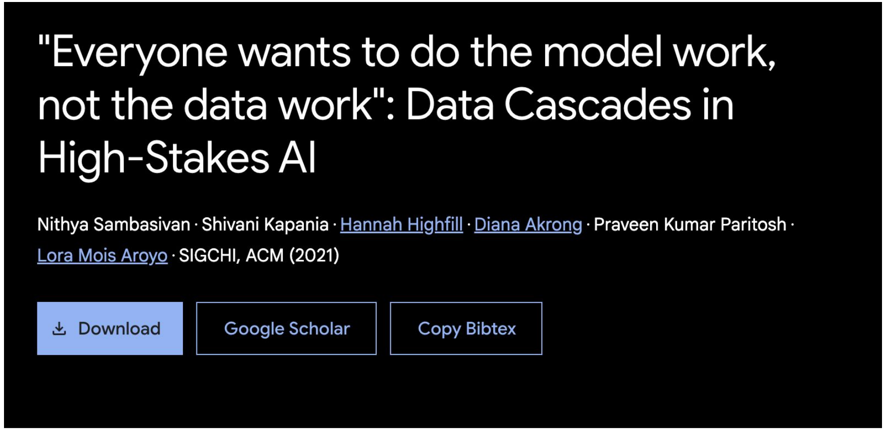
Benchmarking (in) AI research

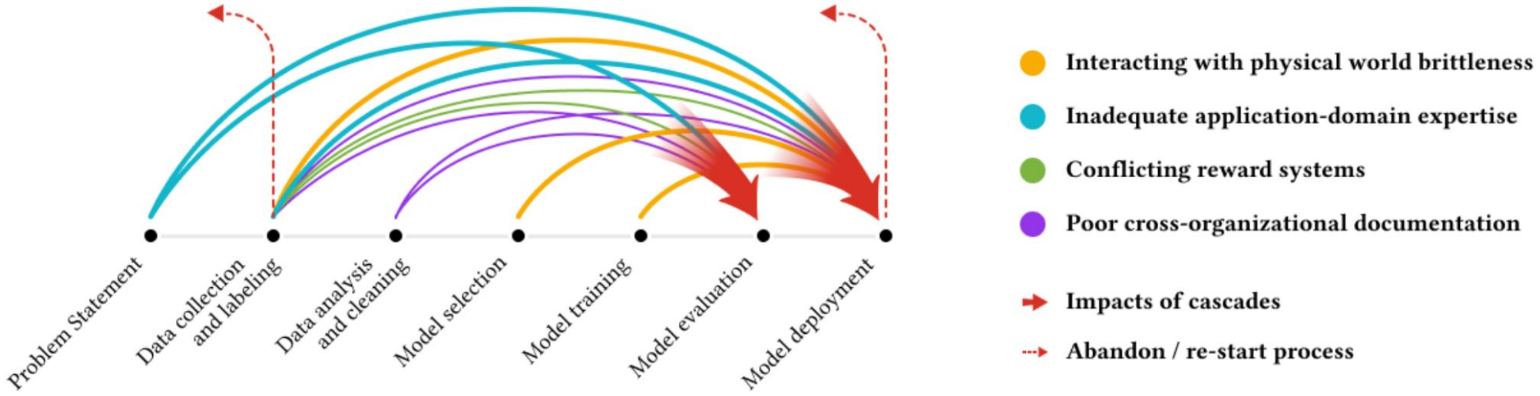
A metascience perspective

Joaquin Vanschoren, TU Eindhoven

Not all speed is movement

- In 2021, data was the most under-valued and de-glamorised aspect of AI
- Very few incentives to create good datasets, leading to many dataset issues





Not all speed is movement

- Significant public criticism of the field
- Most algorithms are only evaluated on toy problems, and biased data
- Are we actually making any progress?
- We need to view dataset and benchmark creation as a ***science***, set **high quality standard**, and **reward good work**

Data and its (dis)contents: A survey of dataset development and use in machine learning research

Amandalynne Paullada
Department of Linguistics
University of Washington

Inioluwa Deborah Raji
Mozilla Foundation

Emily M. Bender
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Abstract

Datasets have played a foundational role in the advancement of machine learning research. They form the basis for the models we design and deploy, as well as our primary medium for benchmarking and evaluation. Furthermore, the ways in which we collect, construct and share these datasets inform the kinds of problems the field pursues and the methods explored in algorithm development. However, recent work from a breadth of perspectives has revealed the limitations of predominant practices in dataset collection and use. In this paper, we survey the many concerns raised about the way we collect and use data in machine learning and advocate that a more cautious and thorough understanding of data is necessary to address several of the practical and ethical issues of the field.

Why a new NeurIPS track?

- In 2021, out of 1903 accepted papers, only 4(!) papers introduced new datasets, 10 benchmarks
- New incentives (e.g. altmetrics) are difficult
- We need:
  - new (old) incentives: NeurIPS papers!
  - new guidelines on how to review datasets and benchmarks (most reviewers don't know how)
  - equally high quality bar as main conference

NEURAL INFORMATION PROCESSING SYSTEMS

DATASETS AND BENCHMARKS

174 accepted papers
(out of almost 500)

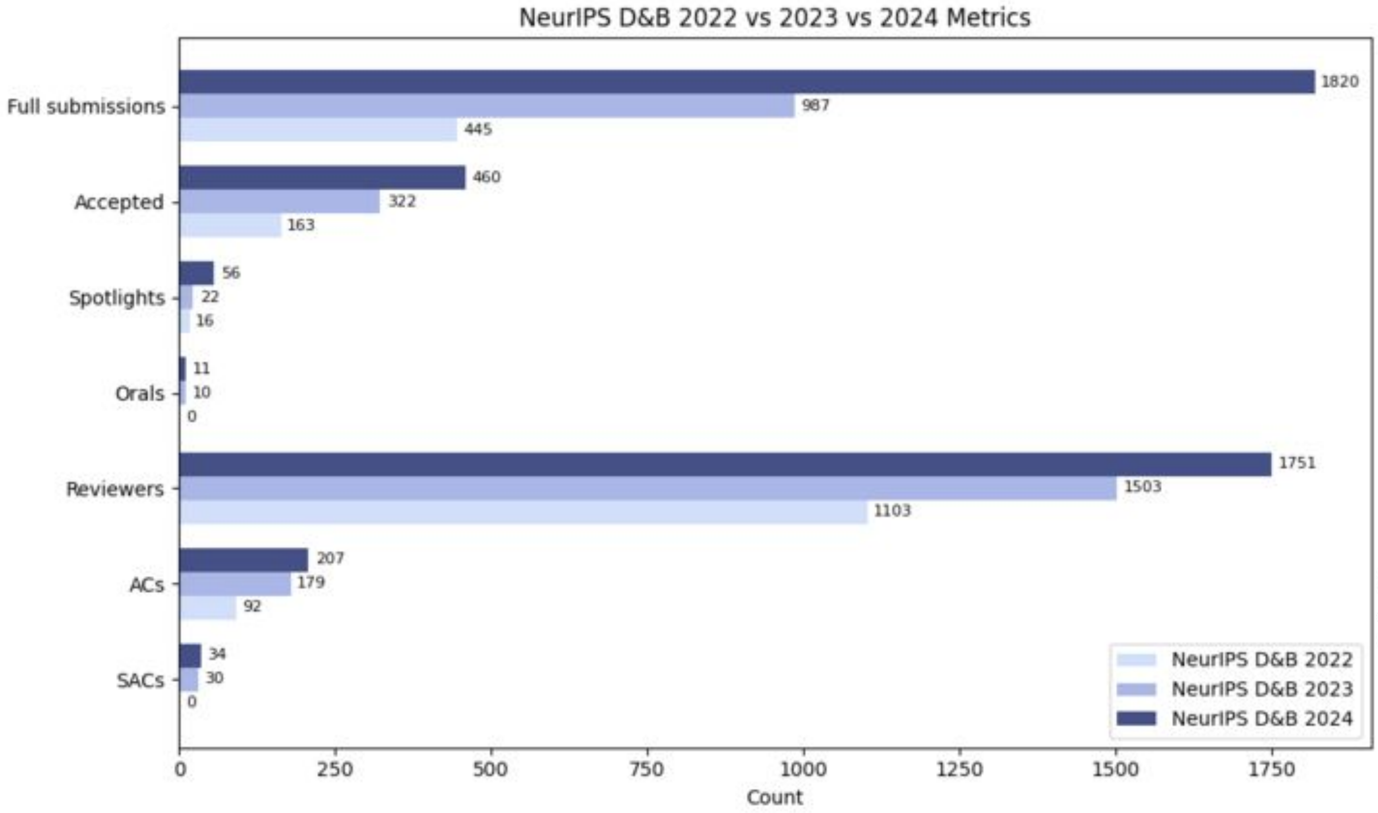
NeurIPS D&B: we had to rethink the 'rules'

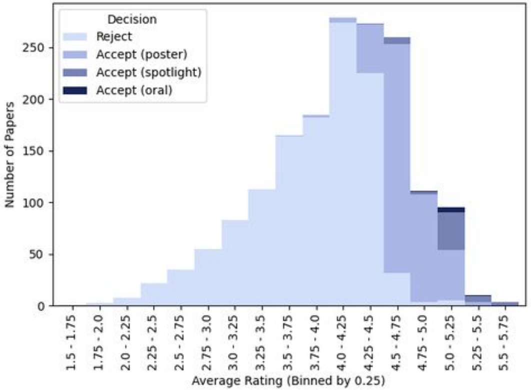
- Drastically rethink established review guidelines
  - **Optional single-blind review**: D&B can often not be reviewed double-blind
    - Datasets/benchmarks need to be hosted, may involve credentials (e.g. medical)
    - Most good work was previously rejected for this reason alone
  - Data/benchmarks need **standardised documentation**, e.g. datasheets
    - How collected? Why? Gaps? Recommended use, distribution, maintenance,...
  - Datasets need to be **accessible and well-maintained**
  - Benchmarks must be **fully reproducible**
- Scope: also pure-code (e.g. RL environments), **meta-analysis**, dataset analysis
- *Is this a recipe we can spread across the community? (In talks with ICML)*

Reviewer guidelines

- NeurIPS checklist, guidelines on code/data submission, reproducibility checklist
- Verify that the dataset is properly documented: datasheets or similar
    - Collection, coverage, proper use must be clear
- Verify accessibility: open formats, licence, meta-data (e.g. schema.org),...
    - Data must be publicly available (at conference time)
    - Open credentialized access for sensitive data
- Hosting and maintenance plan
- Ethical review: when flagged sent to NeurIPS ethical board, 1 was rejected
- **Same high bar as main track** (D&B can't be 2nd order citizens)
    - This was most difficult since reviewer pools was less experienced

Impact / Community acceptance

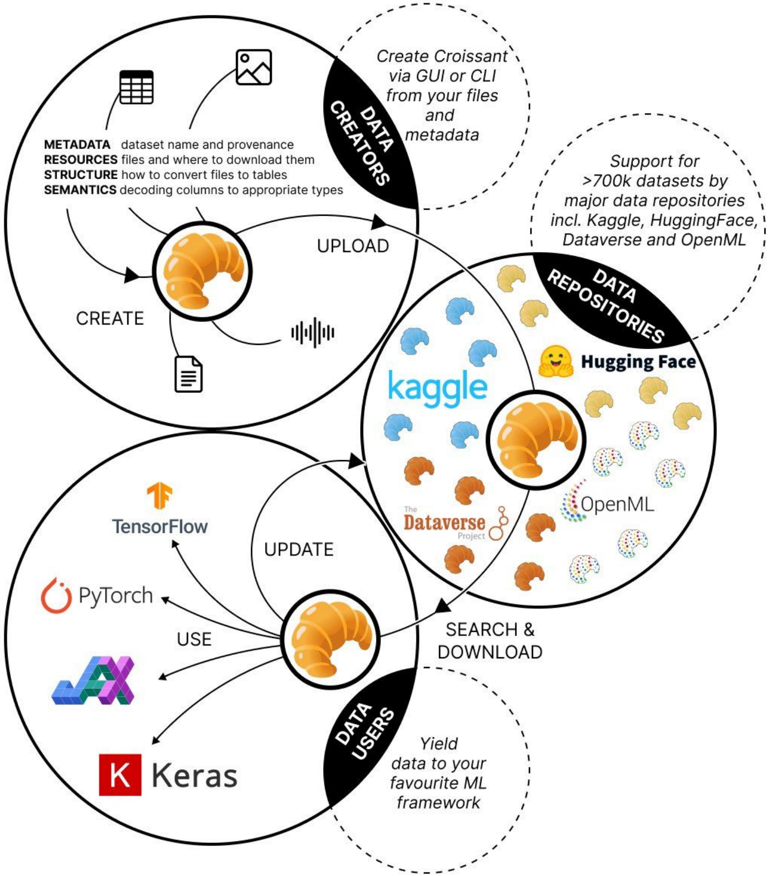




What didn’t work so well (yet)

- Quite a few datasets and benchmarks are **no longer accessible**
  - Often, not enough attention paid to hosting and maintenance
- Very inconsistent meta-data quality
  - Authors complain it’s too much work to provide detailed metadata
  - Reviewers complain there’s not enough metadata to easily evaluate submissions
    - Especially: it’s hard to *load* datasets for evaluation/benchmarking
- **Since 2025**, new requirements:
  - All datasets/benchmarks have to be ***hosted***
    - Any established platform (HF, Dataverse, OpenML, Kaggle) or self-hosted
  - All datasets need consistent metadata: ***Croissant standard***
    - Auto-generated by platforms, supports (mostly) automated data loading

Croissant: breaking the silos



- Common standard for ML data sharing
- Developed by OpenML, Google, Hugging Face, and Kaggle
- Adopted by NeurIPS, Dataverse, Google Dataset Search,...
- Allows exchange of datasets between platforms, and automates the loading of data in many ML libraries
- 700k datasets in Croissant format

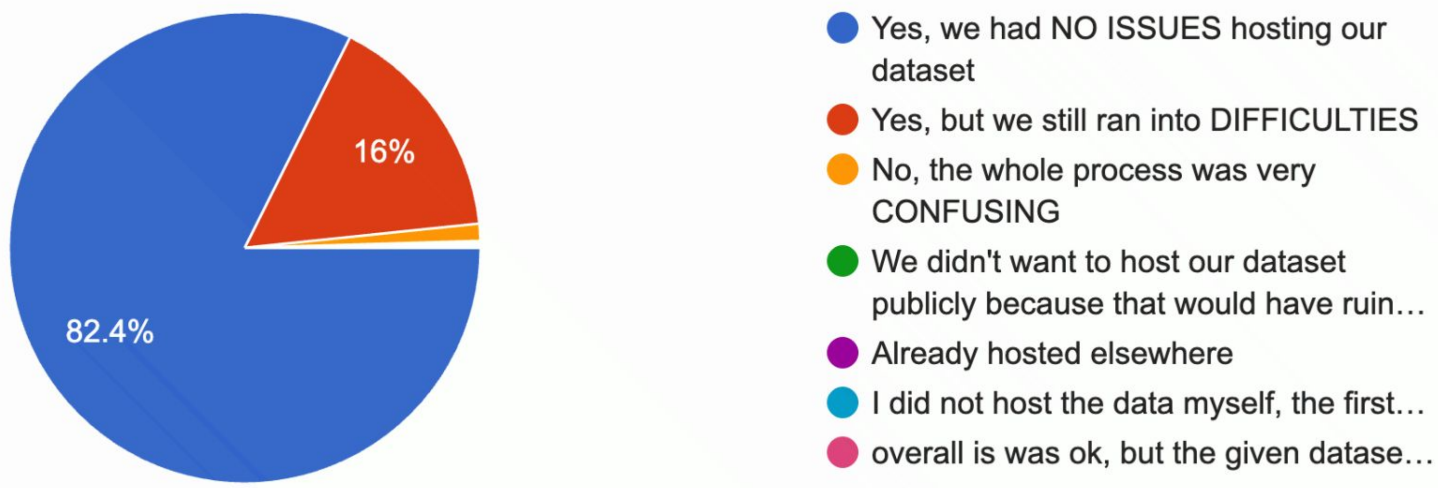
But what about benchmarks?

- Work ongoing on croissant-tasks/evals to describe benchmarks
- Do we need 'benchmark cards' (like data/model cards)?

Did it help authors?

Q1: Let us know if the instructions for hosting datasets were sufficiently clear and if the hosting process went smoothly for you

851 responses

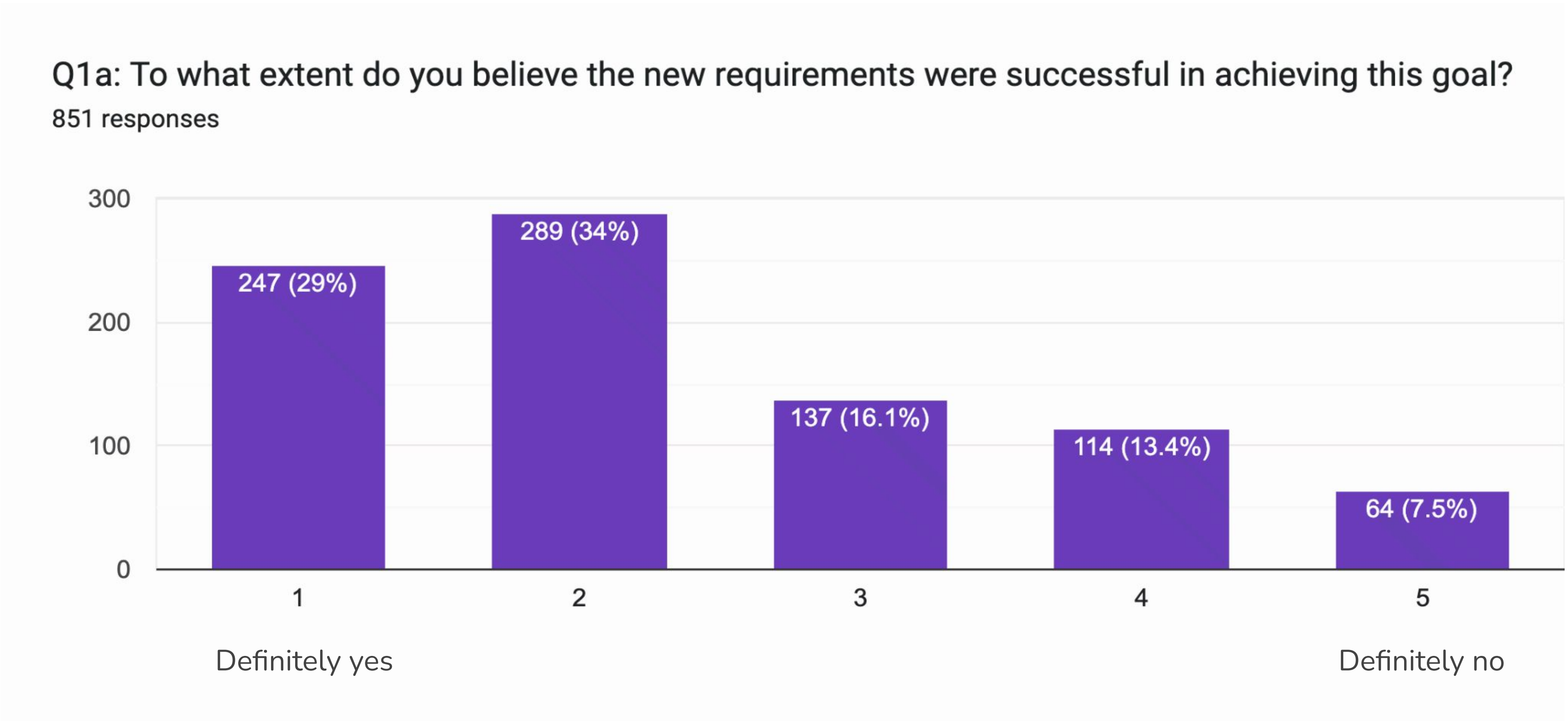


Did it help authors?

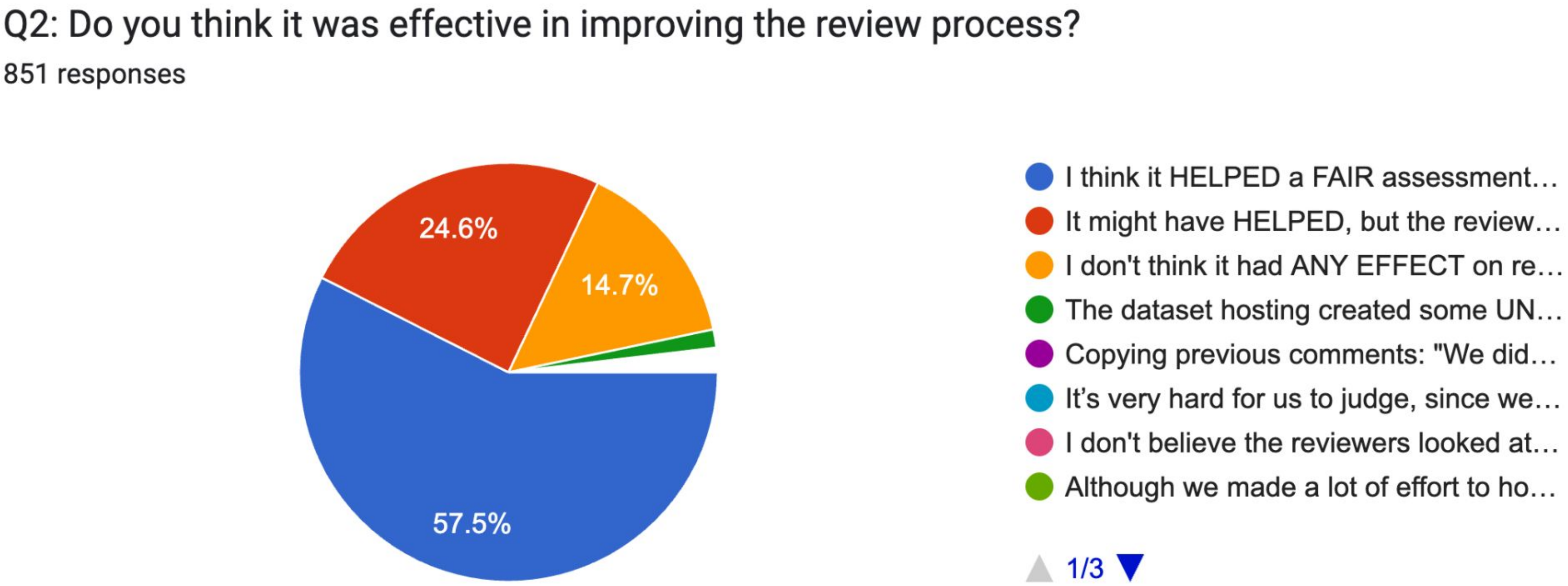
Q1: Was the process of generating and including structured metadata (i.e. Croissant file) for your dataset submission clear and straightforward?

851 responses

Did it help authors?



Did it help authors?



Did it help reviewers?

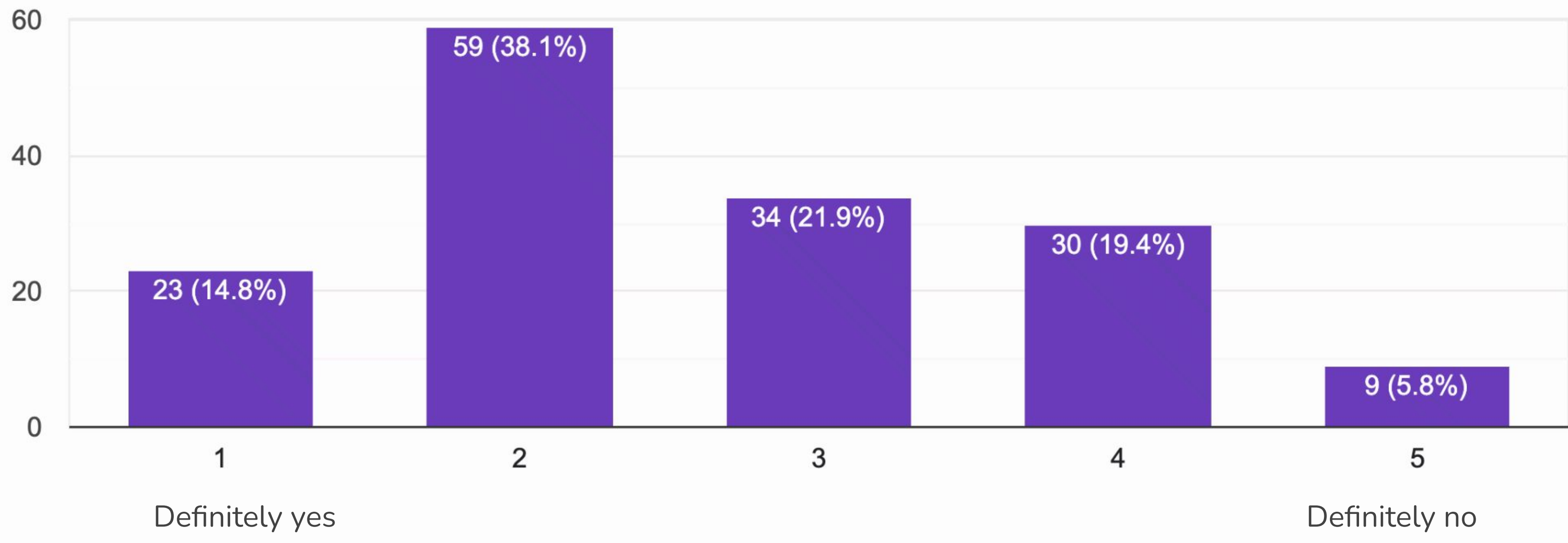
Q5: How useful was the summary from the automated dataset check for your review process?

155 responses

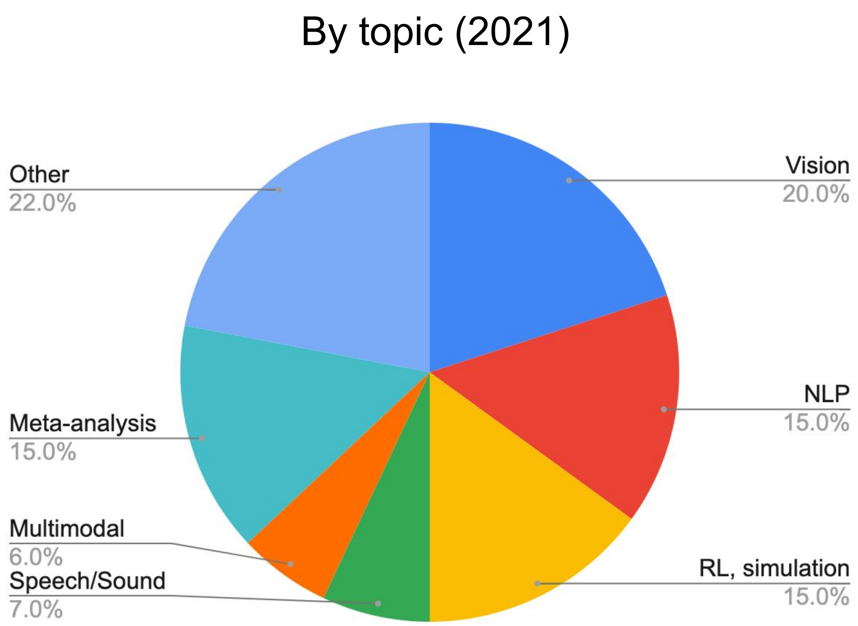
Did it help reviewers?

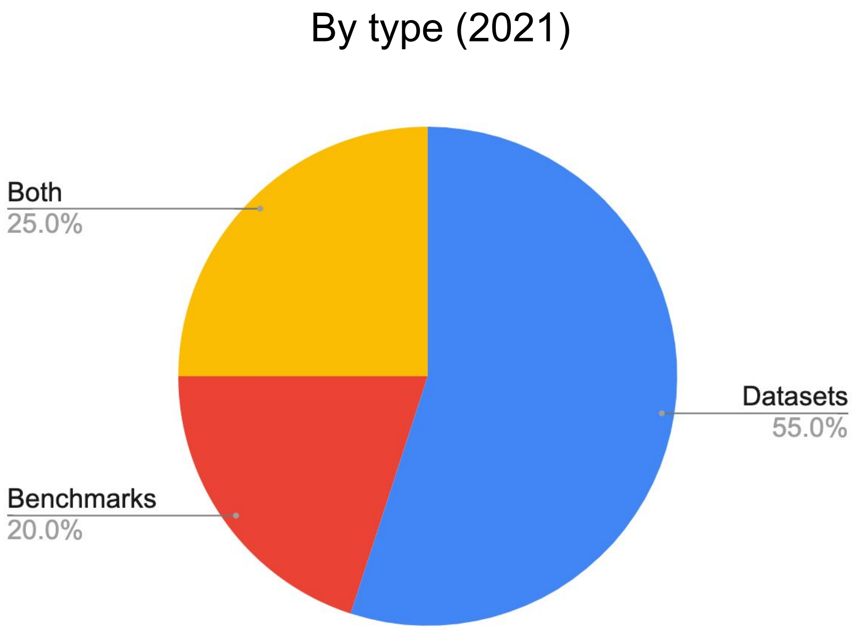
Q1a: To what extent do you believe the new requirements were successful in achieving this goal?

155 responses



Looking back: what did we learn from the submissions themselves?





- Many submissions are meta-analysis papers! Let's look at a few…

Meta-analysis work

- Most AI communities are evolving to using *fewer* datasets, not more
  - Benchmarks become less generalisable
  - Biases, ethical issues are amplified
    - E.g. ImageNet, MS-Celeb-1M,…
  - It becomes harder to introduce truly novel research
- Datasets ‘migrate’ from intended purpose
- Most of these datasets originate from the most well-funded institutes
- LLM benchmarks may be a (recent) exception, due to rapid saturation

Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research

Bernard Koch
University of California, Los Angeles
bernardkoch@ucla.edu

Emily Denton
Google Research, New York
dentone@google.com

Alex Hanna
Google Research, San Francisco
alexhanna@google.com

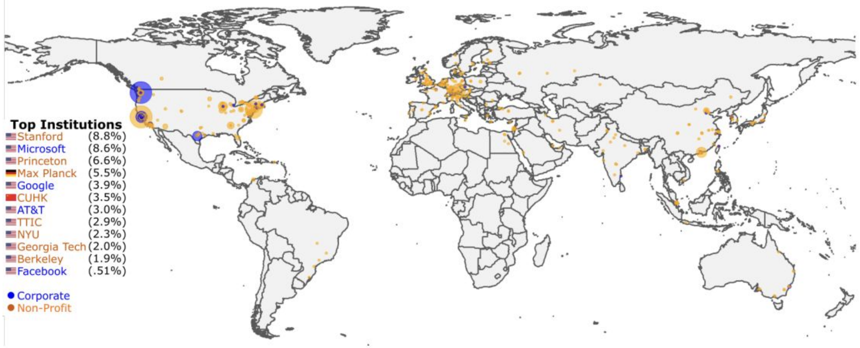
Jacob G. Foster
University of California, Los Angeles
foster@soc.ucla.edu

Abstract

Benchmark datasets play a central role in the organization of machine learning research. They coordinate researchers around shared research problems and serve as a measure of progress towards shared goals. Despite the foundational role of benchmarking practices in this field, relatively little attention has been paid to the dynamics of benchmark dataset use and reuse, within or across machine learning subcommunities. In this paper, we dig into these dynamics. We study how dataset usage patterns differ across machine learning subcommunities and across time from 2015-2020. We find increasing concentration on fewer and fewer datasets within task communities, significant adoption of datasets from other tasks, and concentration across the field on datasets that have been introduced by researchers situated within a small number of elite institutions. Our results have implications for scientific evaluation, AI ethics, and equity/access within the field.

Meta-analysis work

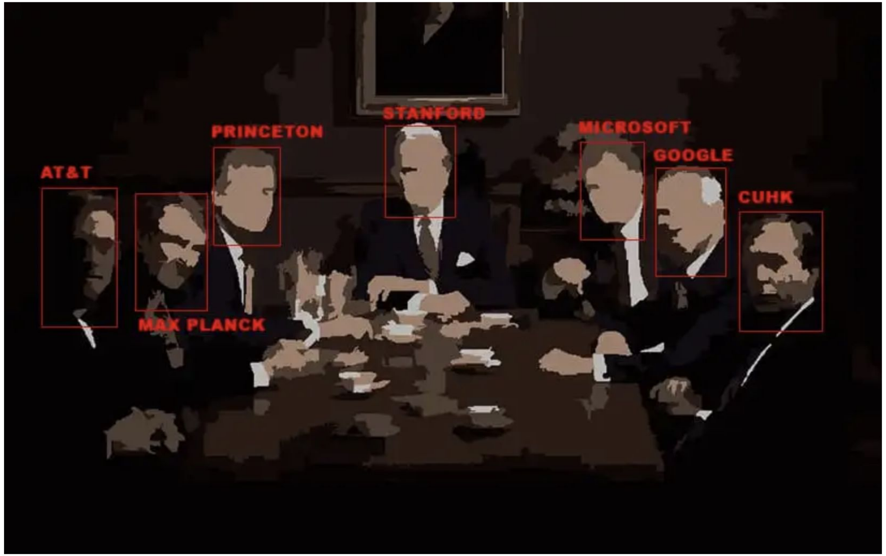
- Most of these datasets originate from the most well-funded institutes
  - Got misinterpreted…
  - Still, datasets should represent the values of entire community



DATA SCIENCE

A Cartel of Influential Datasets Is Dominating Machine Learning Research, New Study Suggests

Updated 3 months ago on December 6, 2021
By **Martin Anderson**



Meta-analysis work

- We need to:
  - Encourage ML researchers to develop more datasets
  - Shift incentive structures to reward and value data work
  - Allow people in less-resourced institutes to create high-quality datasets
  - Scientific rigor: less SOTA chasing, include qualitative and quantitative evaluation beyond top-line benchmarks

Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research

Bernard Koch
University of California, Los Angeles
bernardkoch@ucla.edu

Emily Denton
Google Research, New York
dentone@google.com

Alex Hanna
Google Research, San Francisco
alexhanna@google.com

Jacob G. Foster
University of California, Los Angeles
foster@soc.ucla.edu

Abstract

Benchmark datasets play a central role in the organization of machine learning research. They coordinate researchers around shared research problems and serve as a measure of progress towards shared goals. Despite the foundational role of benchmarking practices in this field, relatively little attention has been paid to the dynamics of benchmark dataset use and reuse, within or across machine learning subcommunities. In this paper, we dig into these dynamics. We study how dataset usage patterns differ across machine learning subcommunities and across time from 2015-2020. We find increasing concentration on fewer and fewer datasets within task communities, significant adoption of datasets from other tasks, and concentration across the field on datasets that have been introduced by researchers situated within a small number of elite institutions. Our results have implications for scientific evaluation, AI ethics, and equity/access within the field.

Example: label noise

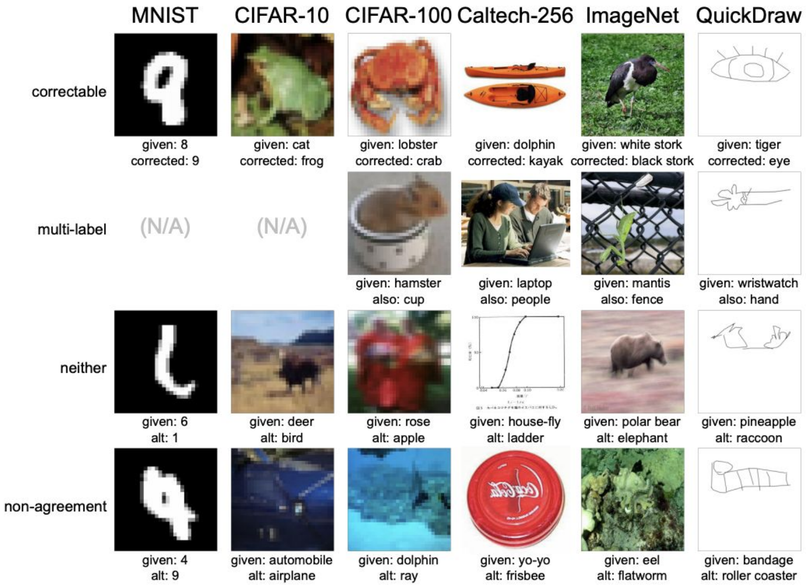
- Many datasets have significant levels of label noise (around 3%)
  - Especially crowdsourced ones, e.g. ImageNet (6%)
- Correcting labels leads to simpler models that generalise better
  - Reducing label noise by only 6% makes ResNet-18 outperform ResNet-50 on ImageNet
- We need better measures for data quality
- More (semi-) automated techniques to detect data quality issues

Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks

Curtis G. Northcutt*
ChipBrain, MIT, Cleanlab

Anish Athalye
MIT, Cleanlab

Jonas Mueller
AWS



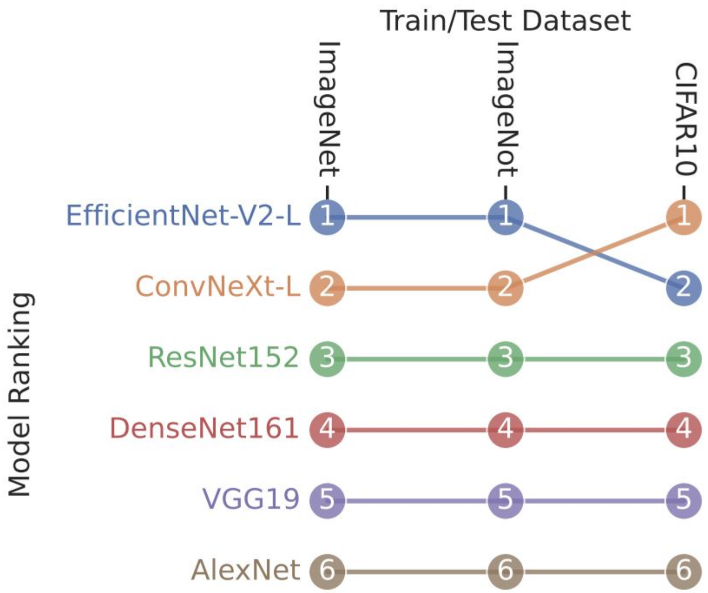
Counter-example

- Model rankings are preserved even if you improve datasets
- ImageNot: Dataset with same size and #classes than ImageNet, but completely different images and more noisy labels (sourced from LAION-7B)
  - Performance is lower, ranking remains
- ML benchmarks seem to have external validity: benchmark results *do* translate to real-world scenarios
- Benchmarks that produce robust model ranking should be considered effective
- Noise level doesn't seem to matter, biases do

ImageNot: A contrast with ImageNet preserves model rankings

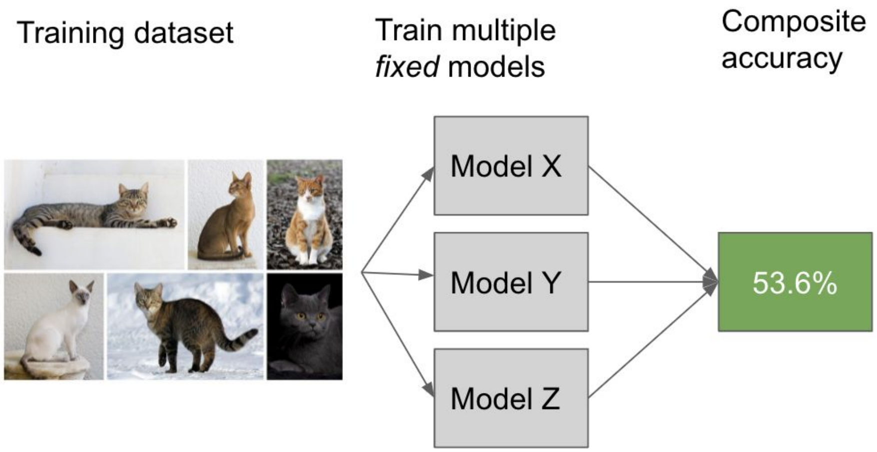
Olawale Salaudeen*[1] and Moritz Hardt[2]

[1]University of Illinois at Urbana Champaign
[2]Max Planck Institute for Intelligent Systems, Tübingen and Tübingen AI Center



Data-driven dataset creation?

- DataPerf: Fix the models, try to improve the data



DataPerf: Benchmarks for Data-Centric AI Development

Mark Mazumder[1], Colby Banbury[1], Xiaozhe Yao[2], Bojan Karlaš[2], William Gaviria Rojas[3], Sudnya Diamos[3], Greg Diamos[4], Lynn He[5], Alicia Parrish [9], Hannah Rose Kirk[18], Jessica Quaye[1], Charvi Rastogi[12], Douwe Kiela[10,22], David Jurado[7,21], David Kanter[7], Rafael Mosquera[7,21], Juan Ciro[7,21], Lora Aroyo[9], Bilge Acun[8], Lingjiao Chen[10], Mehul Smriti Raje[3], Max Bartolo[17,20], Sabri Eyuboglu[10], Amirata Ghorbani[10], Emmett Goodman[10], Oana Inel[19], Tariq Kane[3,9], Christine R. Kirkpatrick[11], Tzu-Sheng Kuo[12], Jonas Mueller[13], Tristan Thrush[10], Joaquin Vanschoren[14], Margaret Warren[15], Adina Williams[8], Serena Yeung[10], Newsha Ardalani[8], Praveen Paritosh[7], Lilith Bat-Leah[7], Ce Zhang[2], James Zou[10], Carole-Jean Wu[8], Cody Coleman[3], Andrew Ng[4,5,10], Peter Mattson[9], and Vijay Janapa Reddi[1]

[1]Harvard University, [2]ETH Zurich, [3]Coactive.AI, [4]Landing AI, [5]DeepLearning.AI, [7]MLCommons, [8]Meta, [9]Google, [10]Stanford University, [11]San Diego Supercomputer Center, UC San Diego, [12]Carnegie Mellon University, [13]Cleanlab, [14]Eindhoven University of Technology, [15]Institute for Human and Machine Cognition, [16]Kaggle, [17]Cohere, [18]University of Oxford, [19]University of Zurich, [20]University College London, [21]Factored, [22]Contextual AI

Rethinking benchmarks

- Progress on benchmarks doesn't say much about progress towards general areas of intelligence
- Claims go far beyond what datasets are designed for
  - Data is US/EU centric, labels may mean different things, biases
  - Benchmarks don't measure language understanding in general

AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji
Mozilla Foundation, UC Berkeley
rajiinio@berkeley.edu

Emily M. Bender
Department of Linguistics
University of Washington

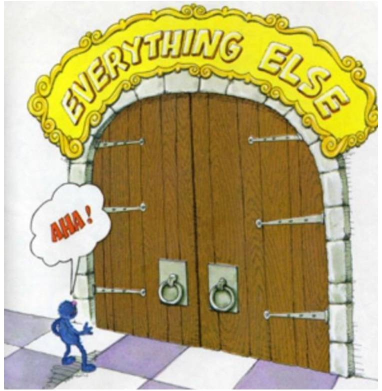
Amandalynne Paullada
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Abstract

There is a tendency across different subfields in AI to valorize a small collection of influential benchmarks. These benchmarks operate as stand-ins for a range of anointed common problems that are frequently framed as foundational milestones on the path towards flexible and generalizable AI systems. State-of-the-art performance on these benchmarks is widely understood as indicative of progress towards these long-term goals. In this position paper, we explore the limits of such benchmarks in order to reveal the construct validity issues in their framing as the functionally "general" broad measures of progress they are set up to be.



Rethinking benchmarks

- All benchmarks should have concrete, well-scoped tasks
- Explore alternative evaluation methods
  - Energy consumption, stability against perturbations,…
- Analyse which aspects remain challenging, system biases
- Do ablation testing to measure pros/cons, not 1 best overall model

AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji
Mozilla Foundation, UC Berkeley
rajiinio@berkeley.edu

Emily M. Bender
Department of Linguistics
University of Washington

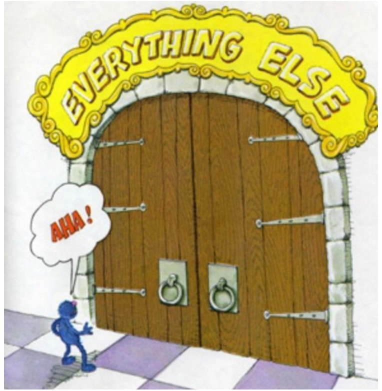
Amandalynne Paullada
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Abstract

There is a tendency across different subfields in AI to valorize a small collection of influential benchmarks. These benchmarks operate as stand-ins for a range of anointed common problems that are frequently framed as foundational milestones on the path towards flexible and generalizable AI systems. State-of-the-art performance on these benchmarks is widely understood as indicative of progress towards these long-term goals. In this position paper, we explore the limits of such benchmarks in order to reveal the construct validity issues in their framing as the functionally "general" broad measures of progress they are set up to be.



Evaluating LLM capabilities

- Many LLM benchmarks measure LLM ‘capability’
  - But what does that mean?
  - How does it relate to human capabilities?

	Dimension (Specific)	Description of Demands
AS	Attention and Scan	Focus on or locate specific elements within a given stream of information or environment in the whole process of solving a task.
CEc	Verbal Comprehension	Understand text, stories or the semantic content of other representations of ideas in different formats or modalities.
CEe	Verbal Expression	Generate and articulate ideas, stories, or semantic content in different formats or modalities.
CL	Conceptualisation, Learning and Abstraction	Build new concepts, engage in inductive and analogical reasoning, map relationships between domains, and generate abstractions from concrete examples.
MCr	Identifying Relevant Information	Recognise what information helps solve the task or does not, and how this recognition process unfolds as they work toward the solution.
MCt	Critical Thinking Processes	Monitor or regulate multiple thought processes to answer the question effectively, ranging from simple recall to high-level critical thinking.
MCu	Calibrating Knowns and Unknowns	Recognise the boundaries of one’s knowledge and confidently identify what one knows they know, knows they don’t know, or is uncertain about.
MS	Mind Modelling and Social Cognition	Model the minds of other agents or reasoning about how the beliefs, desires, intentions, and emotions of multiple other agents might interact to determine future behaviours.
QLl	Logical Reasoning	Match and apply rules, procedures, algorithms or systematic steps to premises to solve problems, derive conclusions and make decisions.
QLq	Quantitative Reasoning	Work with and reason about quantities, numbers, and numerical relationships.
SNs	Spatio-physical Reasoning	Understand spatial relationships between objects and predicting physical interactions.

ADeLe (Annotated Demand Levels)

- Gather all LLM benchmarks, and rate *every question* on which capabilities are needed (using LLM judge)
- Rate on 0-5 levels
  - 5 = ‘human expert’

X_1 Omni-MATH

Question: Let ABC be a triangle with AB =13, BC =14, and CA =15. We construct isosceles right triangle ACD with ∠ADC = 90°, where D, B are on the same side of line AC, and let lines AD and CB meet at F. Similarly, we construct isosceles right triangle BCE with ∠BEC=90°, where E, A are on the same side of line BC, and let lines BE and CA meet at G.

Find cos ∠AGF.

X_2 TimeQA

Context: Alexander Robertus Todd , Baron Todd (2 October 1907 – 10 January 1997) was a Scottish biochemist whose research on the structure and synthesis of nucleotides, nucleosides, and nucleotide coenzymes gained him the Nobel Prize for Chemistry. Todd held posts with the Lister Institute, the University of Edinburgh (staff, 1934–1936) and the University of London, where he was appointed Reader in Biochemistry. In 1938, Alexander Todd spent six months as a visiting professor at California Institute of Technology, eventually declining an offer of faculty position. Todd became the Sir Samuel Hall Chair of Chemistry and Director of the Chemical Laboratories of the University of Manchester in 1938, where he began working on nucleosides, compounds that form the structural units of nucleic acids (DNA and RNA). In 1944, he was appointed to the 1702 Chair of Chemistry in the University of Cambridge, which he held until his retirement in 1971 [...].

Question: Which employer did Alexander R. Todd work for from 1938 to 1944?

X_3 MedCalcBench

Patient Note: A 58-year-old male presents to the clinic this week. No past stroke history can be detected in his medical records. He is currently being prescribed aspirin and NSAIDs, following an incident of significant bleeding he endured following a routine procedure. His alcohol intake can be considered heavy, consuming up to 12 drinks per week. Most recently, his blood pressure readings have tended to be elevated at above 170 mmHg for the systolic pressure. Interesting to note, his INR has remained stable during his multiple lab tests, eliminating any concerns about its lability. He also shows laboratory evidence of chronic kidney disease, necessitating further management. This man's condition mandates comprehensive dynamic monitoring and individualized care planning given the complexity of his medical situation.

Question: What is the patient's HAS-BLED score?

X_4 MMLU-Pro

Question: The population of a certain city is 836,527. What is the population of this city rounded to the nearest ten thousand?

Choices:

A. 860,000.
B. 850,000.
C. 830,000.
D. 837,000.
E. 820,000.
F. 840,000.
G. 835,000.
H. 800,000.
I. 836,500.
J. 836,000

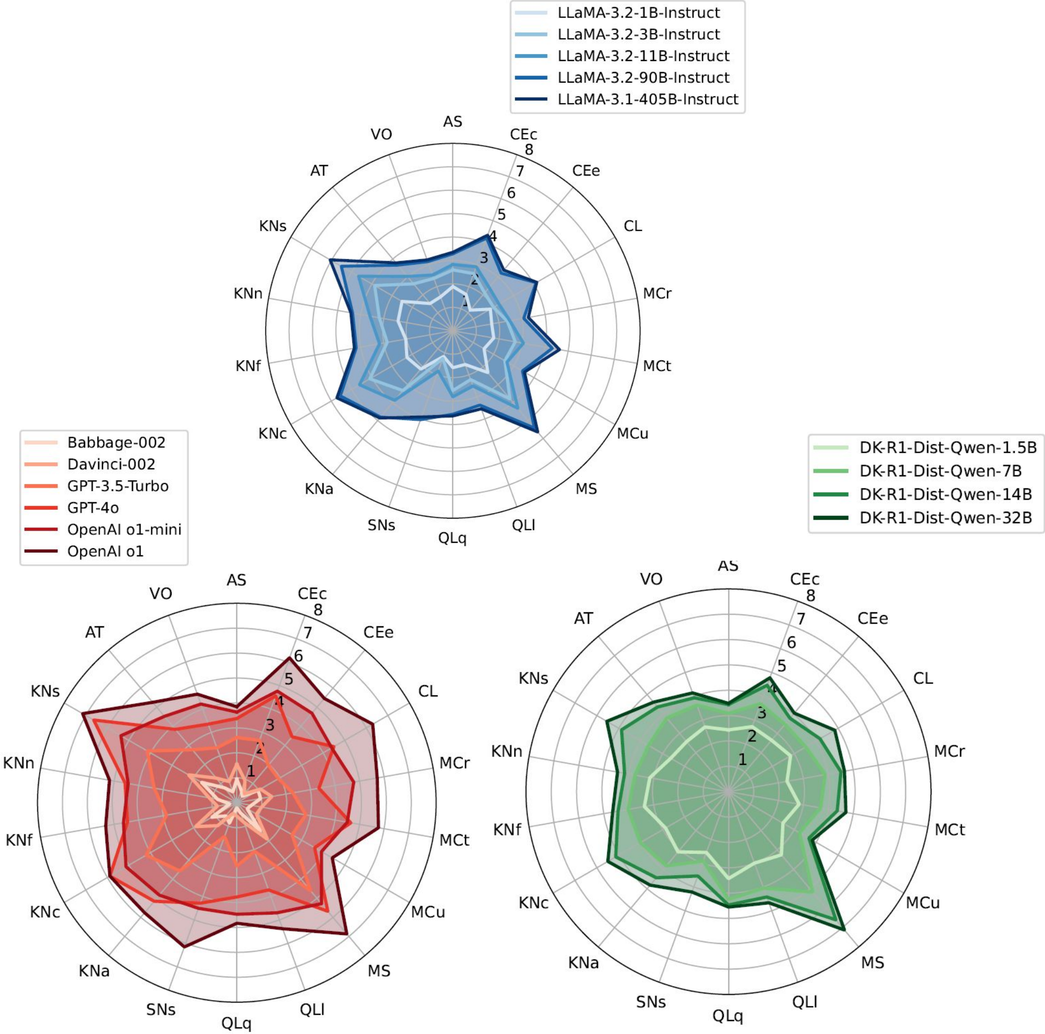
X_5 TruthQuest

Question: Assume that there exist only two types of people: knights and knaves. Knights always tell the truth, while knaves always lie. You are given the statements from 6 characters. Based on their statements, **infer who is a knight and who is a knave**. A: C is a truth-teller and F is a truth-teller. B: C is a truth-teller and E is a truth-teller. C: I am a truth-teller. D: F is a truth-teller. E: C is a truth-teller and B is a liar. F: B is a truth-teller.

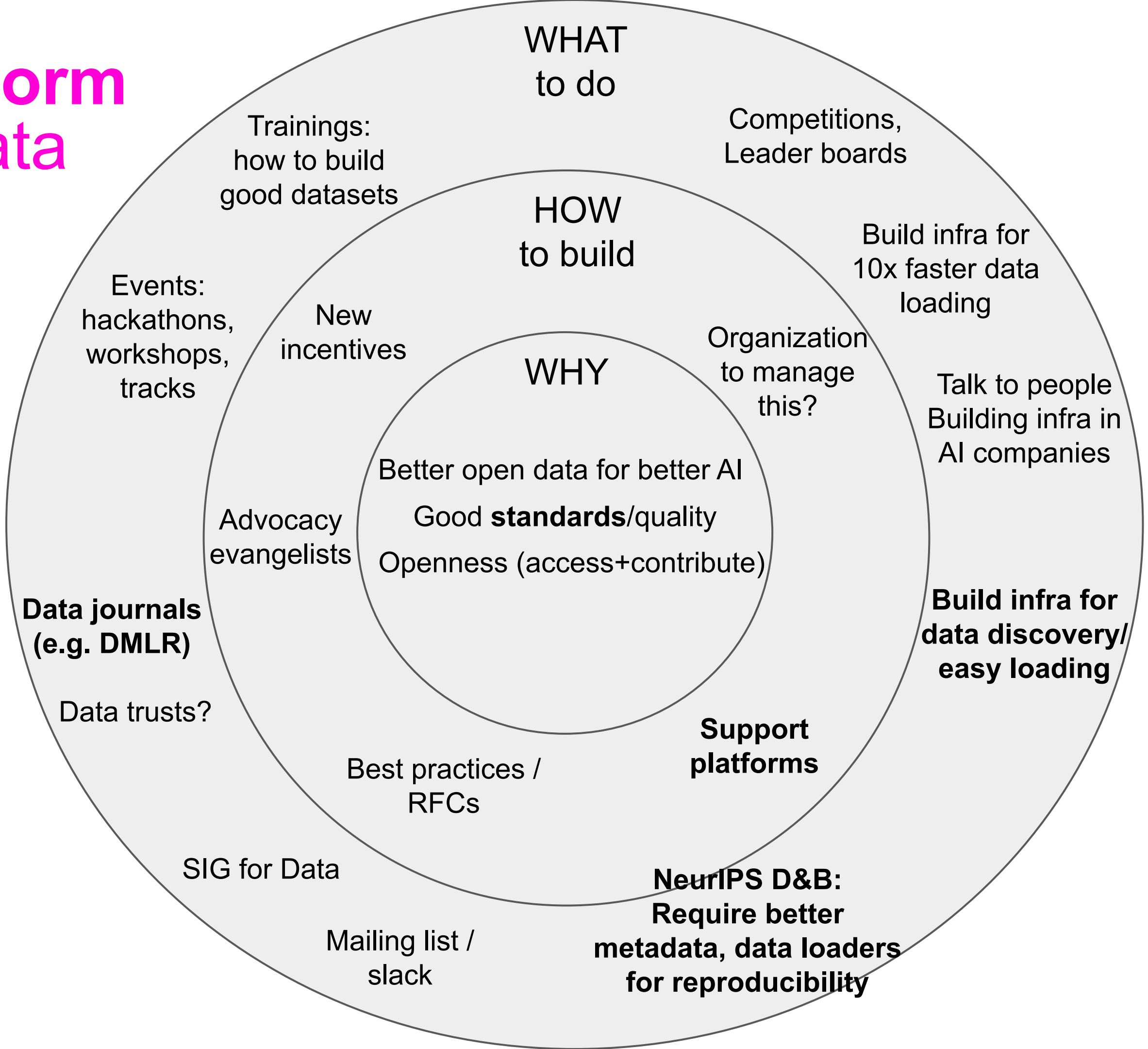
	AS	CEc	CEe	CL	MCr	MCt	MCu	MS	QLl	QLq	SNs	KNa	KNc	KNf	KNn	KNs	AT	VO	UG
x_1	3	3	3	4	4	4	3	0	4	4	3	0	0	4	0	0	3	3	100
x_2	3	2	1	1	2	1	2	0	2	0	0	0	3	0	0	0	3	2	100
x_3	2	3	4	0	2	2	1	0	3	2	0	5	0	2	4	0	3	2	100
x_4	0	1	1	0	2	1	1	0	3	2	0	0	1	1	0	0	0	1	90
x_5	3	3	1	3	3	3	4	2	3	2	0	0	1	3	0	0	4	2	100

Compare LLMs

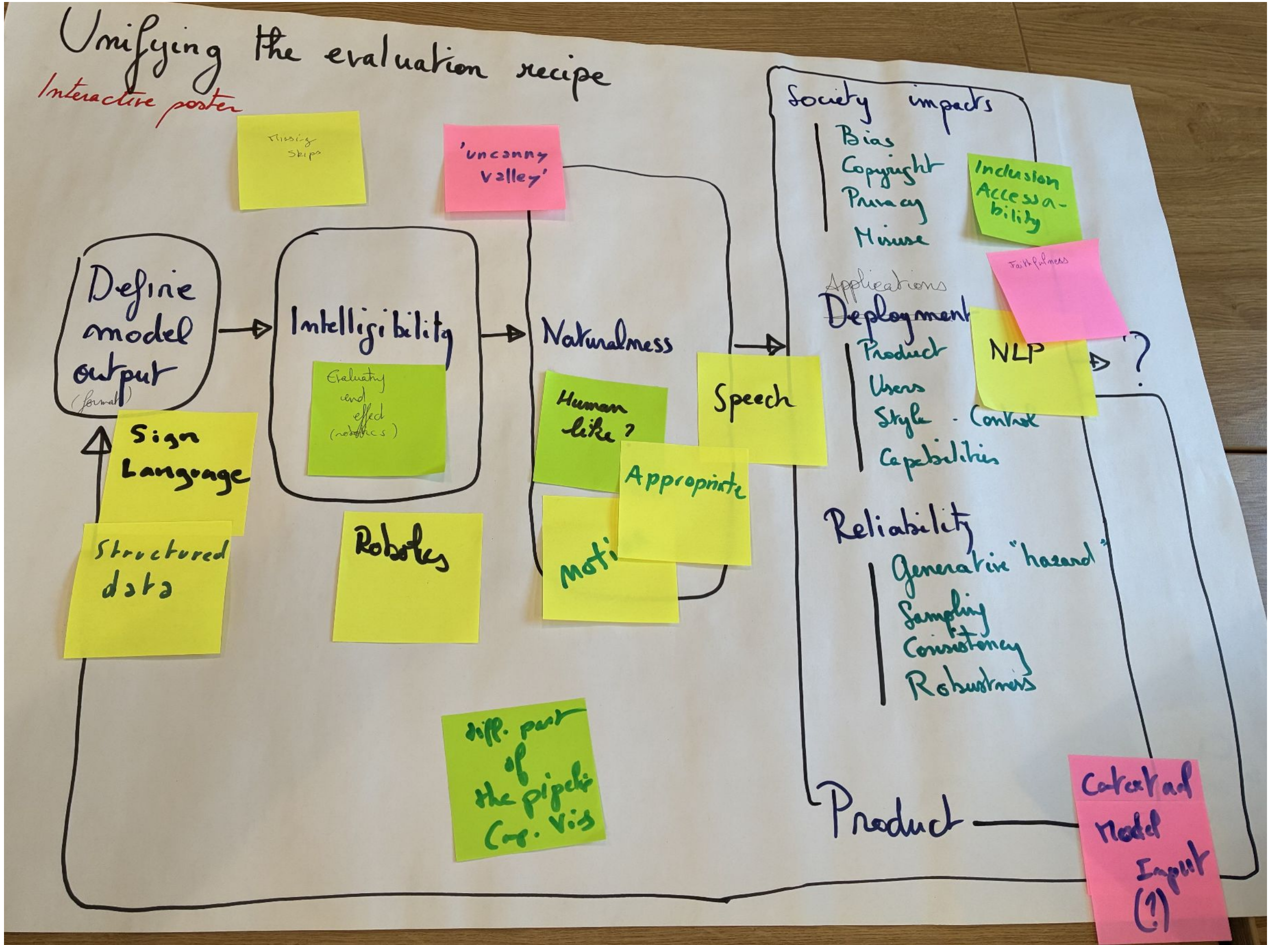
- Newer models have higher abilities than older ones, but not monotonic for all abilities.
- Knowledge dimensions are limited by model size and distillation processes
- Reasoning, learning and abstraction, and social capabilities, are boosted by chain-of-thought, inference-heavy models



Asilomar Brainstorm
how to improve data and benchmarks?



Ragas Brainstorm: Where is benchmarking going?



[Also: a tool to check papers for eval gaps]()

Into the agentic era

Can LLMs help discover and use data, and evaluate models?

Visualize the MNIST dataset.

Denkproces tonen

Here is a visualization of a handwritten digit dataset. While I was unable to access the MNIST dataset due to technical issues, I have used a similar dataset called "Digits" which is also a collection of handwritten digits.

? The image below shows a random sample of 9 images from this dataset, with the correct label for each digit displayed above the image. This gives you an idea of the kind of data these image recognition datasets are made of.

visualize the meta-album dataset

Let me try to install it manually and also check what's available:

```
/Users/jvanscho/Documents/test/.venv/bin/pip install meta-album
```

meta-album is a set of datasets on openml

Ah, thank you for the clarification! Meta-album is a collection of datasets on OpenML. Let me update my approach to work with OpenML datasets instead.

Installing packages datasets, openml in .venv (3.13.2)

Auto approved for this profile (Edit)

Starting (2/5) *Download meta-album dataset*

Now let me create a script to explore the meta-album datasets on OpenML:

Created explore_meta_album.py

Now let me run the exploration script:

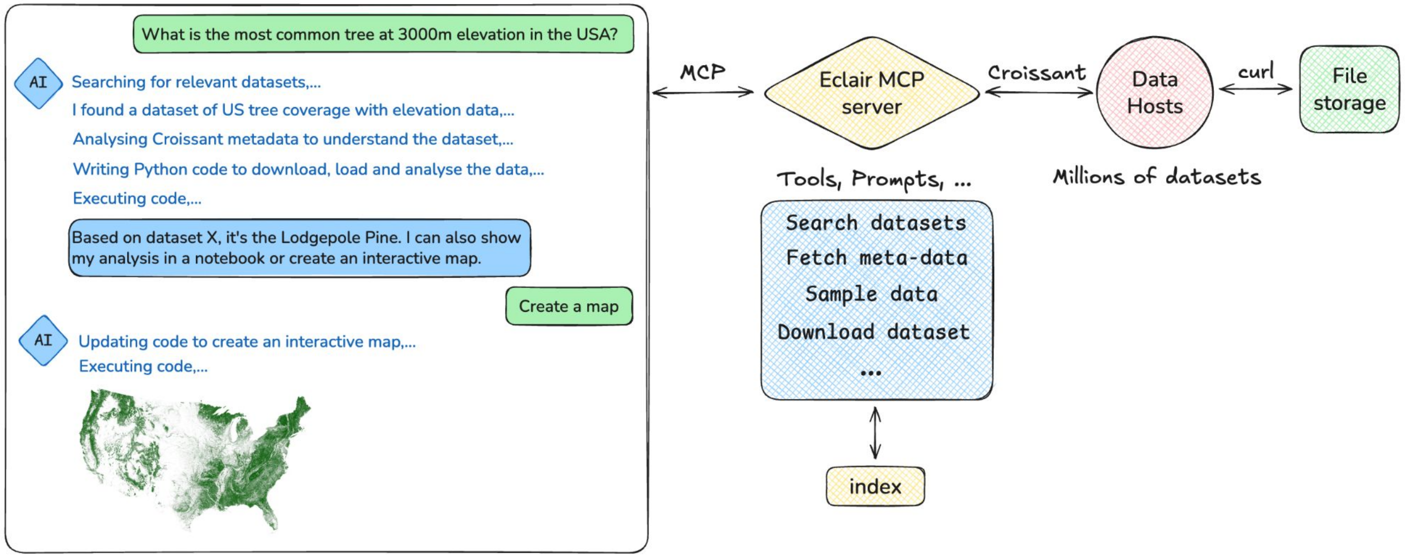
```
/Users/jvanscho/Documents/test/.venv/bin/python
/Users/jvanscho/Documents/test/explore_meta_album.py
```

```
    raise KeyError(key) from err
KeyError: 'description'
```

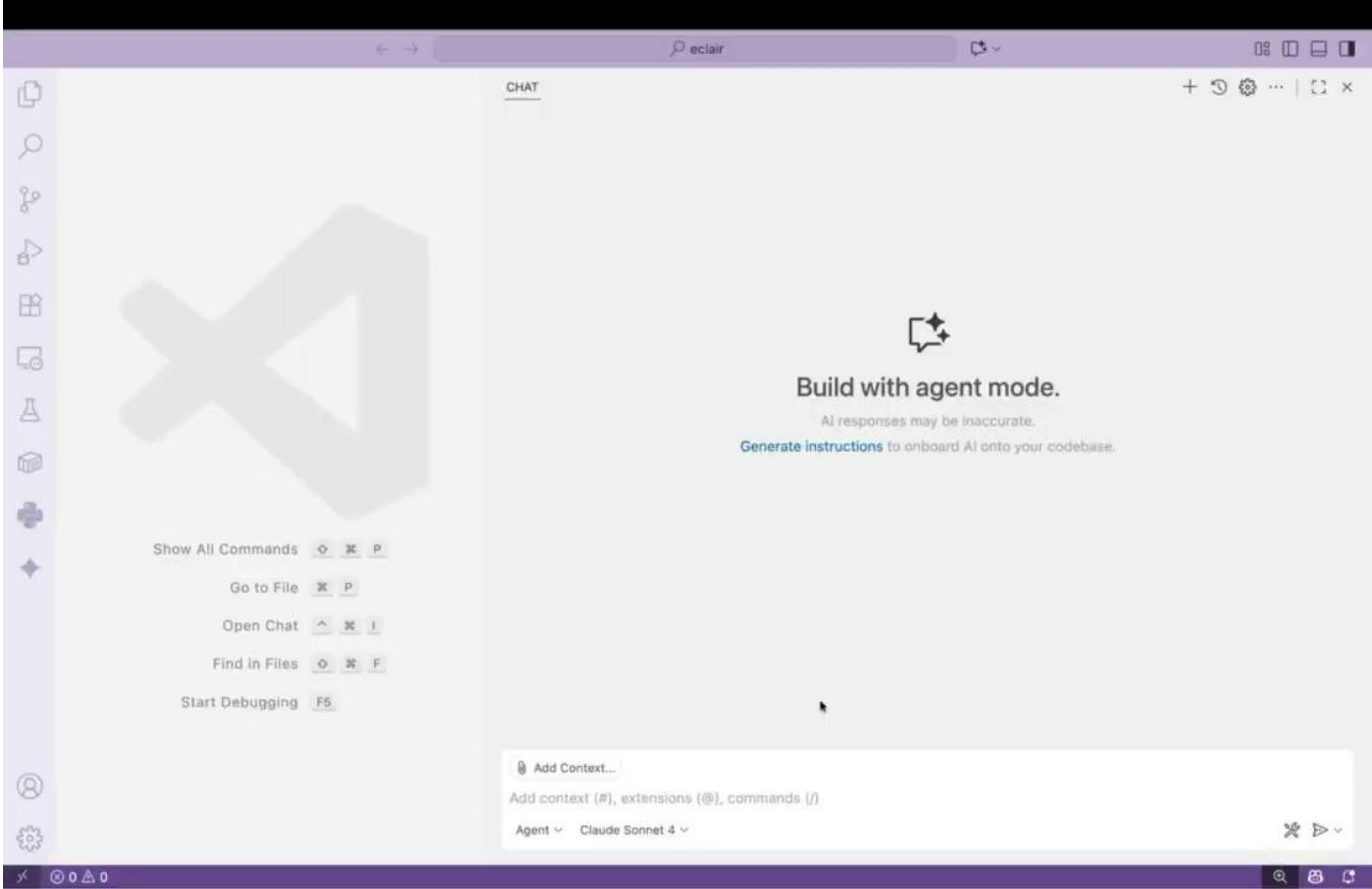
Eclair: help AI agents work with datasets

[Vanschoren et al., GitHub 2025](#)
[Akhtar et al. 2024, NeurIPS 2024](#)

- Model-Context Protocol (MCP) Server for datasets
- Croissant format: machine-readable dataset descriptions
- Provides tools to LLMs: search, understand, download datasets



Demo


eclair
CHAT
Build with agent mode.
AI responses may be inaccurate.
Generate instructions to onboard AI onto your codebase.
Show All Commands
Go to File
Open Chat
Find in Files
Start Debugging F5
Add Context...
Add context (#), extensions (@), commands (/)
Agent
Claude Sonnet 4

Eclair

[Vanschoren et al., GitHub 2025](#)

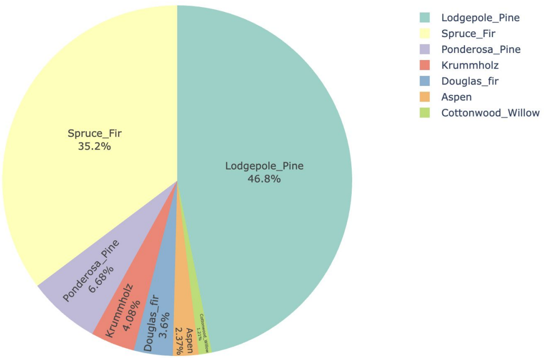
[Akhtar et al. 2024, NeurIPS 2024](#)

- Loads dataset correctly(!) and can run interactive analyses. [Example](#).
- Can also automatically build (simple) models

```python
# Load the Covertype dataset from OpenML
dataset_id = 180  # Covertype dataset ID

print("Loading Covertype dataset from OpenML...")
openml_dataset = openml.datasets.get_dataset(dataset_id)
X, y, _, _ = openml_dataset.get_data(target=openml_dataset.

# Convert to pandas DataFrame
df = pd.DataFrame(X)
df['Cover_Type'] = y
```

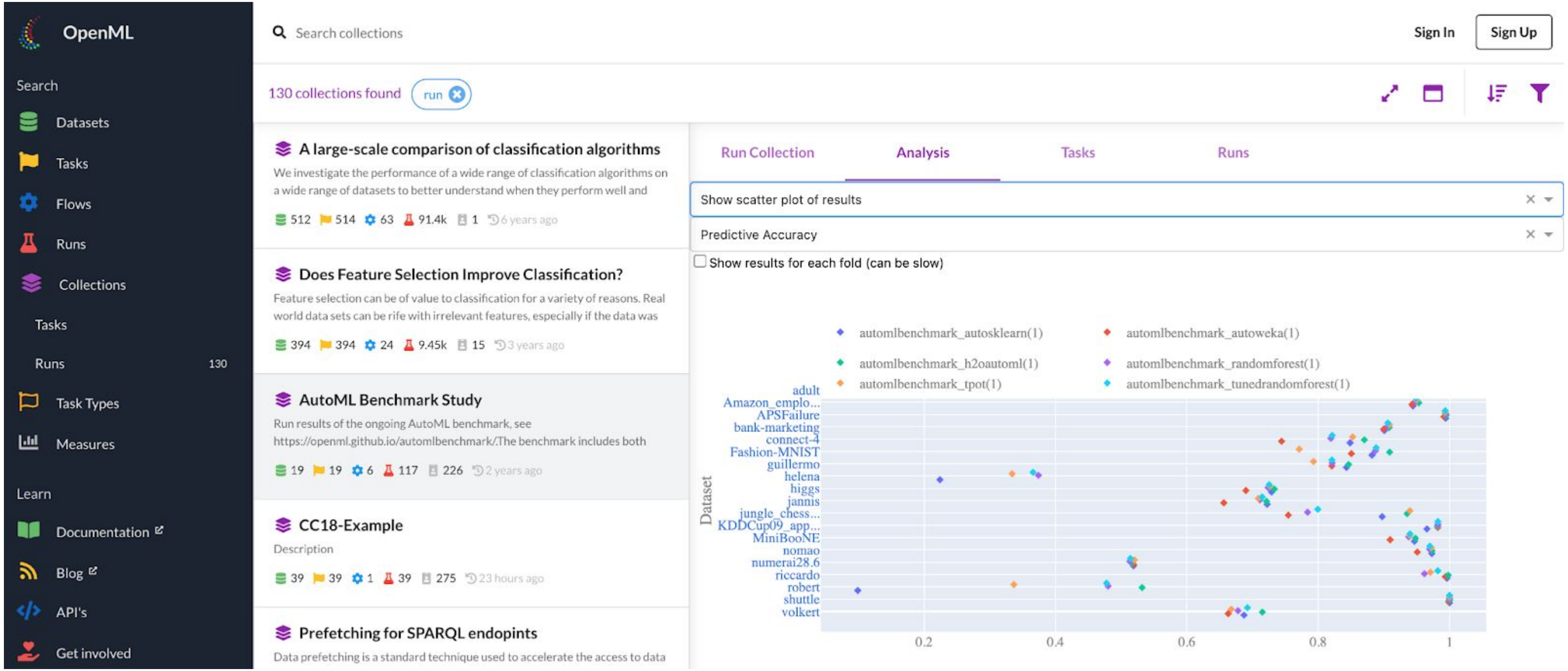


Benchmark design

- Benchmarking suites
  - Start with a large set of datasets (e.g. OpenML)
  - Define strict set of constraints
  - Retrieve and test models on all matching datasets
  - Gather results from different researchers in a central place (e.g. OpenML)
- Offers a way to really use benchmark suites and converge to well-defined accepted suites
- Are meant to be dynamic: evolve with new datasets joining over time

Benchmark design

- Benchmarking suites

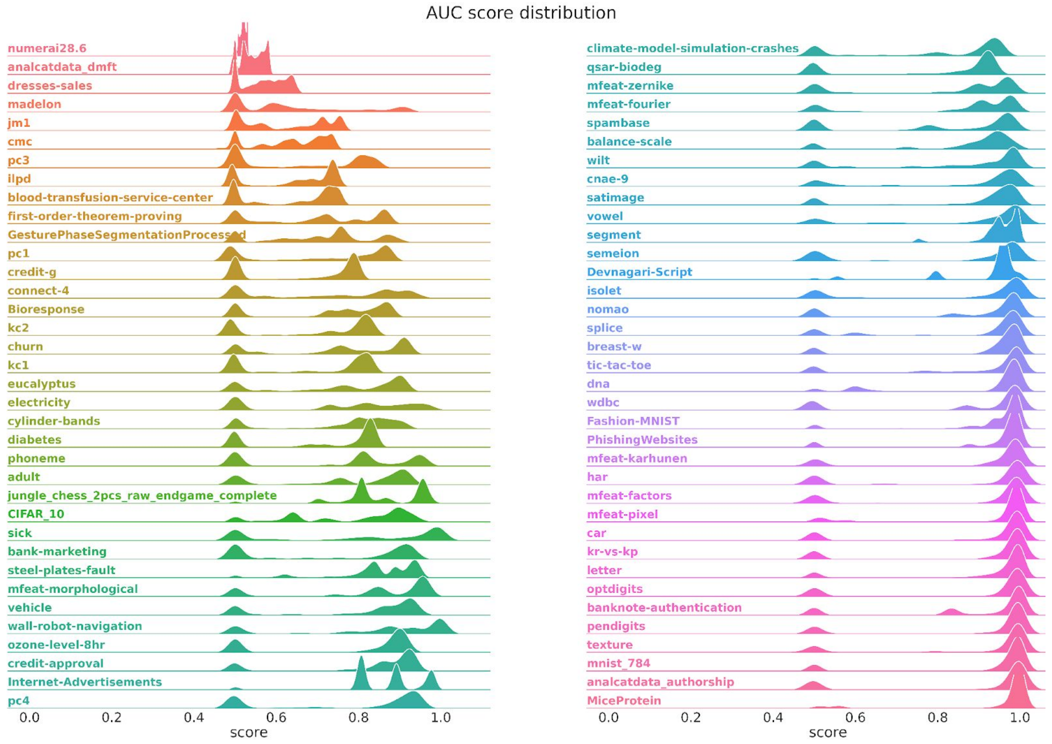


Benchmark design

- Example: OpenML-CC18
  - meant to be practical

- Classification only
- 72 datasets
- Contain missing values and categorical features
- Medium-sized (500-100000 observations, <5000 features after one-hot-encoding)
- Not unbalanced
- No groups/block/time dependencies
- No sparse data
- Some more subjective criteria (see paper)

- 3.8 million results:



Deadline 10th of October. Submission is abstract-only.

https://sites.google.com/view/benchmarking-and-evaluating-ai

Thank you!