Testing and Evaluating Large AI models: Current Trends and Future Work

Emily Casleton

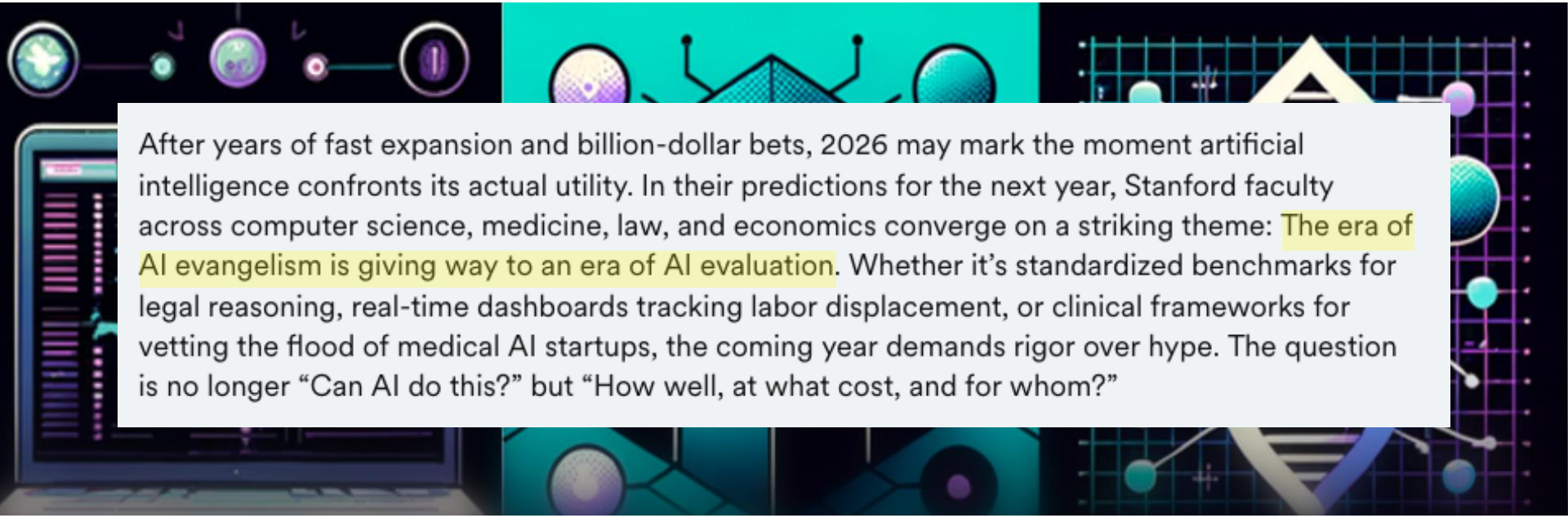
February 18, 2026

LA-UR-26-21000

About Me: Statistician who likes being outdoors

- **What statistics training develops:**
  - Healthy skepticism toward claims lacking evidence
  - Reflexive questioning of causal claims from correlational data
  - Low tolerance for statistical innumeracy in public discourse
- **How this might be perceived:**
  - Pointing out flaws in studies = being negative
  - Questioning claims = being difficult
  - Insisting on precision = being pedantic
  - Seeing uncertainty everywhere = being pessimistic
- **Some statisticians do become genuinely cynical—not because of the math, but from repeatedly seeing:**
  - Misuse of statistics to mislead
  - Cherry-picked data in important decisions
  - Resistance to evidence when it's inconvenient

The importance of Testing and Evaluation is finally being recognized



After years of fast expansion and billion-dollar bets, 2026 may mark the moment artificial intelligence confronts its actual utility. In their predictions for the next year, Stanford faculty across computer science, medicine, law, and economics converge on a striking theme: The era of AI evangelism is giving way to an era of AI evaluation. Whether it's standardized benchmarks for legal reasoning, real-time dashboards tracking labor displacement, or clinical frameworks for vetting the flood of medical AI startups, the coming year demands rigor over hype. The question is no longer "Can AI do this?" but "How well, at what cost, and for whom?"

https://hai.stanford.edu/news/stanford-ai-experts-predict-what-will-happen-in-2026
Released December 15, 2025

The importance of Testing and Evaluation is finally being recognized

Foundation models represent a potentially transformative technology for progressing scientific discovery and innovation. However, their rapid adoption has raised questions and concerns about their reliability, validity, and reproducibility. In 2024, the Department of Energy (DOE) requested that the National Academies of Sciences, Engineering, and Medicine conduct a study to consider current foundation models' capabilities, and future possibilities and challenges.

The National Academies established the Committee on Foundation Models for Scientific Discovery and Innovation to conduct this study. The study compares foundation models with more traditional computational methods, addresses exemplar use cases of foundation models, specifies strategic considerations, and outlines challenges for the development and use of foundation models. The full statement of the committee's task is shown in Appendix A.

The committee met in person in March 2025 and met virtually 15 times to receive briefings from experts and stakeholders (for a list of presentations, see Appendix B), review relevant reports and technical literature, deliberate, and develop this report.

The committee is grateful for the support of DOE's Office of Science, Office of Biological and Environmental Research, and National Nuclear Security Administration. The committee also extends its sincere thanks to the following National Academies' staff for their assistance throughout the study: Blake Reichmuth, Thơ Nguyễn, Erik Svedberg, Sam Koretsky, Jon Eisenberg, and Michelle Schwalbe.

Dona Crawford, *Chair*
Committee on Foundation Models for Scientific Discovery and Innovation
October 2025

National Academies of Sciences, Engineering, and Medicine. (2025). *Foundation models for scientific discovery and innovation: Opportunities across the Department of Energy and the scientific enterprise*. The National Academies Press. https://doi.org/10.17226/29212

But the needs for testing and evaluation for the US Department of Energy (DOE) are unique

Applying foundation models within the Department of Energy's (DOE's) missions presents a multilayered set of technical and operational challenges. These models, which emerged from success in domains such as natural language processing and vision, struggle to transfer directly into DOE's computational science workflows that require physical consistency, mesh- or geometry-aware representations, and scalable inference across high-dimensional, multiscale partial differential equation systems (Pyzer-Knapp et al. 2025). DOE applications

cant burdens on DOE facilities (Koch et al. 2025). Addressing these challenges will require coordinated investments in energy-efficient and sustainable foundation model development, physically informed architectures, domain-specific VVUQ methodologies, and infrastructure for transparent, traceable, and reproducible deployment across DOE's science and national security missions (Teranishi et al. 2025).

National Academies of Sciences, Engineering, and Medicine. (2025). *Foundation models for scientific discovery and innovation: Opportunities across the Department of Energy and the scientific enterprise*. The National Academies Press. https://doi.org/10.17226/29212

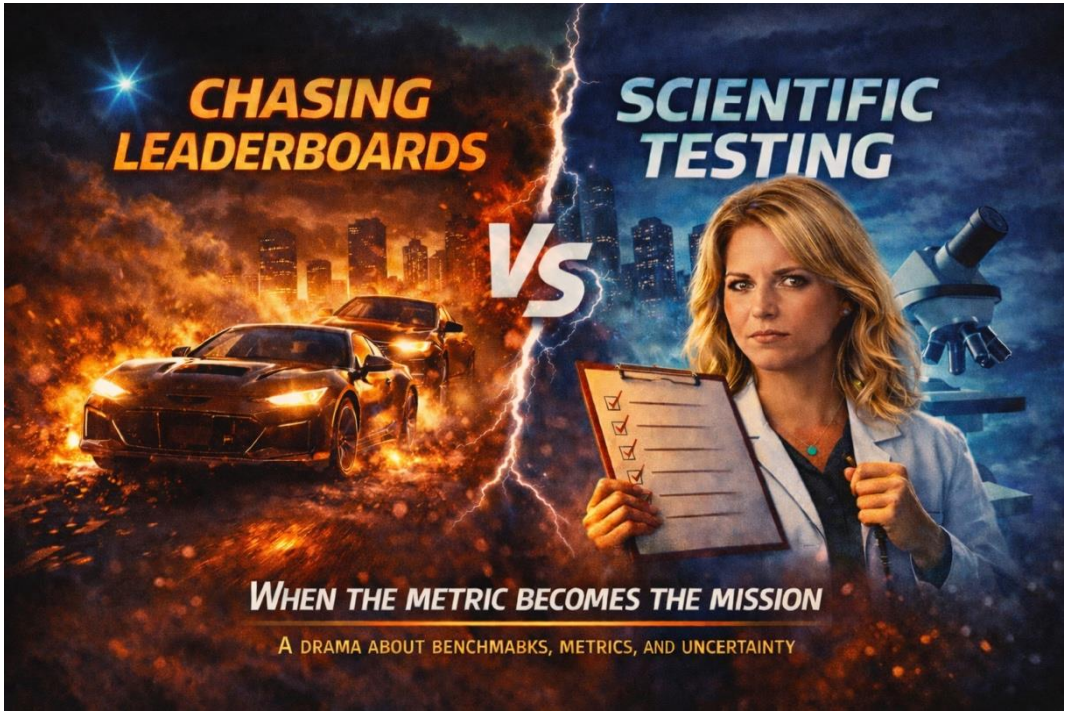
What is Evaluation? Core Aspect of Evaluation is Prediction

- We evaluate to ***predict*** if the system is suitable for its intended purpose.
- Model evaluation is predicting that the system will perform at an acceptable level in the real world on a potentially new task, on average.

"Evaluating AI Evaluation: Perils and Prospects" by John Burden https://arxiv.org/pdf/2407.09221

AI as a field has over-optimized for better benchmarking results or topping a leaderboard

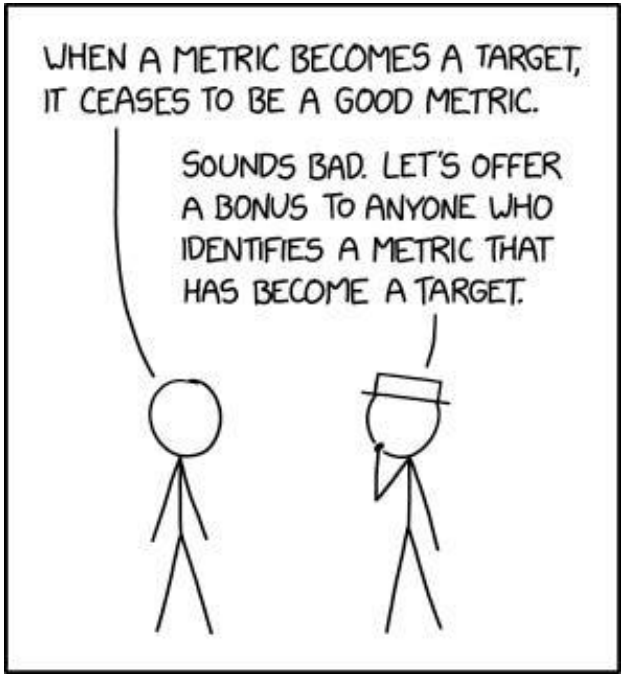
- Chasing to top of a leaderboard vs. scientific testing
- Metrics that do not inform on the scientific question of interest and lack of uncertainty in metric values or attribution of performance to meta-data can lead to a false perception of model dominance



New Metrics for Scientific Applications

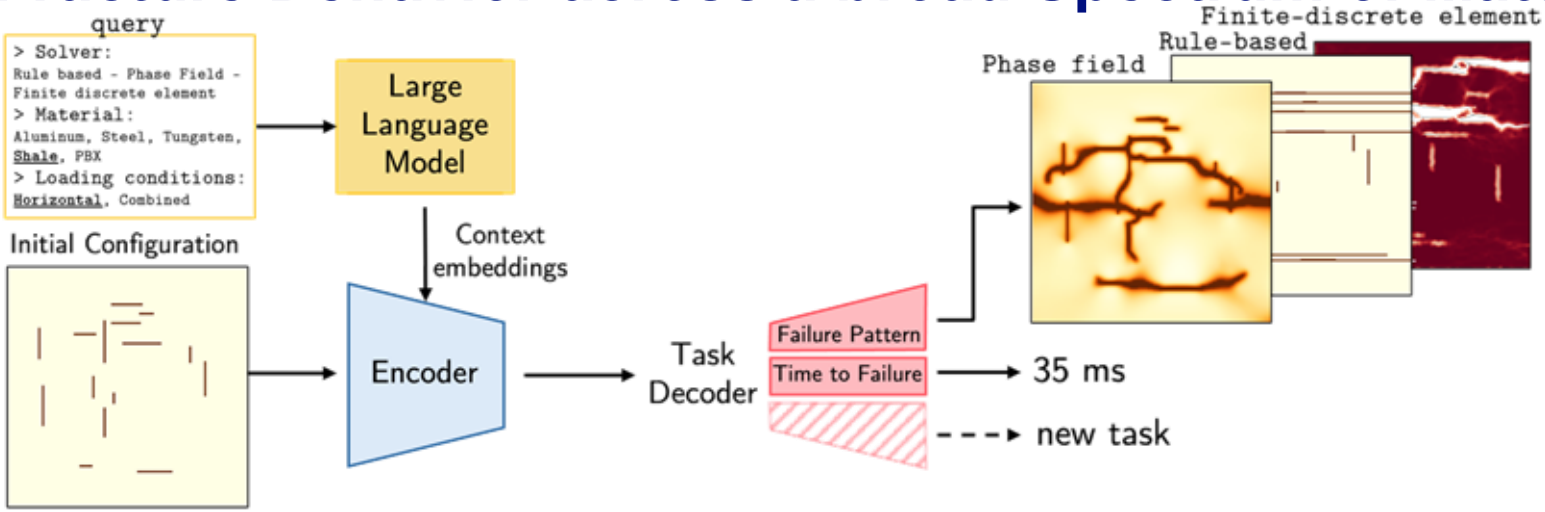
Bespoke metrics for more informed results and more efficient training

- MSE as a default. Little thought is often put towards the metrics that are used to test large AI models or the cost function used to train them
- The process of defining better metrics force the user to define the question(s) of interest and then allow a user to assess ***if a model is appropriately answering those questions***. It can also provide insight into failure modes.
- This requires **intimate collaboration between T&E and model builders and subject matter experts.**



https://xkcd.com/2899/

TellusFM Foundation Model Designed to Predict Fracture Behavior across a broad Spectrum of Materials

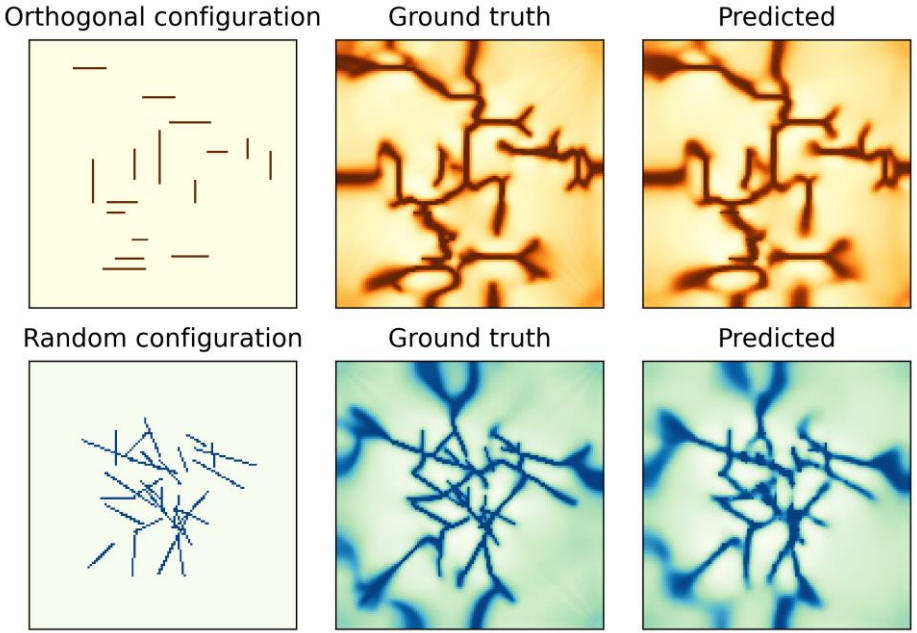


- **Inputs**: initial fracture + LLM-based context embeddings
- **Outputs**: predicted fracture pattern and time-to-failure

- Encoder-decoder transformer architecture
- Handles both structured and unstructured grids

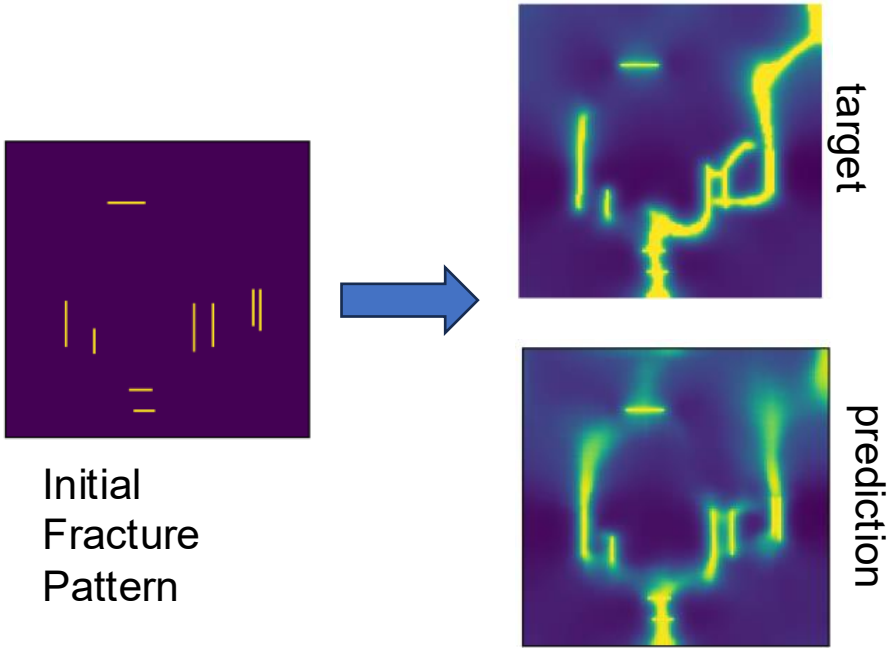
Hosted on HuggingFace:
https://huggingface.co/datasets/smartFRACs/material_fracturing/tree/main

Some example simulation and model outputs for various initial configurations



Slide credit: Aleksandra Pachalieva, LANL

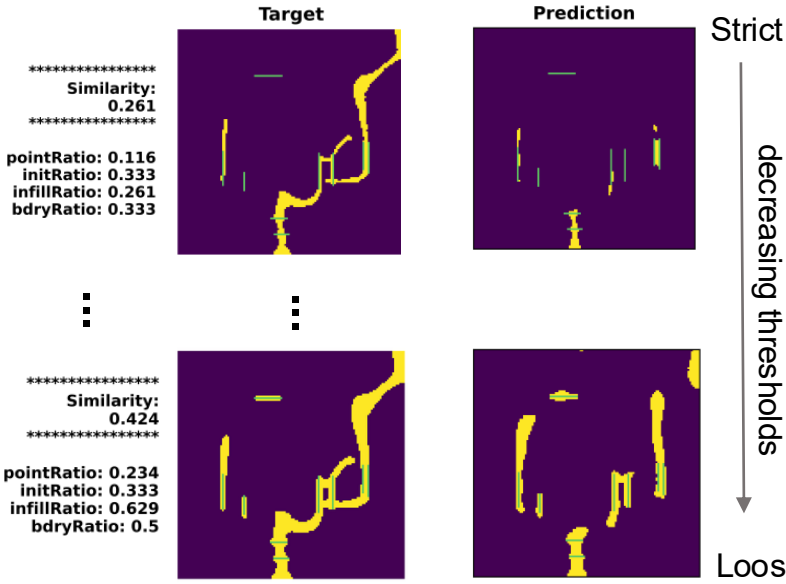
TopoGeo Similarity Metric for fracture patterns captures geometric and topological features



Note: Prediction has low MSE (0.016) but does not represent the target with a spanning crack that breaks the patch

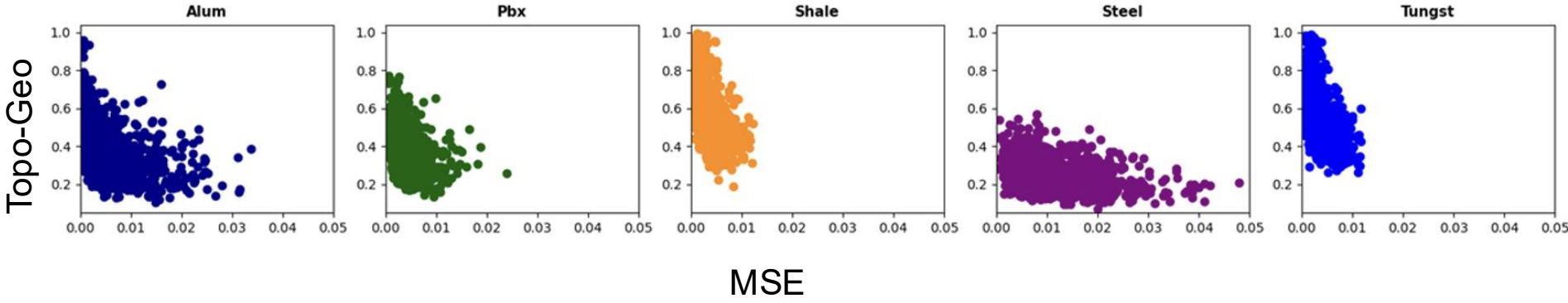
Topo-Geo compares largest connected crack component in the target (CC_T) and the prediction (CC_P)

1. Percentage of points in CC_T that are also in CC_P [pointRatio]
2. Percentage of initial cracks absorbed into CC_T that are also absorbed into CC_P [initRatio]
3. Percentage of points in CC_T that must be added somewhere into the prediction to recreate CC_T [infillRatio]
4. Percentage of boundaries contacted by CC_T that are also contacted by CC_P average with the proportion contacted by any point in the prediction [boundaryRatio]



Overall Similarity is computed as the average of 4 components

Topo-Geo similarity gives the insight that training on MSE does not always force similarity



Auto-regressive rollout: Surrogate for computationally expensive partial differential equations (PDEs)

- Definition: Given the initial state of the simulation, progress the simulation in time
- Promise: Computational speedup
- Issues:
  - Accumulate errors during prediction
  - Need to show can generalize beyond training data, not memorize
  - Need a metric that can distinguish performance

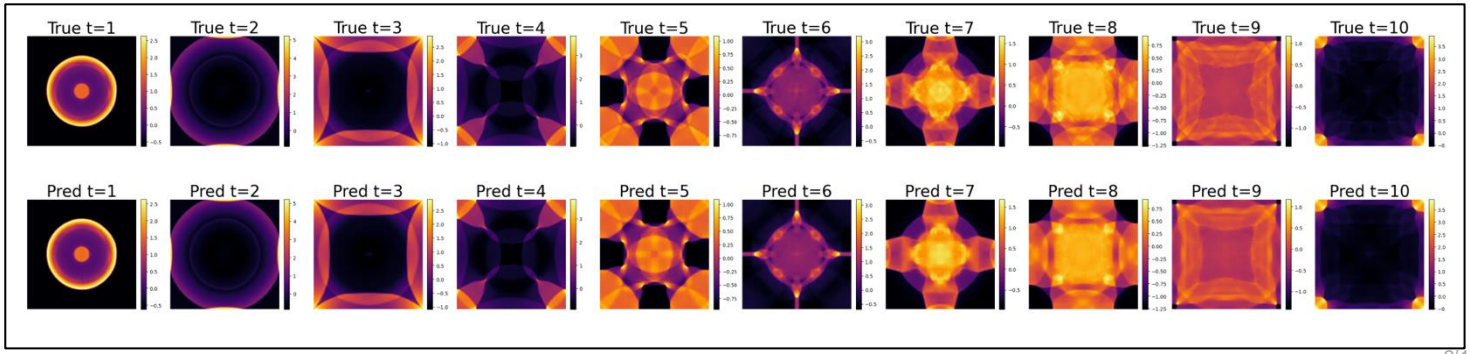
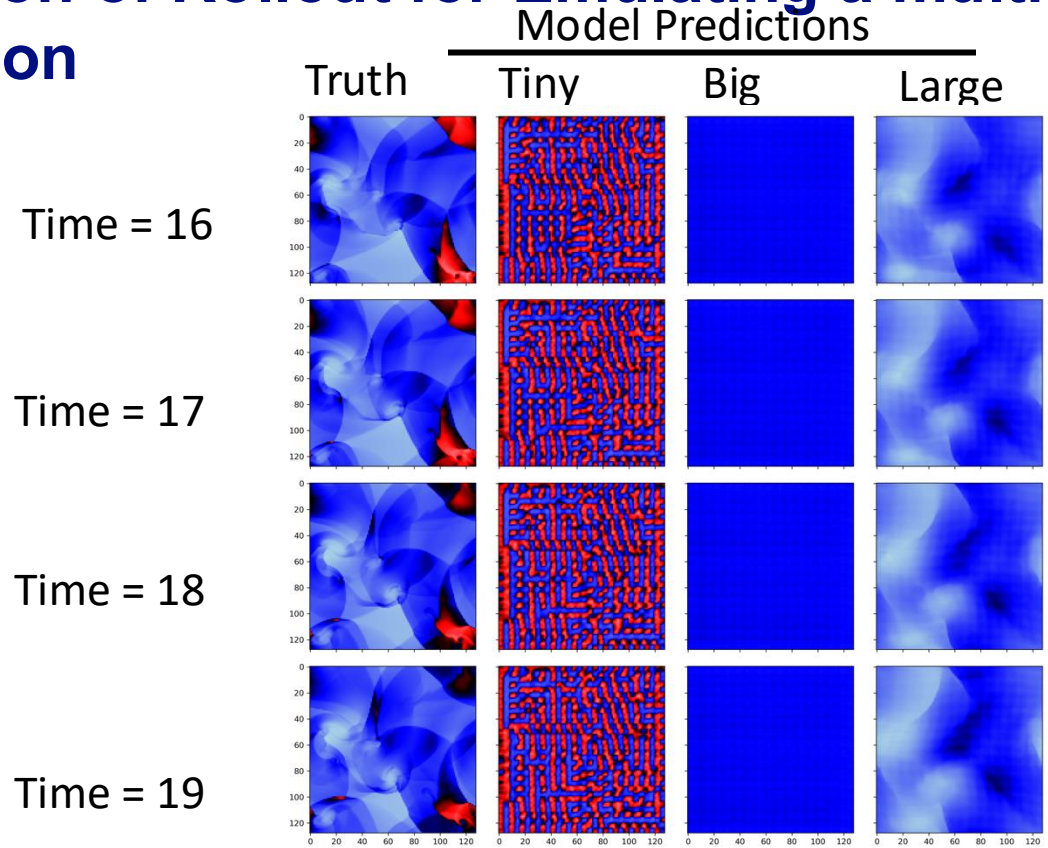


Image credit: Mahindra Rautela, LANL

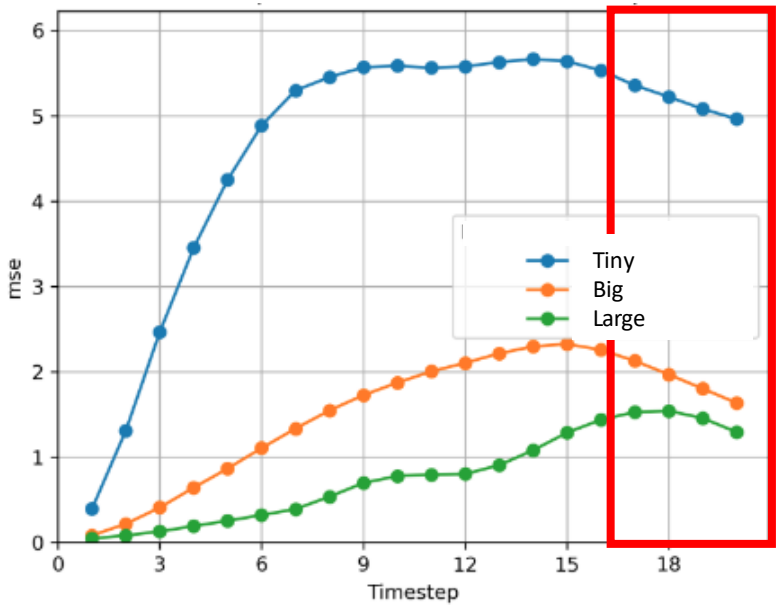
Evaluation of Rollout for Emulating a multi-physics Simulation



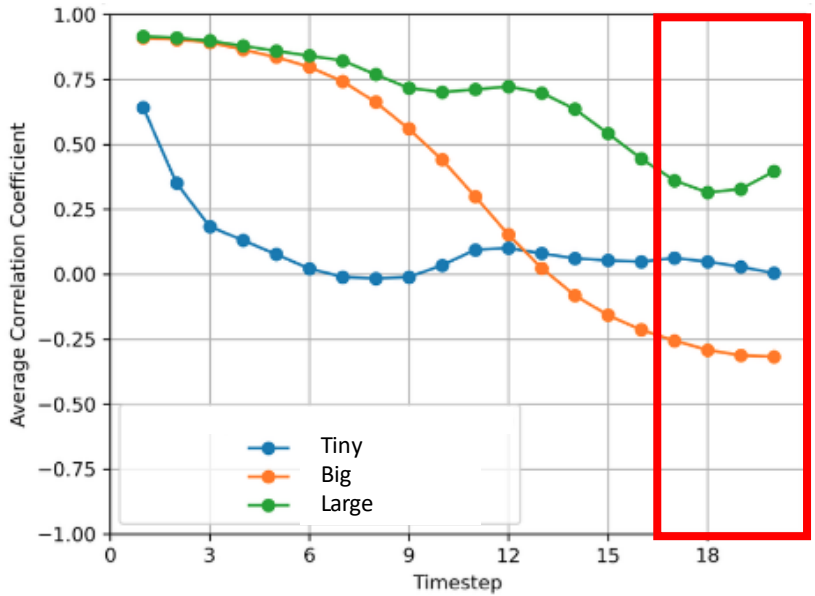
Herde, M., Raonic, B., Rohner, T., Käppeli, R., Molinaro, R., de Bézenac, E., & Mishra, S. (2024). Poseidon: Efficient foundation models for PDEs. *Advances in Neural Information Processing Systems*, *37*, 72525-72624.

Using pointwise correlation coefficient across multiple scales better matches intuition than MSE.

MSE

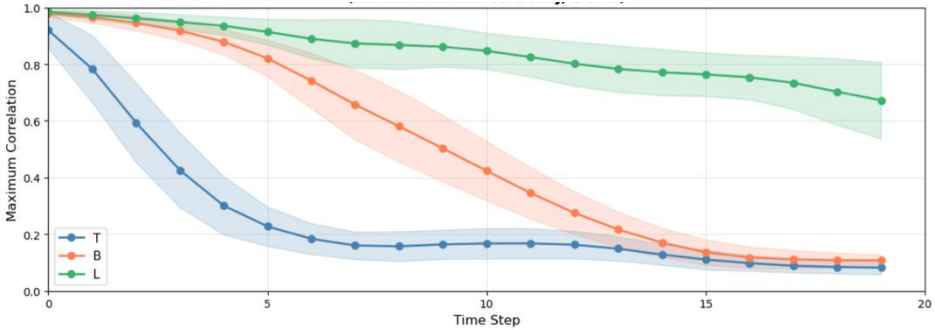


Correlation Coefficient

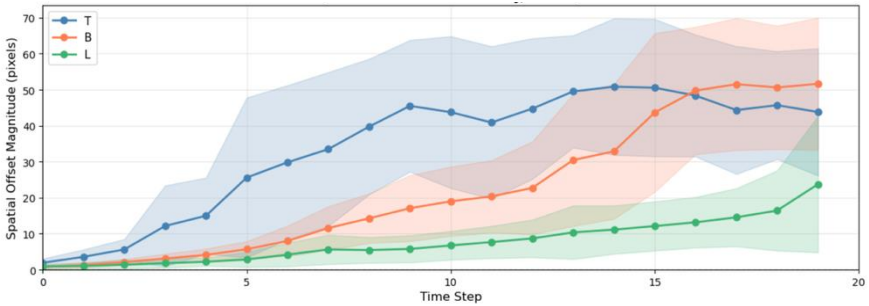


Three 2d-correlation Metrics for Rollout Validation

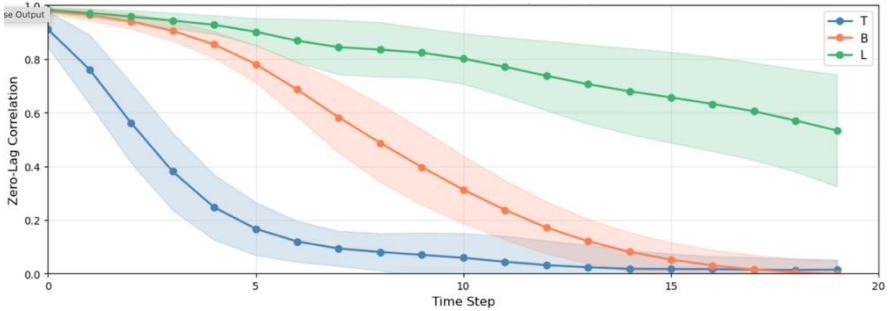
Maximum Correlation (ρ_{max})



Spatial Offset Magnitude ($|\Delta|$)



Zero-Lag Correlation (ρ_o)



The three 2d-correlation metrics have physical interpretation and model suggestions

Condition	Physical Interpretation	Action
$\rho_{max} \approx \rho_0 \approx 1$	**Excellent:** Right physics, right position	Continue
$\rho_{max} \approx 1$, $\rho_0 << \rho_{max}$	**Phase error:** Right physics, wrong timing	Correct advection/timing
$\rho_{max} < 0.9$, $\lvert\Delta\rvert$ small	**Amplitude error:** Wrong intensity/dissipation	Adjust diffusion/viscosity
$\rho_{max} < 0.7$	**Model failure:** Missing physics	Revise model

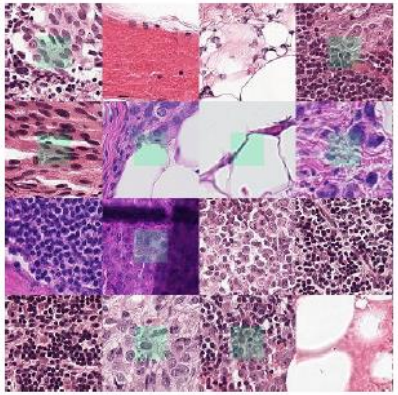
Uncertainty Quantification of Metrics

Visual Task Adaptation Benchmark (VTAB) – popular computer vision benchmark

Natural

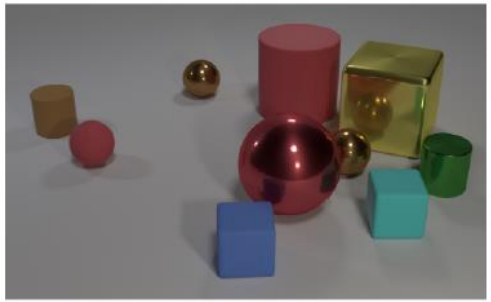
Parkhi et al., 2012

Specialized



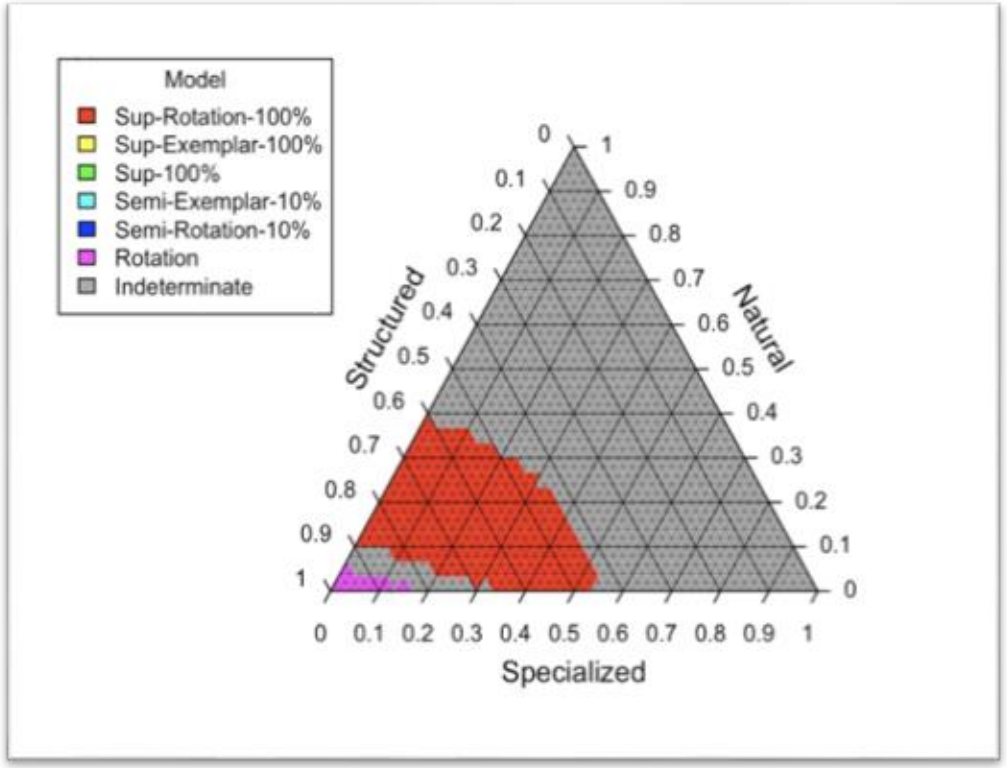
Veeling et al., 2018

Structured



Johnson et al., 2017

Uncertainty on Metric Values and Meta-Data Breakdown of Results

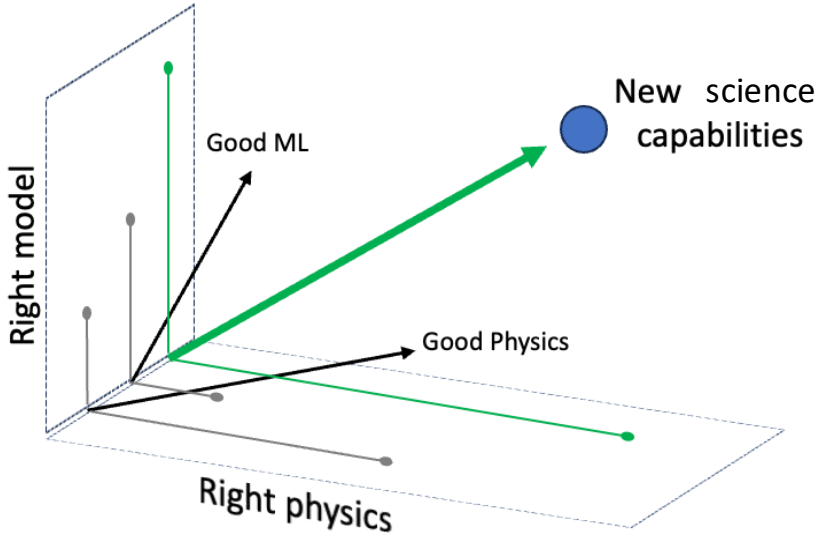


"Statistical Uncertainty Quantification for Aggregate Performance Metrics in Machine Learning Benchmarks" by Longjohn et.al. https://arxiv.org/pdf/2501.04234

We need both traditional ML metrics and physics-informed metrics

- Traditional ML scores tell us the model pipeline works
  - Crucial for SOTA deployments
  - Necessary step on the road to deployment
- Physics fidelity and rollout predictions tell us the science works
  - Often very expensive to implement as a cost function
  - Key to successful development of new science

Advanced T&E enables advanced science that is robust and trustworthy.



The Future of Evaluation: More emphasis on defining specified capabilities and holistic evaluations

- There is no one best way to evaluate all models.
- Correctness is just the tip of the iceberg.
- Evaluation should answer the question: ***"is this model appropriate for my application?"***

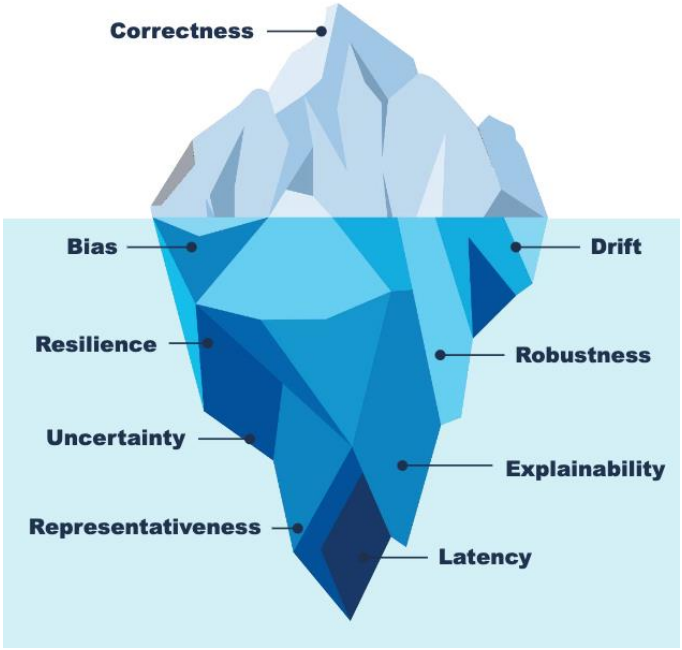


Figure from "Test and Evaluation of Artificial Intelligence Models" by CDAO, April 2024

Thank you!

Emily Casleton
ecasleton@lanl.gov