

WHAT BENCHMARKS REALLY MEASURE: FROM METHODS TO MODELS IN AI EVALUATION

Olawale Salaudeen, PhD

Postdoc, MIT and the Broad Institute of MIT and Harvard

AI Institute Fellow in Residence, Schmidt Sciences

ABOUT ME

- AI Institute Fellow, Schmidt Sciences
- Postdoc EECS, MIT and the Broad Institute of Harvard and MIT
- Ph.D. in Computer Science, UIUC (visiting Ph.D. @ Stanford 2022-2024)
- B.S. in Mechanical Engineering, Texas A&M University

Previously interned at Google Deepmind (w/ Alex D'Amour), Max Planck Institute for Intelligent Systems (w/ Moritz Hardt), Sandia National Laboratories, Cruise LLC, ...

Misc: On the job market this cycle.

OUTLINE

Note: I have attempted to make this talk mostly high-level and conceptual to enable and encourage discussion and/or debate. Please interrupt to ask questions, make a comment, disagree, etc., as much as you want!

- A brief point on the history of benchmarking norms
- **Methods x External Validity**
- On Evaluating **Models vs. Evaluating Methods**
- **Validity**
- Concluding remarks and future directions

A BRIEF POINT HISTORY POINT ON BENCHMARKS

Liberman, M. (2012, July). <https://www.ling.upenn.edu/courses/ling052/LibermanNAS072012a.pdf>

Hardt, Moritz, and Benjamin Recht. Patterns, predictions, and actions: Foundations of machine learning. Princeton University Press, 2022.



MACHINE TRANSLATION

- **1933 France.** George Artsrouni invented the first automated translation system.
- Considered the first attempt at automatic translation and processing of information.

The U.S. government got super excited and dumped money on this topic!

The Committee cannot judge what the total annual expenditure for research and development toward improving translation should be. However, it should be spent hardheadedly toward important, realistic, and relatively short-range goals.

It is important, therefore, to achieve some objective evaluation of accuracy and quality. Work toward practical useful tests, such as that described in Appendix 10, is of the greatest importance.

1985. MACHINE TRANSLATION REVIVAL – “COMMON TASK METHOD”

Emerging trends: A tribute to Charles Wayne

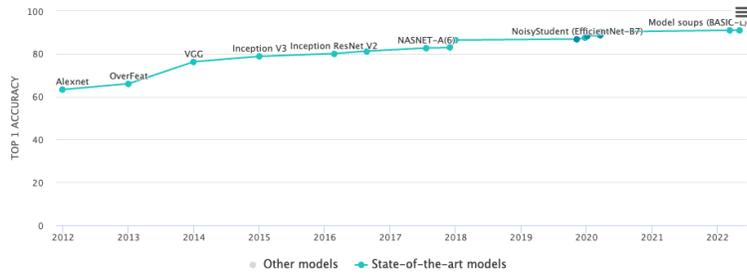
Published online by Cambridge University Press: 09 October 2017

He'll design a speech recognition research program that

- protects against “**glamour and deceit**”
 - because there is a well-defined, objective evaluation metric
 - applied by a neutral agent (NIST)
 - on shared data sets; and
- and ensures that “**simple, clear, sure knowledge is gained**”
 - because participants must reveal their methods
 - to the sponsor and to one another
 - at the time that the evaluation results are revealed

Sharing data and problems

- lowers costs and barriers to entry
- creates intellectual communities
- speeds up replication and extension
- and guards against “glamour and deceit”
(...as well as simple confusion)



Filter: ImageNet-1k only Transformer ResNet CNN ImageNet-22k EfficientNet JFT-300M MLP

ResNeXt Reversible Neighborhood Attention NAT Transformer PatchConvnet FPN MoE Early Exit

Dynamic Model Arch CNN+Transformer CLIP data No Extra Data SNN IG-1B Swin-Transformer

ALIGN Teacher-22k Vision Transformer JFT-3B FLD-900M Pure CNN YFCC-15M Lalon-400M

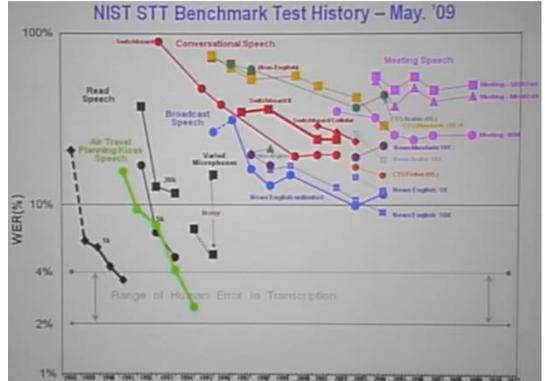
Contrastive ConvNeXt Self-Supervised Learning RegNet Conv+Transformer Mixer Memory-Centric

CLIP Pre-trained CrossCovarianceAttention untagged Hardware Burden Operations per network pass

Robustness reports

Edit Leaderboard

Rank	Model	Top 1 Accuracy	Number of params	GFLOPs	Number of parameters (M)	Top 5 Accuracy	Extra Training Data	Paper	Cod
1	CoCa (finetuned)	91.0%	2100M					CoCa: Contrastive Captioners are Image-Text Foundation Models	
2	Model soups (BASIC-L)	90.98%	2440M					Model soups: averaging weights of multiple fine-tuned models	
3	Model soups (ViT-G/14)	90.94%	1843M						
4	DaViT-G	90.4%	1437M	1038					



model	IN-1K acc.	GFLOPs	model	IN-1K acc.	GFLOPs
ResNet-50 (PyTorch [1])	76.13	4.09	ResNet-200 [29]	78.20	15.01
ResNet-50 (enhanced recipe)	78.82 ± 0.07	4.09	ResNet-200 (enhanced recipe)	81.14	15.01
stage ratio	79.36 ± 0.07	4.53	stage ratio and increase width	81.33	14.52
"patchify" stem	79.51 ± 0.18	4.42	"patchify" stem	81.59	14.38
depthwise conv	78.28 ± 0.08	2.35	depthwise conv	80.54	7.23
increase width	80.50 ± 0.02	5.27	increase width	81.85	16.76
inverting dimensions	80.64 ± 0.03	4.64	inverting dimensions	82.64	15.68
move up depthwise conv	79.92 ± 0.08	4.07	move up depthwise conv	82.04	14.63
kernel size → 5	80.35 ± 0.08	4.10	kernel size → 5	82.32	14.70
kernel size → 7	80.57 ± 0.14	4.15	kernel size → 7	82.30	14.81
kernel size → 9	80.57 ± 0.06	4.21	kernel size → 9	82.27	14.95
kernel size → 11	80.47 ± 0.11	4.29	kernel size → 11	82.18	15.13
ReLU → GELU	80.62 ± 0.14	4.15	ReLU → GELU	82.19	14.81
fewer activations	81.27 ± 0.06	4.15	fewer activations	82.71	14.81
fewer norms	81.41 ± 0.09	4.15	fewer norms	83.17	14.81
BN → LN	81.47 ± 0.09	4.46	BN → LN	83.35	14.81
separate d.s. conv (ConvNeXt-T)	81.97 ± 0.06	4.49	separate d.s. conv (ConvNeXt-B)	83.60	15.35
Swin-T [45]	81.30	4.50	Swin-B [45]	83.50	15.43

Table 3. Results on all tested subpopulation benchmarks, when attributes are *unknown* in both training and validation set. Full results for each dataset and other settings are in Appendix E. Methods that re-train classifier using a two-stage strategy are in gray.

Algorithm	Waterbirds	CelebA	CivilComments	MultiNLI	MetaShift	ImageNetBG	NICO++	MIMIC-CXR	MIMICNotes	CXRMultisite	CheXpert	Living17	Avg
ERM	69.1 ± 4.7	57.6 ± 0.8	63.2 ± 1.2	66.4 ± 2.3	82.1 ± 0.8	76.8 ± 0.9	35.0 ± 4.1	68.6 ± 0.2	80.4 ± 0.2	50.1 ± 0.9	41.7 ± 3.4	27.7 ± 1.1	59.9
Mixup	77.5 ± 0.7	57.8 ± 0.8	65.8 ± 1.5	66.8 ± 0.3	79.0 ± 0.8	76.9 ± 0.7	30.0 ± 4.1	66.8 ± 0.6	81.6 ± 0.6	50.1 ± 0.9	37.4 ± 3.5	29.8 ± 1.8	60.0
GroupDRO	73.1 ± 0.4	68.3 ± 0.9	61.5 ± 1.8	64.1 ± 0.8	83.1 ± 0.7	76.4 ± 0.2	31.1 ± 0.9	67.4 ± 0.5	83.7 ± 0.1	59.2 ± 0.3	74.7 ± 0.3	31.1 ± 0.0	64.5
CVaRDRO	75.5 ± 2.2	60.2 ± 3.0	62.9 ± 3.8	48.2 ± 3.4	83.5 ± 0.5	74.8 ± 0.8	27.8 ± 2.3	68.0 ± 0.2	65.6 ± 1.5	50.2 ± 0.9	50.2 ± 0.9	27.3 ± 1.6	57.8
JTT	71.2 ± 0.5	48.3 ± 1.5	51.0 ± 4.2	65.1 ± 1.6	82.6 ± 0.4	77.0 ± 0.4	30.6 ± 2.3	64.9 ± 0.3	83.8 ± 0.1	57.9 ± 2.1	60.4 ± 4.8	28.3 ± 1.1	60.1
LfF	75.0 ± 0.7	53.0 ± 4.3	42.2 ± 7.2	57.3 ± 5.7	72.3 ± 1.3	70.1 ± 1.4	28.8 ± 2.0	62.2 ± 2.4	84.0 ± 0.1	50.1 ± 0.9	13.7 ± 9.8	26.4 ± 1.3	52.9
LISA	77.5 ± 0.7	57.8 ± 0.8	65.8 ± 1.5	66.8 ± 0.3	79.0 ± 0.8	76.9 ± 0.7	30.0 ± 4.1	66.8 ± 0.6	81.6 ± 0.6	50.1 ± 0.9	37.4 ± 3.5	29.8 ± 1.8	60.0
ReSample	70.0 ± 1.0	74.1 ± 2.2	61.0 ± 0.6	66.8 ± 0.5	81.0 ± 1.7	77.7 ± 1.1	30.6 ± 2.3	67.5 ± 0.3	82.6 ± 0.6	55.0 ± 0.2	74.3 ± 0.4	31.4 ± 0.6	64.3
ReWeight	71.9 ± 0.6	69.6 ± 0.2	59.3 ± 1.1	64.2 ± 1.9	83.1 ± 0.7	76.8 ± 0.9	25.0 ± 0.0	67.0 ± 0.4	84.0 ± 0.1	61.4 ± 1.3	73.7 ± 1.0	27.7 ± 1.1	63.6
SqrtReWeight	71.0 ± 1.4	66.9 ± 2.2	68.6 ± 1.1	63.8 ± 2.4	82.6 ± 0.4	76.8 ± 0.9	32.8 ± 3.5	68.0 ± 0.4	83.1 ± 0.2	61.2 ± 0.6	68.5 ± 1.6	27.7 ± 1.1	64.2
CBLoss	74.4 ± 1.2	65.4 ± 1.4	67.3 ± 0.2	63.6 ± 2.4	83.1 ± 0.0	76.8 ± 0.9	31.7 ± 3.6	67.6 ± 0.3	84.0 ± 0.1	50.2 ± 0.9	74.0 ± 0.7	27.7 ± 1.1	63.8
Focal	71.6 ± 0.8	56.9 ± 3.4	61.9 ± 1.1	62.4 ± 2.0	81.0 ± 0.4	71.9 ± 1.2	30.6 ± 2.3	68.7 ± 0.4	70.9 ± 9.8	50.0 ± 0.9	42.1 ± 4.0	26.9 ± 0.6	57.9
LDAM	70.9 ± 1.7	57.0 ± 4.1	28.4 ± 7.7	65.5 ± 0.8	83.6 ± 0.4	76.7 ± 0.5	31.7 ± 3.6	66.6 ± 0.6	81.0 ± 0.3	50.1 ± 0.9	36.0 ± 0.7	24.3 ± 0.8	56.0
BSoftmax	74.1 ± 0.9	69.6 ± 1.2	58.3 ± 1.1	63.6 ± 2.4	82.6 ± 0.4	76.1 ± 2.0	35.6 ± 1.8	67.6 ± 0.6	83.8 ± 0.3	58.6 ± 1.8	73.8 ± 1.0	28.6 ± 1.4	64.4
DFR	89.0 ± 0.2	73.7 ± 0.8	64.4 ± 0.1	63.8 ± 0.0	81.4 ± 0.1	74.4 ± 1.8	38.0 ± 3.8	67.1 ± 0.4	80.2 ± 0.0	60.8 ± 0.4	75.8 ± 0.3	26.3 ± 0.4	66.2
CRT	76.3 ± 0.8	69.6 ± 0.7	67.8 ± 0.3	65.4 ± 0.2	83.1 ± 0.0	78.2 ± 0.5	33.3 ± 0.0	68.1 ± 0.1	83.4 ± 0.0	61.8 ± 0.1	74.6 ± 0.4	31.1 ± 0.1	66.1
ReWeightCRT	76.3 ± 0.2	70.7 ± 0.6	64.7 ± 0.2	65.2 ± 0.2	85.1 ± 0.4	77.5 ± 0.7	33.3 ± 0.0	67.9 ± 0.1	83.4 ± 0.0	53.1 ± 2.3	75.1 ± 0.2	33.1 ± 0.1	65.4

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics



Overview Data Code Models Discussion **Leaderboard** Rules

Leaderboard

Raw Data Refresh

Search leaderboard

This leaderboard is calculated with all of the test data.

#	Team	Members	Score	Entries	Last	Join
1	Sanyam_Sharma24		1.00000	1	2mo	
2	mos3 santos		1.00000	347	3d	
3	Veniamin_Nelin		1.00000	51	16h	
4	Neha #11		1.00000	1	2mo	
5	AbstracLpwq		1.00000	4	2mo	
6	RIDHAMGAR		1.00000	1	2mo	
			1.00000	3	2mo	
			1.00000	4	2mo	
			1.00000	48	1d	
			1.00000	9	2mo	
			1.00000	4	2mo	

**ONE COULD ARGUE THAT BENCHMARKS
HAVE PRIMARILY SERVED AS A TOOL FOR
DOING SCIENCE.**

Many do!

BENCHMARKS ENABLE FRICTIONLESS REPRODUCIBILITY

Transformation 1: Wrappers

- Adopt a given existing workflow as a starting point.
- Prepend a new workflow element to preprocess the inputs
- Append a new workflow element to postprocess the outputs
- Obtain a new workflow

Transformation 2: Tweaks

- Adopt a given existing workflow as a starting point.
- Assuming this workflow exposes internal hyperparameters that were previously unexplored:
 - Systematically vary those unexplored hyperparameters
 - Extract the optimal hyperparameter combination
 - Obtain a new workflow

Donoho, David. "Data science at the singularity." *arXiv preprint arXiv:2310.00865* (2023).

Why Does Frictionless Reproducibility Work?

Inspired by Thomas Kuhn (1962), we can think of the scientific and engineering process as a massively parallel genetic algorithm. If we want to improve upon the systems we currently have, we might try a small perturbation to see if we get an improvement. If we can find a small change that improves some desired outcome, we could change our systems to reflect this improvement. If we continually search for these improvements and work hard to demonstrate their value, we may head in a better direction over time.

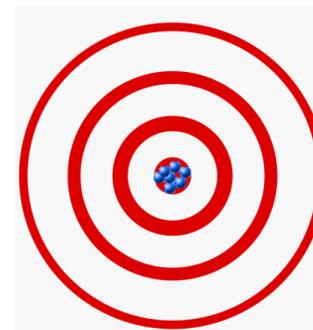
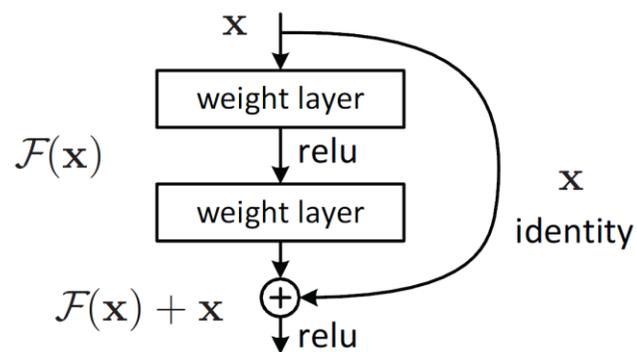
Recht, Benjamin. "The mechanics of frictionless reproducibility." *Harvard Data Science Review* 6.1 (2024).

EXTERNAL VALIDITY

ImageNot: A Contrast with ImageNet Preserves Model Rankings (*Salaudeen & Hardt 2024*) ---

<https://arxiv.org/abs/2404.02112>

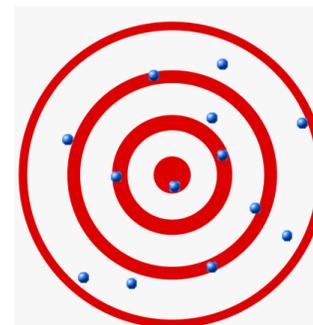
IDEAL SETTING



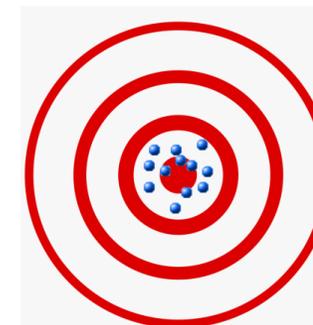
A: accurate and precise



B: precise, but not accurate



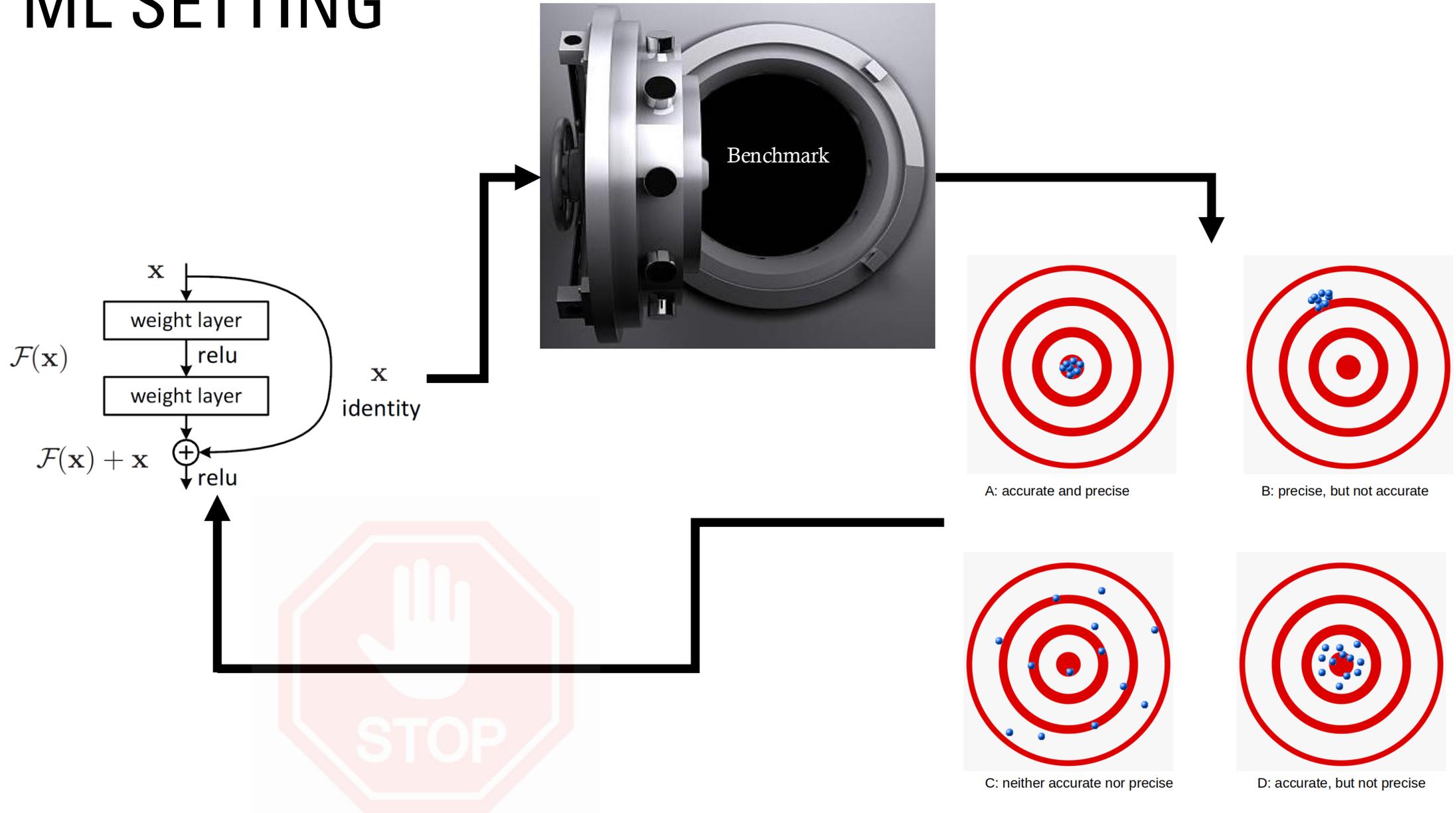
C: neither accurate nor precise



D: accurate, but not precise



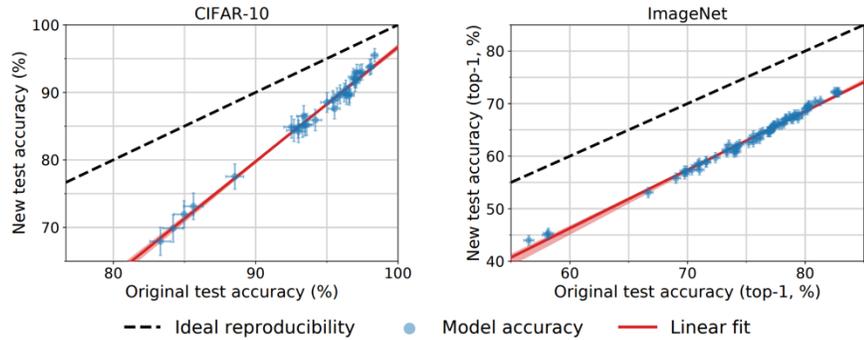
ML SETTING



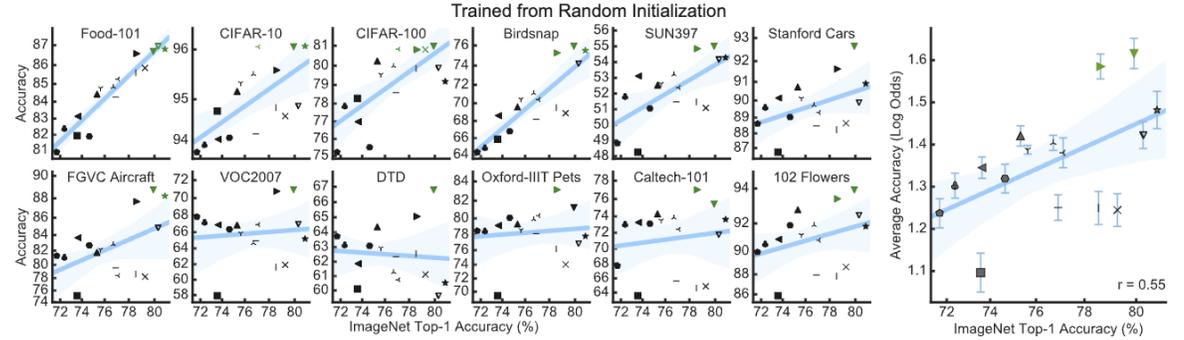


Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht* Rebecca Roelofs Ludwig Schmidt Vaishaal Shankar
 UC Berkeley UC Berkeley UC Berkeley UC Berkeley



		ImageNet Top-1				
Orig. Rank	Model	Orig. Accuracy	New Accuracy	Gap	New Rank	Δ Rank
1	pnasnet_large_tf	82.9 [82.5, 83.2]	72.2 [71.3, 73.1]	10.7	3	-2
4	nasnetalarge	82.5 [82.2, 82.8]	72.2 [71.3, 73.1]	10.3	1	3
21	resnet152	78.3 [77.9, 78.7]	67.0 [66.1, 67.9]	11.3	21	0
23	inception_v3_tf	78.0 [77.6, 78.3]	66.1 [65.1, 67.0]	11.9	24	-1
30	densenet161	77.1 [76.8, 77.5]	65.3 [64.4, 66.2]	11.8	30	0
43	vgg19_bn	74.2 [73.8, 74.6]	61.9 [60.9, 62.8]	12.3	44	-1
64	alexnet	56.5 [56.1, 57.0]	44.0 [43.0, 45.0]	12.5	64	0
65	fv_64k	35.1 [34.7, 35.5]	24.1 [23.2, 24.9]	11.0	65	0



ImageNot: A contrast with ImageNet preserves model rankings

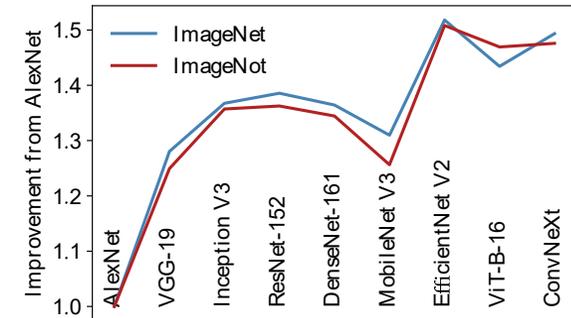
Olawale Salaudeen¹ and Moritz Hardt²

¹University of Illinois at Urbana Champaign

²Max Planck Institute for Intelligent Systems, Tübingen and Tübingen AI Center

ImageNot TLDR:

- Sourced from LAION-5B; web-crawled
- Disjoint ImageNet Class
- RoBERTa & CLIP annotation



- Recht, Benjamin, et al. "Do imagenet classifiers generalize to imagenet?." *International conference on machine learning*. PMLR, 2019.
- Kornblith, Simon, Jonathon Shlens, and Quoc V. Le. "Do better imagenet models transfer better?." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

ON EVALUATING MODELS VS. METHODS

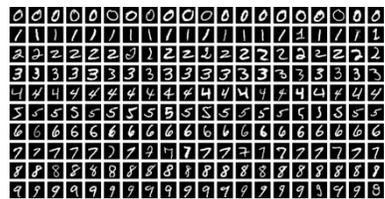
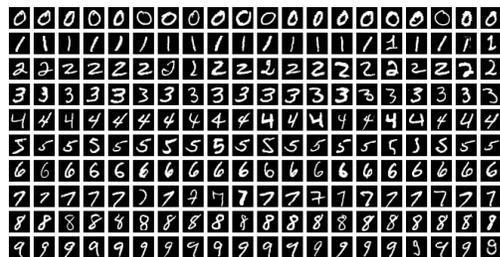
Salaudeen, Olawale Elijah, Florian E. Dorner, and Peter Hase. "On Evaluating Methods vs. Evaluating Models." *(Oral)* NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

1. Classical Paradigm

Core Idea. Develop new algorithms from scratch.

Evaluation. Performance on a single, specific train/test task.

Evaluation Inference. Is this algorithm better than others for this task?



2. Pretraining and Finetuning

Core Idea. Adapt a pre-trained model for a new task.

Evaluation. Performance on downstream tasks *after* fine-tuning.

Evaluation Inference. Is this pre-trained model a good foundation?



3. Prompting, ...

Core Idea. Prompt a single, frozen model to perform any task.

Evaluation. Performance across a wide suite of tasks *without* fine-tuning.

Evaluation Inference. Is this model a general-purpose reasoner?

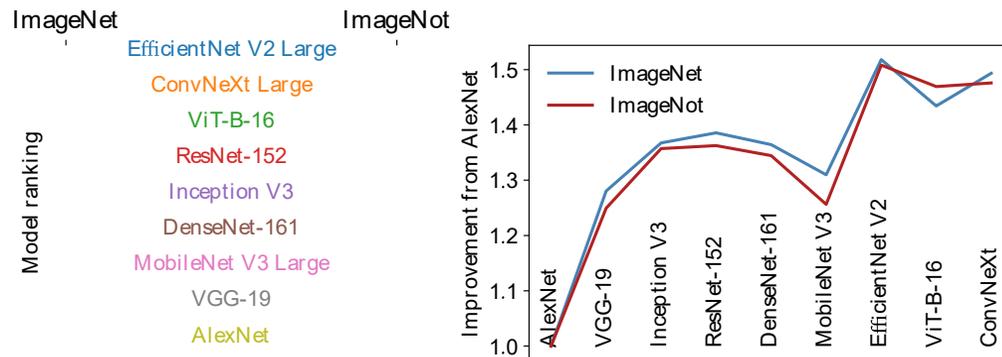


WHAT INFERENCE ARE WE MAKING FROM OUR EVALUATIONS?

Evaluating *Methods*

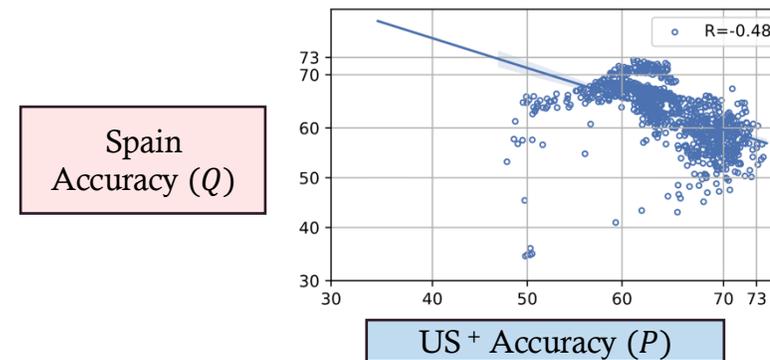
Evaluating methods can be pretty far from use case and still generalize (about scientific advances)

- Inference is ordinal, which is better --- ImageNet shows that relative gains can even be robust

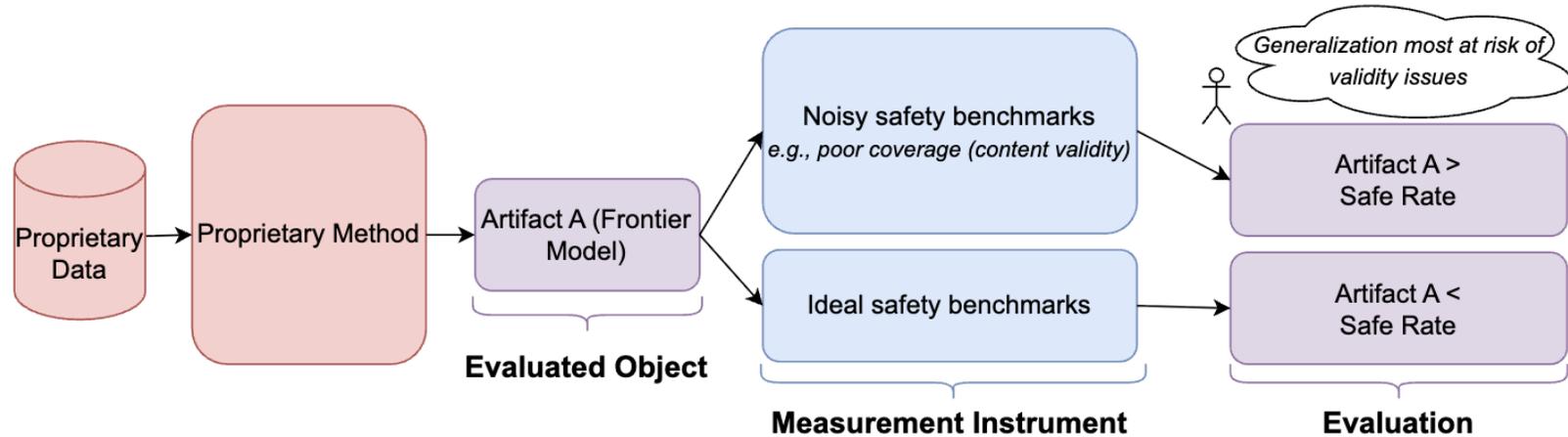
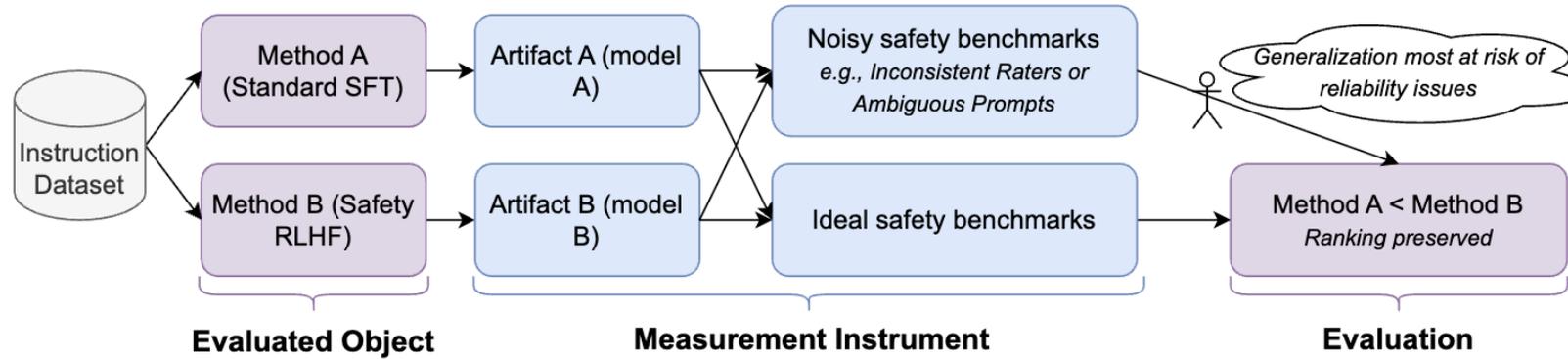


Evaluating *Models*

- Evaluating models could be arbitrarily misleading unless aligned with use case (about task utility)
 - Inference is absolute; how good is a model? Irrelevant interventions can break things.



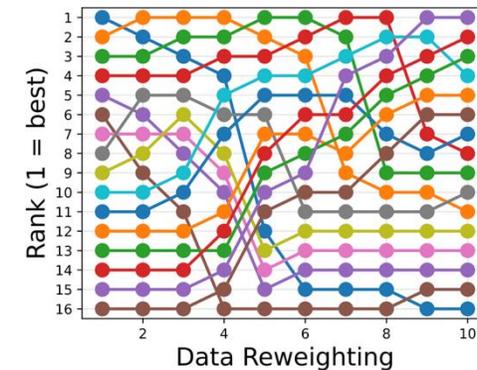
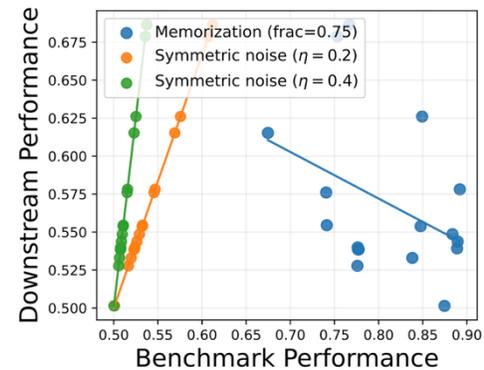
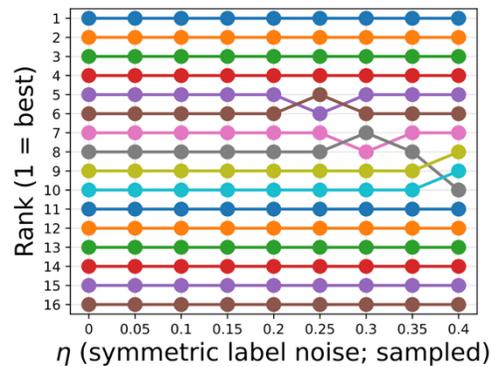
Evaluating Methods: Claim (substantiated). Improvement on standard safety benchmarks tells us that safety RLFH improves harm refusal rates.



Evaluating Models: Claim (unsubstantiated). A model that scores > X score on standard safety benchmarks can be used prominently for safety-critical tasks.

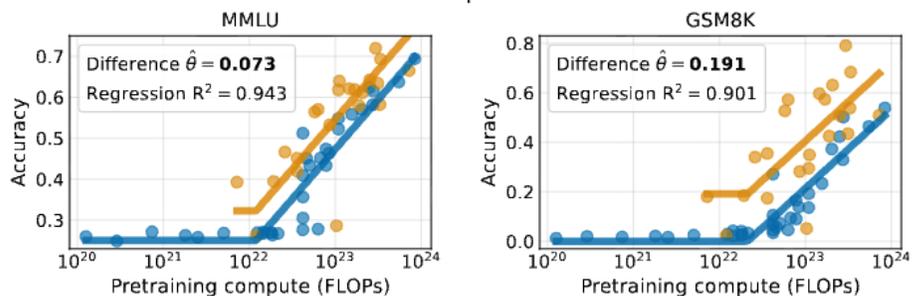
IMPLICATIONS

- Model-independent permutations preserve method rankings.
- For absolute performance, arbitrary changes (non-IID) can arbitrarily change scores.

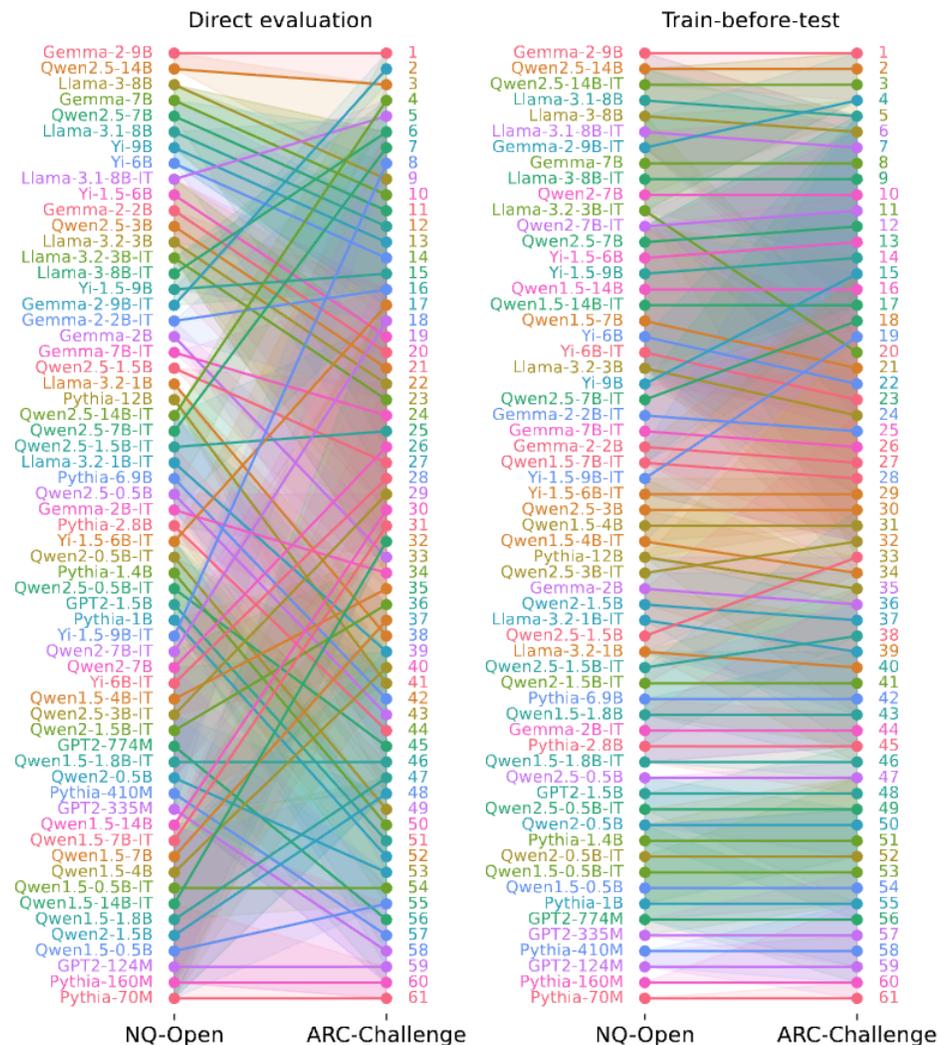
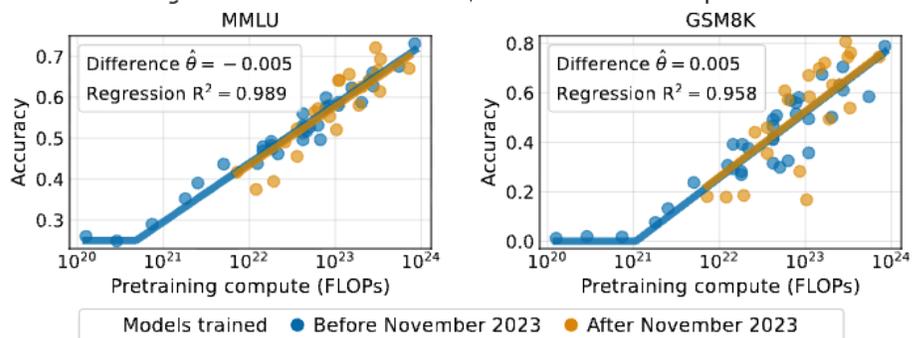


Today, interpreting rankings is tricky b/c we don't know enough to make methods inferences.

Base models trained after November 2023 outperform those trained before November 2023



After fine-tuning all models on the test task, differences in model performance vanish



Accuracy	Calibration	Robustness	Fairness	Efficiency	General information	Bias	Toxicity	Summarization metrics
Model	Mean win rate	MMLU - EM	BoolQ - EM	NarrativeQA - F1				
Llama 2 (70B)	0.944	0.582	0.886	0.77				
LLaMA (65B)	0.908	0.584	0.871	0.755				
text-davinci-002	0.905	0.568	0.877	0.727				
Mistral v0.1 (7B)	0.884	0.572	0.874	0.716				
Cohere Command beta (52.4B)	0.874	0.452	0.856	0.752				
text-davinci-003	0.872	0.569	0.881	0.727				
Jurassic-2 Jumbo (178B)	0.824	0.48	0.829	0.733				
Llama 2 (13B)	0.823	0.507	0.811	0.744				
TNLG v2 (530B)	0.787	0.469	0.809	0.722				
gpt-3.5-turbo-0613	0.783	0.391	0.87	0.625				
LLaMA (30B)	0.781	0.531	0.861	0.752				
Anthropic-LM v4-s3 (52B)	0.78	0.481	0.815	0.728				
gpt-3.5-turbo-0301	0.76	0.59	0.74	0.663				
Jurassic-2 Grande (17B)	0.743	0.475	0.826	0.737				

A BROADER CHALLENGE: THERE IS A LARGE NUMBER OF DOWNSTREAM TASKS.

Does HELM's 'fairness metric' capture (and predict) fairness w.r.t. all the downstream contexts and tasks where we care about?

1. If so, how do we know?
2. If not, do we need a unique evaluation of each of these contexts x tasks?
3. Is there a middle ground? what is it?

Be more careful about the link between what we measure and the inferences or claims we derive from it: “make sure claims are *valid* given the evaluation.”

VALIDITY

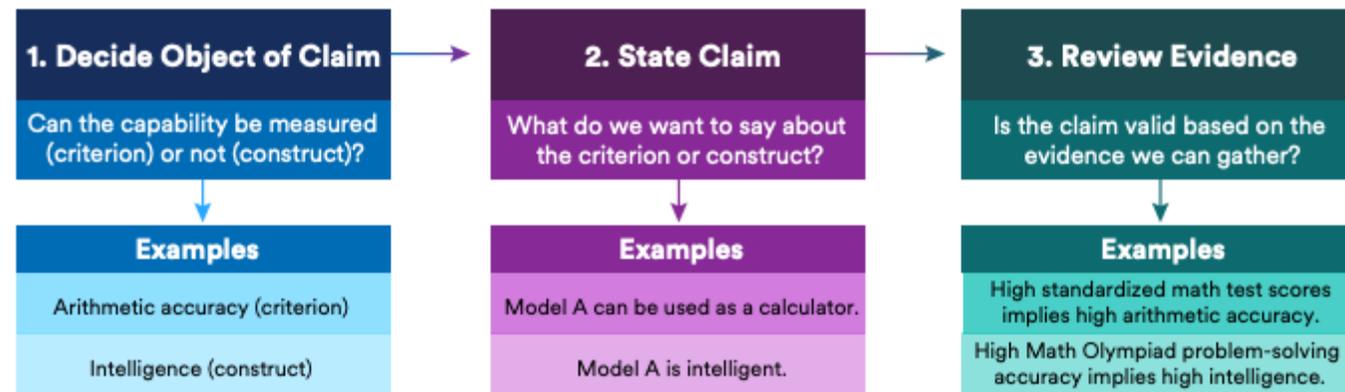
This is actually quite hard...harder than many acknowledge!

WHAT IS OUR GOAL?

We want to make claims and make evaluations to support those claims, e.g.,

- Model M can accurately answer specialized scientific questions.
- Deploying model M at a hospital improves health outcomes.
- The model has unlearned a harmful concept C or private information.

WHEN DO EVALUATIONS SUPPORT OUR CLAIMS



Criterion. *A directly measurable or observable concept (e.g., “textbook linear algebra question answering accuracy”).*

Construct. *An abstract concept not directly measurable (e.g., “mathematical reasoning” or “trustworthiness”).*

EVIDENCE: VALIDITY

How well does your evaluation measure what it is intended to measure?

Criterion Validity

Does your evaluation correlate with a known validated standard?

Construct Validity

Does your evaluation measure the intended construct?

Content Validity

Does your evaluation cover all relevant cases?

External Validity

Does your evaluation generalize across different environments or settings?

Consequential Validity

Does your evaluation consider the real-world impact of test interpretation and use?

Question writing (by question writer) 

Question and choices
Methylcyclopentadiene was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed [...] How many chemically distinct isomers make up the final product (not counting stereoisomers)?
(a) 2 (b) 16 (c) 8 (d) 4

Correct answer (b)

Explanation Methylcyclopentadiene exists as an interconverting mixture of 3 isomers [...] if there are 4 dienes, and 4 different directions of approach the dienophile can take to each of them, there are $4 \times 4 = 16$ possible products.

Expert validation #1 (by expert validator #1)

Part 1: answer Q (correct answer & explanations not shown)
My answer is (a). Here's my explanation: [...] 

Part 2: provide feedback on the following dimensions (correct answer & explanations shown to the validator)

- Post-hoc agreement: Is the answer uncontroversial?
- Did you understand Q fully, now that you see the explanations?
- Is your background sufficient to answer correctly?
- Detailed feedback
- Q difficulty
- Q & answer choice revisions

Post-hoc agreement : I agree that the correct answer is (b), after seeing writer's explanation.
Feedback / revision: [...] I got confused with the sentence 'not counting isomers'. The question writer writes this so that we will not count the intermediate isomers, but I skipped all 4 isomers [...] it is my personal mistake.

Expert validation #2 (by expert validator #2)

Part 1: answer Q (correct answer & explanations not shown)
My answer is (b). Here's the explanation: [...] 

Part 2: provide feedback (correct answer & explanations shown)

Post-hoc agreement : I agree that the correct answer is (b), after seeing writer's explanation.
Feedback / revision: It's difficult and takes a long time [...] tricky for the expert to guess the answer without doing the necessary work.

Question revision (by question writer) 

Revised question and choices
Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed [...] How many chemically distinct isomers make up the final product (not counting stereoisomers)?
(a) 2 (b) 16 (c) 8 (d) 4

Correct answer & explanation [Same as before]

Non-expert validation
(by non-expert validators who are experts in other domains; at least 15 min, avg ~37 min, **allowing Google**)

Non-expert #1 (c) 
Non-expert #2 (b) 
Non-expert #3 (a) 

Include this Q in the **DIAMOND** set because
(1) 2 out of 2 expert validators agree*
(2) ≤ 1 out of 3 non-expert validators answers correctly

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning <i>GPQA, Diamond</i>	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	—	—
Undergraduate level knowledge <i>MMLU</i>	88.7%** 5-shot	86.8% 5-shot	—	85.9% 5-shot	86.1% 5-shot
	88.3% 0-shot CoT	85.7% 0-shot CoT	88.7% 0-shot CoT	—	—

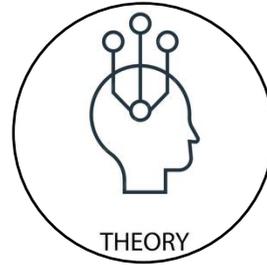
Claims from Graduate-Level Google-Proof Question Answering (GPQA) Benchmark Accuracy Report Card

Claims	Content	Criterion	Construct	External	Consequential
1. AI systems can accurately answer <i>graduate-level specialized multiple-choice questions</i> in biology, physics, and chemistry.					
2. AI systems can accurately answer <i>graduate-level specialized questions</i> in specialized scientific domains.					
3. AI systems can exhibit <i>general reasoning abilities</i> that can transfer beyond current human specialization.					

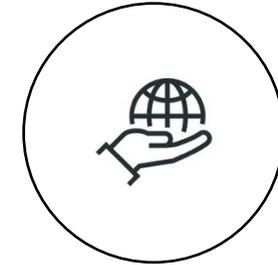
LISSITZ AND SAMUELSEN'S VIEW OF TEST PURPOSES



Utility Determination:
Criterion Validity



Theory Determination:
Construct Validity



Impact Evaluation:
Consequential Validity





Google-Proof Question Answering.

1. Criterion-Aligned Evidence

Measurement supports a claim about a directly related criterion.

GPQA scores *predicts* accuracy on new questions on the same subjects.

2. Criterion-Adjacent Evidence

Measuring supports a claim about a non-directly related criterion.

GPQA scores *predicts* accuracy in specialized scientific domains QA

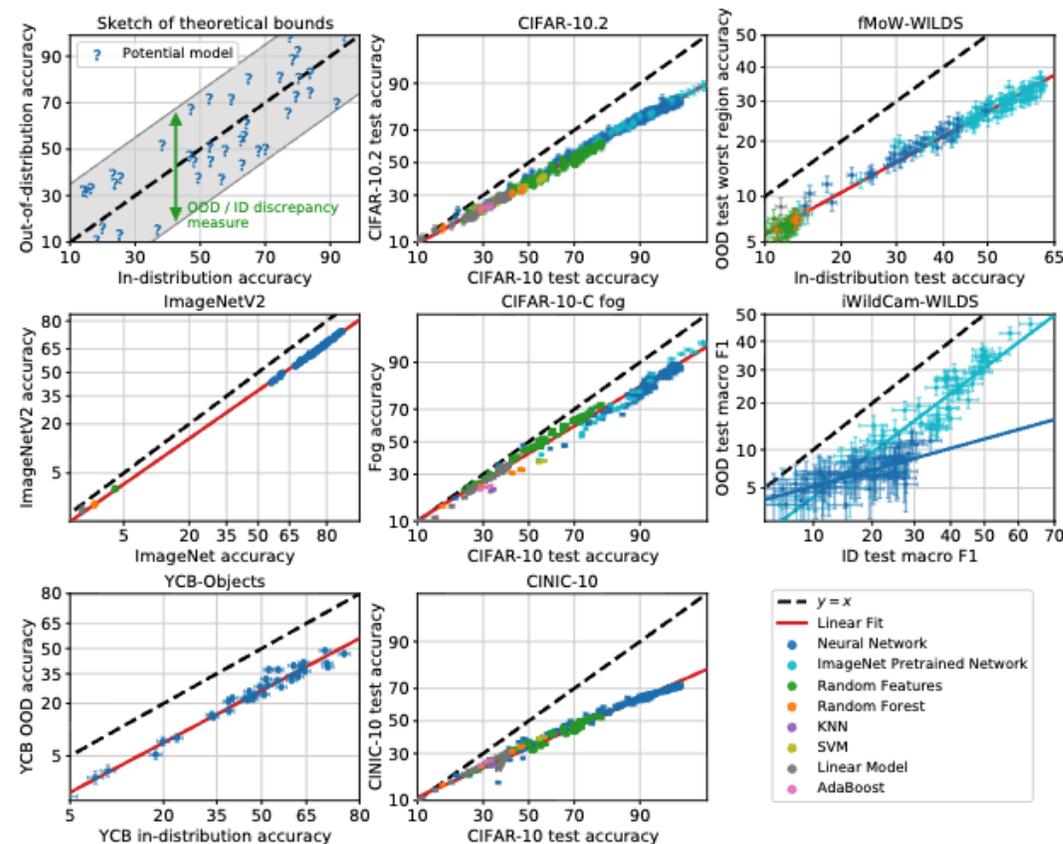
3. Construct Targeted

Measurement supports a claim about a construct.

GPQA *measures* human-like reasoning ability.

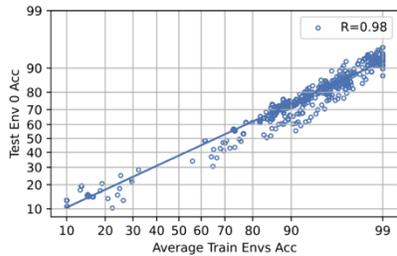
CRITERION VALIDITY

EXAMPLE WITH ACCURACY ON THE LINE

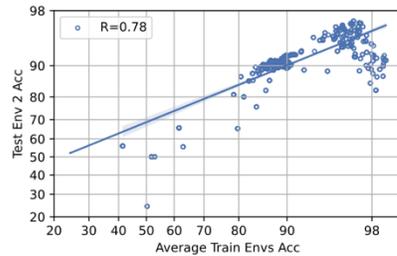


Müller, John P., et al. "Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization." *International conference on machine learning*. PMLR, 2021.

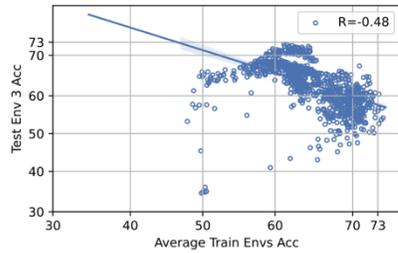
LIMITATIONS OF ACCURACY ON THE LINE



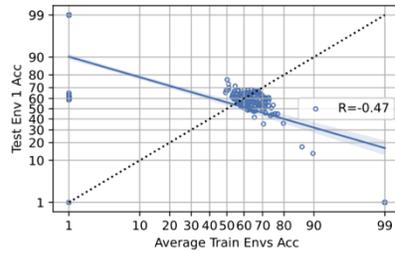
(a) PACS (Env 0)



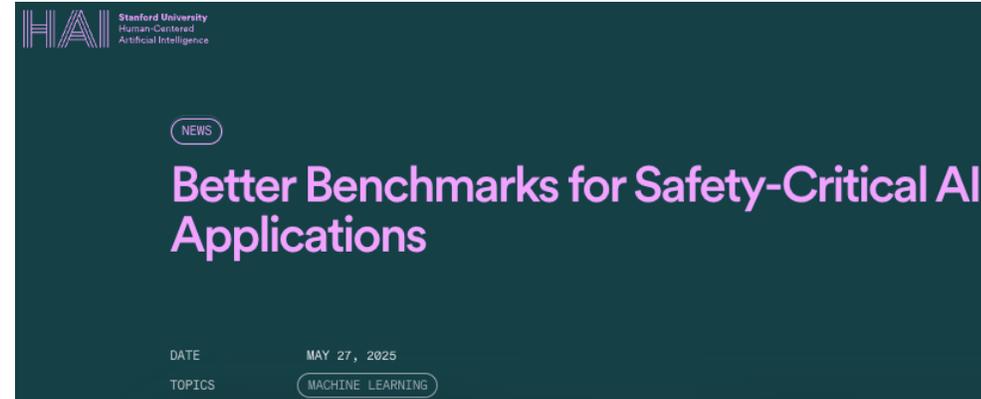
(b) Camelyon (Env 2)



(c) Covid-CXR (Env 3)



(d) CivilComments (Env 1)



Theorem 3 (Benchmarks with Accuracy on the Line are Misspecified Almost Everywhere). *Define*

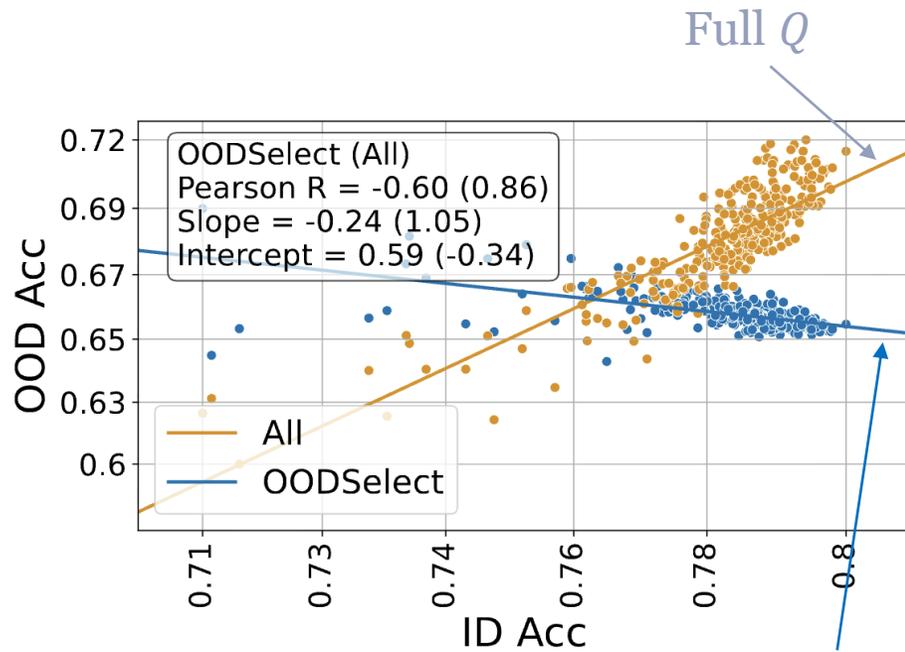
$$\mathcal{W}_\epsilon = \left\{ \begin{array}{l} w_{spu}^\top (M \mu_{spu}) + \sqrt{2(L_\phi \kappa)^2 \|w_{spu}\|_2^2 \log(1/\delta)} < 0 \quad (\text{well-specified; Theorem 1}), \\ M \in \mathbb{R}^{l \times l} : \text{ and} \\ \left| \Phi^{-1}(\text{acc}_{P_{\text{OOD}}}(f_X)) - a \Phi^{-1}(\text{acc}_{P_{\text{ID}}}(f_X)) \right| \leq \epsilon \quad (\text{accuracy on the line; Definition 6}) \end{array} \right\} \quad (12)$$

Then:

- (i) \mathcal{W}_0 has Lebesgue measure zero in $\mathbb{R}^{l \times l}$ (i.e., probability zero for a random draw).
- (ii) For any $0 \leq \epsilon_i \leq \epsilon_j$, we have $\mathcal{W}_{\epsilon_i} \subseteq \mathcal{W}_{\epsilon_j}$.

CONSEQUENTIAL VALIDITY

MIMIC Chest X-Ray Accuracy (Q)
New Patient Population



MIMIC Chest X-Ray Accuracy (P)

$$Q_1 \subset Q$$

$$|Q_1| > 0.7|Q|$$

On Group Sufficiency Under Label Bias

NeurIPS 2025.

Haoran Zhang Olawale Salaudeen Marzyeh Ghassemi
Massachusetts Institute of Technology
{haoranz, olawale, mghassem}@mit.edu

Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness

NeurIPS

2025

Stephen R. Pfohl
Google Research
spfohl@google.com

Natalie Harris
Google Research

Chirag Nagpal
Google Research

David Madras
Google DeepMind

Vishwali Mhasawade
New York University

Olawale Salaudeen
Massachusetts Institute of Technology

Awa Dieng
Google DeepMind

Shannon Sequeira
Google Research

Santiago Arciniegas
The Hospital for Sick Children

Lillian Sung
The Hospital for Sick Children

Nnamdi Ezeanochie
Google

Heather Cole-Lewis
Google

Katherine Heller
Google Research

Sanmi Koyejo
Stanford University

Alexander D'Amour
Google DeepMind

**MAYBE BENCHMARKS ARE
JUST INSUFFICIENT FOR SOME
GOALS**

WHAT ARE OTHER OPTIONS

Validity	Common risks	Investigation Tools	Evidence Exemplar
Content Validity	<ul style="list-style-type: none"><input type="checkbox"/> Coverage deficiency<input type="checkbox"/> Construct irrelevance<input type="checkbox"/> Imbalanced mixture of content	<ul style="list-style-type: none"><input type="checkbox"/> Expert review<input type="checkbox"/> Red-teaming / adversarially designed evaluations<input type="checkbox"/> Synthetic data generation or edge cases	<ul style="list-style-type: none"><input type="checkbox"/> Documentation of how test items comprehensively cover the construct<input type="checkbox"/> Explicit mapping of test content to abstract frameworks or industry standards<input type="checkbox"/> Coverage analysis

WHAT ARE OTHER OPTIONS

Validity	Common risks	Investigation Tools	Evidence Exemplar
Criterion Validity	Predictive and Concurrent Validity <ul style="list-style-type: none"><input type="checkbox"/> Criterion contamination<input type="checkbox"/> Criterion deficiency<input type="checkbox"/> Restricted range<input type="checkbox"/> Temporal/other shifts	<ul style="list-style-type: none"><input type="checkbox"/> Real-world longitudinal studies<input type="checkbox"/> Real-world behavioral testing<input type="checkbox"/> Scaling-law predictive models<input type="checkbox"/> Validated criterion studies<input type="checkbox"/> Periodic post-	<ul style="list-style-type: none"><input type="checkbox"/> Correlation with an existing validated benchmark or gold standard<input type="checkbox"/> Evidence that higher scores in evaluation metrics predict real-world utility

CONSTRUCT VALIDITY

Example Gallery

Please feel free to copy the prompts and test with your own VLMs.

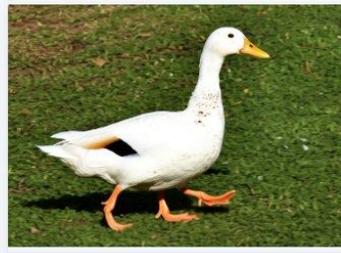


How many legs does this animal have? Answer with a number in curly brackets, e.g., {9}.

Animals

Copy Prompt

Copy Image



How many legs does this animal have? Answer with a number in curly brackets, e.g., {9}.

Animals

Copy Prompt

Copy Image



How many legs does this animal have? Answer with a number in curly brackets, e.g., {9}.

Animals

Copy Prompt

Copy Image



How many legs does this animal have? Answer with a number in curly brackets, e.g., {9}.

Animals

Copy Prompt

Copy Image

How many legs do the animals have?

Animals

Logos

Flags

Chess Pieces

Game Boards

Optical Illusions

Patterned Grids

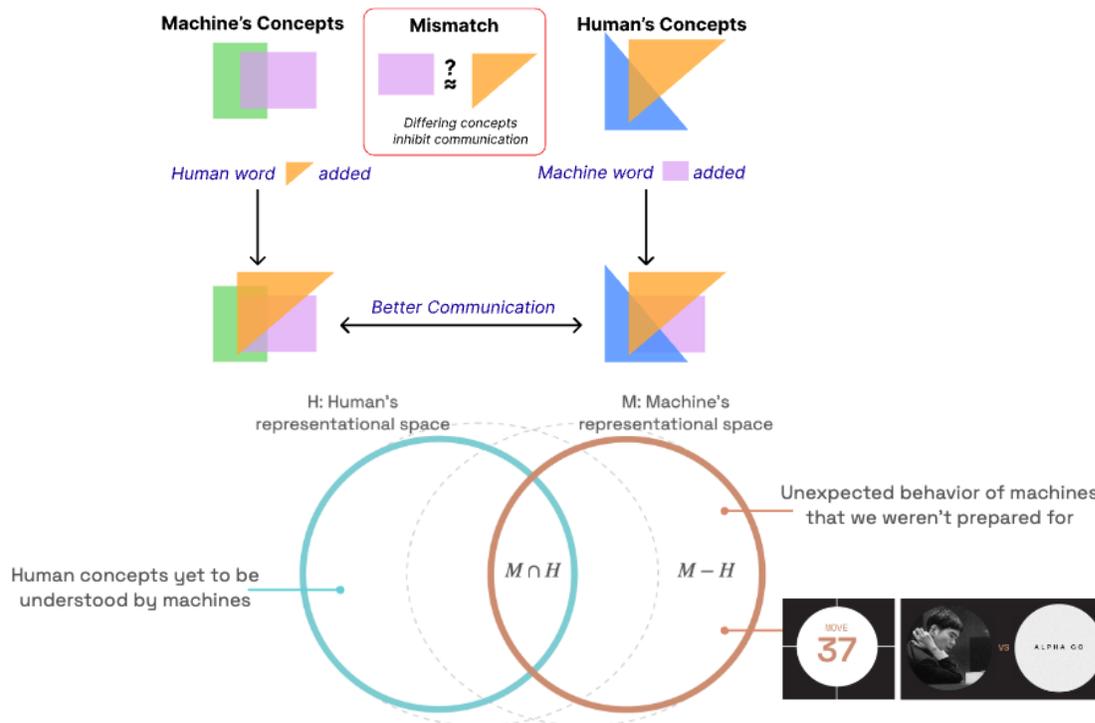
Assumed Mechanism:
visual reasoning.

Actual Mechanism: pattern
recognition and knowledge
(?).

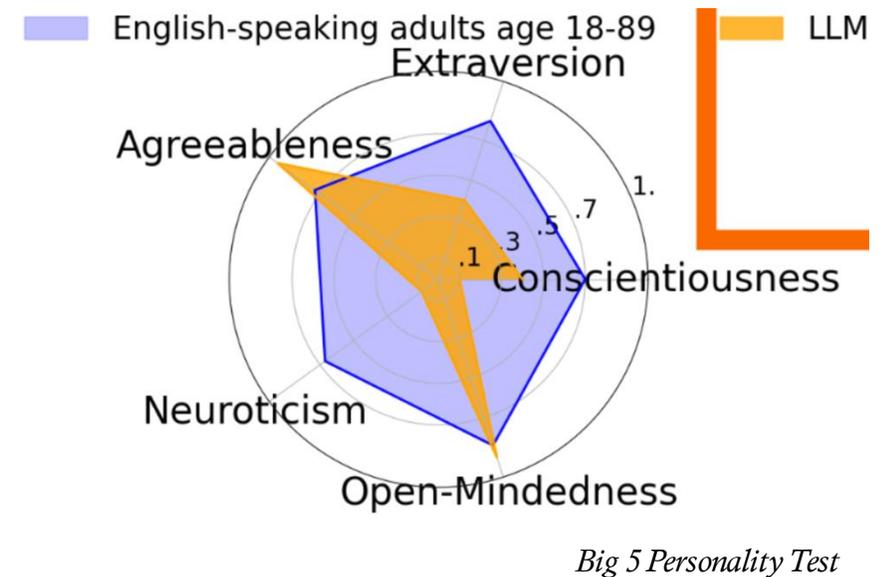
Vo, An, et al. "Vision Language Models are Biased." *arXiv preprint arXiv:2505.23941* (2025).

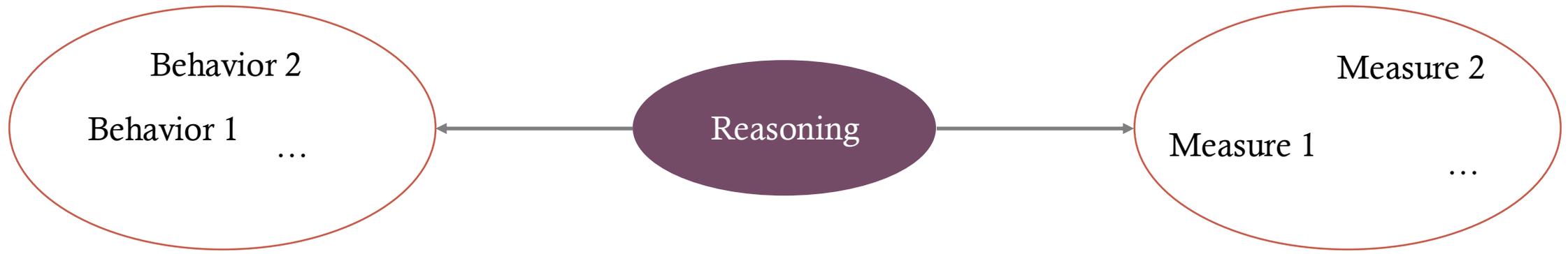
RECONCILING THE HUMAN PRIOR

Constructs & Concepts



Measurements





Claim. There is a difference between the following:

Intrinsic Constructs (discovery) vs.

- *Internally defined, we group a set of smooth, covarying behaviors and name*
 - E.g., reasoning, intelligence,

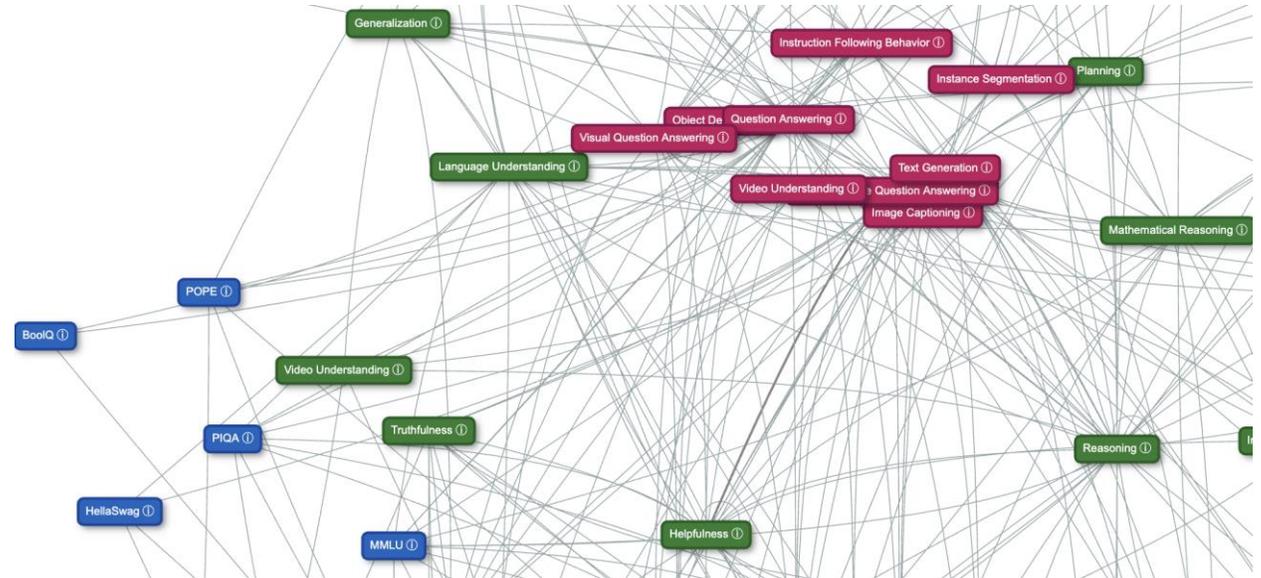
Extrinsic Constructs (systemization)

- *Externally defined, we group a set of behaviors we want and name them*
 - E.g., safety, fairness, ...

FUTURE WORK: AI CONSTRUCT LEXIS

Some questions I am interested in:

1. What are the stable intrinsic constructs for AI?
2. What defines a sample and a population for AI?
3. Are there latent classes for AI systems that need different measurements for the same intrinsic constructs?
4. What new statistical tools do we need for rigorous analysis?



Salaudeen et al., 2026. In Prep



<https://aiconstructlexis.github.io/>

Current implicit structure from the
research community.

FUTURE DIRECTIONS

QUESTIONS?

Acknowledgements

- **Sanmi Koyejo.** *Assistant Professor, Stanford University;* **Moritz Hardt.** *Director, Max Planck Institute for Intelligent Systems (MPI-IS);* **Marzyeh Ghassemi.** *Assistant Professor, MIT;* **Sara Beery.** *Assistant Professor, MIT;* **Angelina Wang.** *Assistant Professor, Cornell Tech; and many more collaborators!*

Contact: olawale@mit.edu

Webpage: <https://olawalesalaudeen.com>