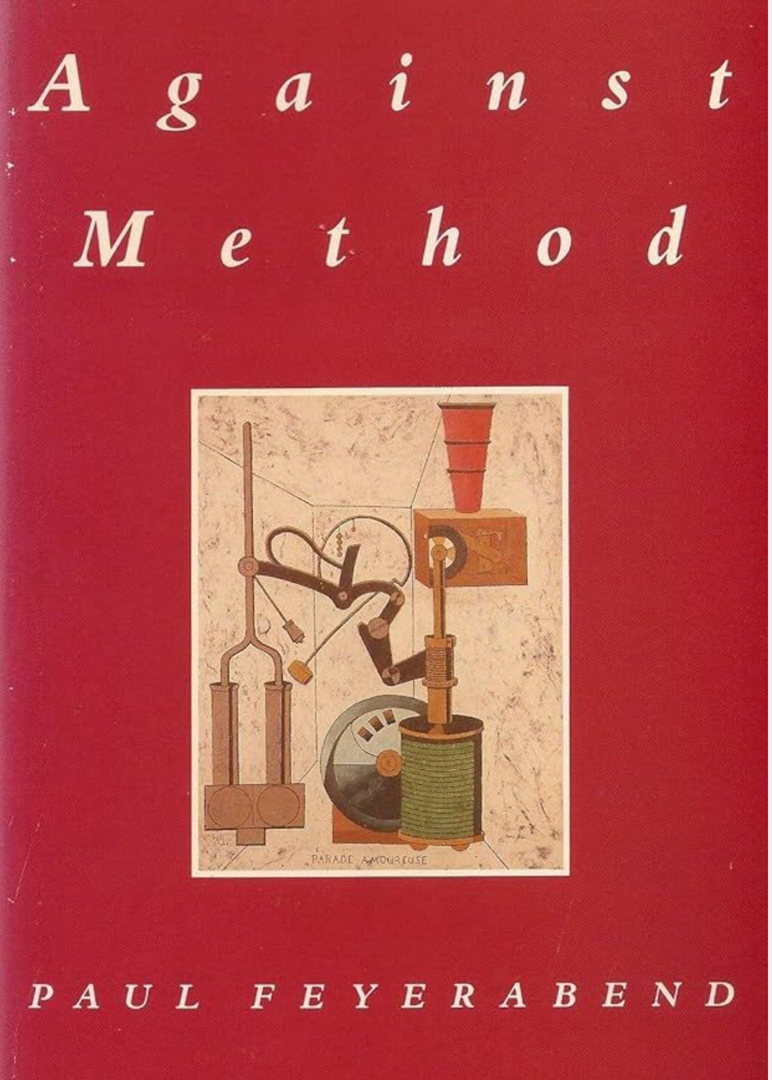
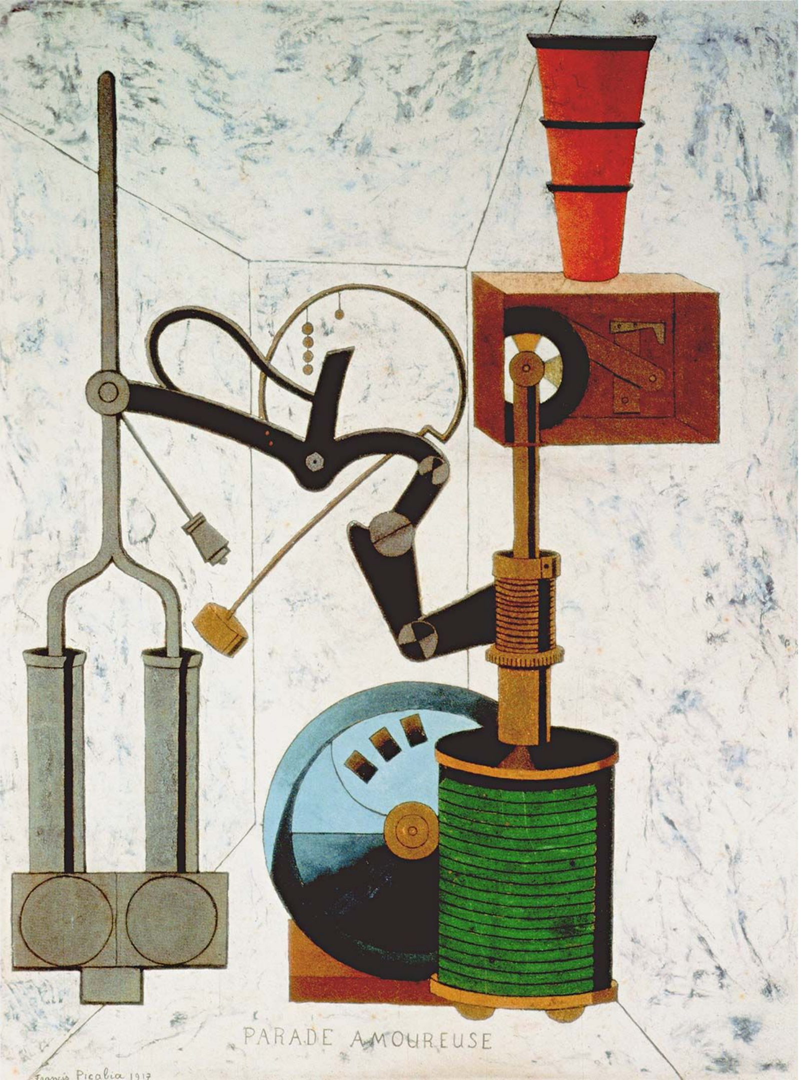
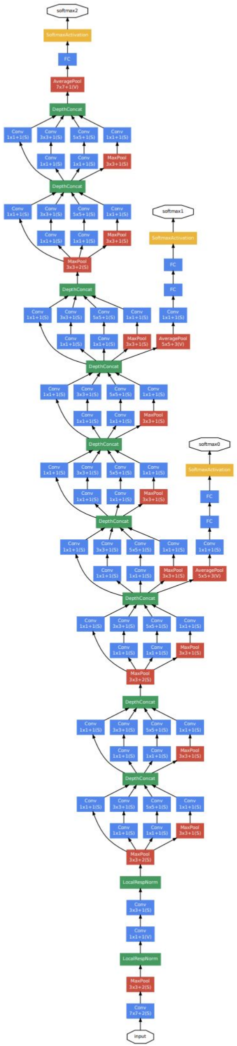
# The emerging science of benchmarks

Moritz Hardt
Max Planck Institute for Intelligent Systems
Tübingen AI Center

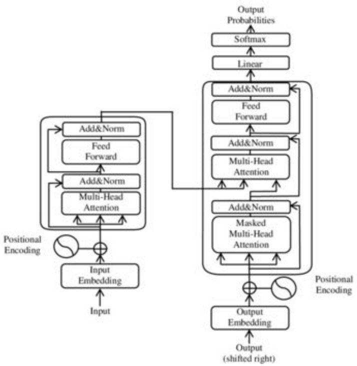


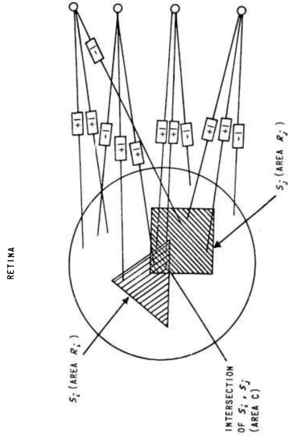

"The only principle that does not inhibit progress is: *anything goes*.

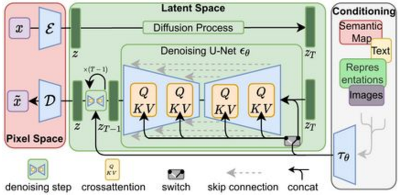


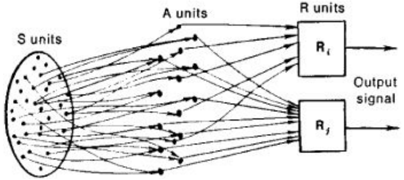



Machine learning has always embraced the *anything goes*.




# Benchmarks: The one rule to tame *anything goes*

The iron rule*: All disputes must be settled by competitive empirical testing.

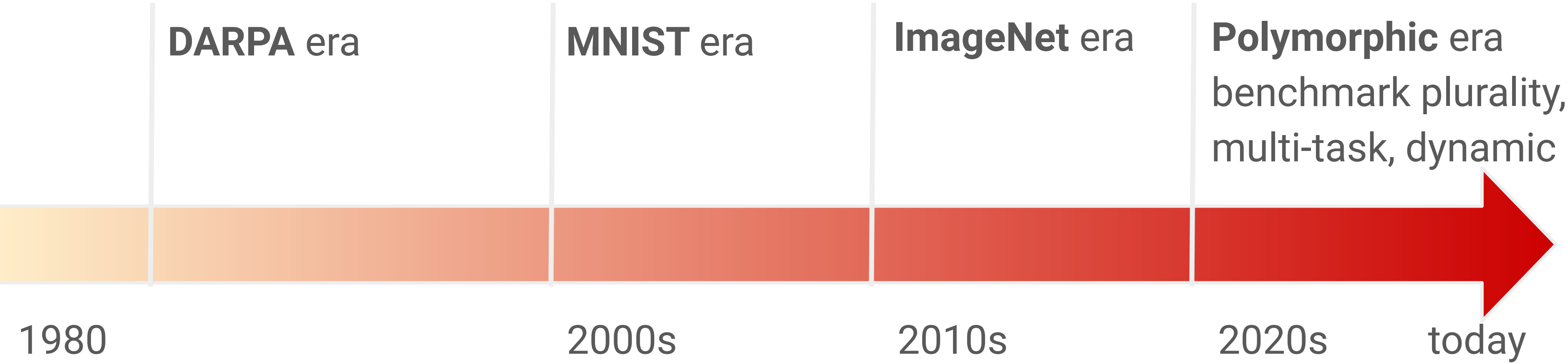
1. Agree on metric
2. Agree on benchmark data
3. Compete

We call this a *benchmark*.

** M. Strevens. Knowledge Machine. 2020.*

# Benchmarks *emerged*

Benchmarks didn't follow any (a priori) theoretical framework




See Liberman's Simons talk (2019), Hardt and Recht (2022) for background

# In this talk

Outline of a science of benchmarks

Scientific takeaways from the ImageNet era

Some challenges in the polymorphic era

Why we need a science of benchmarks

# The beginnings of a science

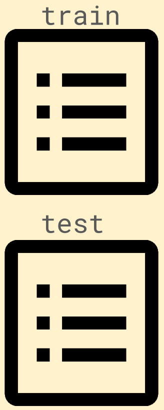
Benchmarks: Just the *holdout method*?

**Fact:** Under vault assumption, test set has *exponential mileage*, i.e., number of testable models is exponential in dataset size *n*.

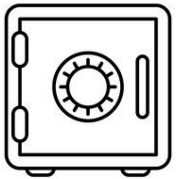
Follows from Hoeffding's inequality + union bound.

<u>Holdout method:</u>

1. Split data
2. Set aside test set
3. Anything goes on train
4. Rank models on test *in the end*



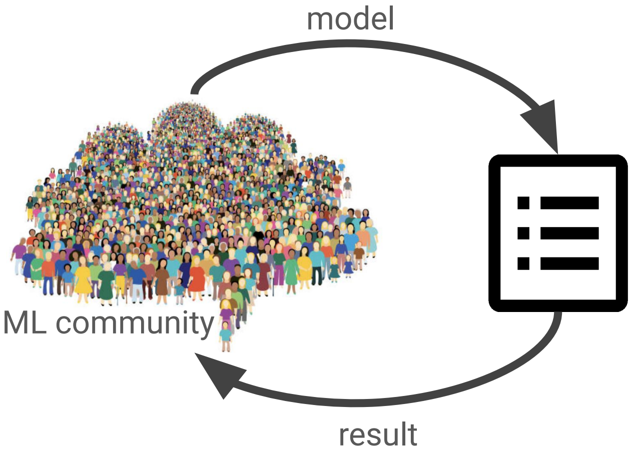

*Ideally, the test set should be kept in a* ***"vault,"*** *and be brought out only at the end of the data analysis.*



– Elements of Statistical Learning. Hastie, Tibshirani, Friedman (2017)

* cais/mmlu
lighteval/mmlu
as of April 25, 2024

# Empirical reality: Test set is *anything but* in a vault!




MMLU* test set 14**K** questions
**5M** downloads on 🤗 *per month*

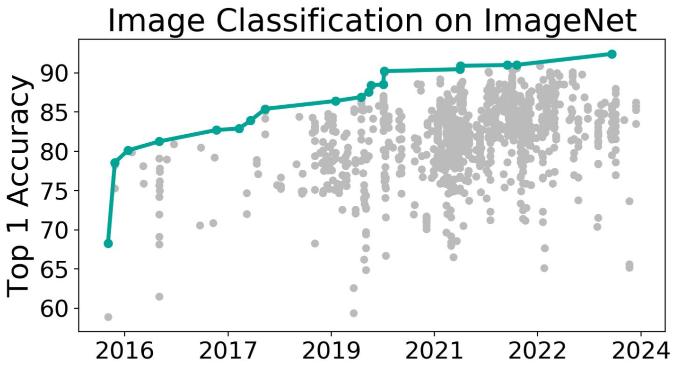
Adaptivity *breaks* all guarantees of the holdout method

*Linear mileage (not exponential)*

Machine learning is ***adaptive*:**
Prior results inform later work, papers, public leaderboards, code

This launched the research area of *adaptive data analysis*.
[Dwork-Feldman-H-Pitassi-Reingold-Roth 2014]

# And, yet: Longevity of benchmarks




Source: [paperswithcode]

ImageNet (ILSVRC2012) supported a decade of active model development

**Question:** Should we trust the model rankings?

Researchers created “fresh” test set:

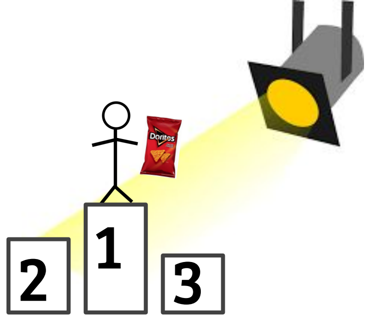
**Model ranking preserved** [Recht-Roelofs-Schmidt-Shankar 2019]

Same for MNIST [Yadav-Bottou 2019], Kaggle [Roelofs et al. 2019], Squad [Miller et al. 2020]

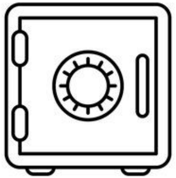
<u>Internal validity of the iron rule:</u>

*Beating the previous best* replicates in similar conditions

# The regularizing force of *competition*




≈



**Iron rule assumption:**
Researchers only care if they beat the previous best.

**Informal Theorem** [Blum-H 2015]:
Assuming iron rule, benchmark data has exponential *mileage*

In other words, *iron rule* (nearly) as good as *iron vault*

**Prescriptive use:**
Implement iron rule as limited feedback mechanism in a benchmark

**Descriptive use:**
Think of iron rule as a postulate about community

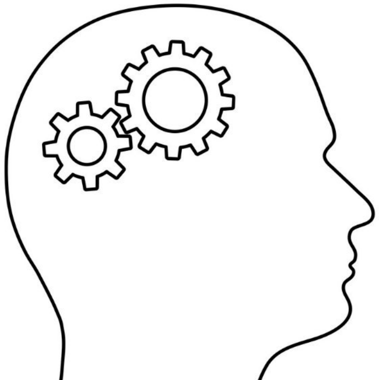
# The sociotechnical forces behind benchmark longevity

Competition [Blum-H 2015]

Collaboration [Mania et al. 2019]

Cognitive and behavioral biases [Zrnic-H 2019]

Dataset artifacts [Feldman-Frostig-H 2019]



All of these promote *internal validity*

See COLT 2019 keynote [“The sociotechnical forces against overfitting”]

So, we know model rankings replicate under *similar* test conditions

**Question:** Do model rankings replicate on radically different test conditions?

## The ImageNot experiment [Salaudeen-H 2024]

**ImageNot:** An *anti-replication* of ImageNet (ILSVRC 2012)

Same scale and diversity, different in every other regard

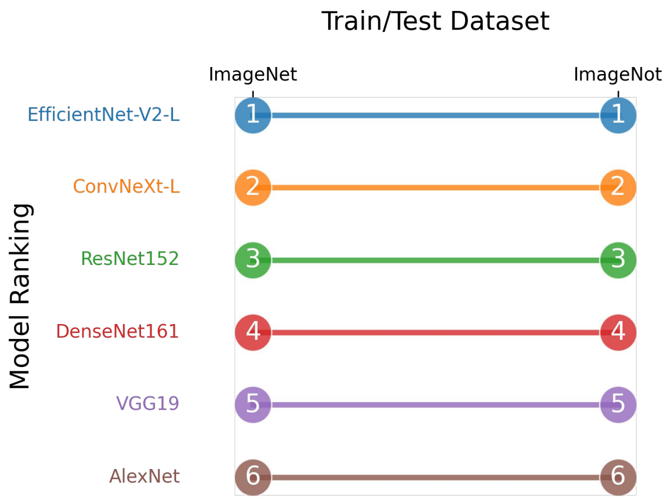
ImageNet: carefully curated by humans, multiple annotators per image

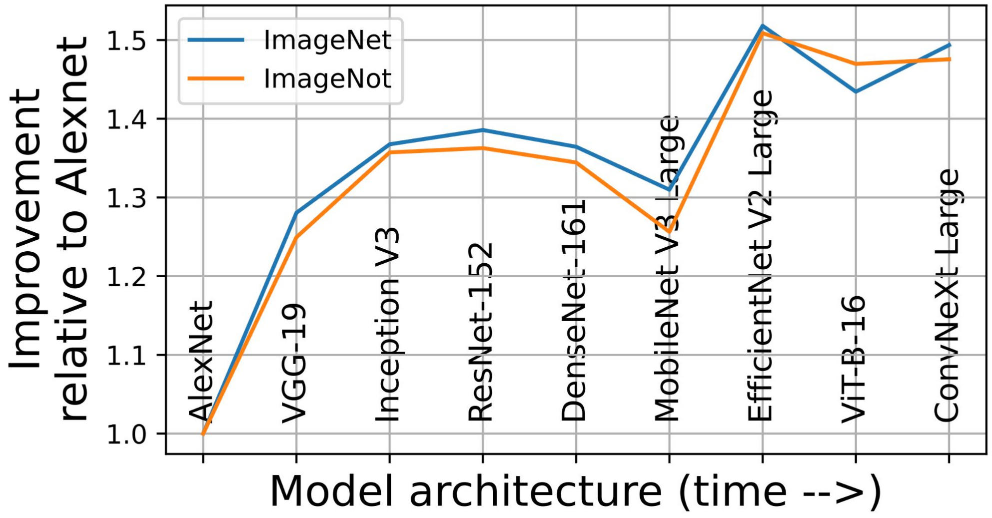
ImageNot: Quick and dirty web crawl selected based on captions

**Experiment:** Retrain key ImageNet era models *from scratch* on ImageNot

*Are the model rankings preserved?*

# ImageNet vs ImageNot

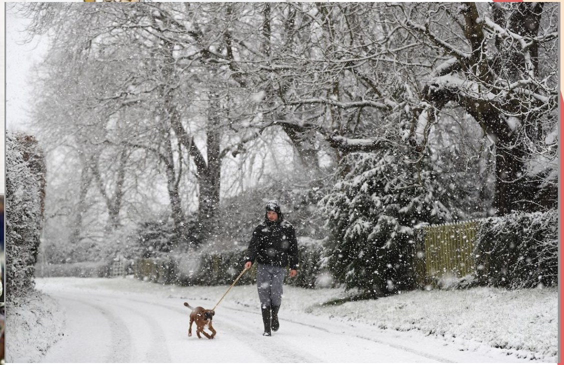


# ImageNot

## Cleats

## Batter




# ImageNet

## Irish terrier

## Blenheim Spaniel

# Let's play the Torralba Efros game

Given an image, name the dataset!

ImageNet vs ImageNot

From: Unbiased Look at Dataset Bias (2011)


| | | | |
|---|---|---|---|
| Caltech101 | Tiny | LabelMe | 15 Scenes |
| MSRC | Corel | COIL−100 | Caltech256 |
| UIUC | PASCAL 07 | ImageNet | SUN09 |

Figure 1. Name That Dataset: Given three images from twelve popular object recognition datasets, can you match the images with the dataset? (answer key below)


ImageNot

ImageNet


ImageNet


DUMPSTER
RULES
• All Trash Must Fit INSIDE of Dumpster with Lid Closed.
• No Furniture.
• No Appliances.
• No Carpet.
• No Construction Materials.
• 24 Hour Video Surveillance.
NO PARKING
in Front of Dumpster

ImageNot

**In fact, trained model gets > 96% accuracy**

And yet, model rankings and relative improvements are the same!

# What we can learn from ImageNot

*External* validity of the iron rule:

If you beat the previous best under sufficiently general conditions,
it will likely replicate elsewhere

Evidence that ImageNet could've been anything of similar scale and diversity

We don't even need clean labels for model ranking!

Let's dive deeper into this claim...

# Benchmarking with noisy labels [Dorner-H 2024]

**Problem:** Given two binary classifiers $f, g$. Which one has higher accuracy?

**Data:** Can draw unlabeled data point $x$ for free, and get label $y$ for **€*1***.

But, label $y$ incorrect with probability $p < ½$.

**Question:** How do we best spend our label budget $n$?

**Common practice:** Sample $n/k$ points, for each $x$ request $k$ labels $y_1, y_2, \ldots, y_k$.
Clean label by taking $y = \texttt{Majority}(y_1, y_2, \ldots, y_k)$.

**Theorem:** It's best to sample $n$ data points with *one* noisy label each.

"All the single labels"


Exit ImageNet


Enter polymorphic era

# The polymorphic era

**Four major challenges:**

*Models have seen the internet*: What if they trained on benchmark related data?

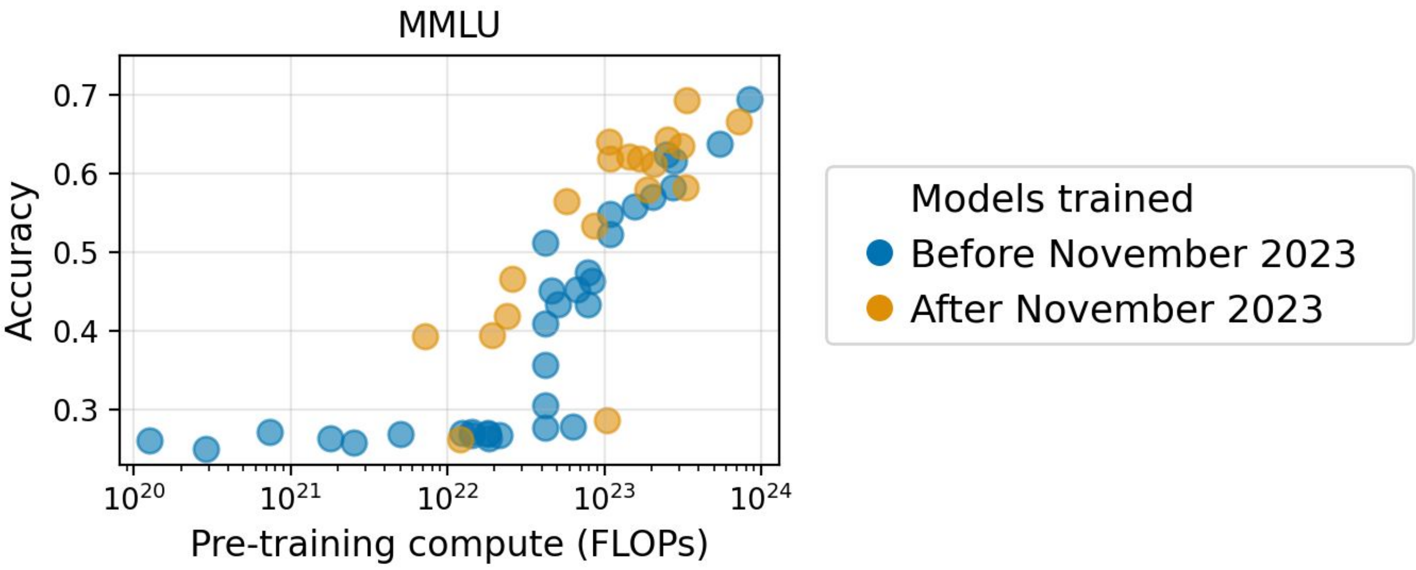
*Models have many capabilities*: Can multi-task benchmarks reliably test them?

*Models evolve rapidly*: Can dynamic benchmarks keep up?

*Models may exceed human expertise*: Can we use models for self evaluation?

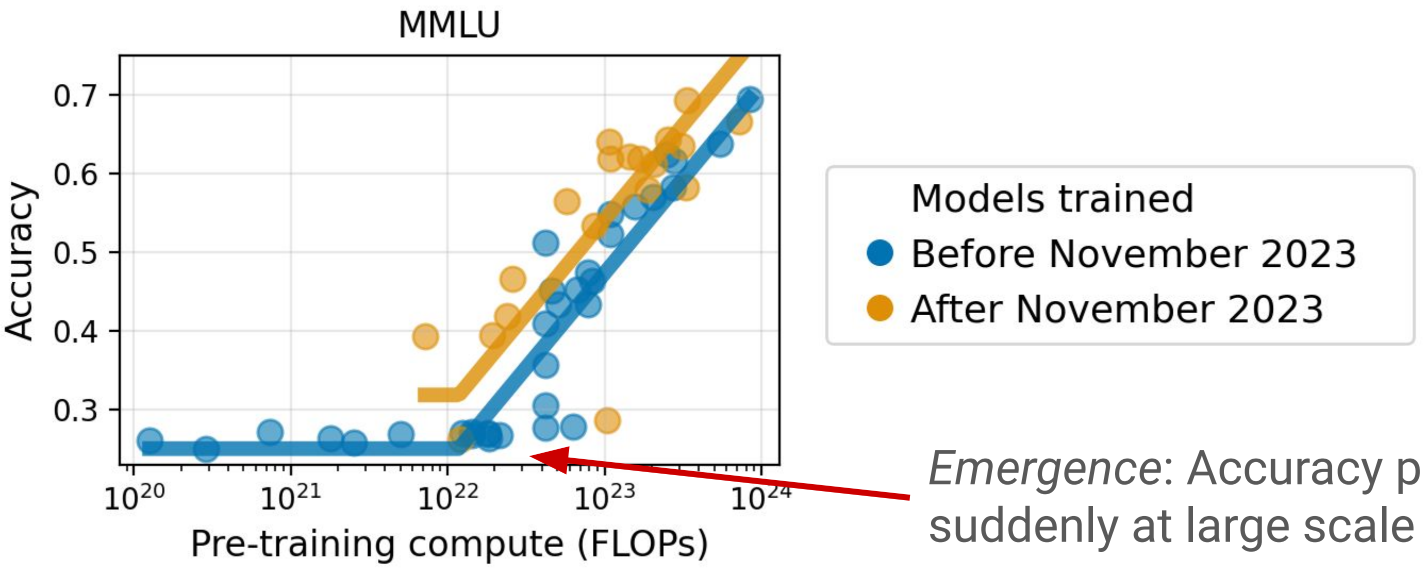
# An empirical puzzle about model evaluation

Newer models appear to better leverage pre-training compute on the MMLU math question answering benchmark.



# An empirical puzzle about model evaluation

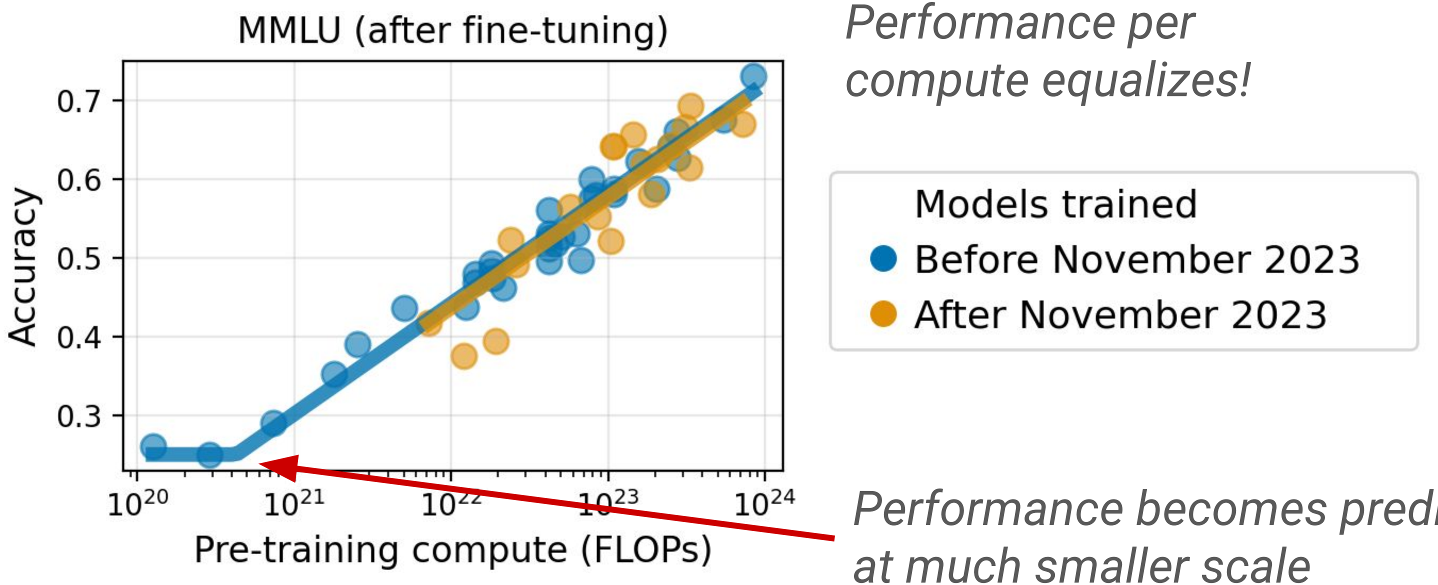
For the same compute, newer models outperform older models by **6.8%** on MMLU




*Emergence*: Accuracy picks up suddenly at large scale

# An observation [Dominguez-Olmedo, Dorner, *H 2024*]

After fine-tuning each model on multiple choice questions similar to MMLU




*Performance per compute equalizes!*

*Performance becomes predictable at much smaller scale*

# Resolving the puzzle [Dominguez-Olmedo, Dorner, *H 2024*]

A small amount of task data can have a large effect on benchmark results.

Newer models models trained more on task relevant data

Include instruction data in pre-training (Qwen, StableLM 2, Olmo, …)

Select pre-training data based on benchmark results (Gemma, Llama 3, …)

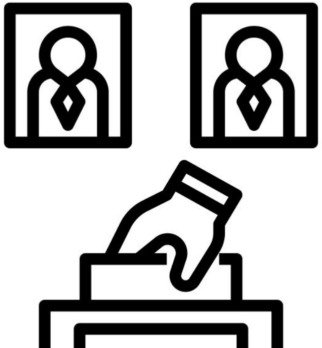
We call this **training on the test task**

**Training on the test task confounds evaluation and emergence**

So, how can we compare models fairly?

*Fight fire with fire:* Give all models the same task specific fine-tuning data

# Multi-task benchmarks and social choice [Zhang-H 2024]

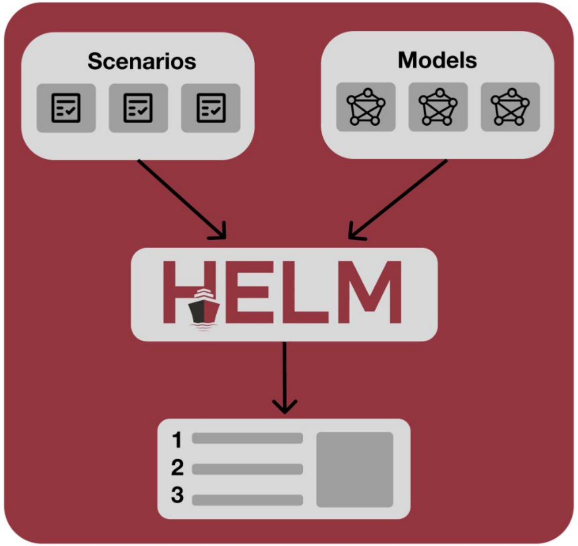


Multi-task benchmarks promise to evaluate models holistically across many tasks

Tasks: Voters

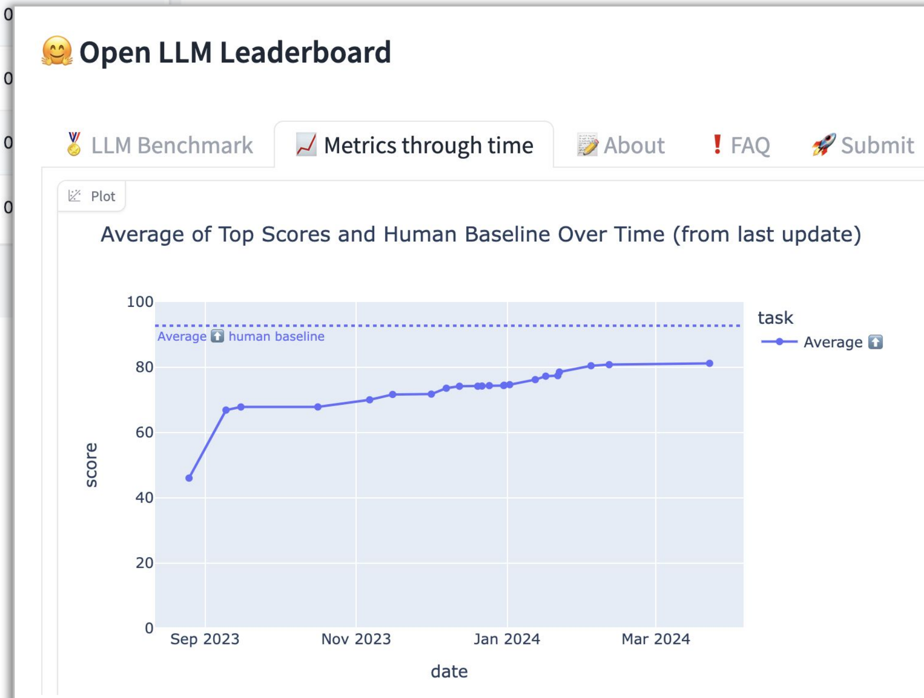
Models: Candidates

Benchmark:
Voting rule aggregating many rankings into one




| Model | Mean win rate |
|---|---|
| GPT-4 (0613) | 0.958 |
| Llama 3 (70B) | 0.903 |
| Mixtral (8x22B) | 0 |
| GPT-4 Turbo (1106 preview) | 0 |
| Palmyra X V3 (72B) | 0 |
| PaLM-2 (Unicorn) | 0 |

Rank by average accuracy across all tasks. Hence, *cardinal*.




Rank models by win rate

Can be computed from individual task *rankings*.

Hence, *ordinal*.

Based on [Eleuther evaluation harness](#)
[Gao-Tow-Abbasi-Biderman 2023]

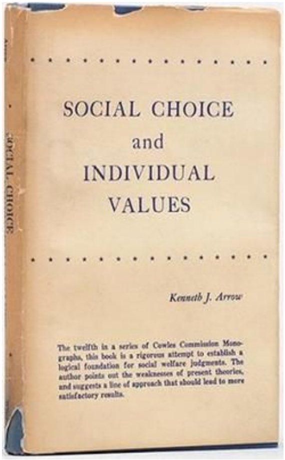
# A stumbling block

Inspired by Arrow's impossibility theorem, we introduce two key properties of a multi-task benchmark:

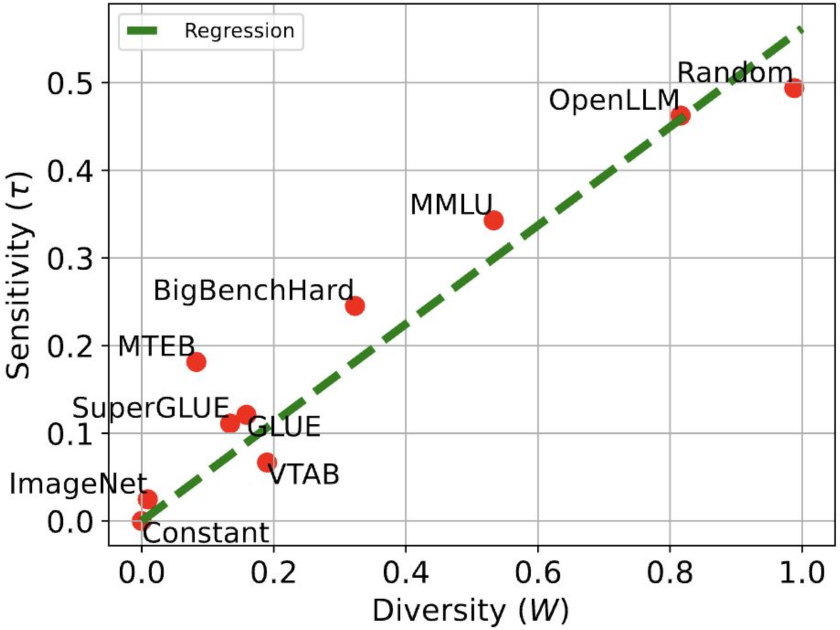
**Diversity:** Variance in rankings (desirable)

**Sensitivity:** How much irrelevant changes to a single task affect the overall ranking. (undesirable)

**Key finding:** The more diverse a multi-task benchmark, the more sensitive it is to irrelevant changes.



# Cardinal benchmarks: Diversity versus sensitivity

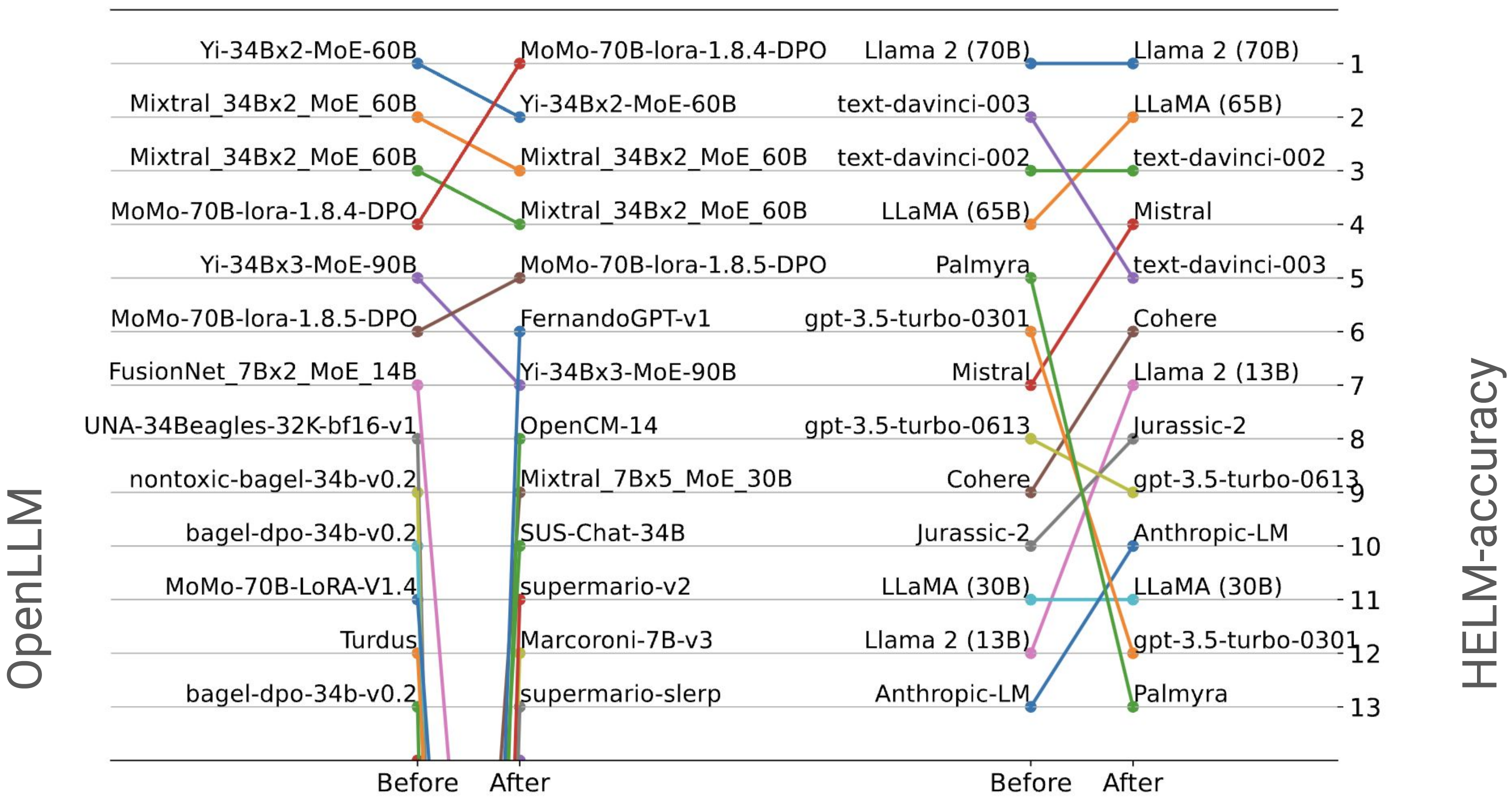



All benchmarks fall on line between constant and random.

Measure of *"multi-taskness"*

**Diversity comes at cost of sensitivity**

# Effect of irrelevant changes on model rankings



# Are dynamic benchmarks the future?

Link:
https://dynabench.org/
Paper: Nie et al. (2020)

## Rethinking AI Benchmarking

Dynabench is a research platform for dynamic data collection and benchmarking. Static benchmarks have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics.

**Initial data** → **Model building** → **Adversarial data collection** → **Model building** → **Adversarial data collection** →

# A theory of dynamic benchmarks [Shirali-Abebe-*H* 2023]

Dynamic benchmark is a DAG with four operations:

- Model building
- Model ensembling
- Data collection
- Data pooling

**Results:**

Progress in standard design can stall after 3 rounds.

More sophisticated designs guarantee strictly more progress. But…


**Standard design:** Directed path alternating model building and adversarial data collection

# Scalable model evaluation at the *frontier*?

**Problem:** Expert evaluation increasingly costly or difficult

Evaluation *frontier*: New models can exceed human expertise

**LLM-as-judge:** Can we use strong models for evaluation new models?

**Major issue:** Models have strong biases (e.g., self-preferencing)

**A solution?** Exciting new debiasing methods promise to combine few expert labels with many model evaluations for unbiased evaluation!

**Theorem** [Dorner-Nastl-*H* 24]: At the frontier, optimal debiasing is no better than using twice the number of expert labels.

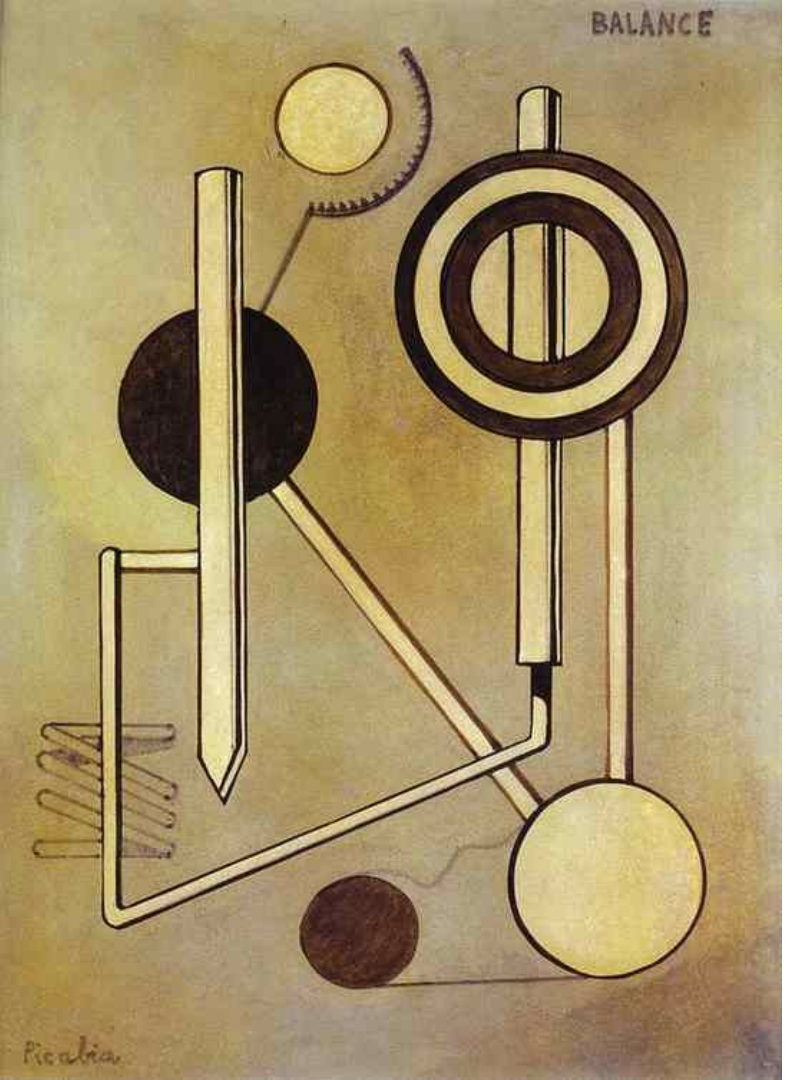
# Summing up

**ImageNet era retrospective taught us a lot about benchmarking**

- The iron rule has both internal and external validity
  - We know more about the former, less about the latter
- Good data not necessary for *ranking models* by accuracy

**Polymorphic era challenges the benchmarking paradigm**

- *Training on the test task* is a confounder we need to adjust for.
- Multi-task benchmark *diversity comes at the cost of stability*.
- Dynamic benchmarks *may stall*.
- LLM-as-judge no better than twice the labels




# The emerging science of benchmarks

ML = *anything goes* + *iron rule*

Simple, powerful engine of scientific progress

We're doing fine on *anything goes*, iron rule less so

We need scientific foundations for the iron rule

Theoretical and empirical program to understand what collective practices promote scientific progress

SOCIAL FOUNDATIONS OF COMPUTATION
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Founded in 2022

2024


Thank you.