ON EVALUATING AI METHODS VS. MODELS

OLAWALE SALAUDEEN

POSTDOC, MIT | INCOMING (2027) ASSISTANT PROFESSOR, UIUC

A CRISIS IN AI EVALUATION

IF I AM SEEING RESULTS OF AN EVALUATION, E.G., A BENCHMARK SCORE OR LEADERBOARD, HOW SHOULD I INTERPRET THE EVALUATION AND WHAT CLAIMS DO THEY SUPPORT?

MORE CONCRETE

SUPPOSE YOU READ

'MODEL A OUTPERFORMS MODEL B ON MIMIC IN-HOSPITAL MORTALITY PREDICTION'

ARE WE EVALUATING:

1. **MODEL A** IS GREAT AT IN-HOSPITAL MORTALITY PREDICTION; THUS, I CAN DEPLOY IT IN MY HOSPITAL **OR**
2. **METHOD A**, WHICH PRODUCES MODEL A, IS BETTER AT PRODUCING IN-HOSPITAL MORTALITY PREDICTORS WHEN APPLIED TO YOUR HOSPITAL

AN IMPORTANT DISTINCTION

MODELS AND METHODS ARE DIFFERENT OBJECTS OF EVALUATION WITH DIFFERENT NEEDS

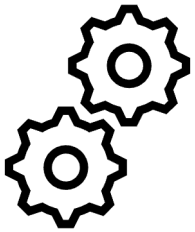
- A **MODEL** IS A CONCRETE, REALIZED ARTIFACT
- A **METHOD** IS A PROCEDURE FOR PRODUCING, ADAPTING, OR SELECTING ARTIFACTS

THE **SAME BENCHMARK SCORE** CAN HAVE TWO ENTIRELY **DIFFERENT INTERPRETATIONS** DEPENDING ON WHAT WE ARE TRYING TO MEASURE!

OUTLINE

- EVALUATING METHODS - BENCHMARKS
  - HOLD-OUT METHOD
  - EXTERNAL VALIDITY
  - ACCURACY ON THE LINE
- EVALUATING METHODS – BEYOND BENCHMARKING
  - VALIDITY
  - LOCAL RETROSPECTIVE VALIDATION, PROSPECTIVE EVALUATION, HUMAN-AI INTERACTION TRIALS, CONTINUOUS MONITORING INFRASTRUCTURES
- RECOMMENDATIONS

System

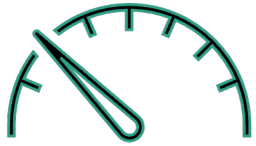


***The object being studied**, e.g., MedPALM, RAG, Workflows*

Measurement Instrument

***The procedure used to generate evidence**, e.g., benchmark, rubric, annotation protocol*

Measurement



***The output produced by the instrument**, e.g., accuracy, score, error rate, human rating*

Evaluation

***Interpretation of measurement**, e.g., relative to a target, criterion, or decision*

Claim

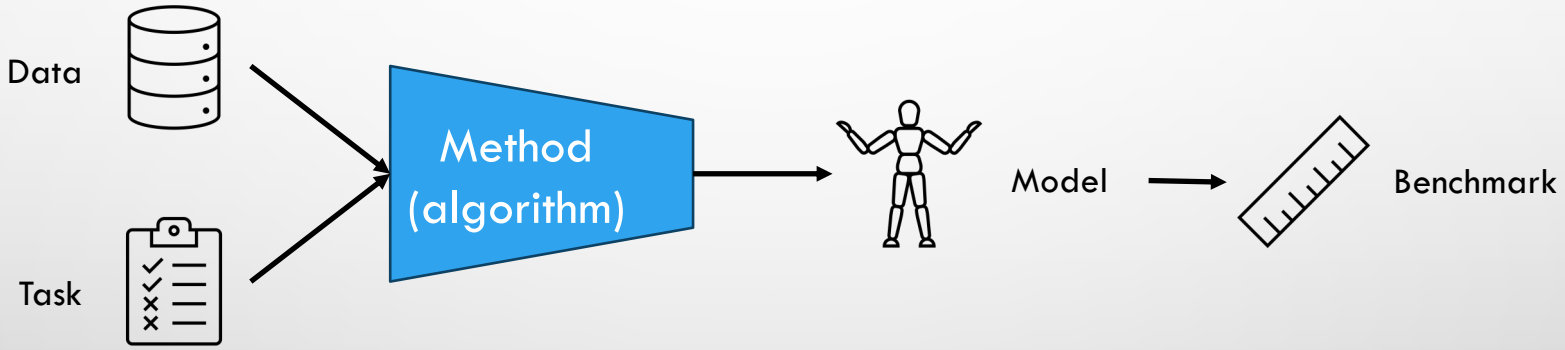
***Conclusion supported by the evaluation**, e.g., robust, useful, fair, deployment-ready*

Salaudeen, Olawale E., Florian E. Dorner, and Peter Hase. "On Evaluating Methods vs. Evaluating Models." *NeurIPS 2025 Workshop on LLM Evaluation* (2025).

METHODS VS. MODELS

WHAT IS A METHOD?

A MAPPING.



Examples.	ERM	Regularization	RAG	RLHF	Data Augmentation	Attention	Finetuning

THE METHOD EVALUATION REGIME

SUPPOSE WE WANT TO COMPARE METHODS F *(RESNET)* AND G *(ALEXNET)*

- OR PREPROCESSING METHODS OR NORMALIZATION OR FEATURES

WE HAVE SOME **EVALUATION WITH DISTRIBUTION P** *(IMAGENET)*

WE ALSO HAVE **THE DISTRIBUTION WE CARE ABOUT *Q (MIMIC CXR)***

- NOTE, P MAY OR MAY NOT BE THE SAME AS Q

THE METHOD EVALUATION REGIME

WE APPLY ***F (RESNET)*** **AND** ***G (ALEXNET)*** **TO** ***P (IMAGENET) AND Q (MIMIC CXR).***

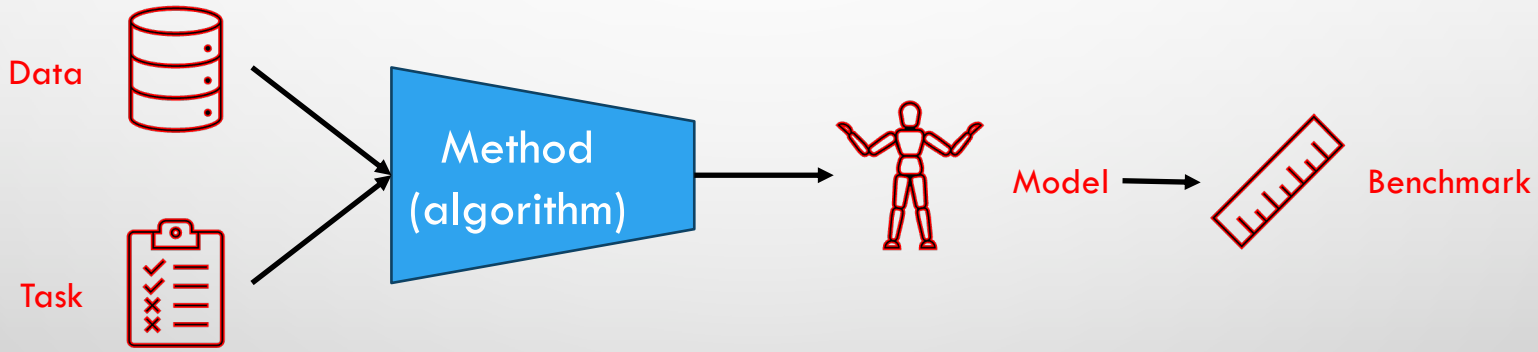
- $F(P,\cdot) \rightarrow \text{MODEL } f_P$ $\quad$ $G(P,\cdot) \rightarrow \text{MODEL } g_P$
- $F(Q,\cdot) \rightarrow \text{MODEL } f_Q$ $\quad$ $G(Q,\cdot) \rightarrow \text{MODEL } g_Q$

WE WANT: $\boldsymbol{f_P > g_P \Rightarrow f_Q > g_Q}$, BUT WHAT WE ARE REALLY SAYING IS THAT $\boldsymbol{F > G}$.

What you are trying to convince stakeholders: you should apply this method to your setting to get the best model for your hospital, relative to competing methods. Note, this is about ***ranking*** *(we will revisit).*

MEASUREMENT INSTRUMENT

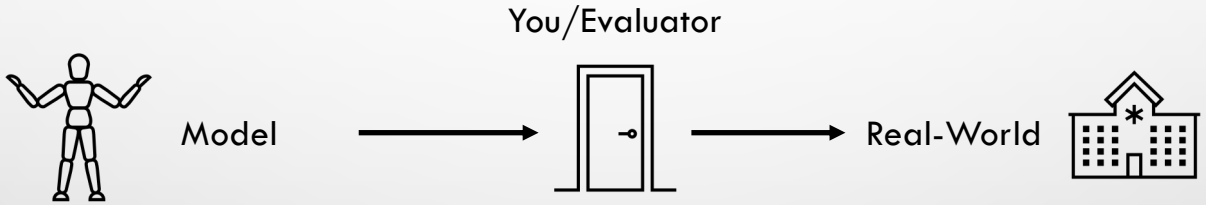
RECALL A MEASUREMENT INSTRUMENT IS DEFINED AS: ***THE PROCEDURE USED TO GENERATE EVIDENCE***



All part of the measurement instrument for the method.

WHAT IS A MODEL?

A MODEL IS A CONCRETE ARTIFACT.



Examples.	Sepsis predictor	CXR-based Classifier	MedPALM	ChatGPT Health

THE MODEL EVALUATION REGIME

SUPPOSE WE WANT TO COMPARE MODELS (SEPSIS PREDICTION MODEL) f (LOGISTIC REGRESSION) AND g (MLP)

- OR f, g WITH DIFFERENT FEATURES OR NORMALIZATION OR PREPROCESSING, …

WE HAVE SOME **EVALUATION WITH DISTRIBUTION P** (PHISIONET DATASET)

WE HAVE SOME **DISTRIBUTION WE CARE ABOUT Q** (YOUR HOSPITAL)

- NOTE, P MAY OR MAY NOT BE THE SAME AS Q

THE MODEL EVALUATION REGIME

WE ARE EVALUATING A **FIXED (POTENTIALLY IMMUTABLE) ARTIFACT** (SEPSIS PREDICTION MODEL) AGAINST **(UN)KNOWN TARGET (Q)** (YOUR HOSPITAL)

THEY COULD HAVE BEEN TRAINED ON SOME P OR SOMETHING ELSE ENTIRELY

EVERY DATASET, CHOICE, AND RANDOM SEED THAT WENT INTO MAKING f AND g MATTERS --- VARIANCE IN THESE COMPONENTS DOESN'T MATTER.

*What you are trying to convince stakeholders: you should deploy this model in your hospital. Note, this is about **absolute performance**, e.g., passes an accuracy threshold…perhaps no models should be deployed.*

THIS DESCRIPTION IS INCREDIBLY IMPORTANT

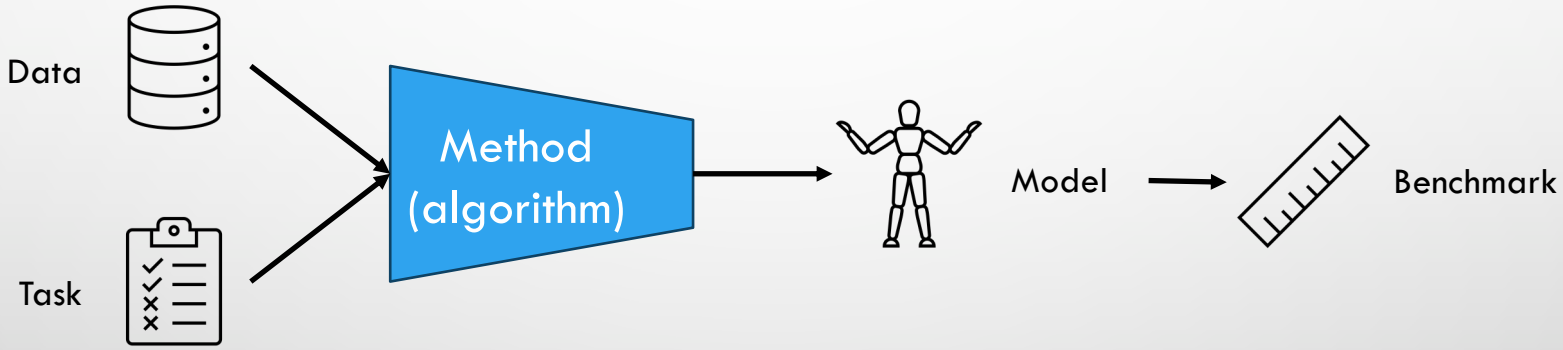
ARE YOU DOING **METHOD DEVELOPMENT**? OR ARE YOU **DEPLOYING SPECIFIC MODELS**?

YOUR EVALUATION BURDEN MAY DIFFER SIGNIFICANTLY!

EVALUATING METHODS

WHAT IS A METHOD?

A MAPPING.



Examples.	ERM	Regularization	RAG	RLHF	Data Augmentation	Attention

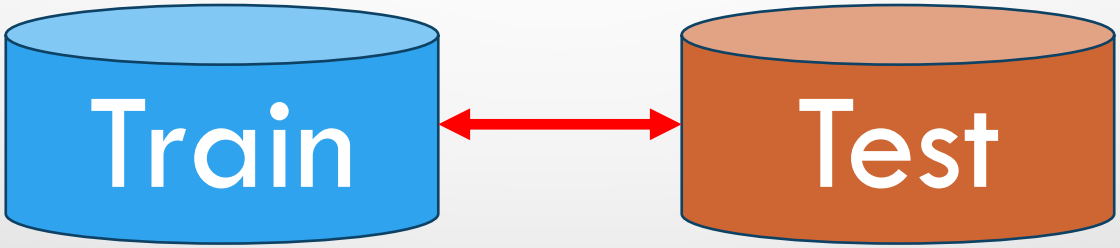
Hardt, Moritz. "The emerging science of machine learning benchmarks." *Manuscript. https://mlbenchmarks. org* (2025).

THE HOLDOUT METHOD

NOT TO BE CONFUSED WITH CROSS-VALIDATION

THE BASIC EMPIRICAL PATTERN IN MACHINE LEARNING DEVELOPMENT. VERY SIMPLE.

SPLIT DATA:



ROUGHLY, YOU CAN DO WHATEVER YOU WANT FOR METHOD DEVELOPMENT, AS LONG AS YOU DON'T TRAIN ON THE TEST SET!

Hardt, Moritz. "The emerging science of machine learning benchmarks." *Manuscript. https://mlbenchmarks. org* (2025).

THE HOLDOUT METHOD

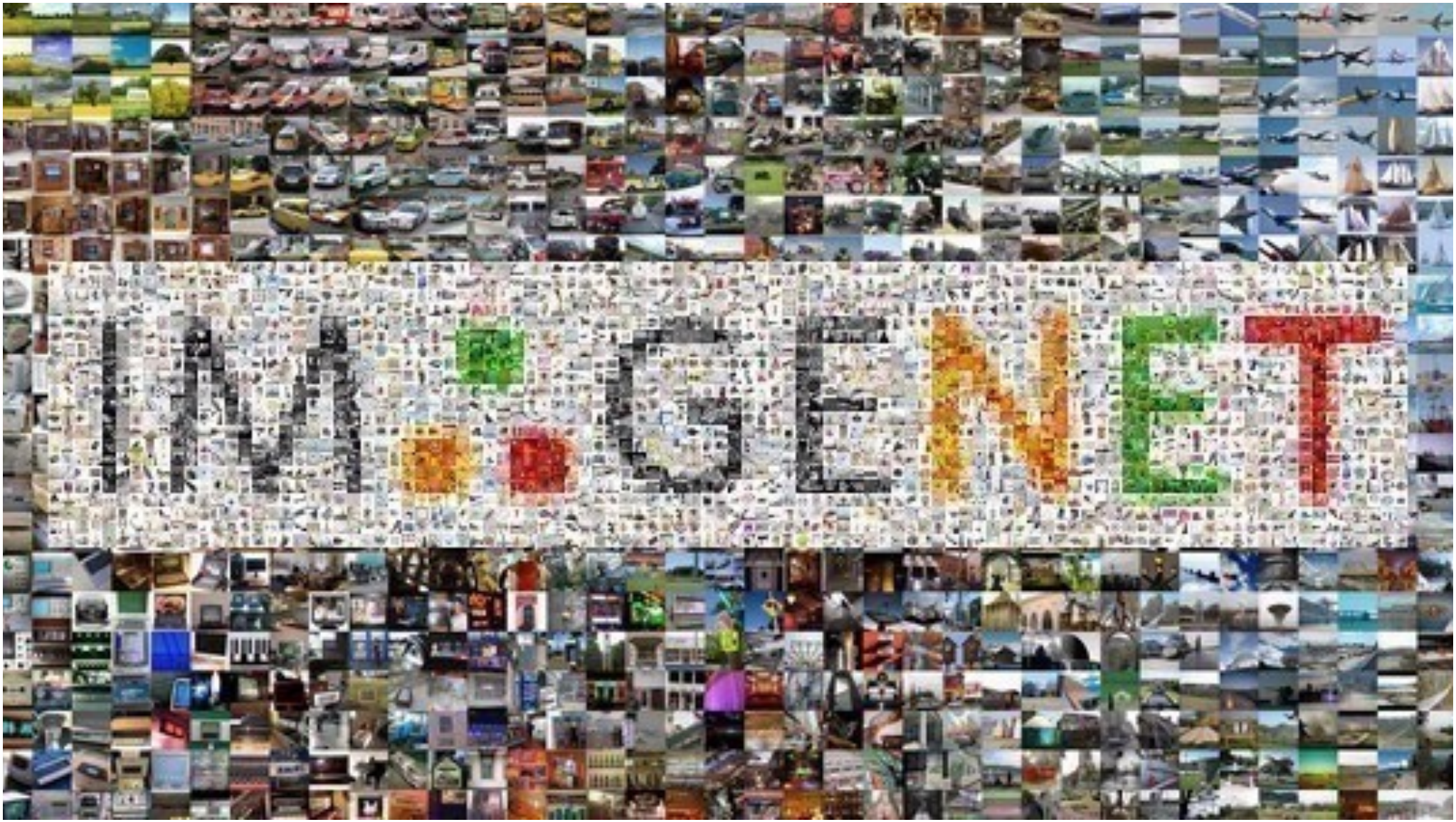
NOT TO BE CONFUSED WITH CROSS-VALIDATION

"ROUGHLY, YOU CAN DO WHATEVER YOU WANT FOR METHOD DEVELOPMENT, AS LONG AS YOU DON'T TRAIN ON THE TEST SET!"

THIS SHOULD SOUND OFF!!

- ITERATING ON THE TEST SET SHOULD BREAK GENERALIZATION – MOST OF OUR THEORY AROUND GENERALIZATION AND BASIC STATISTICS SAYS SO

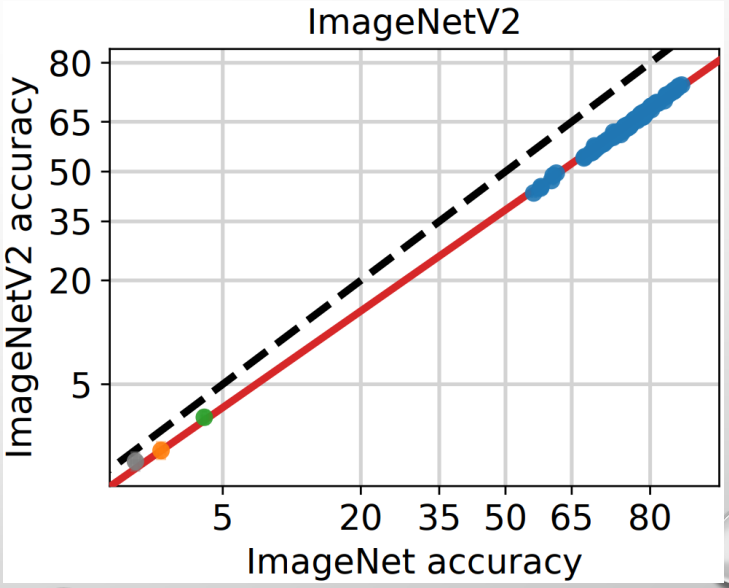
BUT IT SEEMS TO BE TRUE! WE HAVE HAD GENERALIZABLE PROGRESS IN MACHINE LEARNING DESPITE THIS PRACTICE.


IMAGENET

Miller, 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 (2021).
Recht, Benjamin, et al. "Do ImageNet classifiers generalize to ImageNet?" International Conference on Machine Learning (2019).

THE IMAGENETV2 EXPERIMENT

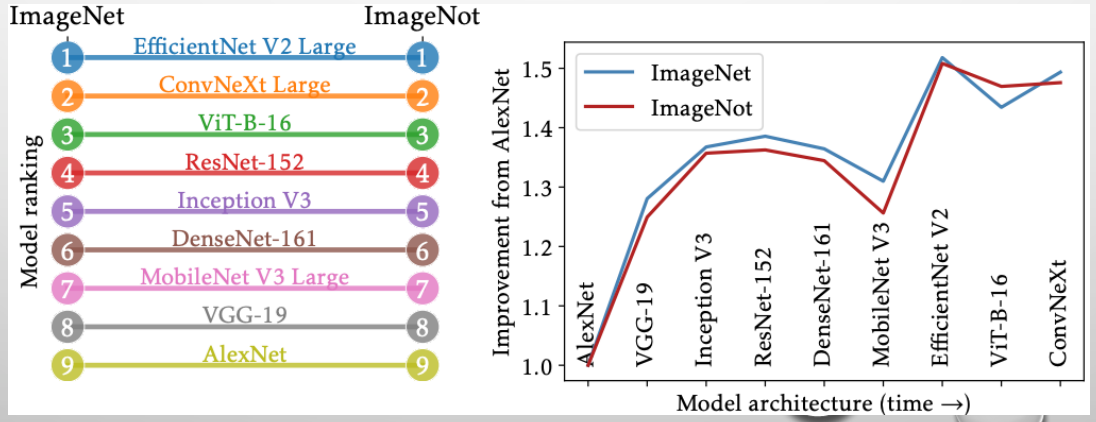
- IMAGENETV2 ASKS WHAT HAPPENS WHEN WE COLLECT A NEW TEST SET USING THE EXACT SAME DATA-DISTRIBUTION PROCESS AS THE ORIGINAL IMAGENET.
- IN THIS EXPERIMENT, THE OBJECTS BEING EVALUATED ARE **FIXED, PRE-TRAINED MODELS.**
  - X-AXIS: A FIXED MODEL EVALUATED ON THE IMAGENET TEST SET.
  - Y-AXIS: THAT *SAME* MODEL EVALUATED ON IMAGENETV2.



Salaudeen, Olawale E., and Moritz Hardt. "ImageNot: A contrast with ImageNet preserves model rankings." Transactions on Machine Learning Research (2026).

THE IMAGENOT EXPERIMENT: RETRAINED METHODS ON A NEW DISTRIBUTION

- INTRODUCE A DIFFERENT DISTRIBUTION WITH ENTIRELY DIFFERENT CLASSES AND IMAGES FOR TRAIN AND TEST, AND **APPLY METHODS FROM SCRATCH**.
- THIS IS A PURE TEST OF THE METHOD: DOES AN ARCHITECTURE AND/OR TRAINING RECIPE THAT WORKS WELL ON IMAGENET ALSO RELIABLY PRODUCE GOOD MODELS IN A NEW ENVIRONMENT?



WHAT IMAGENETV2 AND IMAGENOT TELL US TOGETHER

- IMAGENETV2 SHOWS THAT **METHOD RANKINGS ARE ROBUST TO SIMPLE TEST-SET RESAMPLING AND DISTRIBUTION SHIFT IN TEST SETS.**
- BUT IMAGENOT SHOWS THAT **METHOD RANKINGS ARE ROBUST TO DISTRIBUTION SHIFTS ON BOTH TRAINING AND TEST SETS.**
- METHODS THAT IMPROVED OVER EARLIER ARCHITECTURES ON IMAGENET ALSO IMPROVED WHEN RETRAINED ON IMAGENOT.
  - *THE ORIGINAL BENCHMARK CAPTURED SOMETHING REAL AND ENDURING ABOUT RELATIVE METHOD (MODEL DEVELOPMENT) PROGRESS,* ***ENTIRELY INDEPENDENT OF ABSOLUTE PERFORMANCE****.*

THE METHOD TAKEAWAY

BENCHMARK PROGRESS TYPICALLY REFLECTS REAL METHOD PROGRESS.

THE CLAIM ISN'T:

- "THE BENCHMARK MEASURES REAL-WORLD ACCURACY."

THE CLAIM IS:

- "THE BENCHMARK PRESERVES RELATIVE COMPARISONS AMONG METHODS UNDER SHIFT."

CAN ONE PUSH THE DISTRIBUTION SHIFT ENOUGH THAT EVEN THIS BREAKS? MAYBE!

GREAT NEW FOR MODEL DEVELOPERS

KEEP HILL CLIMBING!!!

THOSE WHO JUST CARE ABOUT DEVELOPING BETTER METHODS FOR BUILDING MODELS!

CAVEAT

THE STORY IS SLIGHTLY MORE COMPLICATED FOR FOUNDATION MODELS, BUT TALK TO ME OFFLINE FOR MORE DISCUSSION.

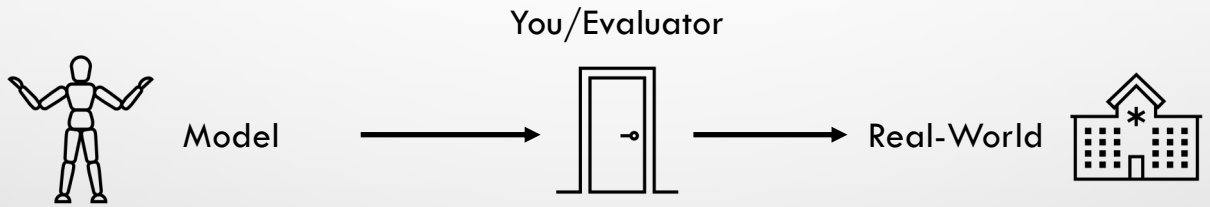
REFERENCES:

- *DOMINGUEZ-OLMEDO, RICARDO, FLORIAN EDDIE DORNER, AND MORITZ HARDT. "TRAINING ON THE TEST TASK CONFOUNDS EVALUATION AND EMERGENCE." INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS. VOL. 2025. 2025.*
- ZHANG, GUANHUA, RICARDO DOMINGUEZ-OLMEDO, AND MORITZ HARDT. "TRAIN-BEFORE-TEST HARMONIZES LANGUAGE MODEL RANKINGS." *ARXIV PREPRINT ARXIV:2507.05195* (2025).
- ZHANG, GUANHUA, AND MORITZ HARDT. "INHERENT TRADE-OFFS BETWEEN DIVERSITY AND STABILITY IN MULTI-TASK BENCHMARKS." *PROCEEDINGS OF THE 41ST INTERNATIONAL CONFERENCE ON MACHINE LEARNING*. 2024.

EVALUATING MODELS

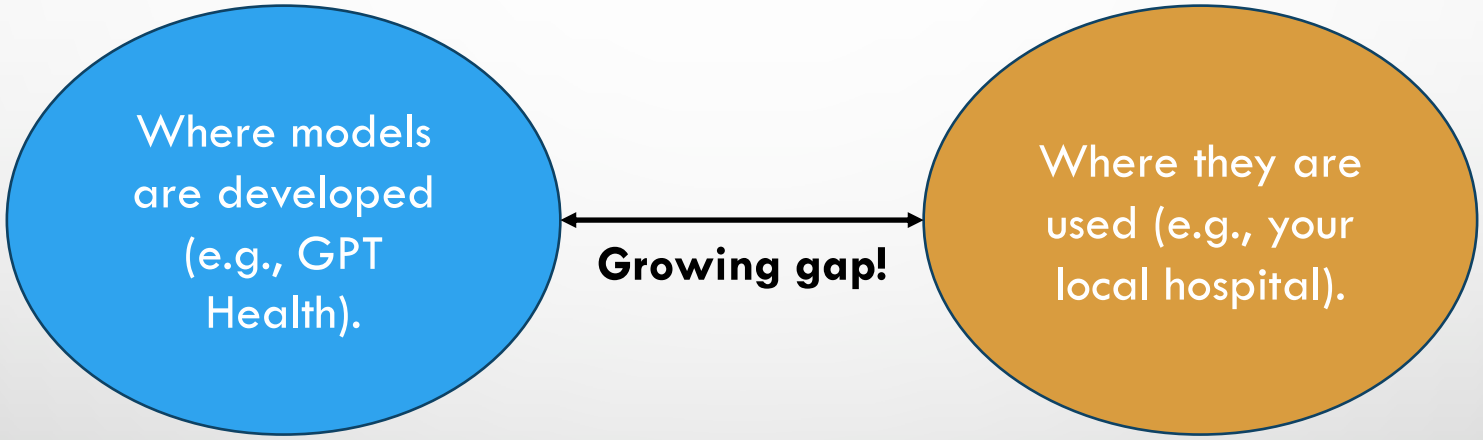
WHAT IS A MODEL?

A MODEL IS A CONCRETE ARTIFACT.



Examples.	Sepsis prediction model	CXR Classifier	MedPALM	ChatGPT Health

NEW REALITY

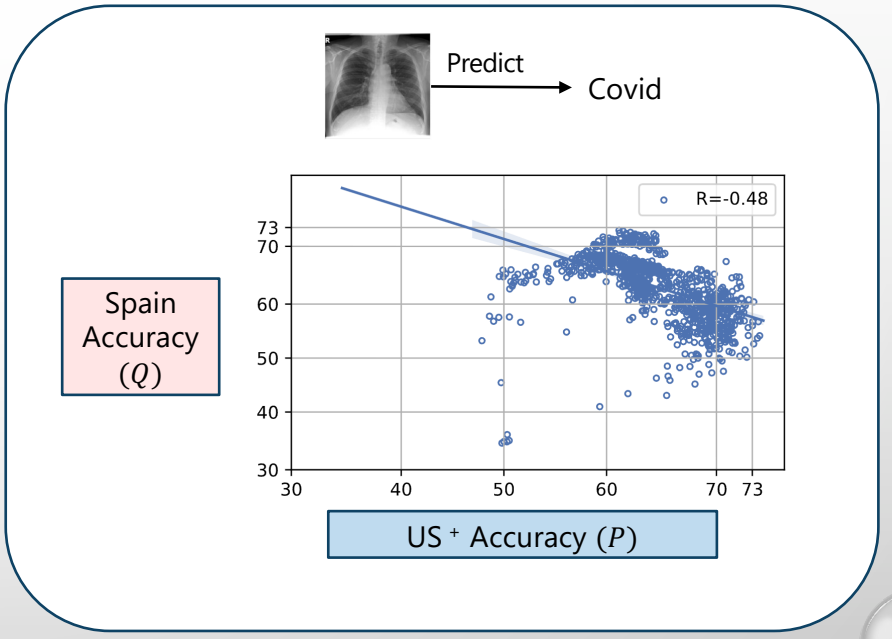

Where models are developed (e.g., GPT Health).
Growing gap!
Where they are used (e.g., your local hospital).

Salaudeen, Olawale, et al. "Are Domain Generalization Benchmarks with Accuracy on the Line Misspecified?." *Transactions on Machine Learning Research* (2025).
Teney, Damien, et al. "ID and OOD performance are sometimes inversely correlated on real-world datasets." Advances in Neural Information Processing Systems 35 (2022).
Sanyal, Amartya, et al. "Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation." International Conference on Artificial Intelligence and Statistics (2025).

CONTEXT

TAKE A FIXED SET OF MODELS THAT IMPROVE ON ONE DISTRIBUTION P, DO THEY ALSO IMPROVE ON A NEW DISTRIBUTION Q

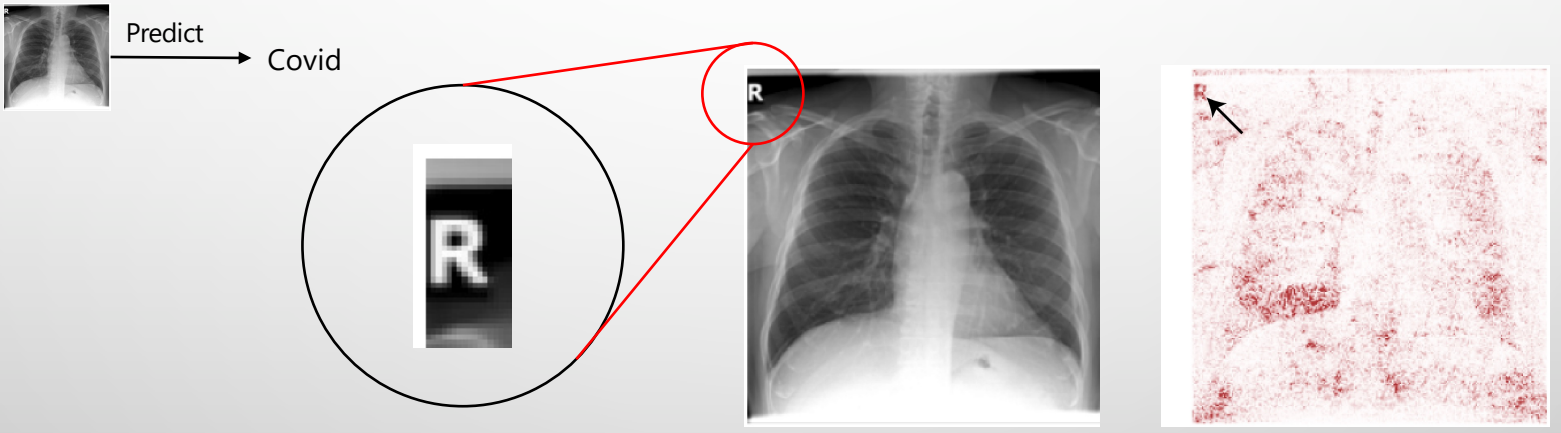
- NOTE, WE ARE NOT APPLYING METHODS, JUST EVALUATING MODELS
- IT DEPENDS ON P AND Q



DeGrave et al. "AI for radiographic COVID-19 detection selects shortcuts over signal." *Nature Machine Intelligence* 3.7 (2021): 610-619.

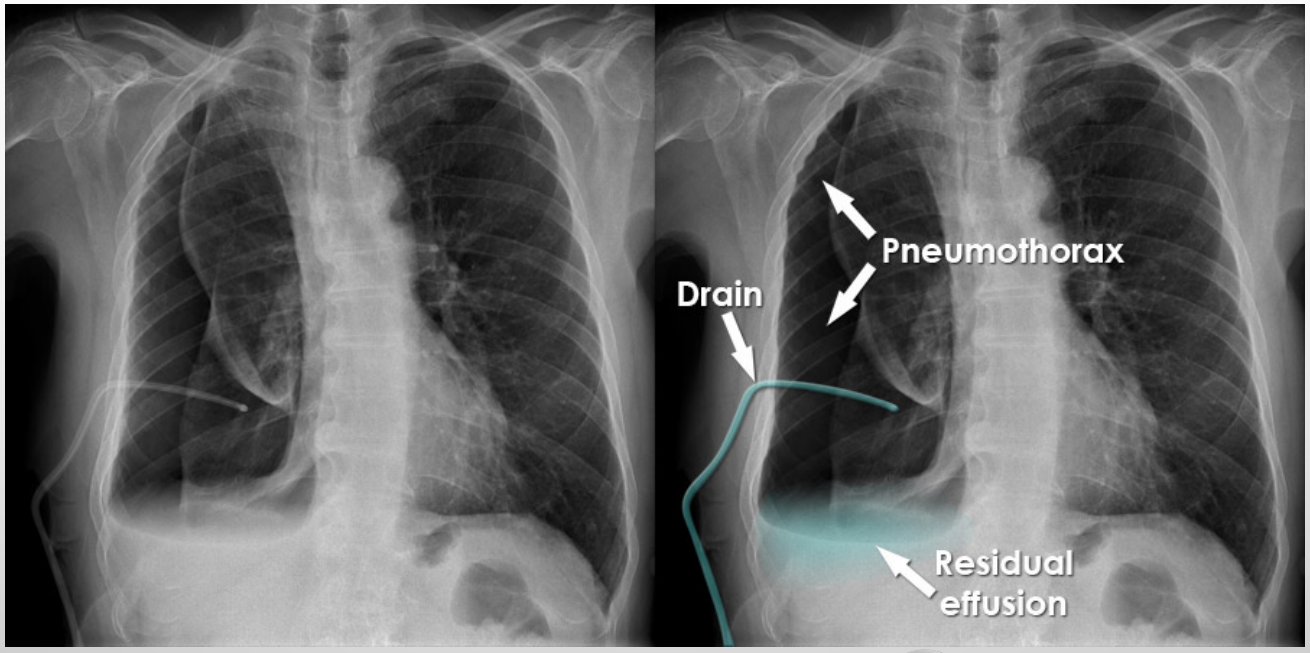
ONE EXAMPLE OF WHEN THIS MIGHT OCCUR

YOUR MODEL IS GOOD ON P FOR THE WRONG REASON.



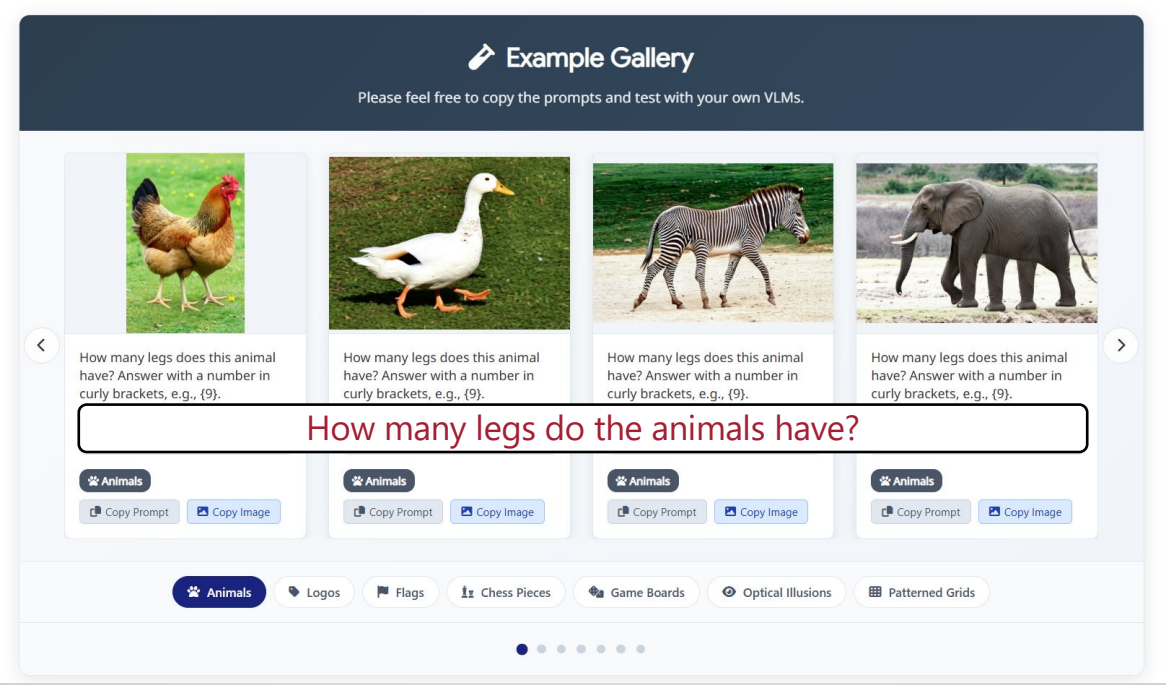
Oakden-Rayner, Luke, et al. "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging." Proceedings of the ACM conference on health, inference, and learning. 2020.

ANOTHER EXAMPLE OF WHEN THIS MIGHT OCCUR



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

YET ANOTHER EXAMPLE



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

Salaudeen, Olawale, et al. "Measurement to meaning: A validity-centered framework for ai evaluation." arXiv preprint arXiv:2505.10573 (2025).

YOUR INTUITION ABOUT WHAT YOU ARE MEASURING IS PROBABLY WRONG!!!

A FRAMEWORK FOR GROUNDING MODEL EVALUATION

WHAT DO YOU WANT TO MEASURE?

Criterion: Observable phenomenon, e.g., *in-hospital mortality prediction accuracy, transcription accuracy, etc.*

Construct: Fuzzy and abstract concepts, e.g., *medical reasoning, patient wellness, quality of life*

IS YOUR MEASUREMENT (INSTRUMENT) SUFFICIENT?

CONTENT VALIDITY

DOES THE EVALUATION ADEQUATELY COVER THE NECESSARY DOMAIN?

A CLINICAL CLAIM USUALLY IMPLIES BROAD USEFULNESS. IF AN EVALUATION DATASET ONLY CONTAINS DATA FROM YOUNG, HEALTHY ADULTS OR LACKS SEVERE DISEASE PRESENTATIONS, IT HAS LOW CONTENT VALIDITY.

ANY EXAMPLES?

ASSESSING CONTENT VALIDITY

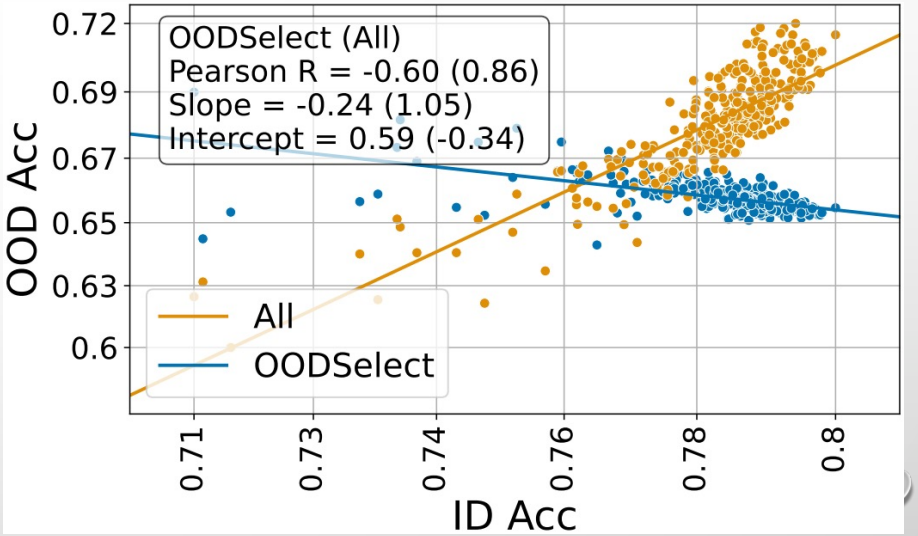
- **EPIDEMIOLOGICAL MAPPING:** COMPARING THE DISTRIBUTION OF THE TEST SET AGAINST THE REAL-WORLD PREVALENCE OF THE DISEASE.
- **DISAGGREGATED METRICS:** SLICING PERFORMANCE ACROSS INTERSECTIONAL DEMOGRAPHICS, COMORBIDITIES, AND SEVERITY LEVELS.
- **EXPERT DOMAIN REVIEW:** HAVING CLINICIANS AUDIT THE INCLUSION AND EXCLUSION CRITERIA OF THE BENCHMARK DATASET.

WHAT ELSE?

Salaudeen, Olawale, et al. "Aggregation hides out-of-distribution generalization failures from spurious correlations." *Advances in Neural Information Processing Systems* 38 (2026): 94160-94199.

AGGREGATION HIDES OOD FAILURES

- **THE OBSERVATION:** AGGREGATE OOD TRENDS ARE DRIVEN BY THE AVERAGE SHIFT ACROSS MANY ENVIRONMENTS.
- **THE FINDING:** WITHIN A SPECIFIC ENVIRONMENT (A SINGLE HOSPITAL OR DEMOGRAPHIC), THE CORRELATION CAN BE ZERO OR NEGATIVE.



CRITERION / PREDICTIVE VALIDITY

DOES THE BENCHMARK PERFORMANCE ACTUALLY PREDICT A VALIDATED MEASURE OF DOWNSTREAM UTILITY?

A MODEL MIGHT HAVE A HIGH AUROC ON A RETROSPECTIVE DATASET, BUT DOES THAT SCORE CORRELATE WITH BETTER CLINICAL DECISIONS IN THE WORKFLOW?

PREDICTIVE VALIDITY REQUIRES EVIDENCE THAT THE MEASURED SCORE TRANSLATES INTO THE INTENDED REAL-WORLD IMPACT (E.G., REDUCED TIME-TO-TREATMENT OR IMPROVED PATIENT OUTCOMES).

ANY EXAMPLES?

ASSESSING CRITERION / PREDICTIVE VALIDITY

- **SHADOW MODE VALIDATION:** RUNNING THE MODEL SILENTLY ON LIVE HOSPITAL DATA TO COMPARE PREDICTIONS AGAINST ACTUAL CLINICIAN DECISIONS.
- **CLINICIAN-IN-THE-LOOP A/B TESTING:** MEASURING WHETHER DOCTORS ASSISTED BY THE AI MAKE BETTER DECISIONS THAN THOSE WITHOUT IT.
- **DECISION CURVE ANALYSIS:** EVALUATING THE NET CLINICAL BENEFIT OF THE MODEL ACROSS DIFFERENT RISK THRESHOLDS.

WHAT ELSE?

EXTERNAL VALIDITY

DOES THE ENTIRE VALIDITY ARGUMENT HOLD ACROSS DIFFERENT SETTINGS?

IF A MODEL IS PROVEN SAFE AND EFFECTIVE IN HOSPITAL A, CAN WE TRANSFER THAT CLAIM TO HOSPITAL B?

EXTERNAL VALIDITY REQUIRES UNDERSTANDING THE SCOPE CONDITIONS OF THE MODEL. WHAT DEMOGRAPHIC, TEMPORAL, OR WORKFLOW SHIFTS CAUSE THE PRECEDING VALIDITY ARGUMENTS TO BREAK DOWN?

ANY EXAMPLES?

ASSESSING EXTERNAL VALIDITY

- **MULTI-CENTER VALIDATION:** TESTING THE MODEL ON DISTINCT GEOGRAPHIC POPULATIONS AND DIFFERENT HOSPITAL SYSTEMS.
- **TEMPORAL VALIDATION:** TRAINING ON HISTORICAL DATA (E.G., 2018-2022) AND EXPLICITLY TESTING ON FUTURE DATA (E.G., 2024-2026) TO CHECK FOR PRACTICE DRIFT.
- **HARDWARE SHIFT TESTING:** EVALUATING PERFORMANCE ACROSS DIFFERENT ACQUISITION DEVICES (E.G., SIEMENS VS. GE SCANNERS).

WHAT ELSE?

CONSTRUCT VALIDITY

ARE WE ACTUALLY MEASURING THE LATENT TRAIT OR THE EXTERNAL CONCEPT WE THINK WE ARE MEASURING?

SHORTCUTS ARE AN EXAMPLE OF A THREAT TO THIS. *IF A MODEL PREDICTS PNEUMONIA BY RELYING ON THE SPECIFIC TYPE OF X-RAY SCANNER USED RATHER THAN LUNG PATHOLOGY, THE BENCHMARK SCORE IS MEASURING A SPURIOUS ARTIFACT, NOT THE CONSTRUCT OF "DIAGNOSTIC CAPABILITY."*

ANY EXAMPLES?

ASSESSING CONSTRUCT VALIDITY

- **EXPERT ASSESSMENT:** ENGAGING CLINICIANS TO REVIEW MODEL EXPLANATIONS, FEATURE IMPORTANCE, AND FAILURE MODES TO VERIFY ALIGNMENT WITH ESTABLISHED PATHOPHYSIOLOGICAL REASONING RATHER THAN SPURIOUS CORRELATIONS.
- **COUNTERFACTUAL TESTING:** ALTERING BACKGROUND ARTIFACTS (LIKE REMOVING A SURGICAL MARKER) TO SEE IF THE PREDICTION FLIPS.
- **SALIENCY MAPS & FEATURE ATTRIBUTION:** VISUALIZING WHAT PIXEL REGIONS DRIVE THE MODEL'S DECISION.
- **CONCEPT BOTTLENECK MODELS:** FORCING THE MODEL TO PREDICT INTERPRETABLE CLINICAL CONCEPTS BEFORE MAKING A FINAL DECISION.

WHAT ELSE?

CONSEQUENTIAL VALIDITY

WHAT ARE THE SYSTEMIC EFFECTS AND POTENTIAL HARMS OF DEPLOYING THE MODEL?

DEPLOYMENT CHANGES THE ENVIRONMENT. EVEN A MODEL WITH HIGH PREDICTIVE VALIDITY CAN FAIL HERE.

DOES THE MODEL CAUSE CLINICIAN ALERT FATIGUE? DOES IT INADVERTENTLY ALLOCATE RESOURCES AWAY FROM VULNERABLE POPULATIONS? CONSEQUENTIAL VALIDITY DEMANDS EVIDENCE THAT THE OVERALL SOCIAL AND SYSTEMIC IMPACT OF USING THE MODEL JUSTIFIES ITS DEPLOYMENT.

ANY EXAMPLES?

ASSESSING CONSEQUENTIAL VALIDITY

- **WORKFLOW AUDITING:** SURVEYING CLINICIANS TO MEASURE COGNITIVE LOAD, AUTOMATION BIAS, AND ALERT FATIGUE POST-DEPLOYMENT.
- **RESOURCE ALLOCATION TRACKING:** MONITORING IF THE ALGORITHM SYSTEMATICALLY SHIFTS CARE ACCESS OR WAIT TIMES AWAY FROM MARGINALIZED GROUPS.
- **POST-MARKET SURVEILLANCE:** CONTINUOUS TRACKING OF LONGITUDINAL PATIENT OUTCOMES TO CATCH DELAYED SYSTEMIC HARMS.

WHAT ELSE?

Alaa, Ahmed, et al. "Position: Medical large language model benchmarks should prioritize construct validity." Forty-second International Conference on Machine Learning Position Paper Track. 2025.

A HEALTHCARE EXAMPLE

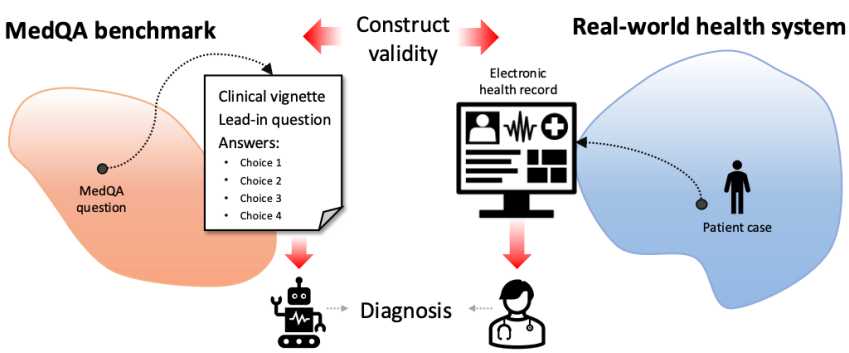


Figure 4. **Empirical validation of the MedQA benchmark using EHR data.** Each MedQA item consists of a clinical vignette, a question, and multiple-choice answers, while each EHR patient case includes a clinical note and a corresponding clinical decision. To assess the validity MedQA, we empirically test whether strong performance on benchmark items reflects the ability of an LLM to encode and apply medical knowledge in real-world practice.

	MedQA	**Real-world data**	
	Accuracy	*Accuracy*	*α*
Llama 3	0.54	0.48	0.56
GPT-4	0.71	0.28	0.29
Chimera Llama	0.60	0.45	0.48
Biomerge	0.57	0.36	0.49
Orpomed	0.49	0.24	0.38
JSL MedLlama	0.61	0.37	0.49
PMY MedLLama	0.75	0.36	0.45

Table 1. Predictive validity of the MedQA benchmark.

α = P(Correct on real−world case | Correct on MedQA)

Salaudeen, Olawale, et al. "Measurement to meaning: A validity-centered framework for ai evaluation." arXiv preprint arXiv:2505.10573 (2025).

A REPORT CARD

GOOGLE PROOF QUESTION ANSWERING (GPQA). A SET OF QUESTIONS IN BIOLOGY, PHYSICS, AND CHEMISTRY, CURATED BY EXPERTS, THAT ONLY OTHER EXPERTS CAN ANSWER.

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 *graduate-level specialized multiple-choice questions* in biology, physics, and chemistry.	OK	OK	OK	OK	⚠
2. AI systems can accurately answer *graduate-level specialized questions* in specialized scientific domains.	⚠	⚠	⚠	⚠	⚠
3. AI systems can exhibit *general reasoning abilities* that can transfer beyond current human specialization.	⚠	✗	✗	✗	⚠

PUTTING IT ALL TOGETHER

VALIDITY AS A BURDEN OF PROOF

- **A MODEL-LEVEL CLAIM:** *"THIS MODEL PREDICTS SEPSIS WITH HIGH AUROC ON SIMILAR DISTRIBUTIONS. (REQUIRES **CONTENT AND EXTERNAL** VALIDITY EVIDENCE)"*
- **A UTILITY CLAIM:** *"THIS MODEL IMPROVES WELL-DEFINED CLINICAL DECISIONS." (REQUIRES **CONTENT, EXTERNAL, PREDICTIVE** VALIDITY EVIDENCE)*
- **A DEPLOYMENT CLAIM:** "THIS MODEL SAFELY IMPROVES PATIENT OUTCOMES ACROSS THE HEALTH SYSTEM." *(REQUIRES **CONTENT, EXTERNAL, PREDICTIVE, CONSTRUCT, AND CONSEQUENTIAL** VALIDITY EVIDENCE)*

THE VALIDITY BURDEN INCREASES WITH THE STRENGTH OF THE CLAIM.

THE PROPER ROLE OF BENCHMARKS

- IF THE VALIDITY BURDEN FOR DEPLOYMENT IS SO HIGH, WHAT ARE BENCHMARKS ACTUALLY GOOD FOR?
- BENCHMARKS FACE AN INCREDIBLY HIGH BAR TO "MEAN" SOMETHING FOR CLINICAL DEPLOYMENT. HOWEVER, THEY ARE ESSENTIAL FOR TWO UPSTREAM TASKS:

1. **HILL-CLIMBING:** ITERATIVE METHOD DEVELOPMENT. BENCHMARKS ARE THE ENGINE FOR OPTIMIZING PROCEDURES AND COMPARING METHODS
2. **TRIAGING:** FILTERING OUT BAD ARTIFACTS. IF A MODEL FAILS ON A WELL-CONSTRUCTED BENCHMARK, IT IS NOT WORTH THE COST AND RISK OF CLINICAL VALIDATION.

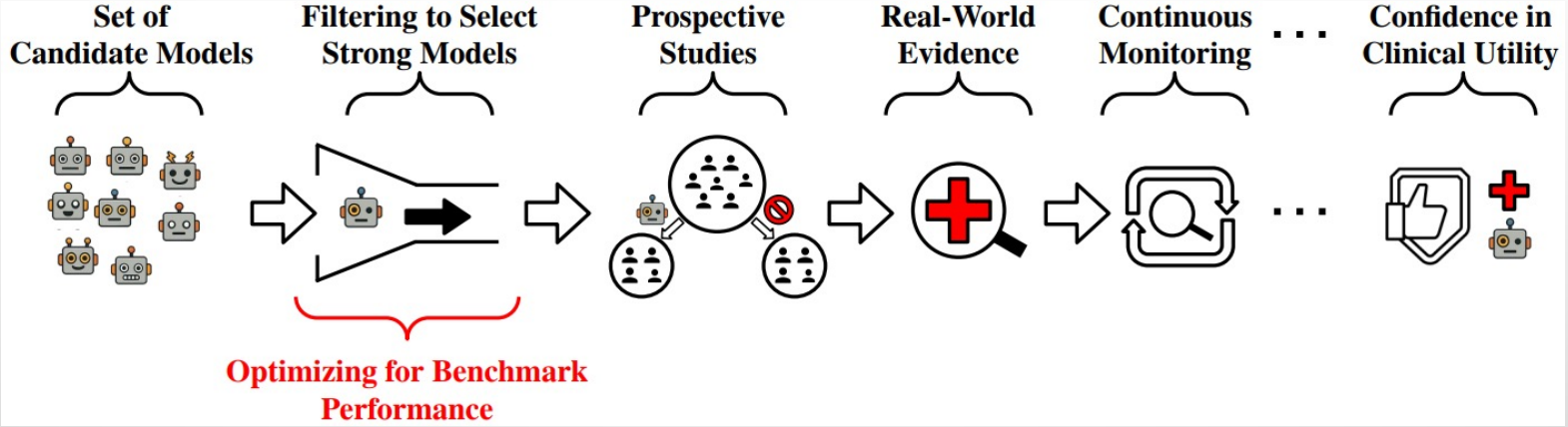
Raji, Inioluwa Deborah, Roxana Daneshjou, and Emily Alsentzer. "It's time to bench the medical exam benchmark." Nejm Ai 2.2 (2025): Ale2401235.

YOU CANNOT BENCHMARK YOUR WAY TO VALID CLAIMS ABOUT CLINICAL USE

- BENCHMARKS ABSTRACT AWAY THE ENVIRONMENT TO CREATE STANDARDIZED MEASUREMENTS. BUT CLINICAL DEPLOYMENT REQUIRES EVALUATING THE ARTIFACT *INSIDE* THE ENVIRONMENT.
- A BENCHMARK CAN TELL YOU WHICH METHODS ARE MOST EFFECTIVE WHEN APPLIED TO YOUR SETTING OR WHICH MODELS YOU SHOULD RULE OUT, IF NEEDED,
- BUT A BENCHMARK **CANNOT** VALIDATE WORKFLOW FIT, CLINICIAN RELIANCE, OR PATIENT OUTCOMES. THOSE REQUIRE EXITING THE BENCHMARK PARADIGM ENTIRELY.

Zhang, Haoran, et al. "Benchmarks Do Not Measure Deployment Readiness in Clinical AI." Forty-third International Conference on Machine Learning Position Paper Track. (2026)

YOU CANNOT BENCHMARK YOUR WAY TO VALID CLAIMS ABOUT CLINICAL USE



BENCHMARK OPTIMIZATION IS ONLY THE FIRST FILTER; CLINICAL UTILITY REQUIRES PROSPECTIVE STUDIES, REAL-WORLD EVIDENCE, AND CONTINUOUS MONITORING.

BEYOND THE BENCHMARK: EVALUATING MODELS

TO VALIDLY EVALUATE MODELS FOR CLINICAL USE, WE NEED THE "OTHER THINGS" OUTSIDE THE HOLDOUT SET:

- **LOCAL RETROSPECTIVE VALIDATION:** TESTING THE ARTIFACT ON THE SPECIFIC HOSPITAL SYSTEM'S HISTORICAL DATA TO ESTABLISH A LOCAL BASELINE.
- **SHADOW MODE / PROSPECTIVE EVALUATION:** RUNNING THE MODEL SILENTLY ON LIVE DATA TO TEST PREDICTIVE VALIDITY AGAINST ACTUAL, REAL-TIME CLINICIAN DECISIONS.
- **HUMAN-AI INTERACTION TRIALS:** MEASURING HOW THE ARTIFACT ACTUALLY ALTERS CLINICIAN BEHAVIOR (CONSEQUENTIAL VALIDITY).
- **CONTINUOUS MONITORING INFRASTRUCTURES:** TRACKING PERFORMANCE POST-DEPLOYMENT TO CATCH THE INEVITABLE DISTRIBUTION SHIFTS.

WHAT ELSE?

YOU BUILD METHODS ON BENCHMARKS. YOU VALIDATE MODELS IN THE REAL WORLD.

TAKEAWAYS

- **DISCOVERY VS. VERIFICATION:** BENCHMARKS ARE FOR FINDING THE BEST *PROCEDURE* (DISCOVERY); CLINICAL EVALUATION IS FOR VERIFYING THE SPECIFIC *ARTIFACT* (VERIFICATION).
- **THE LIMITS OF THE HOLDOUT SET:** YOU CANNOT BENCHMARK YOUR WAY TO CLINICAL UTILITY. TASK PERFORMANCE IS A NARROW CRITERION; CLINICAL USEFULNESS IS A BROAD CONSTRUCT THAT INCLUDES WORKFLOW AND HUMAN INTERACTION.
- **THE VALIDITY BURDEN:** AS A CLAIM MOVES FROM "IT PERFORMS WELL ON A TEST SET" TO "IT SHOULD BE USED ON PATIENTS," THE BURDEN OF PROOF SHIFTS FROM SCORES TO A MULTI-FACETED VALIDITY ARGUMENT.

WHAT ELSE?

CLOSING: MOVING BEYOND THE SCORE

BEFORE INTERPRETING ANY AI RESULT, WE MUST SHIFT THE CONVERSATION FROM "WHAT IS THE ACCURACY?" TO "WHAT IS THE OBJECT AND THE CLAIM?"

- ARE WE EVALUATING A **METHOD** (FOR PROGRESS) OR A **MODEL** (FOR USE)?
- IS THE BENCHMARK USED FOR **HILL-CLIMBING**, OR IS IT BEING OVER-INTERPRETED AS **CLINICAL USE READINESS PROOF**?
- DOES THE EVIDENCE SUPPORT A NARROW **CRITERION** OR A BROAD **CLINICAL CONSTRUCT**?

AGAIN: YOU BUILD METHODS ON BENCHMARKS. YOU VALIDATE MODELS IN THE REAL WORLD.

THANK YOU! **QUESTIONS AND COMMENTS?**

OLAWALE SALAUDEEN

OLAWALE@MIT.EDU; HTTPS://WWW.OLAWALESALAUDEEN.COM