Emergent world representations:
Exploring a sequence model trained on a synthetic task.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg.






Harvard John A. Paulson School of Engineering and Applied Sciences
MIT CSAIL
MIT EECS
Northeastern University Khoury College of Computer Sciences

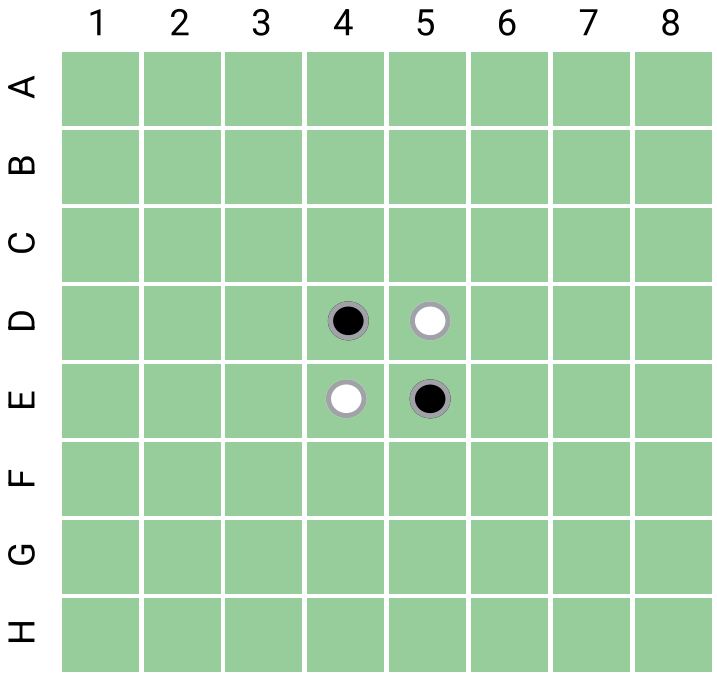
Are models learning something meaningful, causal, about the world...

Are models learning something meaningful, causal, about the world...
Or are they just memorizing data (then regurgitating it)?

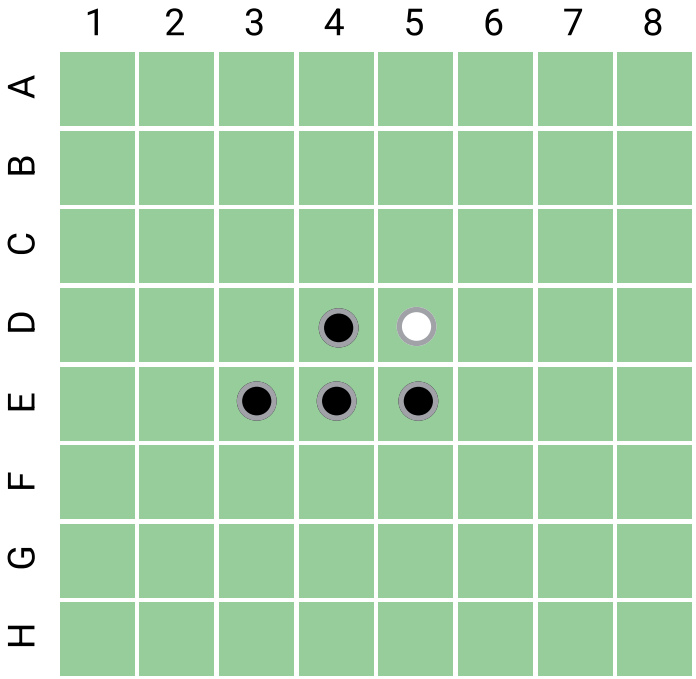
Looking for meaningful representations in large models is like looking for a needle in a haystack

Let's look at tiny pile of hay

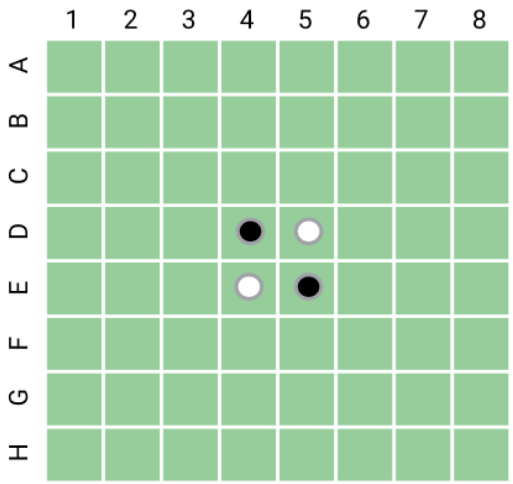
Game of Othello: A Toy Model

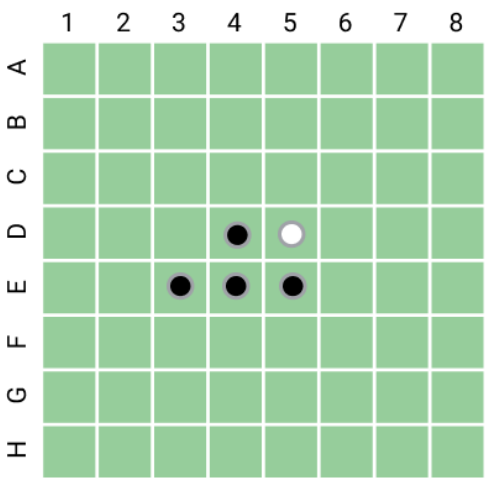


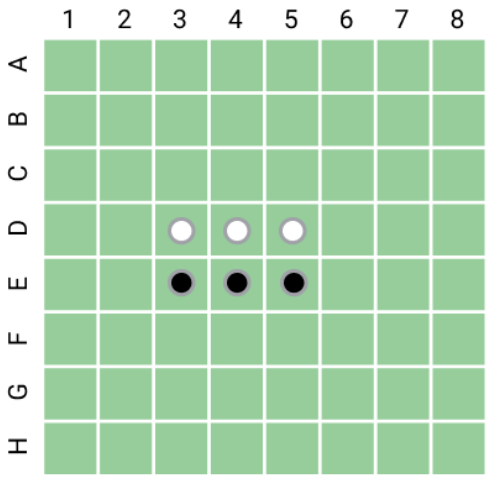
Othello



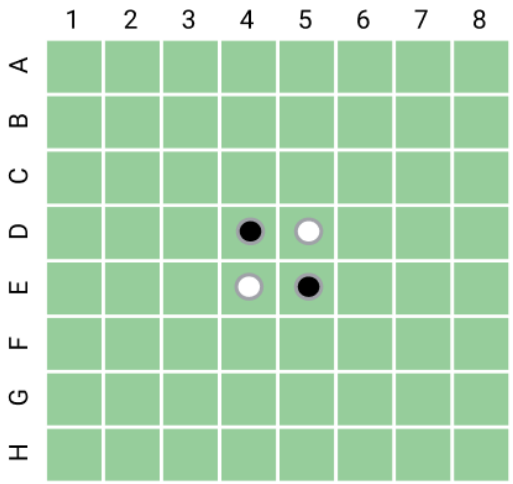
Game of Othello: A Toy Model

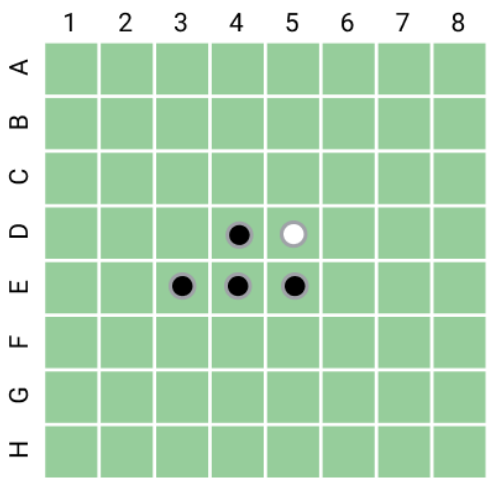


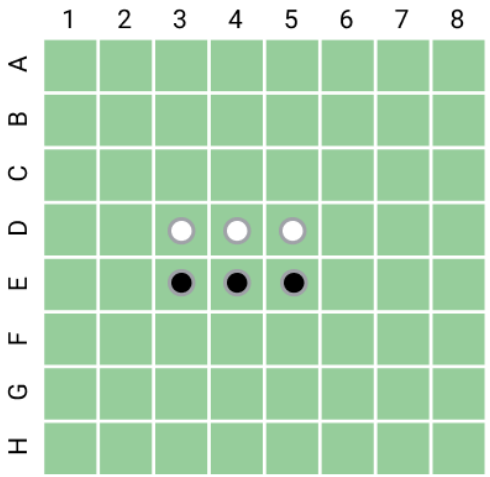




Game of Othello: A Toy Model





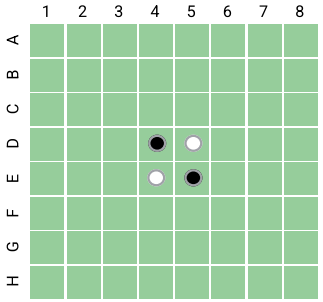


Toy model captures aspects of the general case:

- Legality is nonlinear function of board state and
- The board state is a nonlinear function of the moves

Trained “GPT-Othello” to predict tokens in transcripts of Othello games

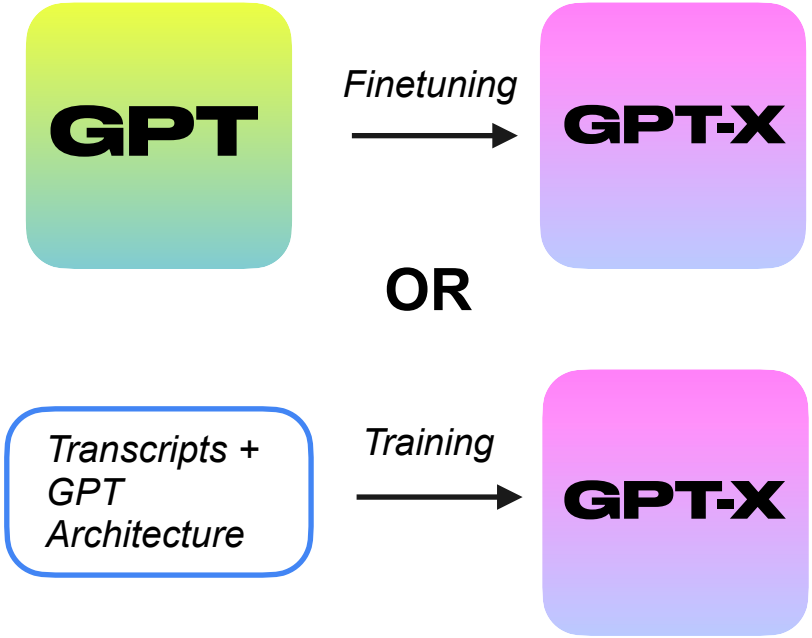
Example task: C4 C3 D3 C5 D6 F4 B4 C6 B5 B3 B6 E3 C2 __?

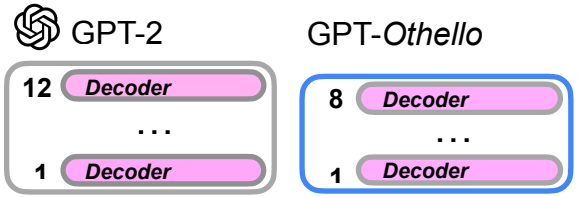


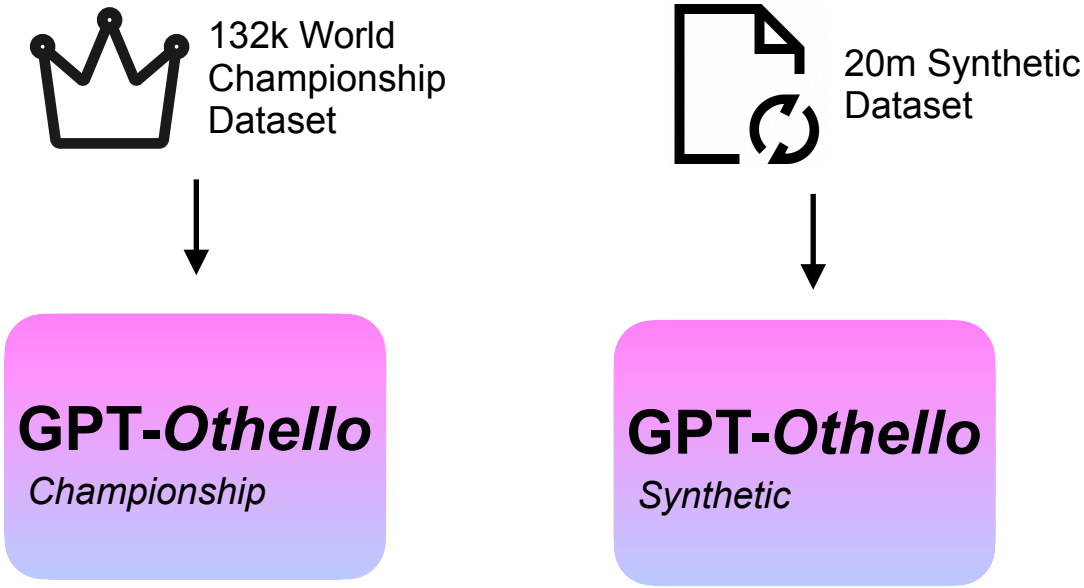
→ E3, D3, C3, … → 32, 26, 18, … → 60 x 512 **Word Embeddings**

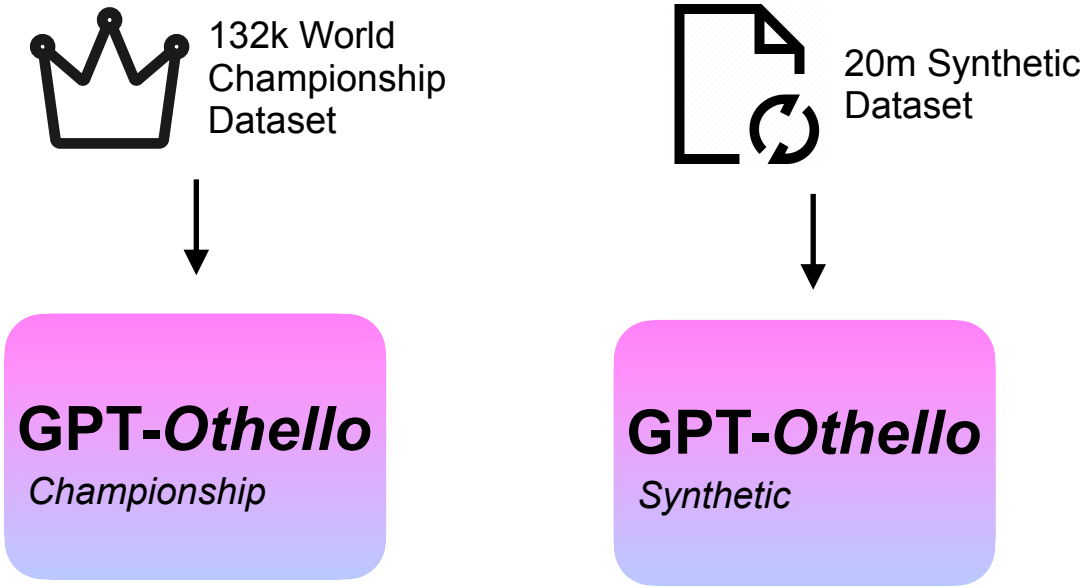
No prior knowledge of game, rules, board.
Just sequences of tokens.

Game Play Transformer






132k World Championship Dataset
20m Synthetic Dataset
GPT-Othello
Championship
GPT-Othello
Synthetic

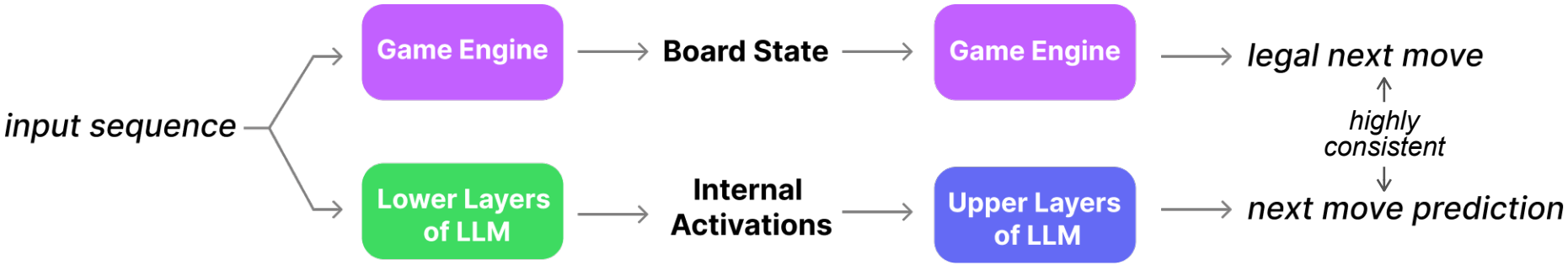


Can GPT-Othello predict a (legal) next move?

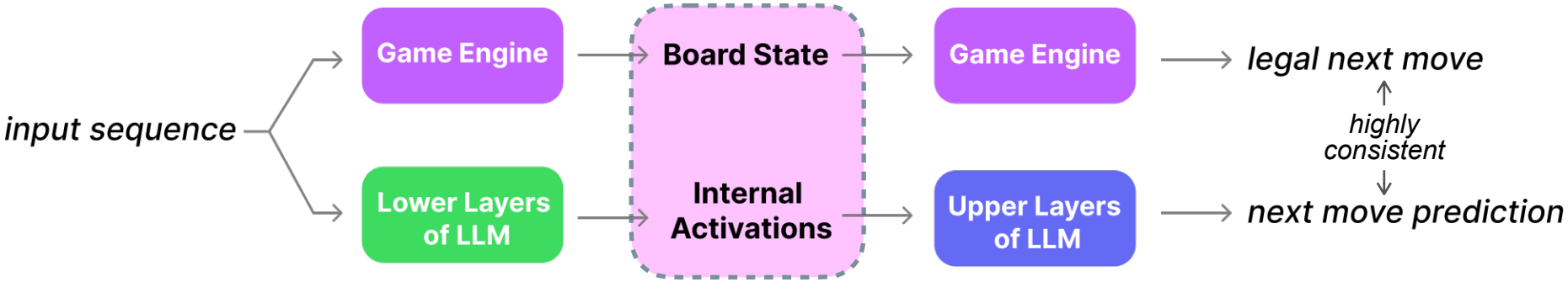
Percentage of illegal top predictions (error) in validation

Synthetic	0.01%
Championship	5.17%
Randomly Initialized	93.29%

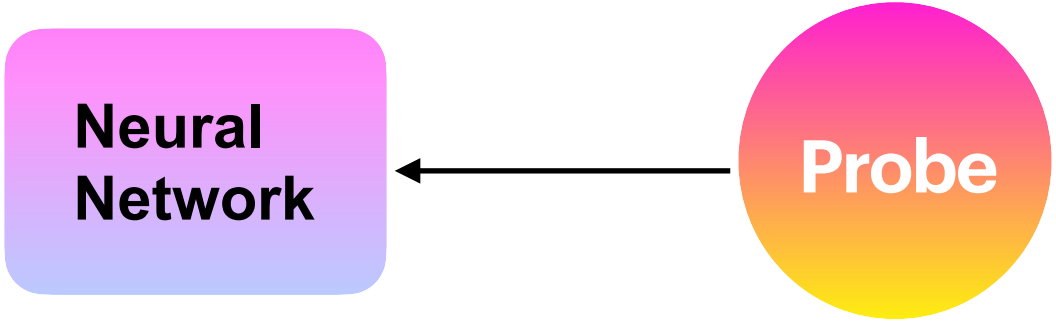
Comparison of Game Engine to GPT



Comparison of Game Engine to GPT



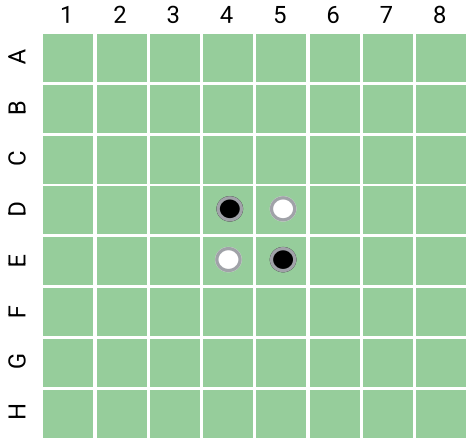
In language models, we do this by training probes—classifiers—on the language model's layer activations.



Alain & Bengio, 2016

Defining Board State Probe for Othello GPT

For each square **s** on the board, can we train a simple classifier f_s, such that $f_s(x_i) = \{\text{white, black, empty}\}$ (where x_i represents the value of concept $\mathbf{C}$ in the input) reflecting whether **s** is white, black, or empty?



$f_{A8}(x_i) = \{\text{empty}\}$

Probe Experiment

Modeling: 3-way classification (black/blank/white) for each square (64 in total)

Features: Layer activations between Transformer blocks

Linear Probe:

$$\mathrm{softmax}(Wx)$$

$$W \in \mathbb{R}^{F \times 3}$$

Two-layer Probe:

$$\mathrm{softmax}(W_1 \mathrm{ReLU}(W_2 x))$$

$$W_1 \in \mathbb{R}^{H \times 3}$$

$$W_2 \in \mathbb{R}^{F \times H}$$

Probe Performance

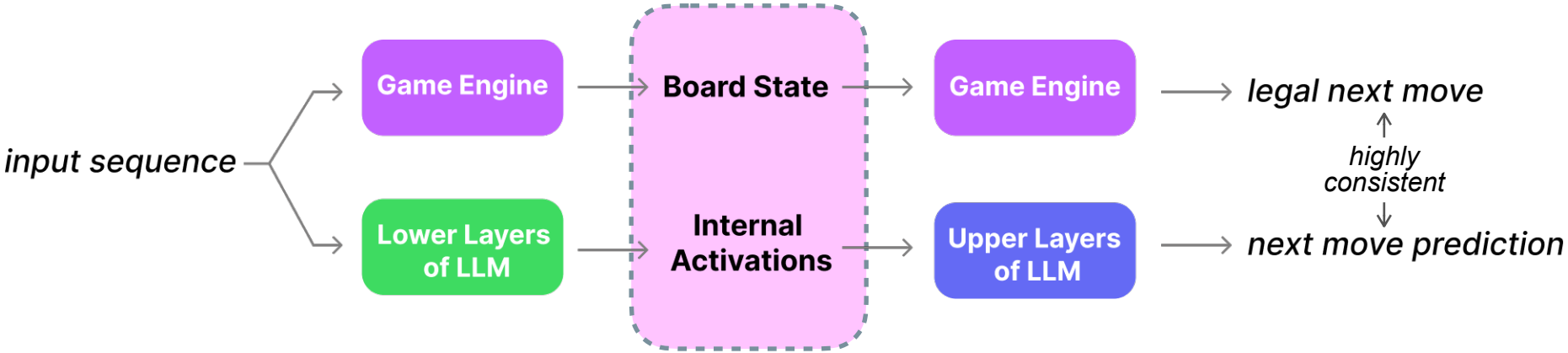
Error rates of linear probe across different layers

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
Randomized	26.7	27.1	27.6	28.0	28.3	28.5	28.7	28.9
Championship	24.2	23.8	23.7	23.6	23.6	23.7	23.8	24.3
Synthetic	21.9	20.5	20.4	20.6	21.1	21.6	22.2	23.1

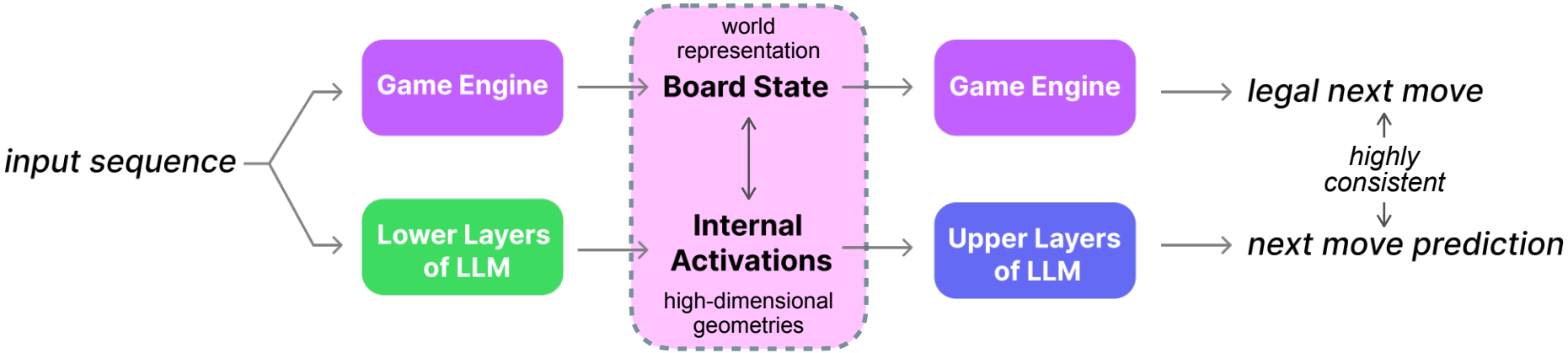
Error rates of nonlinear probe across different layers

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
Randomized	25.5	25.4	25.5	25.8	26.0	26.2	26.2	26.4
Championship	12.8	10.3	9.5	9.4	9.8	10.5	11.4	12.4
Synthetic	11.3	7.5	4.8	3.4	2.4	1.8	1.7	4.6

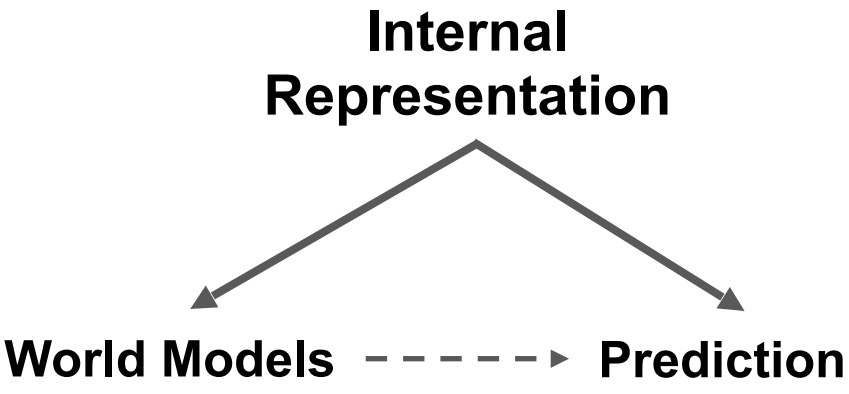
Probe provides evidence of board state model

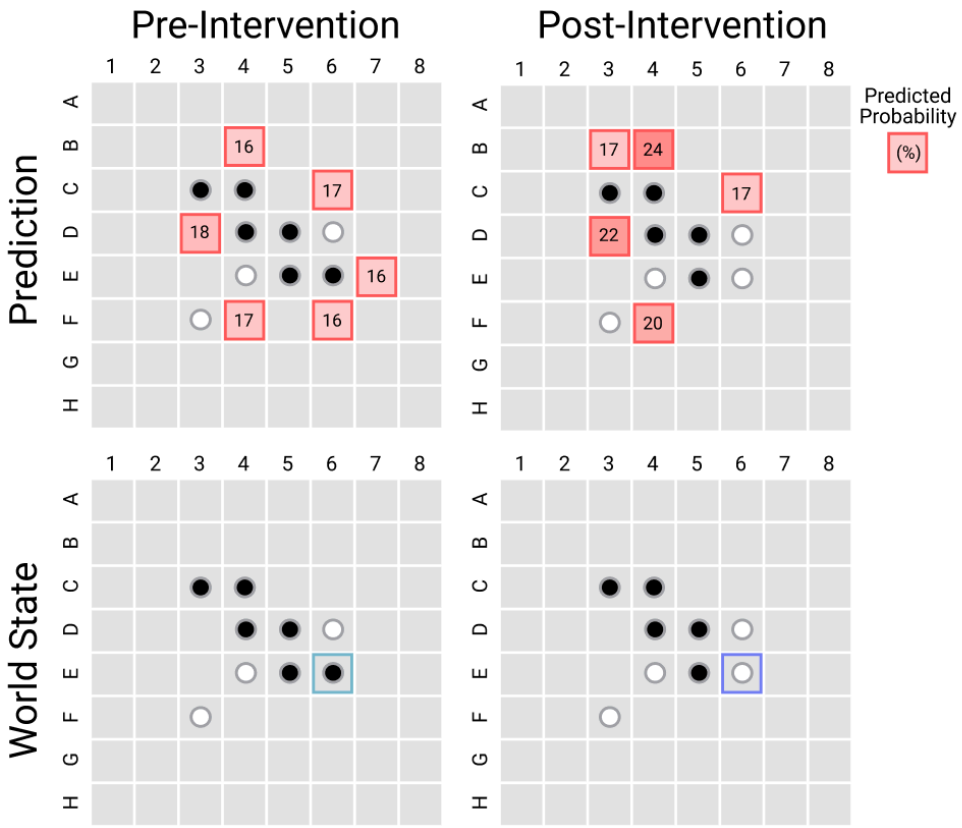


Probe provides evidence of board state model

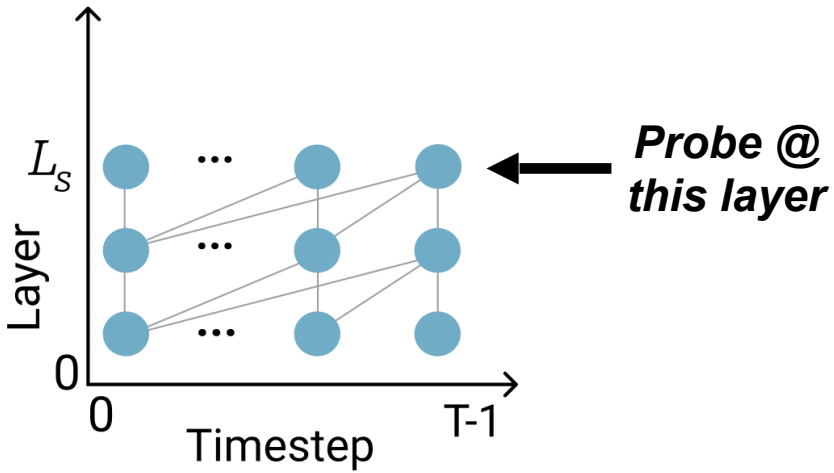


Controllable World Model



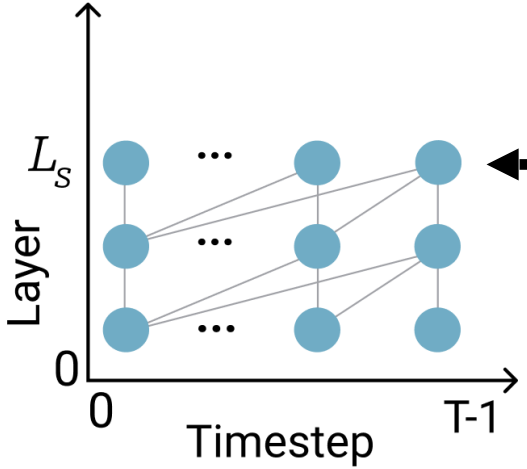


Intervention

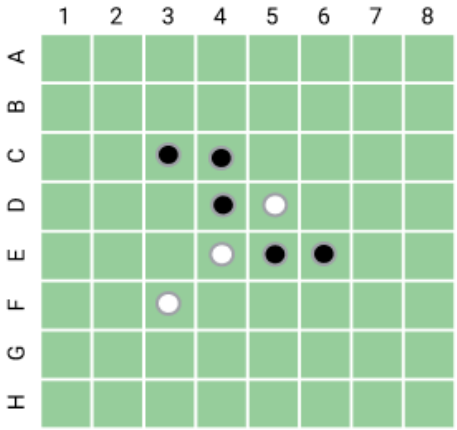


: Factual feature

Intervention

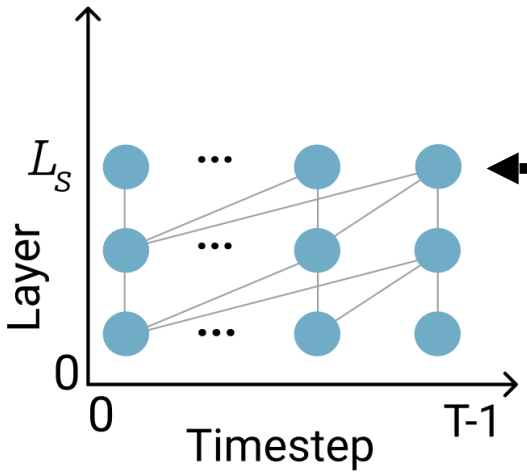


Probe @ this layer

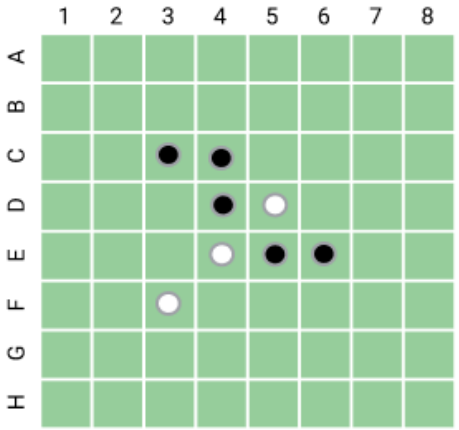


: Factual feature

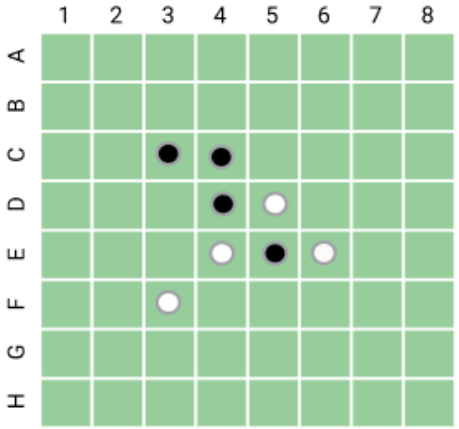
Intervention



Probe @ this layer

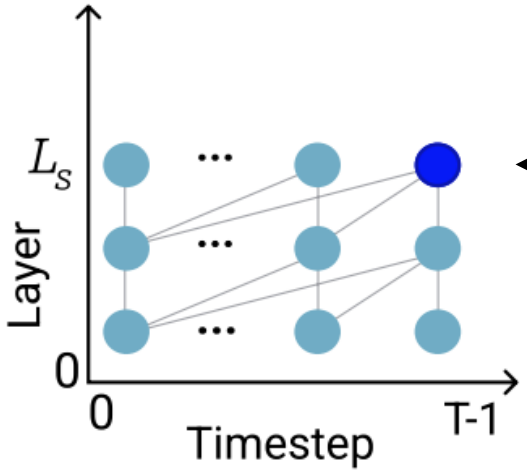


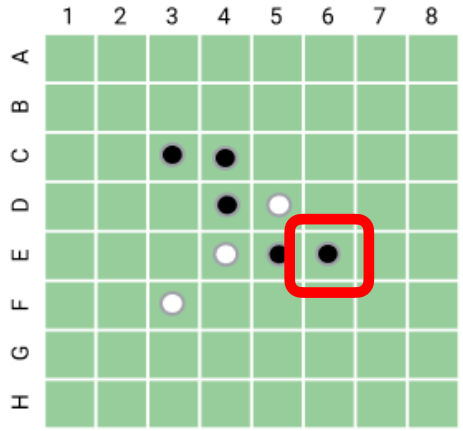
Intervene with gradient descent



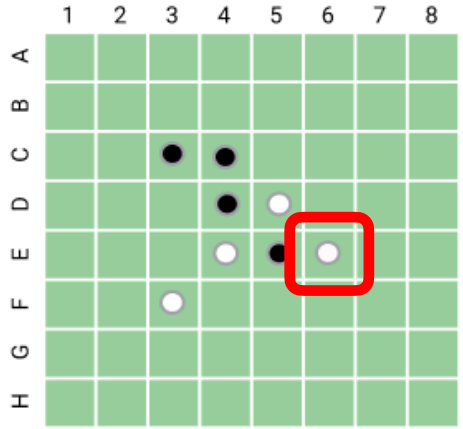
: Factual feature

Intervention





Intervene with gradient descent



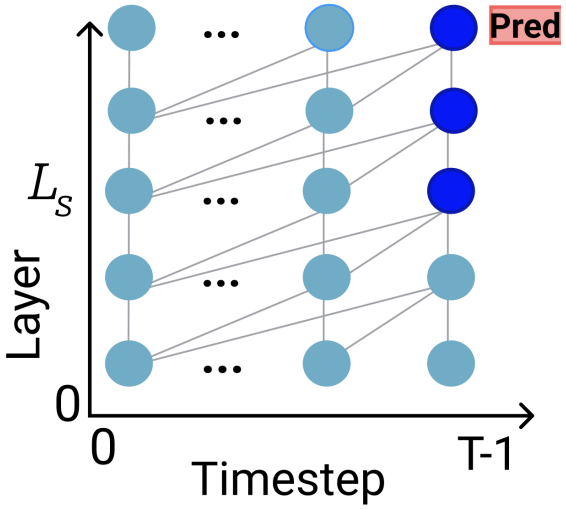
Copy back

: Factual feature

: Intervened feature

pause, hack, put-back (continue)

Intervention



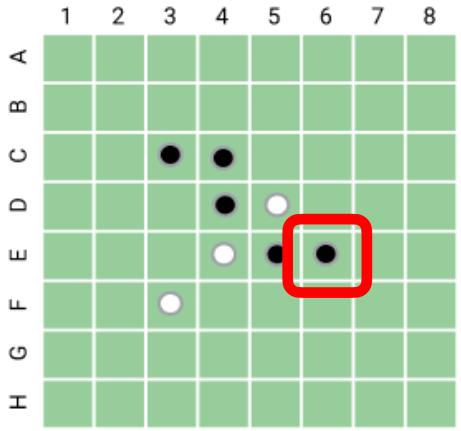
: Factual feature

: Intervened feature

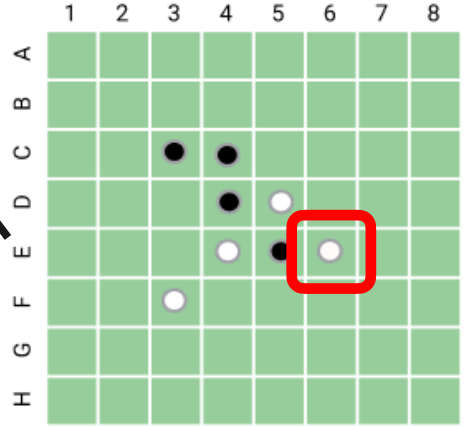
: Self-corrected

Probe @ this layer

Copy back

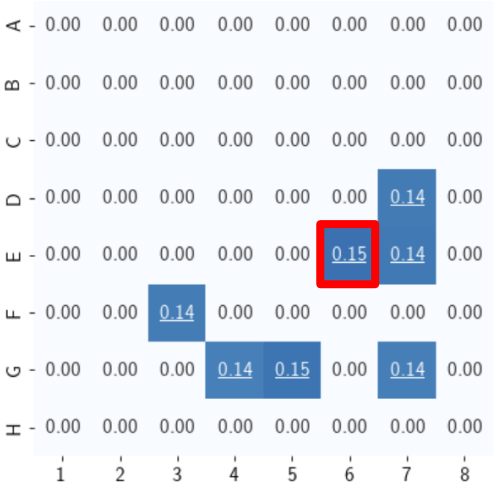


Intervene with gradient descent



Attribution via intervention

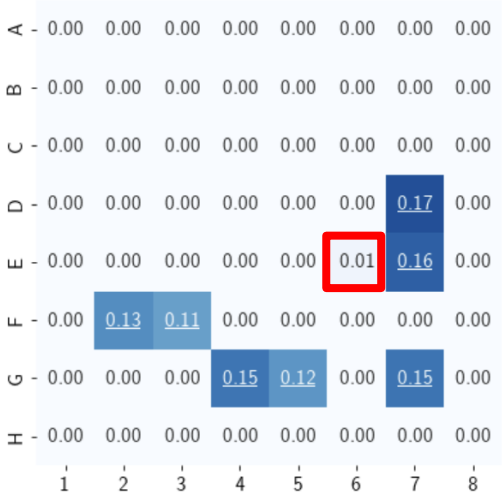
Logit map ***before***
flipping E3 world representation



0.15

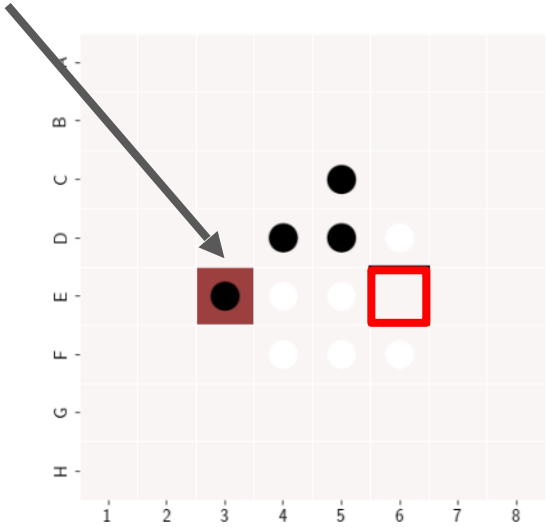
-

Logit map ***after***
flipping E3 world representation



0.01

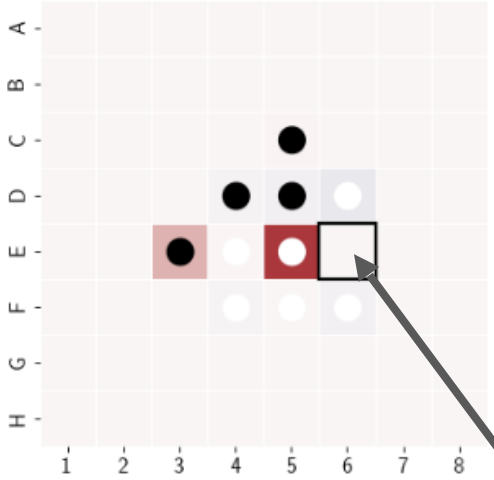
=



0.14

: E6, the square we attribute

Latent saliency maps



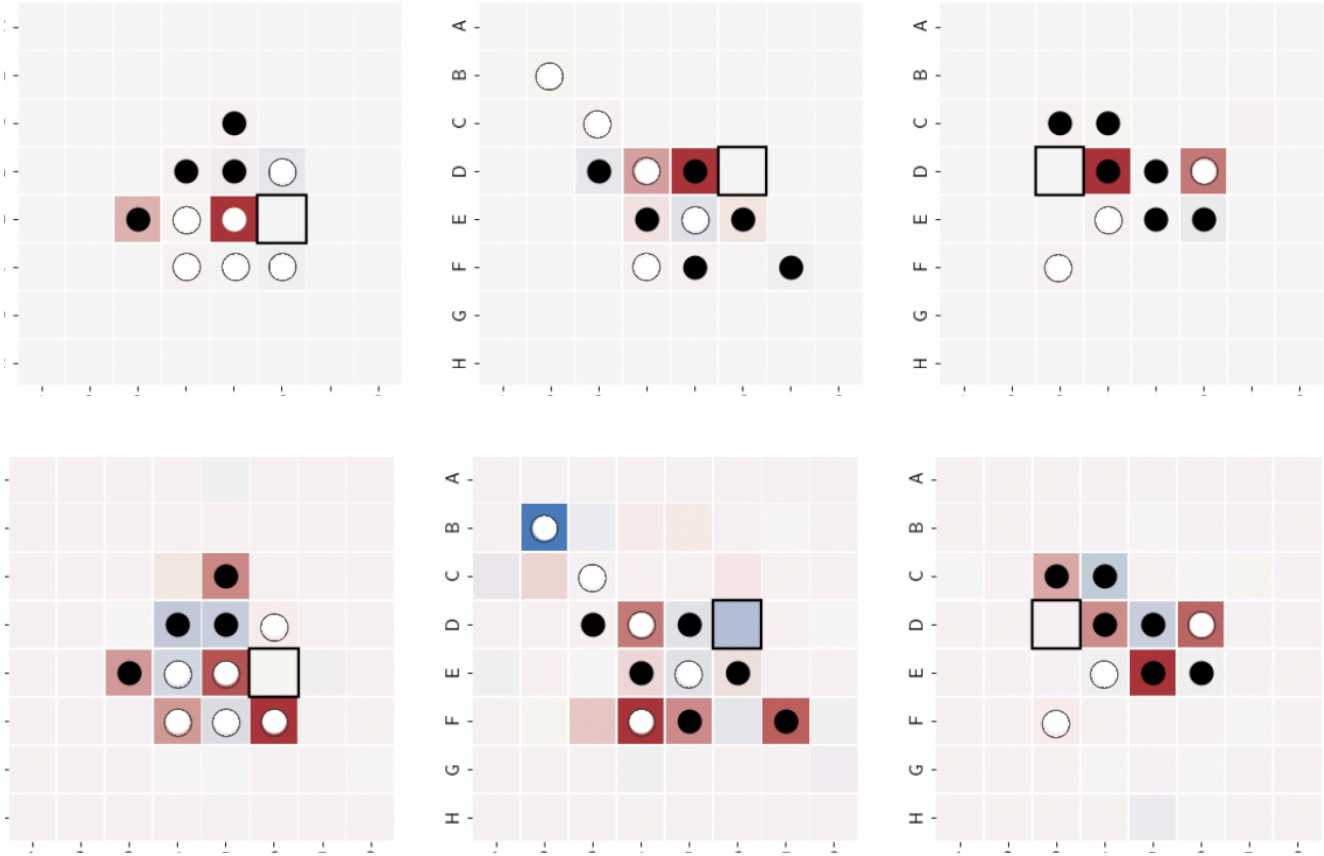
By holding the world model of the colored tile as it is, Othello-GPT thinks it is…

Negatively **Positively**

…contributing to the prediction of the next step.

Enclosed E6: a legal next step we attribute

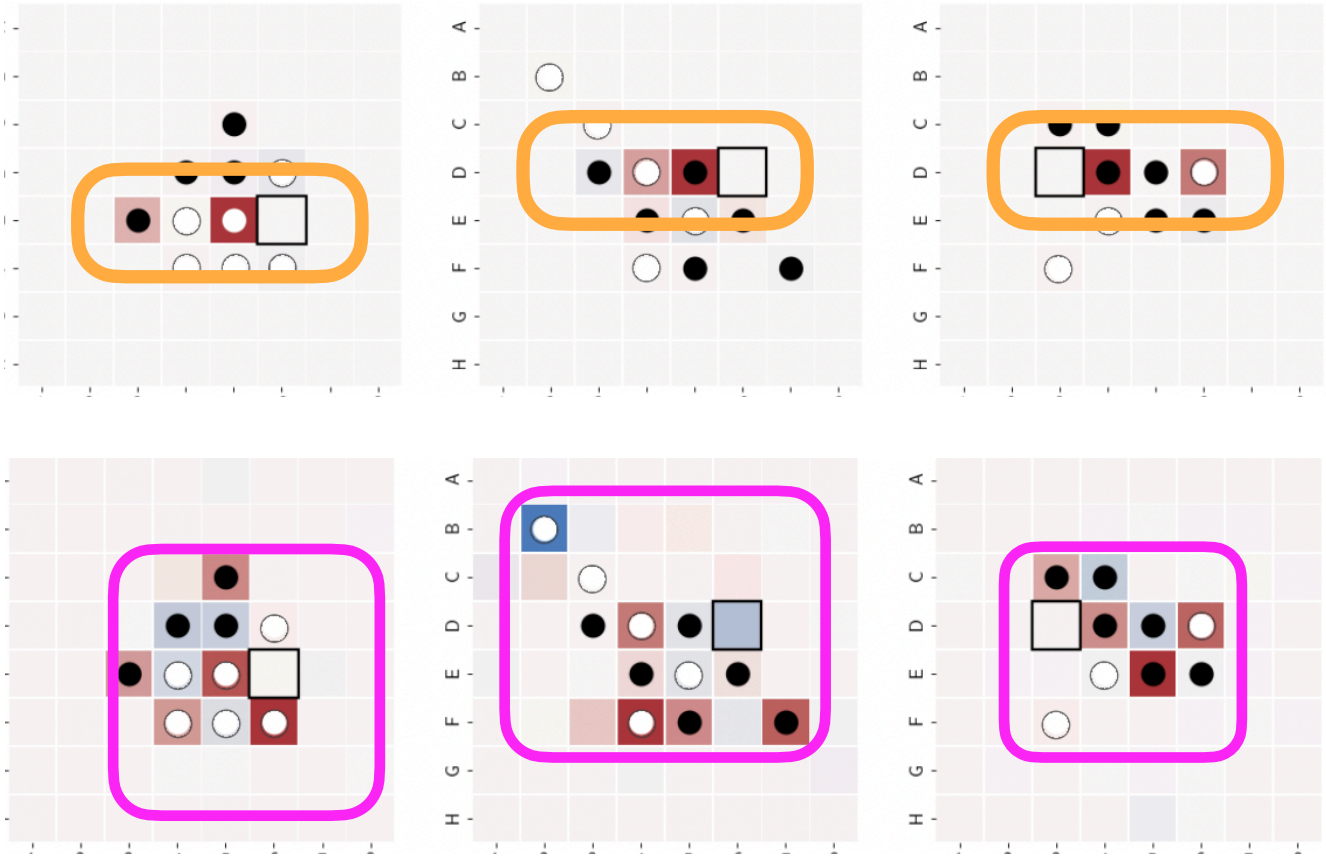
Which row is strategic? Which is just “legal”



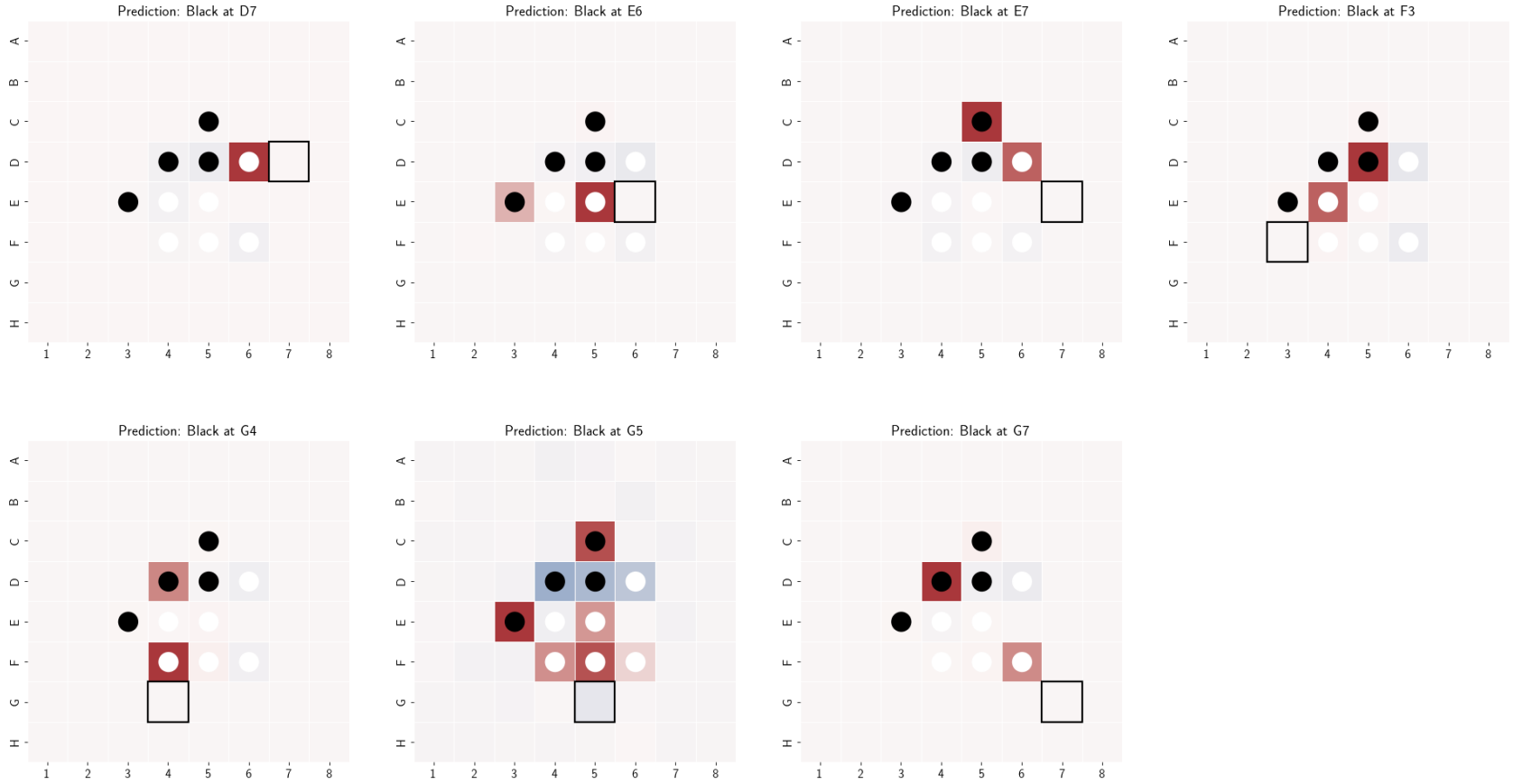
Which row is strategic? Which is just “legal”

Legal

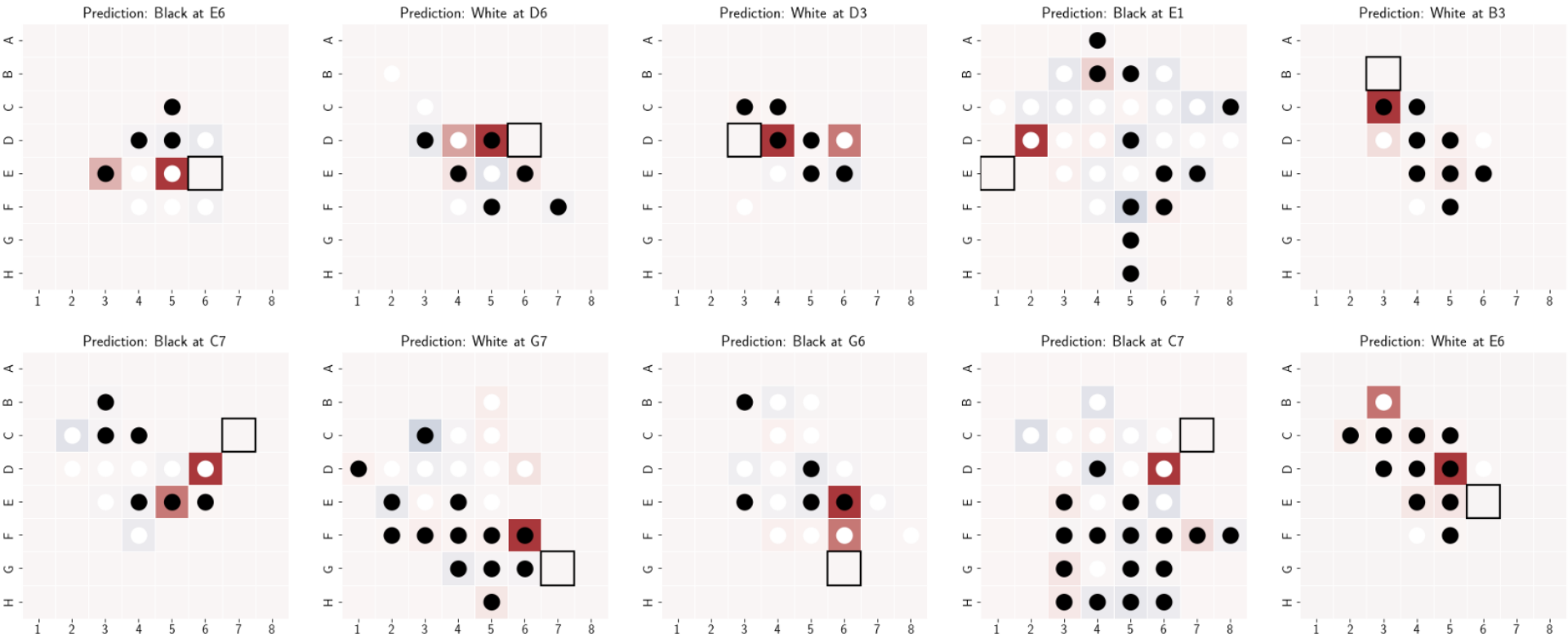
Strategic



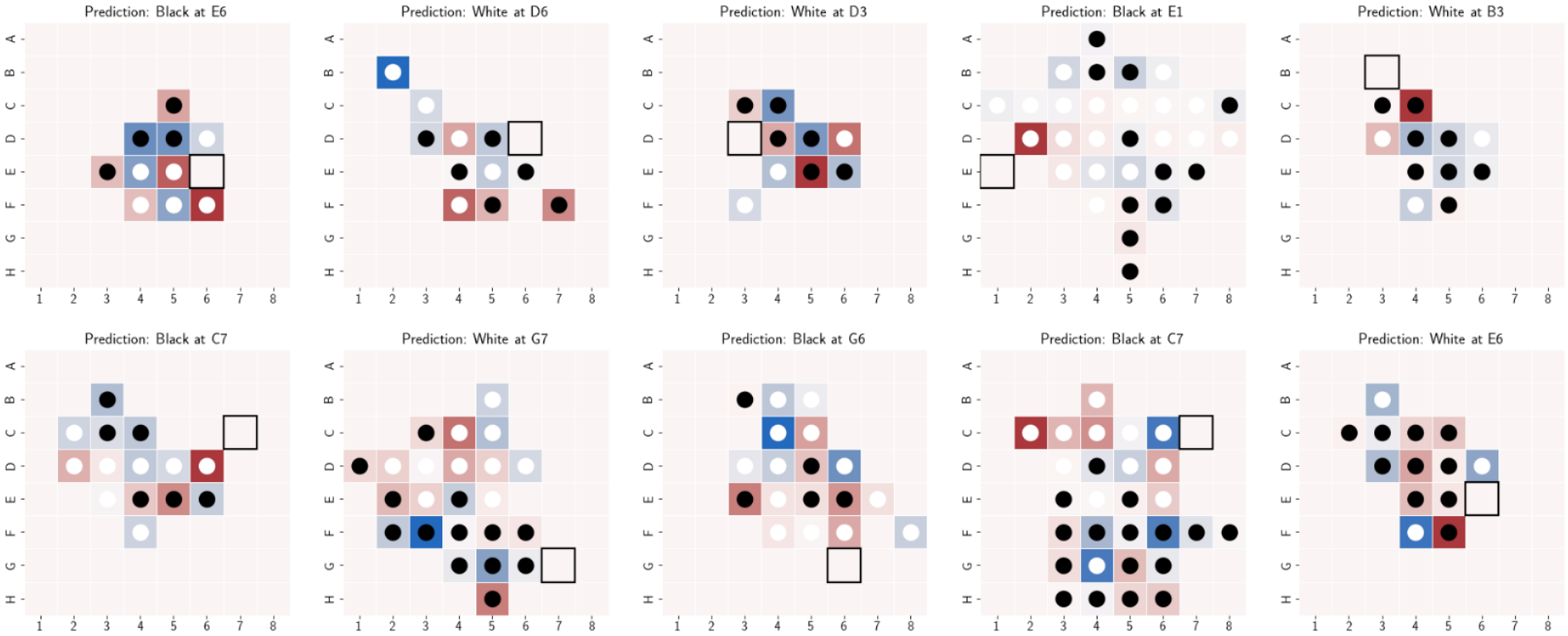
Attributing other legal moves



More cases on synthetic Othello-GPT



More cases on championship Othello-GPT



1 *Development and evaluation of GPT model,* ***GPT-Othello***

2 *Comparison of* ***linear*** *and* ***non-linear*** *probing*

3 *Novel* ***intervention*** *technique*

4 *Novel* ***latent saliency maps*** *built using intervention*

Actually, Othello-GPT Has A Linear Emergent World Representation

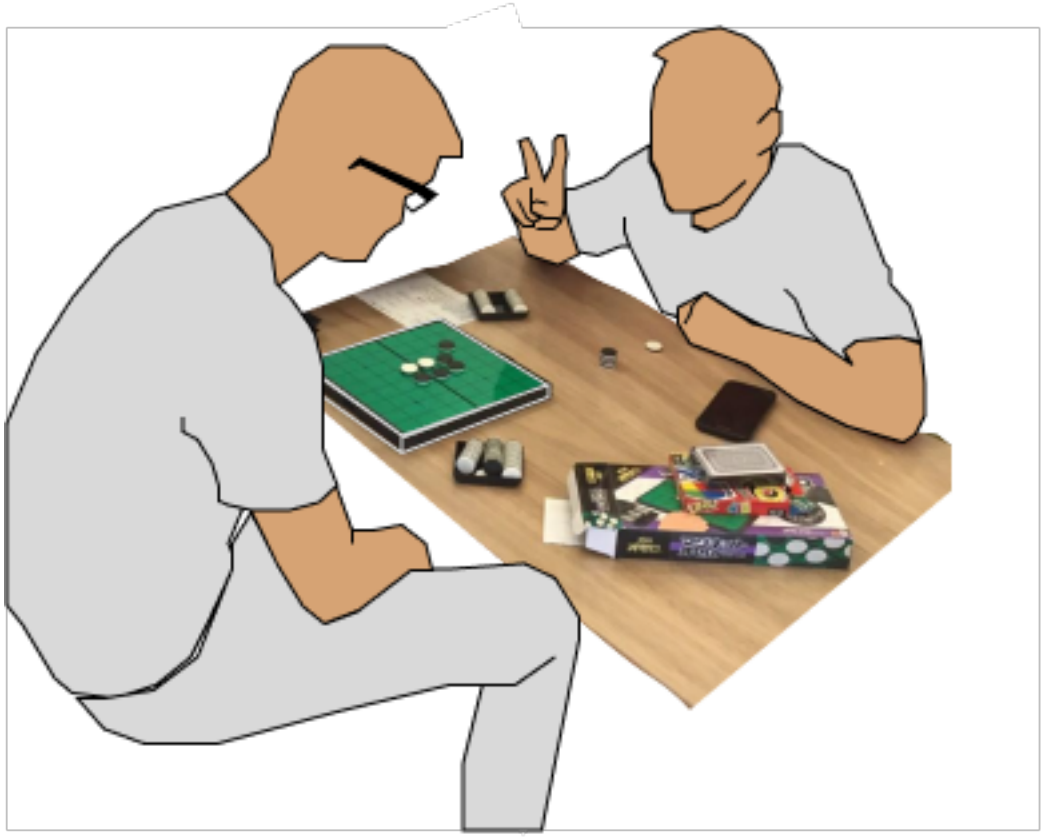
by **Neel Nanda** 23 min read 29th Mar 2023 17 comments

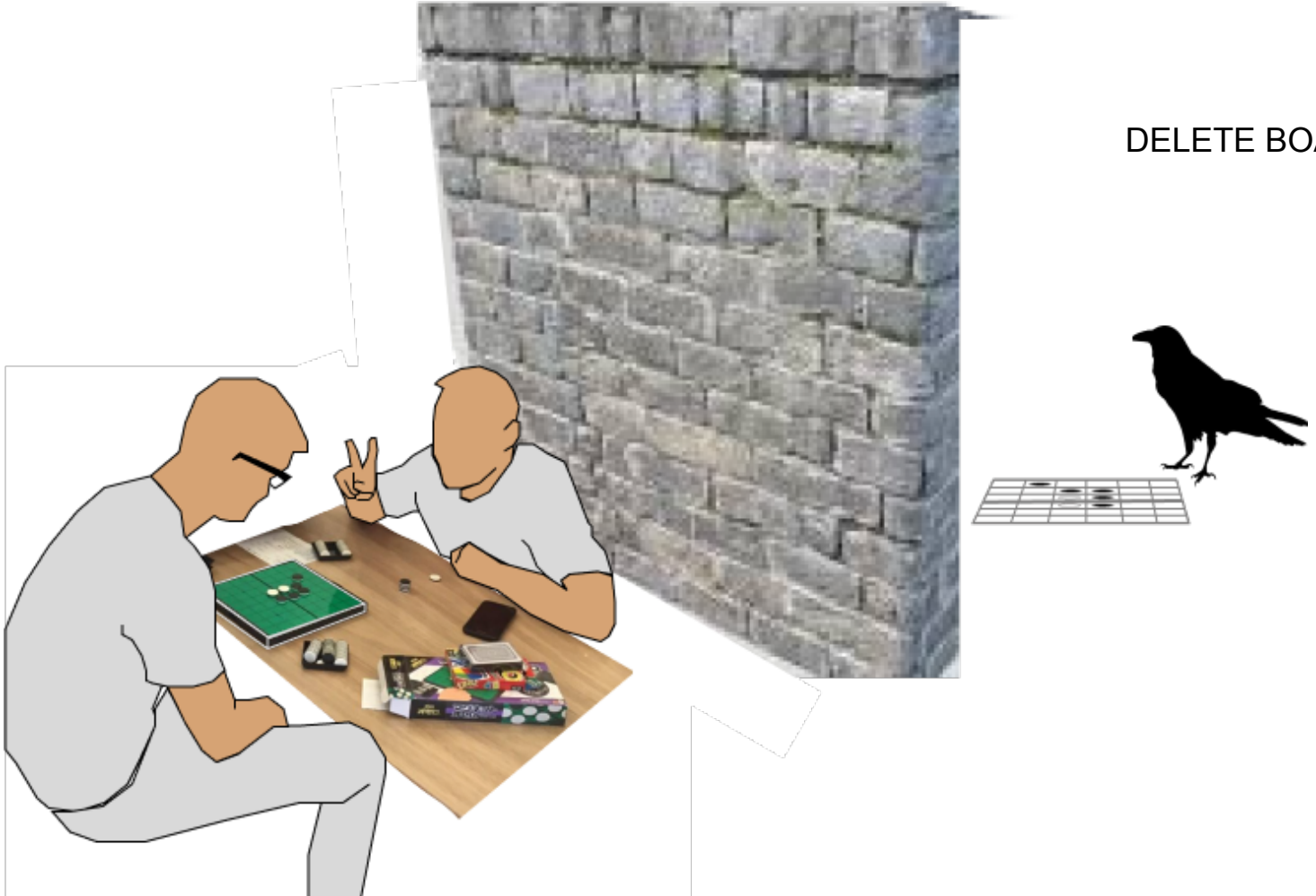
Interpretability (ML & AI) AI Frontpage

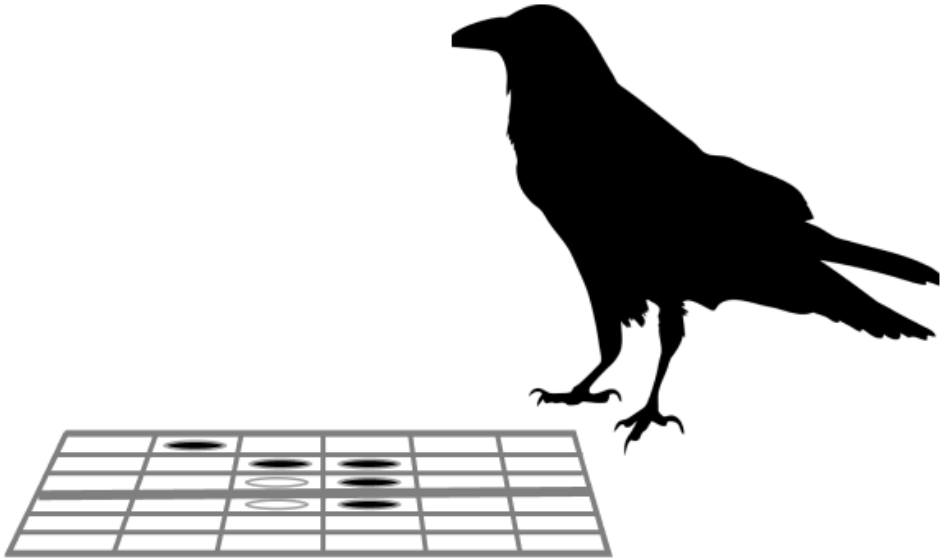
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is a linkpost for https://neelnanda.io/mechanistic-interpretability/othello

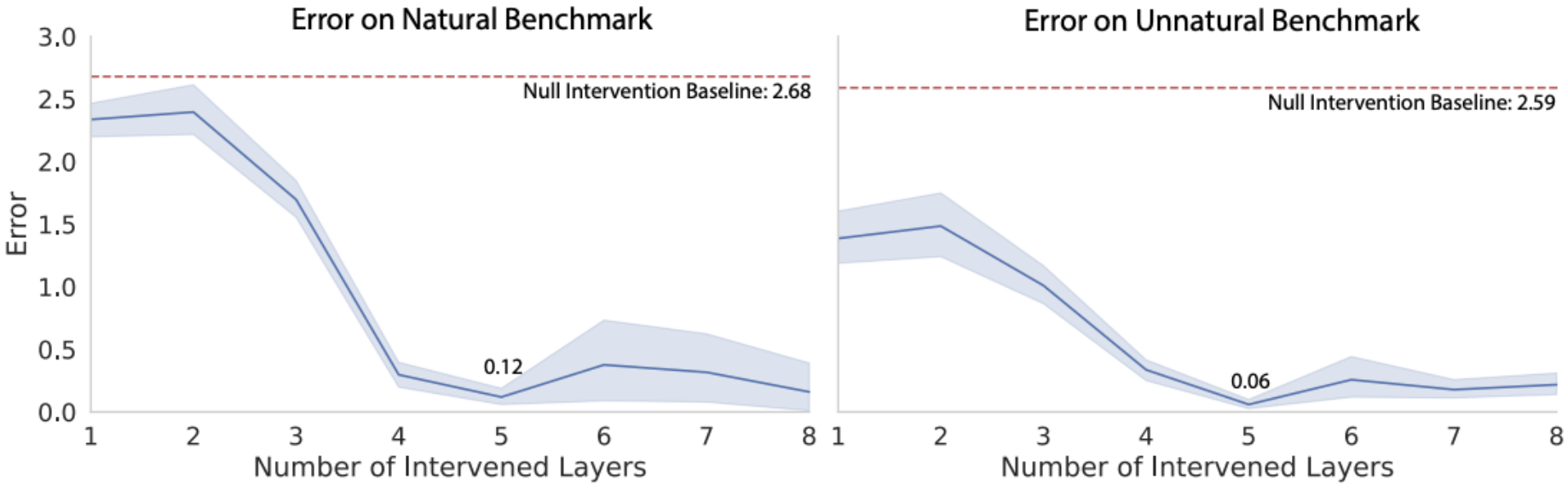

Thank you




DELETE BOARD



Comparing top-N prediction to the N ground truths



Superficial Statistics **vs** ***Meaningful Representations***

(World Models)

Geometry of probe weights