Poster

What Has a Foundation Model Found? Inductive Bias Reveals World Models

Keyon Vafa · Peter Chang · Ashesh Rambachan · Sendhil Mullainathan

West Exhibition Hall B2-B3 #W-1106

[Abstract] [Lay Summary]

[OpenReview]

Wed 16 Jul 11 a.m. PDT — 1:30 p.m. PDT

Michał Włodarczyk
20.10.2025r.

Authors

Keyon Vafa
Postdoc @ Harvard
PhD in CS @ **Columbia**
advised by **David Blei**
Interned at
- Google AI
- Facebook AI Research

Peter G. Chang
PhD Student @ MIT EECS, advised by **Sendhil Mullainathan**

Ashesh Rambachan
PhD Student @ MIT, advised by **Sendhil Mullainathan**

Sendhil Mullainathan
Professor at MIT, Spec. in Economics, Electrical Engineering and Computer Science

Launching
The Bike Shop @ MIT

First Author's Background

World Models

He want them to learn accurate world models. “While generative models are often trained to make accurate predictions, we often hope that models recover structure about the real world.”

Foundation Models for Statistical Estimation

He works on adapting foundation models and developing new fine-tuning procedures to address these goals.

Behavioral Machine Learning

He works on behavioral machine learning: incorporating insights from the behavioral sciences into formal, computational models in order to evaluate and improve.

His paper have been widely adapted and used as resources in New York Times and Nature articles.

Keyon Vafa
Postdoc @ Harvard
PhD in CS @ **Columbia**
advised by **David Blei**
Interned at
- Google AI
- Facebook AI Research

Citations: 575
Since 2020: 562

Prior Works

Evaluating the World Model Implicit in a Generative Model

Keyon Vafa
Harvard University

Justin Y. Chen
MIT

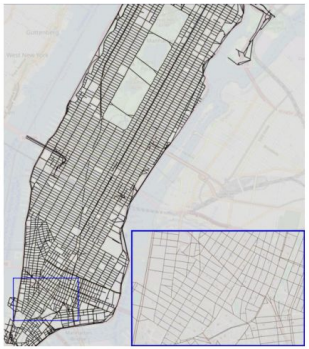
Ashesh Rambachan
MIT

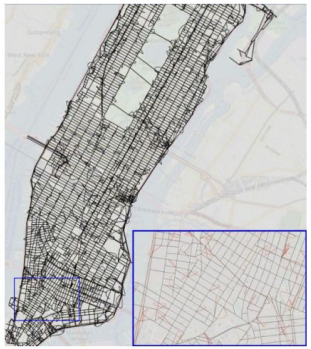
Jon Kleinberg
Cornell University

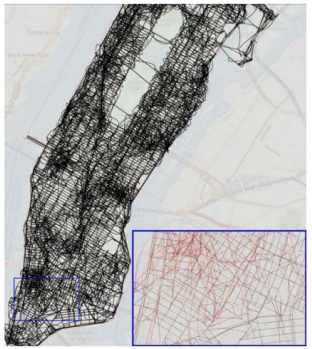
Sendhil Mullainathan
MIT

Abstract

Recent work suggests that large language models may implicitly learn world models. How should we assess this possibility? We formalize this question for the case where the underlying reality is governed by a deterministic finite automaton. This includes problems as diverse as simple logical reasoning, geographic navigation, game-playing, and chemistry. We propose new evaluation metrics for world model recovery inspired by the classic Myhill-Nerode theorem from language theory. We illustrate their utility in three domains: game playing, logic puzzles, and navigation. In all domains, the generative models we consider do well on existing diagnostics for assessing world models, but our evaluation metrics reveal their world models to be far less coherent than they appear. Such incoherence creates fragility: using a generative model to solve related but subtly different tasks can lead to failures. Building generative models that meaningfully capture the underlying logic of the domains they model would be immensely valuable; our results suggest new ways to assess how close a given model is to that goal.







(a) World model **(b)** World model with noise **(c)** Transformer

They suggests theoretically grounded metrics for assessing the world models implicit inside generative models.
Their results suggest that LLMs can perform some of these tasks well (such as finding shortest paths between two points on a map) without having a coherent world model.
The work has applications to maps, games, and logic puzzles.

Last Author's Background

the Bike Shop @ MIT

“Imagine algorithms that can do what people do. We think that’s a pretty uninspiring vision. Imagine, instead, algorithms that can take us whole new places we could never reach – or dream of reaching. That’s what we build at the Bike Shop @ MIT: bicycles for the mind.”

Co-founder of Dandelion

“I’m also a co-founder of Dandelion, a company that catalyzes AI in healthcare. Dandelion combines high fidelity data with a streamlined platform to allow innovators to develop AI models in a fraction of the time and effort it currently requires”

Sendhil Mullainathan
Professor @ MIT
PhD in Econ @ **Harvard**
Affiliations
- Harvard
- MIT
- University of Chicago

Citations: 99566
Since 2020: 52654

Last Author's Background

Dissecting racial bias in an algorithm used to manage the health of populations Z Obermeyer, B Powers, C Vogeli, S Mullainathan Science 366 (6464), 447-453	6904	2019
Some consequences of having too little AK Shah, S Mullainathan, E Shafir Science 338 (6107), 682-685	2103	2012
Are CEOs rewarded for luck? The ones without principals are M Bertrand, S Mullainathan The Quarterly Journal of Economics 116 (3), 901-932	2590	2001
Scarcity: Why having too little means so much S Mullainathan, E Shafir Macmillan	4007 *	2013
Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination M Bertrand, S Mullainathan American economic review 94 (4), 991-1013	8466	2004

Sendhil Mullainathan
Professor @ MIT
PhD in Econ @ **Harvard**
Affiliations
- Harvard
- MIT
- University of Chicago

Citations: 99566
Since 2020: 52654

What Has a Foundation Model Found? Inductive Bias Reveals World Models

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, Sendhil Mullainathan

Promise of foundation models:

Good predictions => deeper understanding/world models

Slides from ICML 2025

Promise of foundation models:

Good predictions => deeper understanding/world models

Centaur: a foundation model of human cognition

Marcel Binz[1*], Elif Akata[1], Matthias Bethge[2], Franziska Brändle[3,5], Fred Callaway[4], Julian Coda-Forno[1], Peter Dayan[2,5], Can Demircan[1], Maria K. Eckstein[6],

Perspective | Open access | Published: 06 March 2025

Foundation models for materials discovery – current state and future directions

Edward O. Pyzer-Knapp, Matteo Manica, Peter Staar, Lucas Morin, Patrick Ruch, Teodoro Laino, John R. Smith & Alessandro Curioni

A foundation model of transcription across human cell types

Xi Fu, Shentong Mo, Alejandro Buendia, Anouchka P. Laurent, Anqi Shao, Maria del Mar Alvarez-Torres, Tianji Yu, Jimin Tan, Jiayu Su, Romella Sagatelian, Adolfo A. Ferrando, Alberto Ciccia, Yanyan Lan, David M. Owens, Teresa Palomero, Eric P. Xing & Raul Rabadan

Nature 637, 965–973 (2025) | Cite thi

Med-PaLM

A large language model from Google Research, designed for the medical domain.

Nucleotide Transformer: building and evaluating robust foundation models for human genomics

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez & Thomas Pierrot

Nature Methods 22, 287–297 (2025) | Cite this article

New idea in one sense, old in another.

New idea in one sense, old in another.

Example: Predicting movements of planets in the night sky.

Johannes Kepler (1571-1630)

- Used geometric properties to **predict** planetary orbits
- Could **not explain** why orbits obeyed these properties

Johannes Kepler (1571-1630)

- Used geometric properties to **predict** planetary orbits
- Could **not explain** why orbits obeyed these properties

Isaac Newton (1643-1727)

- Developed general (Newtonian) mechanics to **predict** orbits
- Could **explain** not only orbits but also other physical problems

Across domains, foundation models don't have inductive biases toward the correct world model.

So what is their inductive bias toward?

Challenge: Defining what it means for a foundation model to have the right world model.

Challenge: Defining what it means for a foundation model to have the right world model.

Approach 1: Mechanistic

Drawbacks: Difficult to understand mechanisms of large models (Olah, 2022), may not reflect how models behave (Caper et al., 2023)

Challenge: Defining what it means for a foundation model to have the right world model.

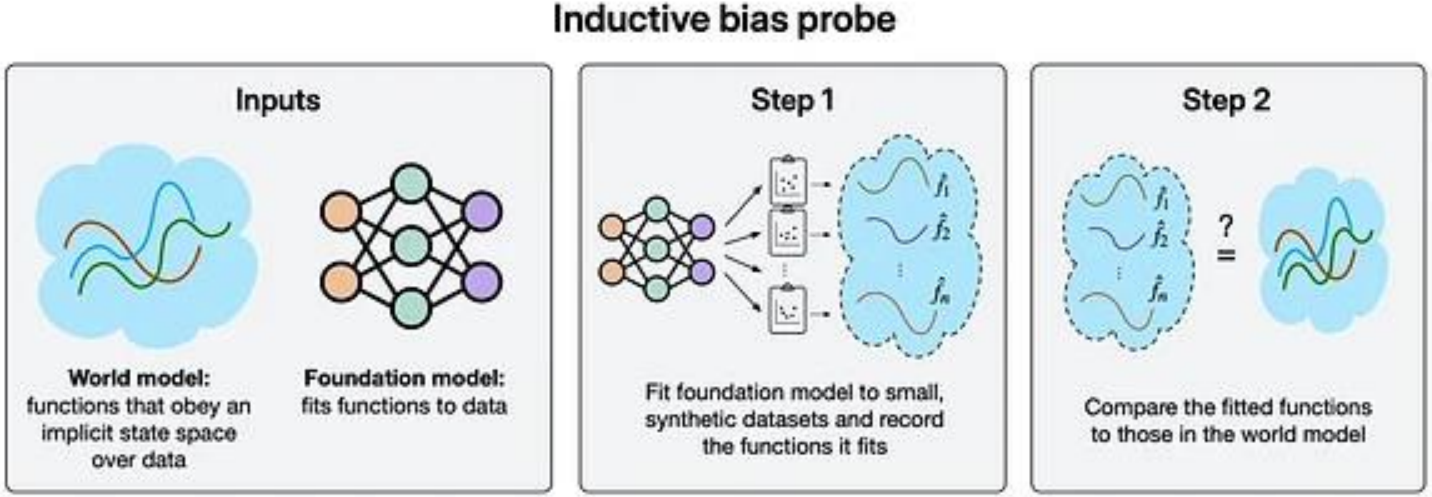
Approach 1: Mechanistic

Drawbacks: Difficult to understand mechanisms of large models (Olah, 2022), may not reflect how models behave (Caper et al., 2023)

Approach 2: Study behavior on single task
(e.g., Li et al., 2021; Toshniwal et al., 2021; Li et al., 2023; Vafa et al., 2024)

Drawbacks: Doesn't capture how foundation models are used in the real world -- as tools for new tasks.

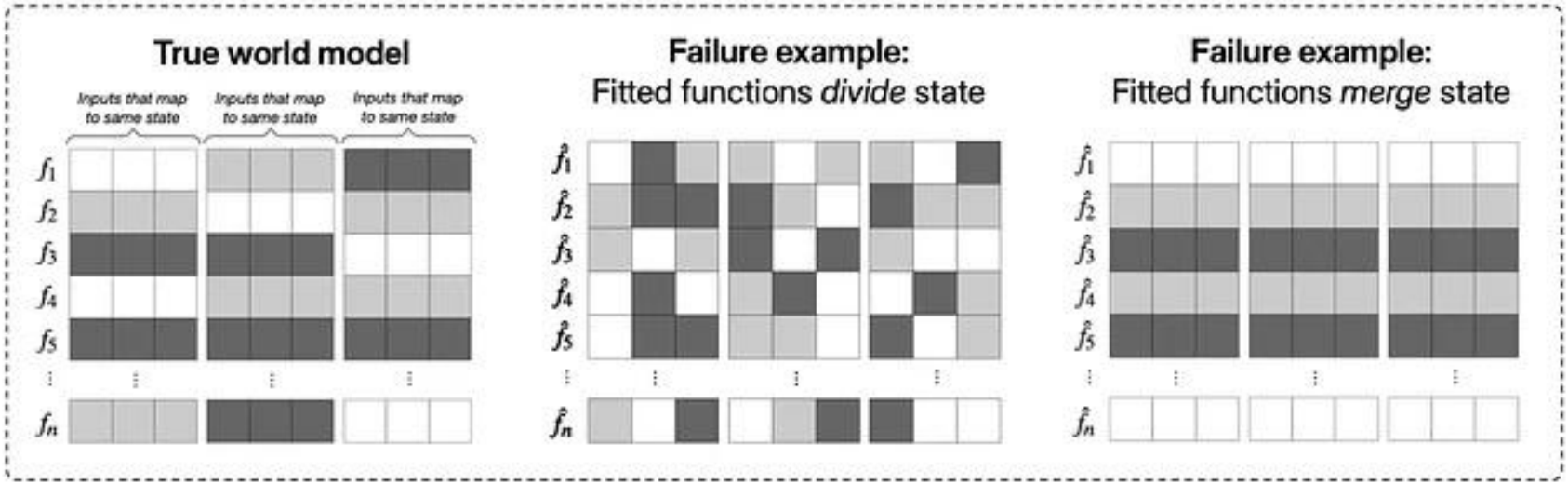
Our approach: Study foundation model's inductive bias as it adapts to new tasks.



How a model behaves on small amounts of data reveals what it believes.

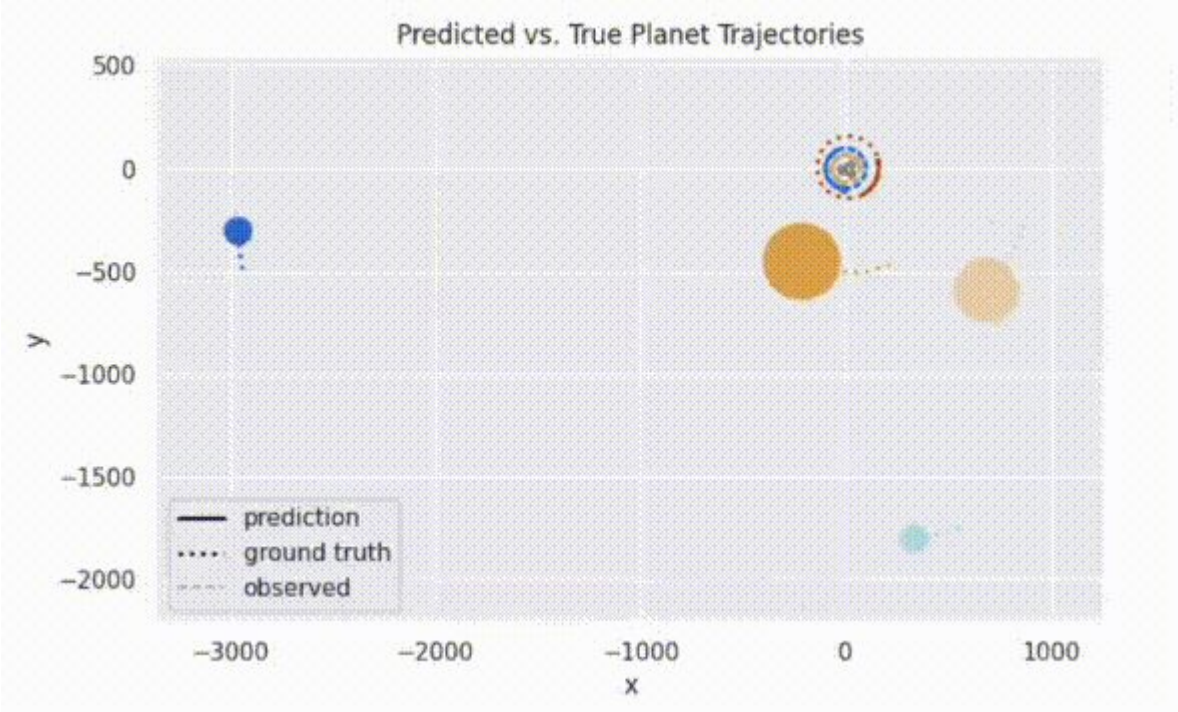
World model: Collection of functions that obey an implicit state space.

Example: Finite state space



Slides from ICML 2025

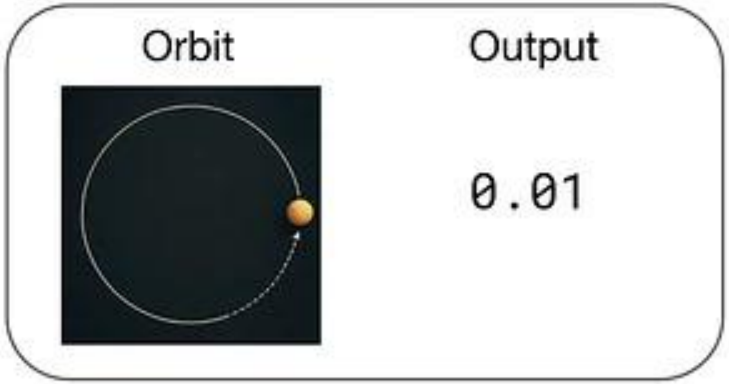
We train a foundation model of planetary orbits.



Has the model learned Newtonian mechanics?

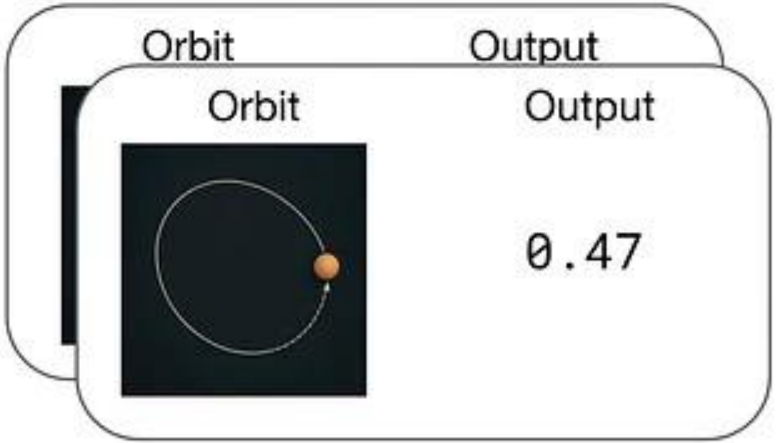
Has the model learned Newtonian mechanics?

Test how the model behaves when it is fine-tuned to new tasks with small amounts of labeled data.



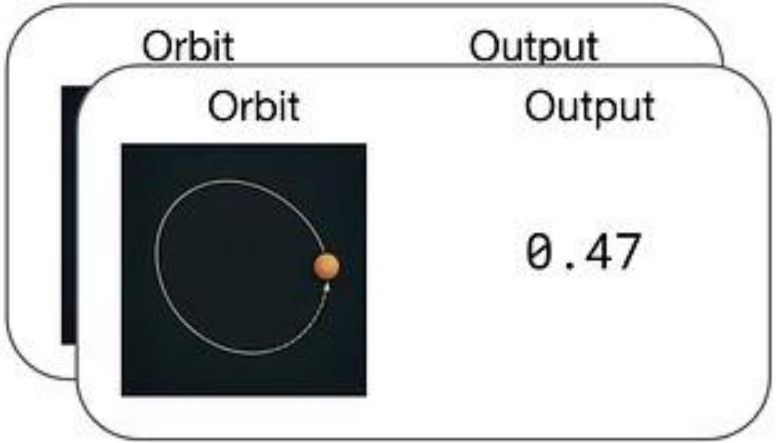
Has the model learned Newtonian mechanics?

Test how the model behaves when it is fine-tuned to new tasks with small amounts of labeled data.



Has the model learned Newtonian mechanics?

Test how the model behaves when it is fine-tuned to new tasks with small amounts of labeled data.

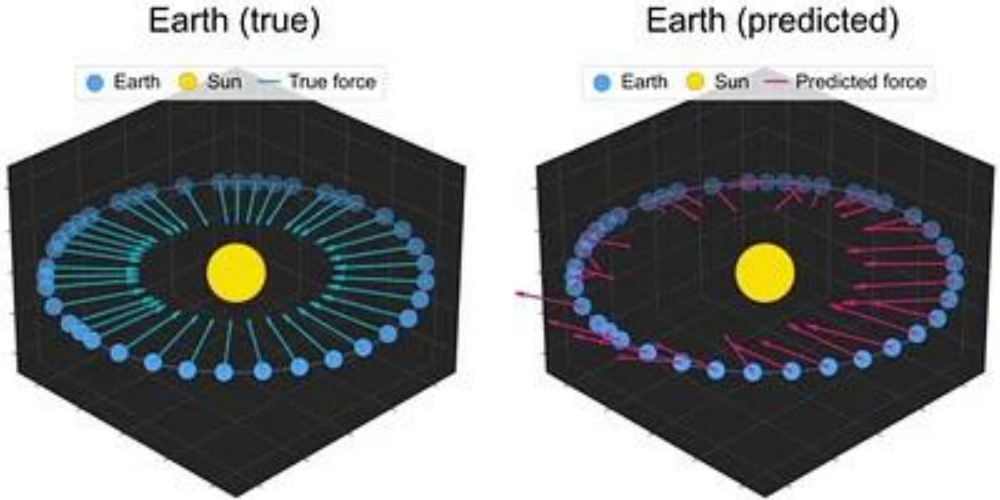


Physicists using Newtonian mechanics would extrapolate based on masses, velocities, etc.

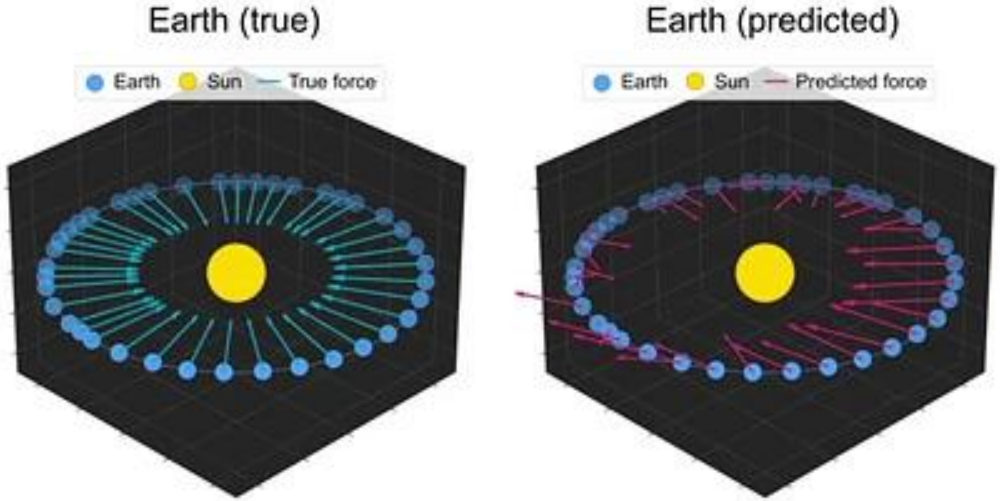
A foundation model using Newtonian mechanics should do the same.

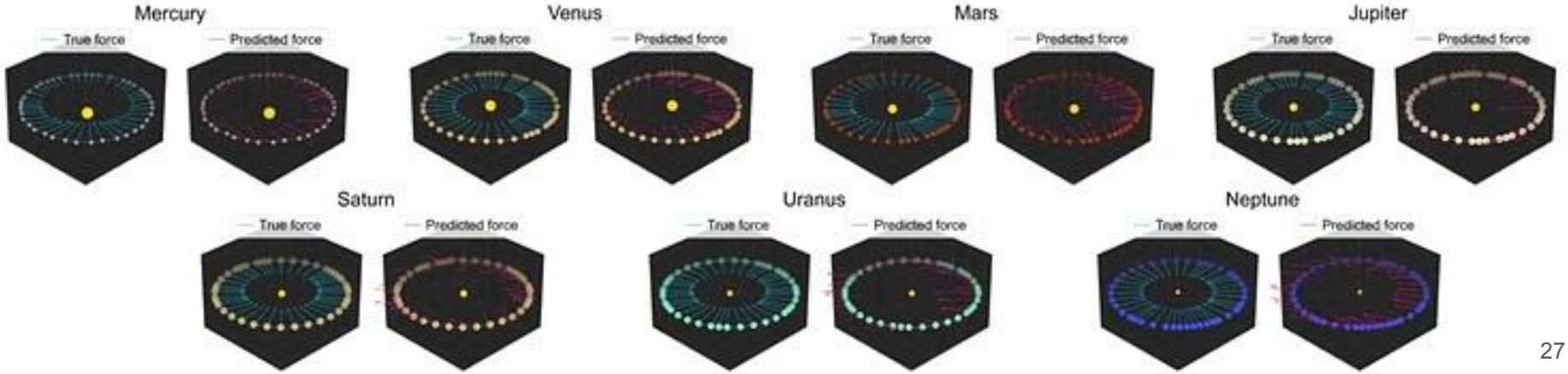
Fine-tune model to predict
force vectors between planets.

Fine-tune model to predict force vectors between planets.



Fine-tune model to predict force vectors between planets.





The true force law is a simple function of Newtonian primitives.

True force law (Newton)

$$F \propto \frac{m_1 m_2}{r^2}$$

The true force law is a simple function of Newtonian primitives.

True force law (Newton)

$$F \propto \frac{m_1 m_2}{r^2}$$

Use symbolic regression to estimate force law the model has recovered.

Recovered force law (transformer)

$$F \propto \exp\left[(1/r) * \sin\left(\sin\left(e^{e^{m_2+1.1}}\right) * r\right)\right]$$

The true force law is a simple function of Newtonian primitives.

True force law (Newton)

$$F \propto \frac{m_1 m_2}{r^2}$$

Use symbolic regression to estimate force law the model has recovered.

Recovered force law (transformer)

$$F \propto \exp\left[(1/r) * \sin\left(\sin\left(e^{e^{m_2+1.1}}\right) * r\right)\right]$$

No universal law: different laws on different galaxies.

Ground-truth law		$F \propto \frac{m_1 m_2}{r^2}$
Estimated laws	Galaxy 1	$F \propto \cos(\cos(-1/m_1))$
	Galaxy 2	$F \propto \sin(e^{m_2+3.8})$
	Galaxy 3	$F \propto \cos(\cos(1/m_2))$
	Galaxy 4	$F \propto \exp\left[\frac{1}{r}\sin(\sin(e^{e^{m_2+1.1}})r)\right]$

Across domains, foundation models don't have inductive biases toward the correct world model.

Across domains, foundation models don't have inductive biases toward the correct world model.

So what is their inductive bias toward?

Across domains, foundation models don't have inductive biases toward the correct world model.

So what is their inductive bias toward?

Hypothesis: models group together sequences that have **similar legal next-tokens**, even if those sequences represent different worlds.

Reasoning: Models are trained to perform next-token prediction, so that's what their inductive bias is toward.

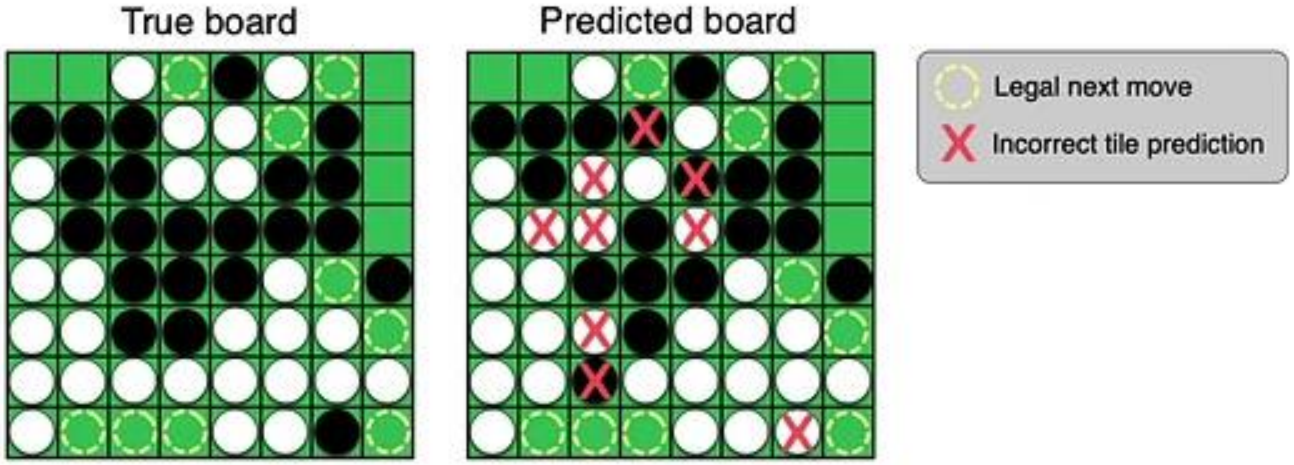
Example: Othello

Fine-tune an "Othello foundation model" to predict boards.

Example: Othello

Fine-tune an "Othello foundation model" to predict boards.

In Othello, two different boards can have the same set of legal next-moves.



Foundation model only recovers “enough of” the board to ensure legal next moves.

Conclusion

Inductive bias probes: A model's inductive bias on small data reveals its world model.

Across domains, good predictions don't lead to inductive biases toward world models.

Inductive biases are stronger toward next-tokens than states in world model.

Poster: Tuesday, July 15 at 11am-1:30pm.

Assumptions

Let $x \in \mathcal{X}$ denote an input and $y \in \mathcal{Y}$ denote some output.

A dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is a collection of n input-output pairs.

A task $f : \mathcal{X} \to \mathcal{Y}$ is a mapping between inputs and outputs.

A foundation model is a learning algorithm which, when given a dataset D, returns a prediction function $\widehat{m}_D : \mathcal{X} \to \mathcal{Y}$ that relates the input to the outputs.

$\widehat{m}_D$ could be some pre-trained model that is fine-tuned on the dataset D, or it can be an LLM that is supplied D in-context.

Assumptions

A postulated world model is summarized by a state space Φ and a mapping $\phi: \mathcal{X} \to \Phi$ that associates each input with some state $\phi(x) \in \Phi$.

A dataset D is consistent with the world model if for each $(x, y) \in D$, the output is a deterministic function of the state, $y = g(\phi(x))$ for some $g : \Phi \to \mathcal{Y}$

Intuition

The inductive bias probe evaluates weather a foundation model's inductive bias is towards a postulated world model.

At a high level, the probe repeatedly applies the foundation model to synthetic datasets consistent with the postulated world model and each time evaluates its predictions on held-out inputs.

If the foundation model's inductive bias is towards the postulated world model, its extrapolations should have two properties.

- First, the foundation model's predictions should respect state: if two inputs map to the same state, the foundation model should have the same predicted outputs when applied across synthetic datasets.

Intuition

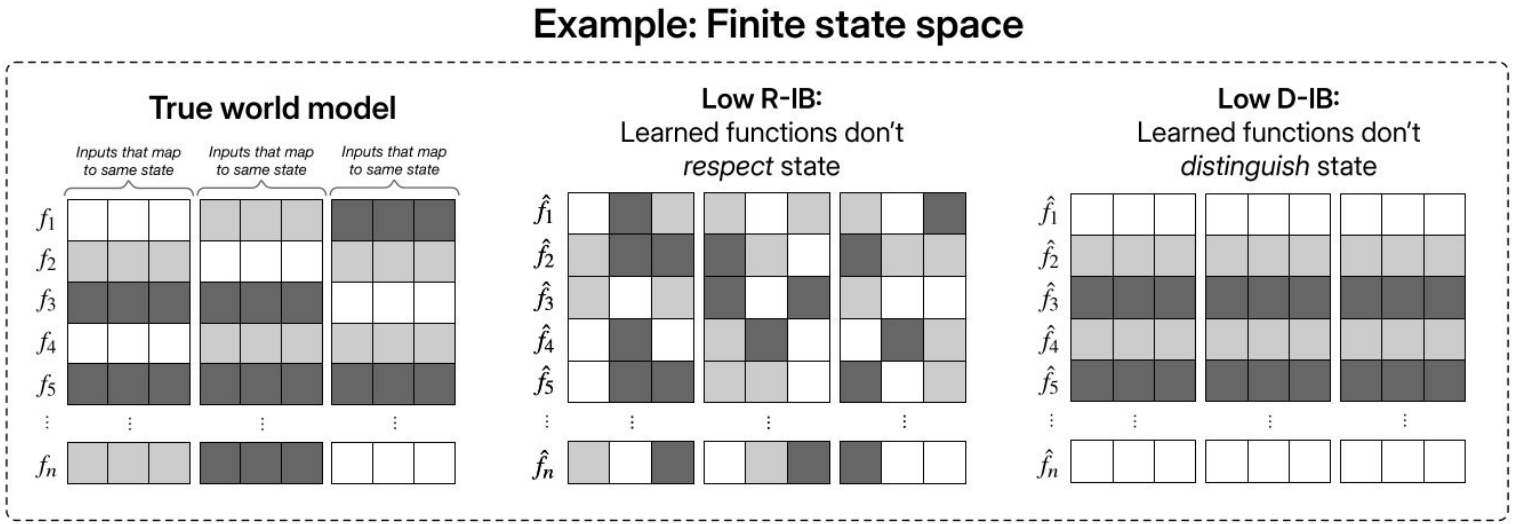
The inductive bias probe evaluates weather a foundation model's inductive bias is towards a postulated world model.

At a high level, the probe repeatedly applies the foundation model to synthetic datasets consistent with the postulated world model and each time evaluates its predictions on held-out inputs.

If the foundation model's inductive bias is towards the postulated world model, its extrapolations should have two properties.

- Second, the foundation model's predictions should distinguish state: if two inputs map to different states, the foundational model should typically have different predicted outputs across synthetic datasets.

Intuition



An illustration of the inductive bias probe when the given world model has a finite state space. Each row represents a function and each column represents an input, with inputs belonging to the same state grouped together. The shading illustrates each function's value at the corresponding input. A foundation model has low R-IB (middle) if it learns functions that divide states, while a foundation model has low D-IB (right) if it learns function that merge states.

Intuition

Let $1(y, y')$ denote the indicator for whether $y = y'$. Let's specify a sampling distribution over consistent datasets $D \sim P_D$ and a sampling distribution over inputs $(X_i, X_j) \sim P_X \times P_X$

The foundation model's inductive bias towards respecting state (R-IB) is

$$\mathbb{E}_{X_i, X_j, D}[1(\widehat{m}_D(X_i), \widehat{m}_D(X_j)) \mid \phi(X_i) = \phi(X_j)]$$

The foundation model's inductive bias towards distinguishing state (D-IB) is

$$1 - \mathbb{E}_{X_i, X_j, D}[1(\widehat{m}_D(X_i), \widehat{m}_D(X_j)) \mid \phi(X_i) \neq \phi(X_j)]$$

Formulation

Let's introduce a collection of admissible functions on state $\mathcal{G}$ that rule the state-space and WM output with each $g \in \mathcal{G} \colon \Phi \to \mathcal{Y}$. A dataset is now consistent with the world model if for each $(x, y) \in D$, $y = g(\phi(x))$ for some $g \in \mathcal{G}$.

Extrapolative predictability. Let's specify a family of predictors $\mathcal{H}$ with $h \in \mathcal{H}$ such that $h : \mathcal{Y} \to \mathcal{Y}$ and the loss function over outputs $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

The explorative predictability between two inputs is .

$$\widehat{I}(x_i, x_j) = -\min_{h \in \mathcal{H}} \mathbb{E}_{D \sim P}[\ell(h(\widehat{m}_D(x_i)), \widehat{m}_D(x_j))]$$

Formulation

Oracle FM. Is a FM that is given access to the true state space Φ and admissible $\mathcal{G}$ functions When applied to consistent dataset D, the oracle foundation model returns

$$m_D^* = \arg\min_{g \in \mathcal{G}} \frac{1}{|D|} \sum_{(x_i, y_i) \in D} \ell(g(\phi(x_i)), y_i)$$

The oracles extrapolative predictability is .

$$I^*(x_i, x_j) = -\min_{h \in \mathcal{H}} \mathbb{E}_{D \sim P}[\ell(h(m_D^*(x_i)), m_D^*(x_j))]$$

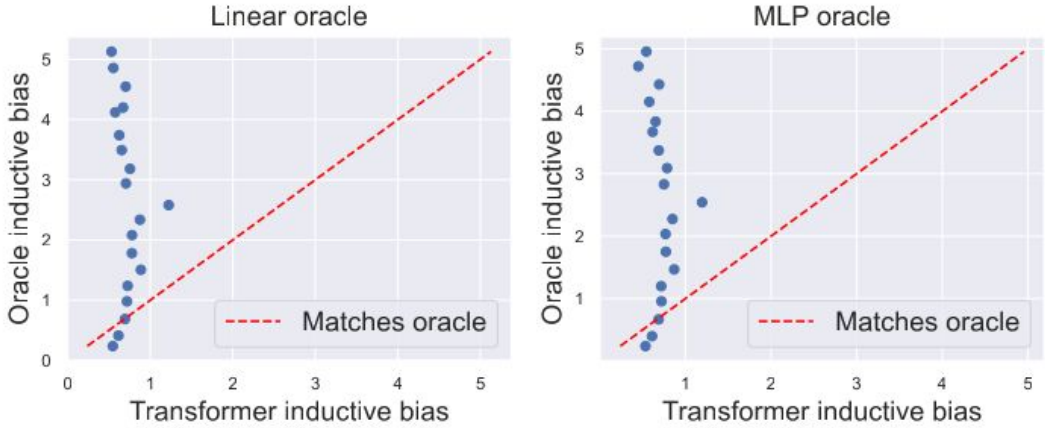
Formulation

The inductive bias probe compares the foundation model's extrapolative predictability to that of the oracle.

The inductive bias towards the world model is defined as, for any $0 \leq \underline{s} \leq \overline{s}$,

$$\mathrm{IB}(\underline{s}, \overline{s}) = \mathbb{E}_{X_i, X_j}[\widehat{I}(X_i, X_j) \mid \underline{s} \leq I^*(X_i, X_j) \leq \overline{s}]$$

Interpretation



Inductive bias probe performance for a transformer pretrained on orbital trajectories. A 45-degree line would indicate perfect inductive bias toward an oracle that extrapolates based on the Newtonian state vector.

In practice the oracle model is a linear map or a 2 layer MLP with 5 nodes in each hidden layer.

Models Compared

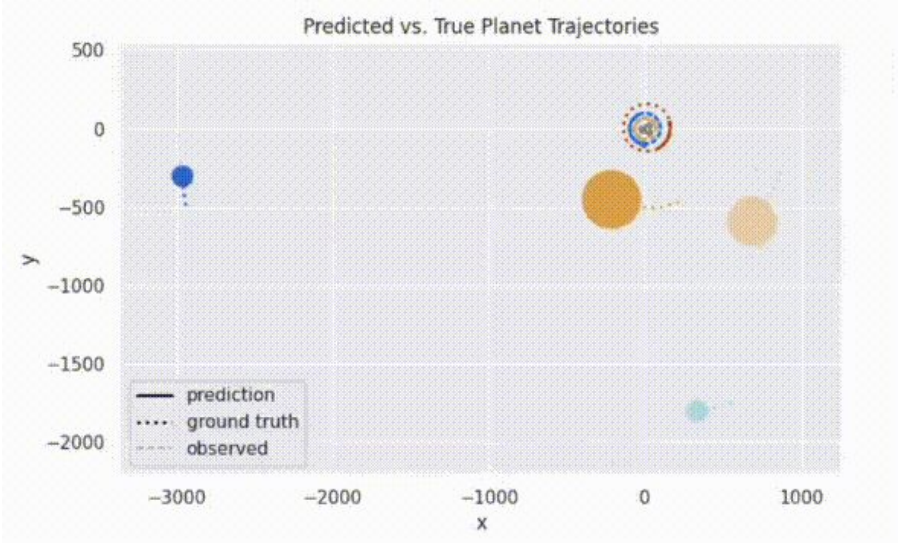
They use the following specifications for each model:

- **RNN** - They use 6 (Othello) or 2 (Lattice) uni-directional RNN layers with 768 embedding dimensions.
- **LSTM**: Same as RNN except layers
- **Transformer**: Transformer decoder architecture, with 12 layers, 12 attention heads, and 768 embedding dimensions.
- **Mamba**: They encode inputs with a 768-dimension embedding layer. Pass inputs through 24 Mamba layers (analog to transformer).
- **Mamba-2**: We use the same architecture as for Mamba except the mixer in each block is a Mamba-2 module.

Each model is fine-tuned to specific tasks, the model can be pre trained on in-distribution data or fine-tuned a random initialization.

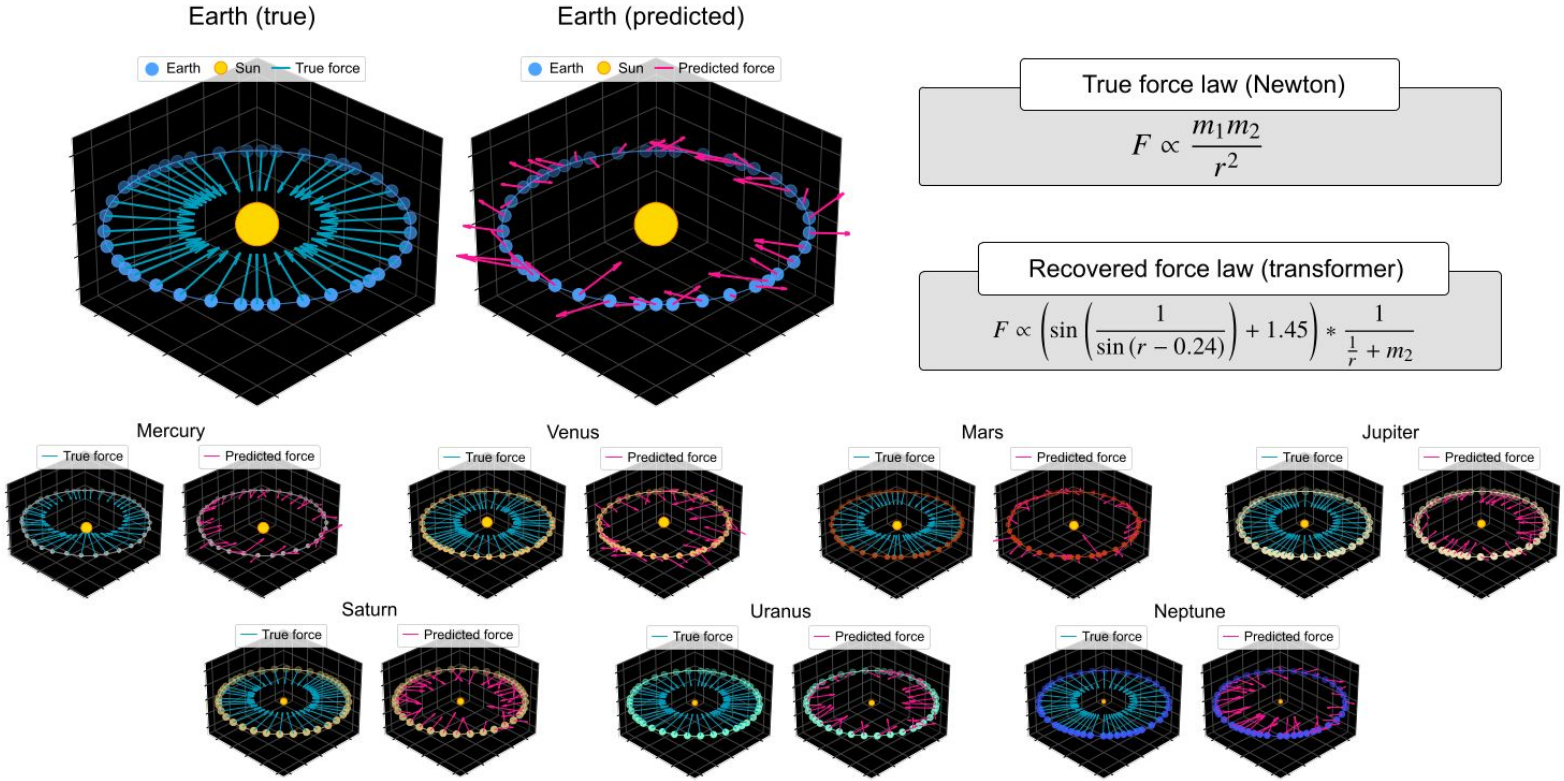
Experiments - Newtonian Mechanics

- **Goal** - Train a model to predict the location of planets across solar systems
- **Dataset** sequences describe planets in motion around the sun
  - Randomly sampled initial conditions.
  - Simulate each planet's trajectory around the sun using Newton's laws of motion (omitting the interactions between planets)
  - Sequences of (x, y) coordinates of each planet across different intervals, sequence of length 1k
  - They used both fixed-length and dynamic intervals between observations

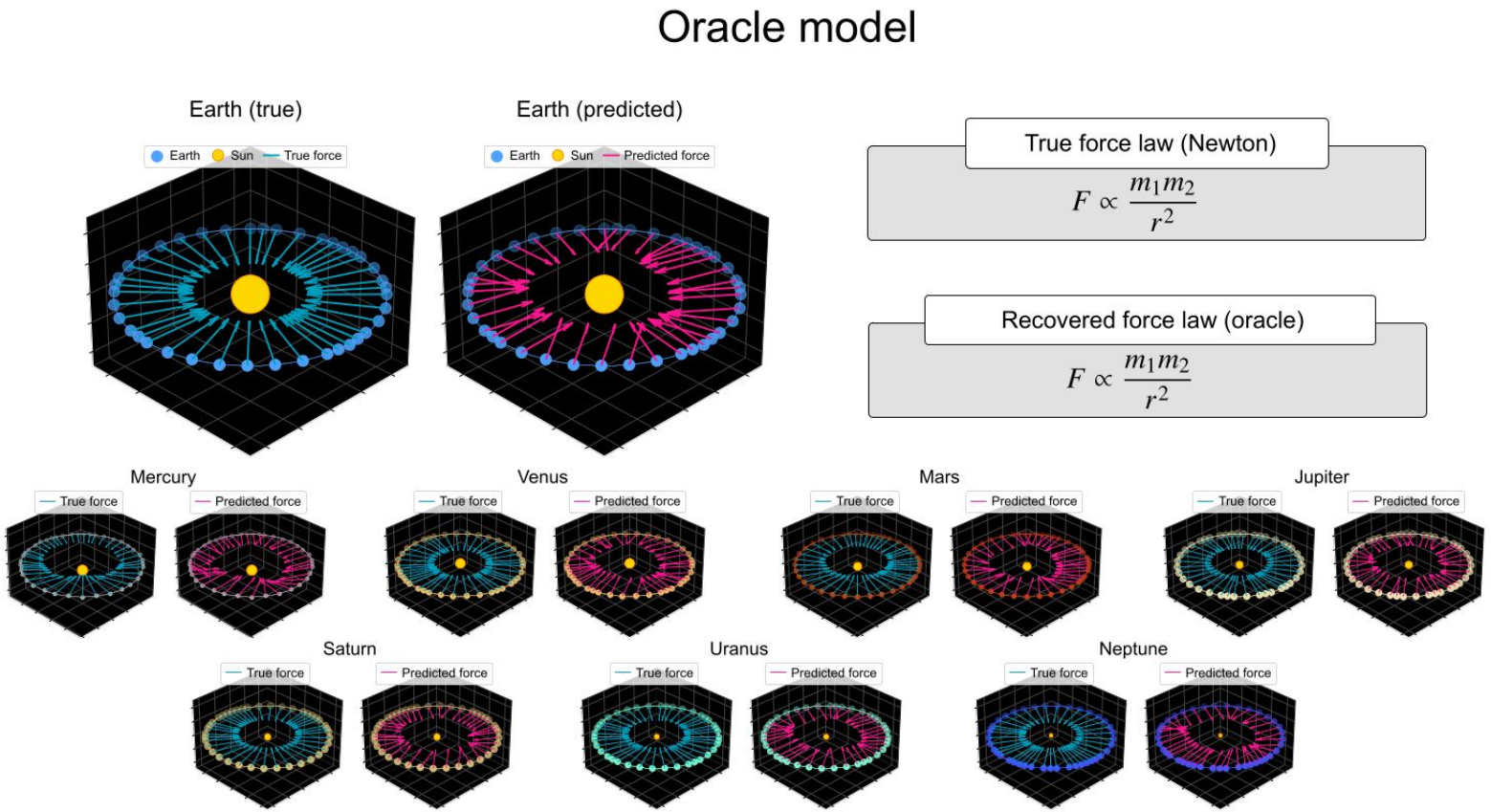


Train set - 10M sequences amounting to 20B tokens

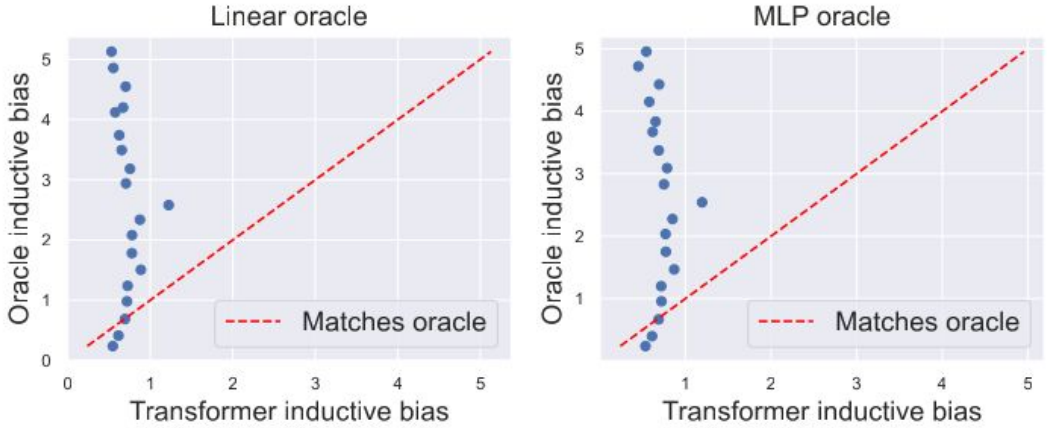
Experiments - Newtonian Mechanics



Experiments - Newtonian Mechanics



Experiments - Newtonian Mechanics



Inductive bias probe performance for a transformer pretrained on orbital trajectories. A 45-degree line would indicate perfect inductive bias toward an oracle that extrapolates based on the Newtonian state vector.

The inductive bias toward simple functions of Newtonian state is poor.

The model's inductive bias is not toward Newtonian state.

When it has to extrapolate, it makes similar predictions for orbits with very different states and different predictions for orbits with very similar states.

Experiments - Newtonian Mechanics

Ground-truth law		$F \propto \frac{m_1 m_2}{r^2}$
Estimated laws	Galaxy 1	$F \propto \left(\sin\left(\frac{1}{\sin(r-0.24)}\right) + 1.45\right) * \frac{1}{\frac{1}{r}+m_2}$
	Galaxy 2	$F \propto \cos\left(\cos(2.19 * m_1)\right)$
	Galaxy 3	$F \propto \cos\left(\sin(\frac{0.48}{m_1})\right)$
	Galaxy 4	$F \propto \sin\left(r + 8569.2 + \frac{1}{m_1}\right)$
	Galaxy 5	$F \propto \cos\left(\cos(e^{m_2})\right)$

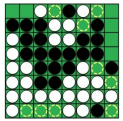
Force equations recovered via symbolic regression of a transformer pretrained on orbital data and fine-tuned to different galaxy samples. The model recovers different equations for each sample, never recovering the true law.

They fine-tune the pretrained transformer to predict the force vector on the planet implied by the state of the orbit.

They then perform symbolic regression to extract the laws from the force predictions.

They find that the extracted laws are in no way similar to the ground-truth law.

Other Experiments

Othello 

- A testbed for evaluating the World Models of sequence models
- Two player board game on a 8x8 size board
- Each game is tokenized into a sequence of at most 60 moves
- Each token indicates which of the 60 squares the most recent tile was placed on
- The state-space corresponds to all 8x8 boards and the mapping converts game sequences into states
- **Train set** - 20M games, 3.8M hold-out

Lattice

- They study a lattice setting that simulates an agent moving along a line segment
- Finite number of possible positions - state-space

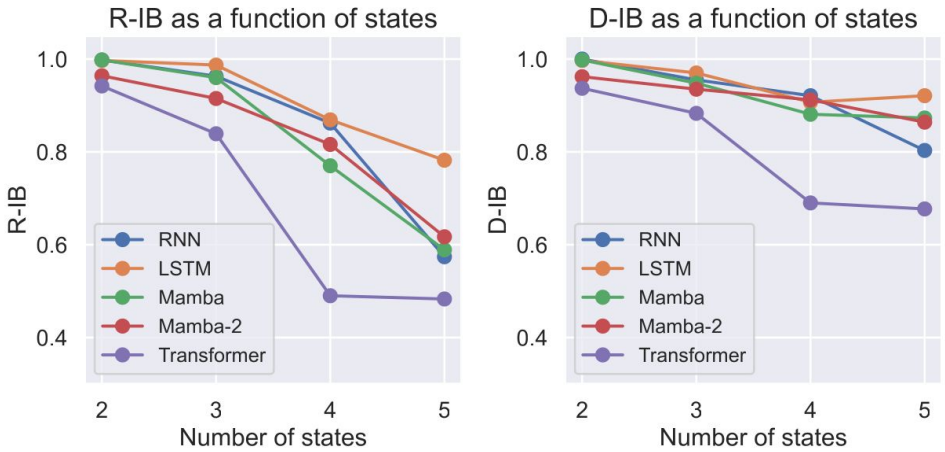
$$\Phi = \{1, 2, \ldots, S\}$$

- The language consists of sequences with three tokens $\Sigma = \{L, \perp, R\}$
- Number of states varies from 2 to 5
- They generate sequences of 100 length
- **Train set** - 10M tokens, 100k hold-out

Experiments - Othello and Lattice

	Lattice	Othello
RNN	1.00	0.992
LSTM	1.00	0.996
Transformer	1.00	0.999
Mamba	1.00	0.999
Mamba-2	1.00	0.999

Results for the next token test for models pre-trained on next-token predictions



Inductive Bias Probe results for lattice problem as a function of the underlying number of states. Pretrained on NTP

Experiments - Othello and Lattice

		Lattice (5 States)		Othello	
	Pre-training	R-IB (↑)	D-IB (↑)	R-IB (↑)	D-IB (↑)
RNN (Elman, 1990)	Untrained	0.346 (0.026)	0.749 (0.027)	0.228 (0.016)	0.990 (0.002)
	NTP trained	0.574 (0.026)	0.803 (0.032)	0.632 (0.023)	0.797 (0.023)
LSTM (Hochreiter, 1997)	Untrained	0.456 (0.028)	0.718 (0.031)	0.438 (0.030)	0.681 (0.031)
	NTP trained	0.782 (0.021)	0.921 (0.030)	0.563 (0.030)	0.610 (0.034)
Transformer (Vaswani et al., 2017)	Untrained	0.268 (0.027)	0.742 (0.028)	0.708 (0.022)	0.843 (0.021)
	NTP trained	0.483 (0.031)	0.677 (0.034)	0.703 (0.025)	0.624 (0.033)
Mamba (Gu & Dao, 2023)	Untrained	0.260 (0.026)	0.771 (0.027)	0.303 (0.016)	0.929 (0.009)
	NTP trained	0.571 (0.023)	0.866 (0.029)	0.682 (0.021)	0.728 (0.027)
Mamba-2 (Dao & Gu, 2024)	Untrained	0.244 (0.026)	0.785 (0.026)	0.468 (0.019)	0.896 (0.016)
	NTP trained	0.617 (0.021)	0.864 (0.029)	0.653 (0.022)	0.694 (0.029)

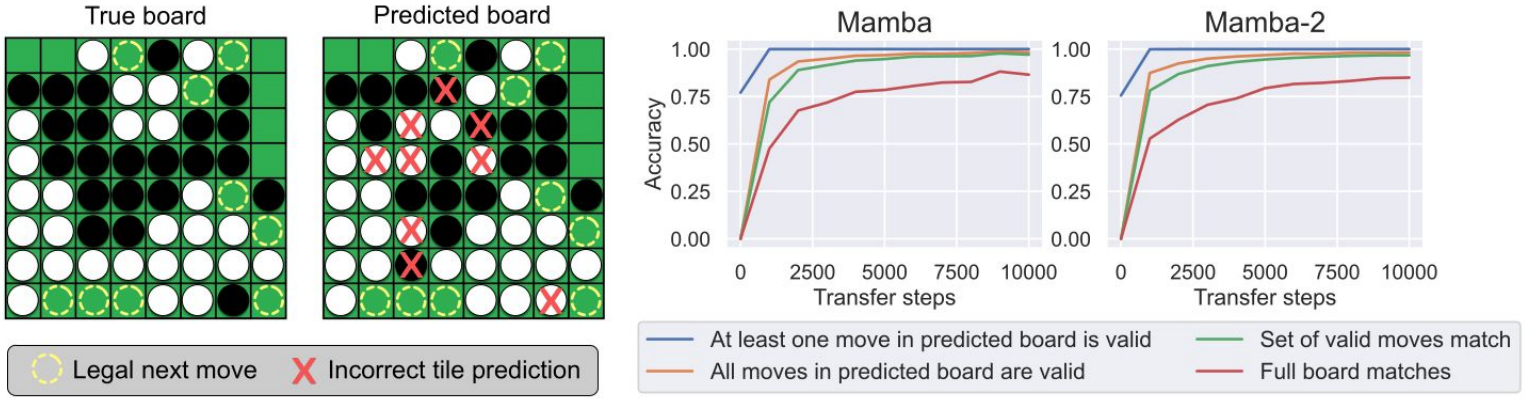
Inductive Bias Probe results for both Lattice and Othello problems. “NTP-trained” represents a model pre-trained on next-token predictions, while “untrained” refers to a model trained on the same synthetic tasks, initialized from scratch.

Experiments - Othello and Lattice

		Majority Tiles		Board Balance		Edge Balance	
	Pretraining	NLL (↓)	ACC (↑)	NLL (↓)	ACC (↑)	NLL (↓)	ACC (↑)
RNN	Untrained	0.492 (0.004)	0.755 (0.003)	0.405 (0.005)	0.806 (0.003)	0.462 (0.002)	0.816 (0.002)
	NTP trained	0.431 (0.004)	0.792 (0.002)	0.302 (0.004)	0.856 (0.002)	0.080 (0.002)	0.964 (0.001)
LSTM	Untrained	0.436 (0.004)	0.786 (0.003)	0.305 (0.004)	0.864 (0.002)	0.105 (0.002)	0.953 (0.001)
	NTP trained	0.232 (0.004)	0.901 (0.002)	0.164 (0.003)	0.927 (0.001)	0.041 (0.002)	0.982 (0.001)
Transformer	Untrained	0.497 (0.004)	0.754 (0.003)	0.340 (0.005)	0.855 (0.002)	0.075 (0.002)	0.967 (0.001)
	NTP trained	0.100 (0.002)	0.956 (0.001)	0.086 (0.002)	0.965 (0.001)	0.013 (0.001)	0.996 (0.000)
Mamba	Untrained	0.377 (0.004)	0.816 (0.002)	0.246 (0.004)	0.888 (0.002)	0.099 (0.002)	0.952 (0.001)
	NTP trained	0.149 (0.003)	0.937 (0.002)	0.158 (0.003)	0.931 (0.002)	0.027 (0.001)	0.989 (0.001)
Mamba-2	Untrained	0.379 (0.004)	0.821 (0.002)	0.258 (0.004)	0.891 (0.002)	0.068 (0.001)	0.969 (0.001)
	NTP trained	0.069 (0.002)	0.970 (0.001)	0.059 (0.002)	0.976 (0.001)	0.012 (0.002)	0.995 (0.001)
IB Correlation	—	0.462	0.477	0.610	0.653	0.970	0.960

Results showing transfer performance across new functions of state. Table contains negative log-likelihoods and accuracy. IB Correlation - (unsigned) correlation between each column of results to the ratios of the inductive bias metrics (R-IB D-IB) from previous table. Transfer learning results are correlated to the inductive bias metrics; models with low inductive bias perform worse at transfer.

Othello: Board predictions



On the left, a true Othello board implied by a sequence, and on the right, the predicted board from a model fine-tuned to predict boards. Although the prediction has errors, the set of predicted next tokens exactly matches the true board. On the right, metrics about board reconstruction during fine-tuning. Consistently, even as Mamba models struggle to recover full boards, they recover them well enough such that the sets of valid next moves match those in the true boards.

Next-token Prediction Hypothesis

Let q denote the next-token coarsening of the state-space such that $q(x) = q(x')$ if and only if $\text{NextTokens}(\phi(x)) = \text{NextTokens}(\phi(x'))$, where NextTokens(s) is a set of valid next tokens for a state s.

D-IB decomposition. Define $\text{Same}(X_i, X_j)$ as the event that $\phi(X_i) \neq \phi(X_j)$ but $q(X_i) = q(X_j)$. Then define

$$\text{D-IB}_{q=} = 1 - \mathbb{E}\left[1(\widehat{m}_D(X_i), \widehat{m}_D(X_j)) \mid \text{Same}(X_i, X_j)\right]$$

Similarly define $\text{Diff}(X_i, X_j)$ as the event that $\phi(X_i) \neq \phi(X_j)$ and $q(X_i) \neq q(X_j)$, Then define

$$\text{D-IB}_{q\neq} = 1 - \mathbb{E}\left[1(\widehat{m}_D(X_i), \widehat{m}_D(X_j)) \mid \text{Diff}(X_i, X_j)\right]$$

Next-token Prediction Hypothesis

	Lattice		Othello	
	D-IB$_{q=}$	D-IB$_{q\neq}$	D-IB$_{q=}$	D-IB$_{q\neq}$
RNN	0.740 (0.042)	0.844 (0.034)	0.521 (0.031)	0.798 (0.023)
LSTM	0.873 (0.051)	0.952 (0.034)	0.519 (0.035)	0.610 (0.034)
Transformer	0.626 (0.037)	0.710 (0.037)	0.458 (0.033)	0.625 (0.033)
Mamba	0.764 (0.040)	0.933 (0.035)	0.485 (0.030)	0.729 (0.027)
Mamba-2	0.778 (0.042)	0.920 (0.033)	0.553 (0.032)	0.694 (0.029)

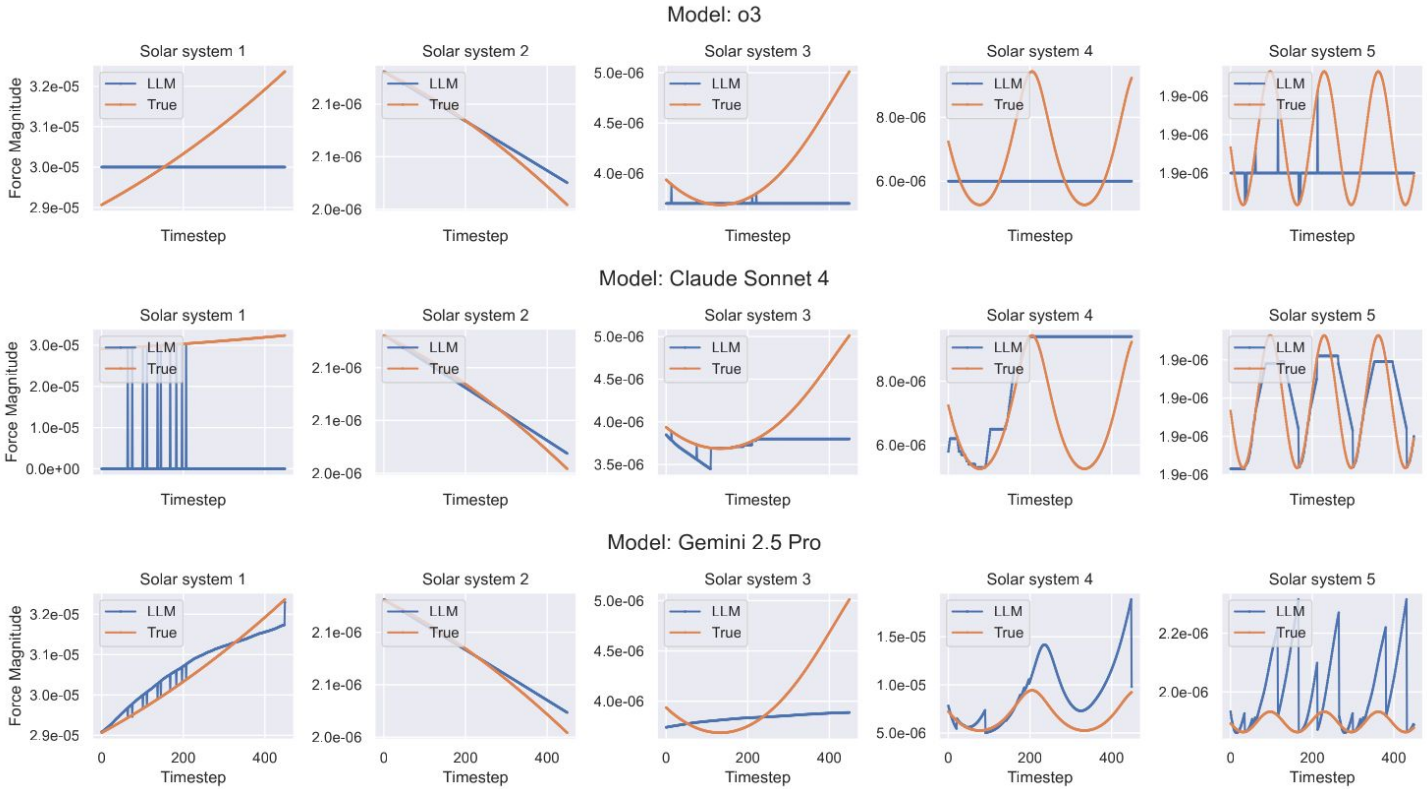
Metrics for assessing whether a model's inductive bias is toward its legal next-token partition. Low values of D-IBq= and high values of D-IBq≠ suggest that failures to differentiate state are driven by the models having an inductive bias toward the legal next-token partition

LLM Physics Experiments

LLM Prompt

```
You are a physics expert. You are given a sequence of coordinates and outcomes. The
coordinates are the positions of a planet in a 2-body solar system. The planet is
orbiting the sun. The sun is at the origin.

Here is a sequence of observations. Some of them are unknown. Your job is to predict
the outcomes for the unknown timesteps.

Timestep: 0, Coordinates: (-26.08, -6.98), Outcome: Unk
Timestep: 1, Coordinates: (-26.08, -6.99), Outcome: Unk
Timestep: 2, Coordinates: (-26.06, -7.01), Outcome: 2.907672751462087e-05
Timestep: 3, Coordinates: (-26.06, -7.04), Outcome: Unk
Timestep: 4, Coordinates: (-26.05, -7.05), Outcome: Unk
Timestep: 5, Coordinates: (-26.04, -7.08), Outcome: 2.9093407647451386e-05
Timestep: 6, Coordinates: (-26.04, -7.09), Outcome: Unk
Timestep: 7, Coordinates: (-26.02, -7.12), Outcome: Unk
Timestep: 8, Coordinates: (-26.02, -7.14), Outcome: Unk
Timestep: 9, Coordinates: (-26.01, -7.16), Outcome: Unk
...
Timestep: 449, Coordinates: (-20.28, -15.66), Outcome: Unk

You can reason all you'd like, but your answer should end with "ANSWER: " followed by
the predicted outcomes for all of the timesteps, even the unknown ones. You should
structure your predictions as a dict, where each key is a timestep and each value is
the prediction. You should make predictions for all of the timesteps, even the ones
that are known.

Here is an example of the output format:
ANSWER: {
   0: 1.0e-8,
   1: 1.0e-8,
   2: 1.0e-8,
   ...
   449: 1.0e-8,
}
```

LLM Physics Experiments



Conclusions

- They develop a framework for evaluating whether a foundation model has learned a postulated world model by measuring its inductive biases when transferring to new tasks.
- They find that these models may be relying on coarsened state representations or non-parsimonious representations.
- They highlight future work should prioritize methods for automatically constructing the world model implicit in the foundation model's behavior.