich

Elliott Ash*
ETH Zurich

Abstract

Recent work has used LLM agents to reproduce empirical social science results with access to both the data and code. We broaden this scope by asking: Can they reproduce results given only a paper's methods description and original data? We develop an agentic reproduction system that extracts structured methods descriptions from papers, runs reimplementations under strict information isolation—agents never see the original code, results, or paper—and enables deterministic, cell-level comparison of reproduced outputs to the original results. An error attribution step traces discrepancies through the system chain to identify root causes. Evaluating four agent scaffolds and four LLMs on 48 papers with human-verified reproducibility, we find that agents can largely recover published results, but performance varies substantially between models, scaffolds, and papers. Root cause analysis reveals that failures stem both from agent errors and from underspecification in the papers themselves.

Trends

- Agentic Reproducibility

Preprint. Under review.

Read the Paper, Write the Code: Agentic Reproduction of Social-Science Results

Benjamin Kohler
ETH Zurich

David Zollikofer
ETH Zurich

Johanna Einsiedler
University of Basel

Alexander Hoyle*
ETH Zurich

Elliott Ash*
ETH Zurich

Abstract

Recent work has used LLM agents to reproduce empirical social science results with access to both the data and code. We broaden this scope by asking: Can they reproduce results given only a paper's methods descrip-
eproduction system that
apers, runs reimplemen-
ts never see the original
, cell-level comparison of
or attribution step traces
y root causes. Evaluating
with human-verified re-
ver published results, but
ls, scaffolds, and papers.
th from agent errors and
s.

REPRO-Bench: Can AI agents Automate Research Reproducibility Assessments?

Daniel Kang Follow 5 min read · Jul 28, 2025

Trends

- Agentic Reproducibility

ARA: Agentic Reproducibility Assessment For Scalable Support Of Scientific Peer-Review

Kevin Riehl[1]

ETH Zürich, IVT & Agentic Systems Lab (ASL) kriehl@ethz.ch

&Andres L. Marin[1]

European Commission, Joint Research Centre & University of Konstanz

E-Mail: andres.laverde-marin@ec.europa.eu

&Nikiforos Zacharof Ideas Forward

nikiforos.zacharof@ideasforward.com

&Fan Wu ETH Zürich, ASL E-Mail: fanwu@ethz.ch

&Patrick Langer ETH Zürich, ASL E-Mail: planger@ethz.ch

&Robert Jakob ETH Zürich, ASL E-Mail: rjakob@ethz.ch

&Anastasios Kouvelas ETH Zürich, IVT kouvelas@ethz.ch

&Georgios Fontaras European Commission, Joint Research Centre

georgios.fontaras@ec.europa.eu

&Michail A. Makridis ETH Zürich, IVT mmakridis@ethz.ch

Preprint. Under review.

Read the Paper, Write the Code: Agentic Reproduction of Social-Science Results

Benjamin Kohler
ETH Zurich

David Zollikofer
ETH Zurich

Johanna Einsiedler
University of Basel

Alexander Hoyle*
ETH Zurich

Elliott Ash*
ETH Zurich

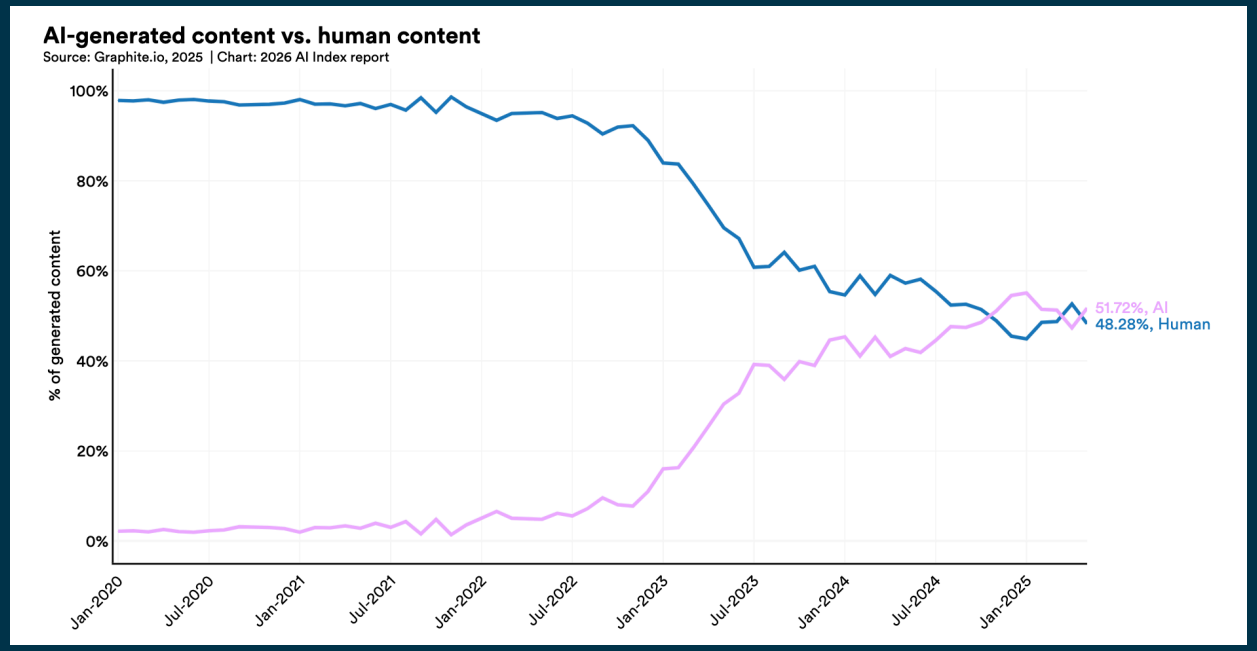
Abstract

Recent work has used LLM agents to reproduce empirical social science results with access to both the data and code. We broaden this scope by asking: Can they reproduce results given only a paper's methods descrip-
eproduction system that
apers, runs reimplemen-
ts never see the original
, cell-level comparison of
or attribution step traces
y root causes. Evaluating
with human-verified re-
ver published results, but
ls, scaffolds, and papers.
th from agent errors and
s.

REPRO-Bench: Can AI agents Automate Research Reproducibility Assessments?

Daniel Kang Follow 5 min read · Jul 28, 2025

What will this mean for coding practices?



AI Index Report 2026,
Human-Centered AI, Stanford University

Key Take-Away

Value Reproducible Code and Open Data

- ✦ Teach students best coding practices, not only CS students
- ✦ Reward time spent on code clean up and software publishing
- ✦ Introduce minimal requirements on code usability

Thank you!

Contact:
elisabeth.wetzer@uit.no