Othello-GPT

Emergent Representations (? (Question Mark))

Lars Tapken, 2024/11/07
Seminar: Mechanistic Interpretability, CL Heidelberg

Contents

- Why?
- Othello-GPT
- Some thoughts
- Sources
- Discussion & FAQs

Why?

Some motivation

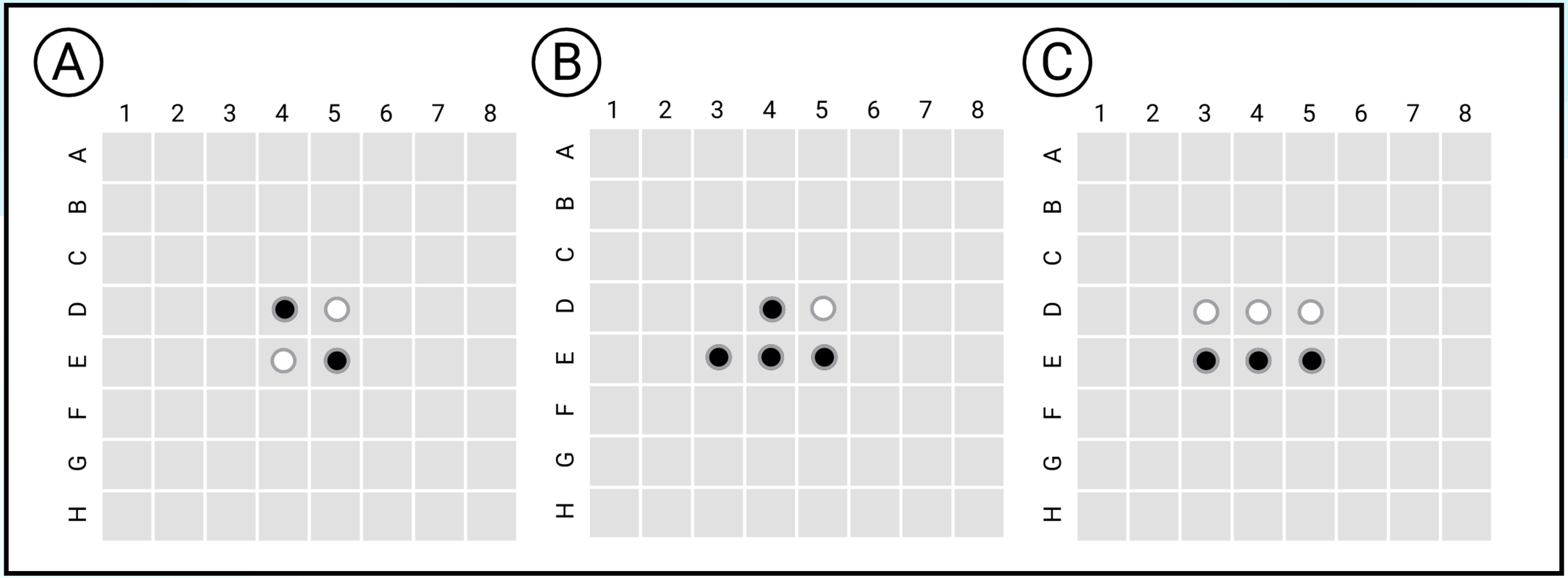
- What are ✨ *F e a t u r e s* ✨*?*
- Are models just parrots?
- Toy models are cool!
- Finding linearity would be good for Interpretability
- …

Disclaimer

- Sources are *(Li et al. 2022)* or *(Nanda et al. 2023)* if not otherwise stated
- My own bias:
  - I’m more of a *“LLMs simply use heuristics”* guy :)

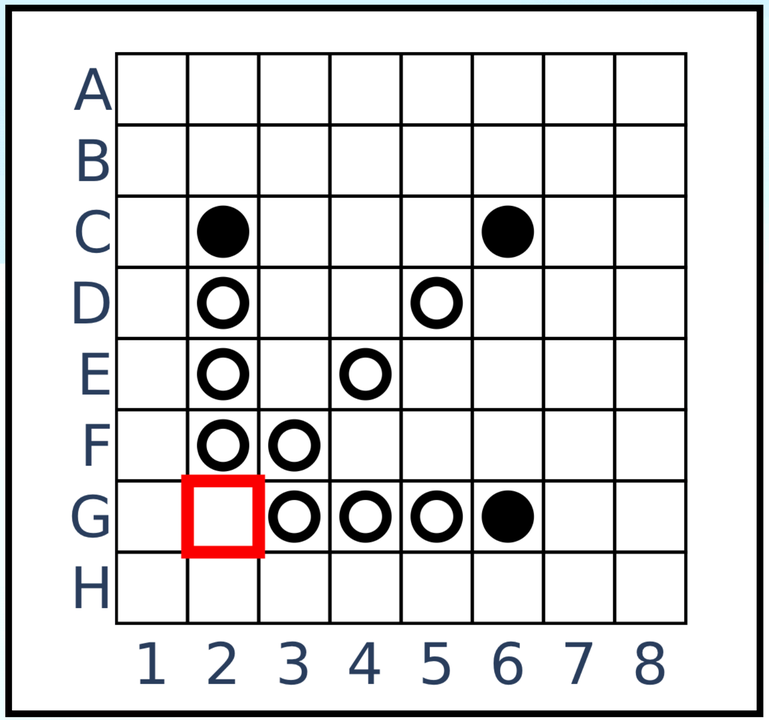
What is Othello?

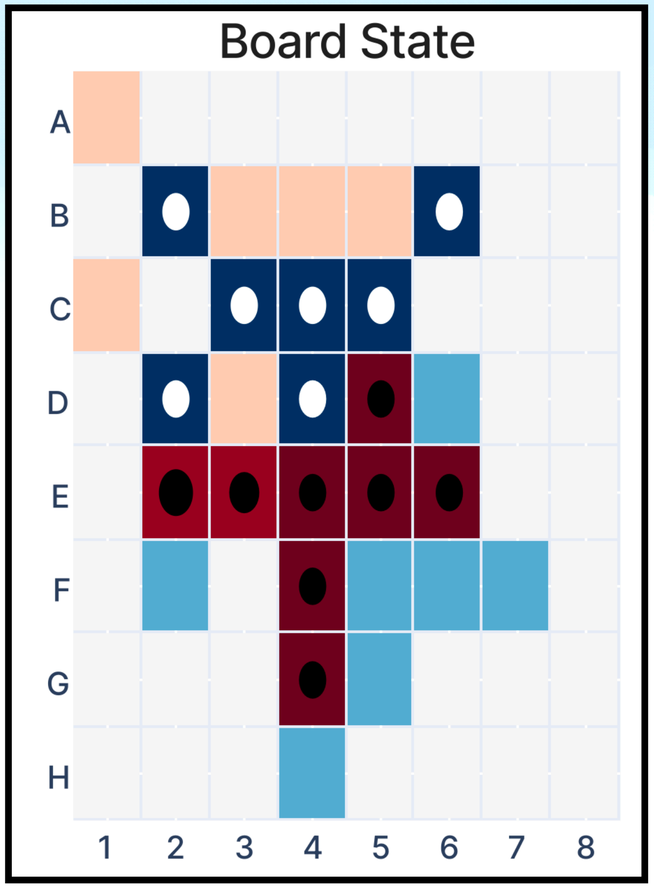
Basic rules



What is Othello?

Two examples



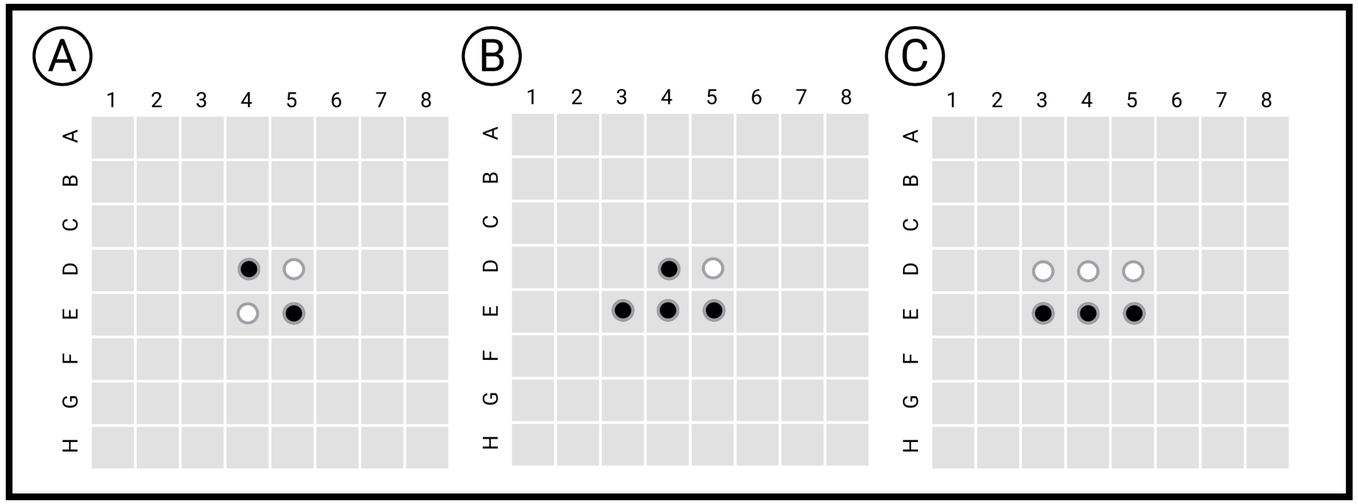


Othello-GPT

Basics

- GPT-2 style, autoregressive Transformer
- 8 layers, 8 heads
- “Next disc prediction”
- Trained on sequences of game moves, e.g.:
  - *<|Start|>, E3, D3*
  - *Tokenize to:*

    _, 33, 27



Othello-GPT

Data

Two Datasets:

1. ***Championship***
   - ~ 140K **actual** game sequences
   - 4:1 split
2. ***Synthetic***
   - ~ 25M games
   - uniformly sampled **legal** game sequences
   - 5:1 split

Othello-GPT

Data

Sequences are text only!

No other information is given!

Othello-GPT

Evaluation

- Measured in Error Rate (= 1 – Accuracy)
- Is the next prediction **legal**?
- Untrained:
  - ER = 93.29%
- Championship:
  - ER = 5.17%
- Synthetic:
  - ER = 0.01%

Othello-GPT

Evaluation

- Memorization?

➡ Remove one quarter of all possible train sequences

  - ER = 0.02% (vs 0.01% before)

Othello-GPT

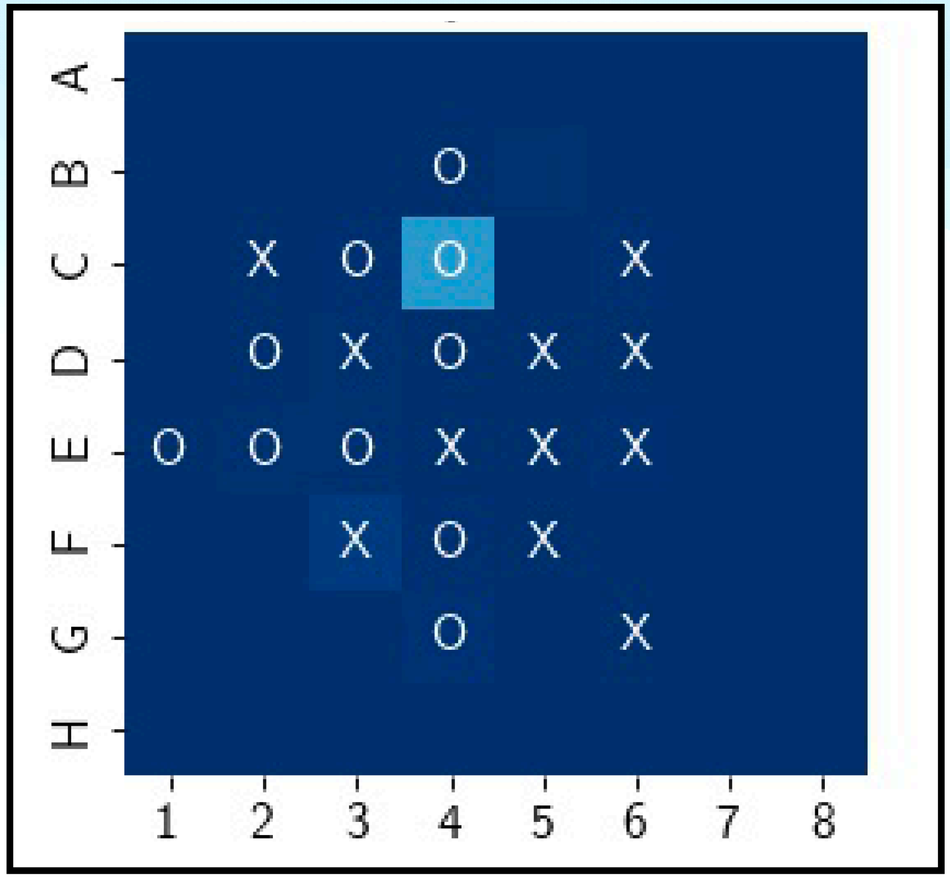
What's inside?

- Internal representations of the game state?
- Probes!
- Train probes on internal activations for each tile:
  - {Black, White, Empty}
- Linear & Non-Linear:
  - $p_\theta(x_t^l) = \text{softmax}(Wx_t^l)$
  - $p_\theta(x_t^l) = \text{softmax}(W_1\text{ReLU}(W_2x_t^l))$

Othello-GPT

Probes

- Board reconstruction example from probes



Othello-GPT

Probes

- ERs of **Linear** probes:

	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

- ERs of **Non-Linear** probes:

	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

Othello-GPT

Probes

If there is an internal representation of the board state,

it does not have a simple linear form …

Othello-GPT

Probes

BEWARE!

Correlation ≠ Causation!

Othello-GPT

Validating Probes

- Activations → Probe(s) → Board States

- What if we manipulate the internal activations of Othello-GPT? …

Othello-GPT

Validating Probes

- Validate causality
- Does changing the internal board state, lead to **different** and **legal** next disc predictions?

Othello-GPT

Validating Probes

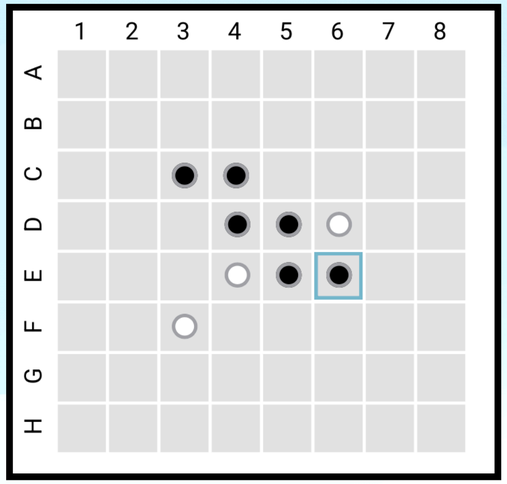
- Change internal activations of board state …
- … until exactly **one (1)** disc changes colour
  - i.e. Black to White

- If new predictions are legal, assume causality.

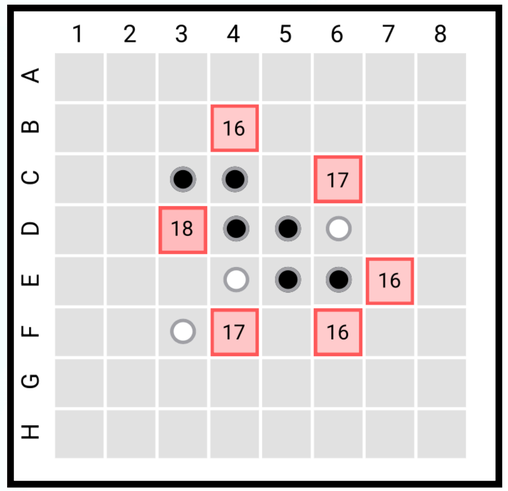
Othello-GPT

Manipulating Activations

Current board state



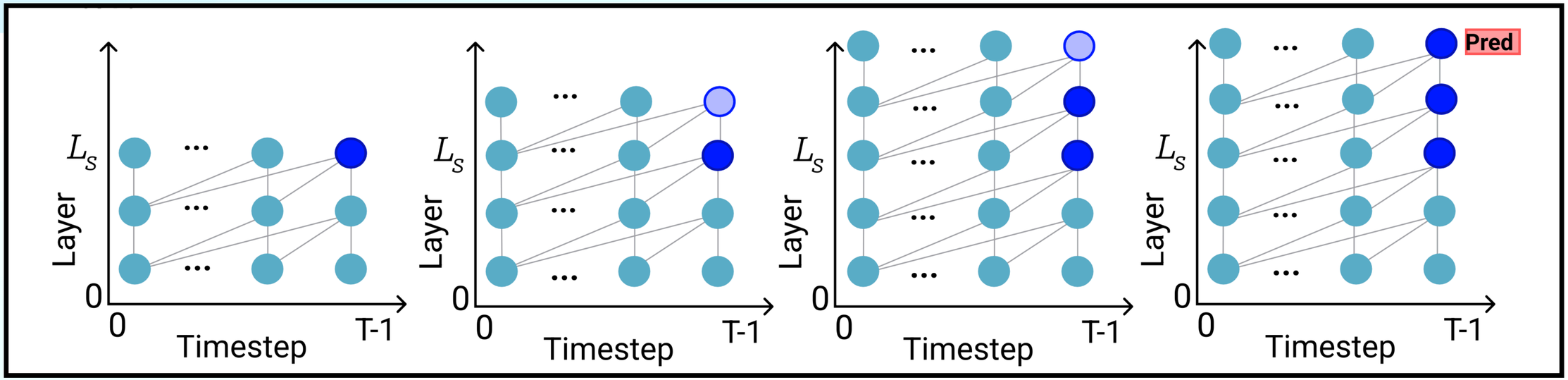
Current predictions



Othello-GPT

Manipulating Activations

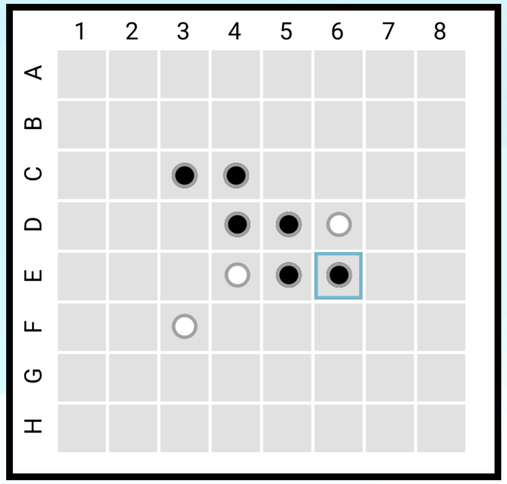
- Change board representation from Layer L_S onwards



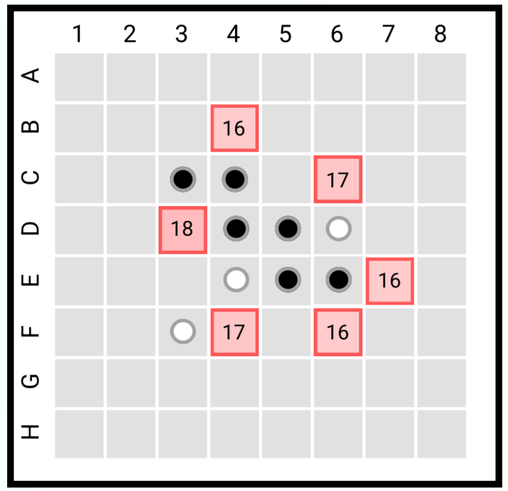
Othello-GPT

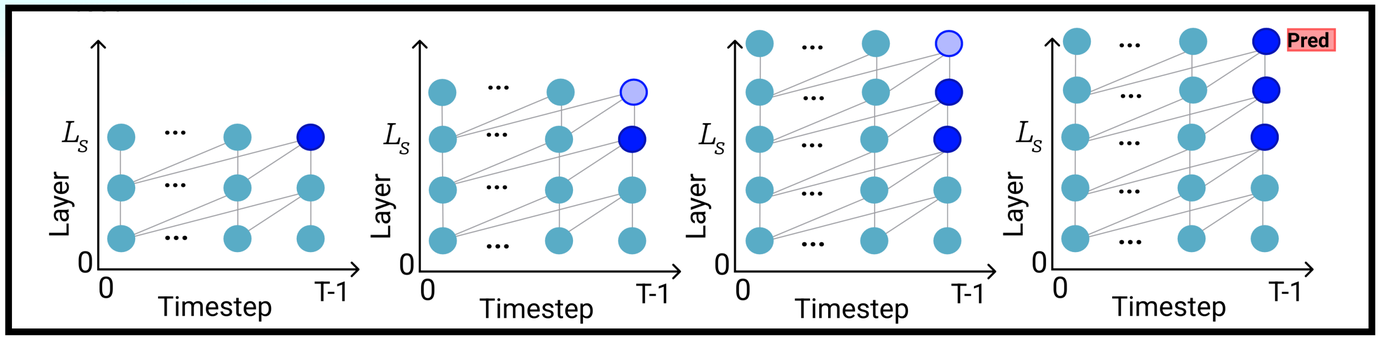
Manipulating Activations

Current board state



Current predictions





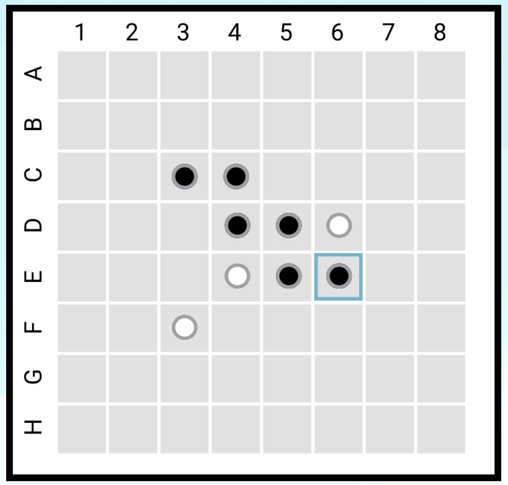
For example:
A1, … D6, E6

(not an actual sequence)

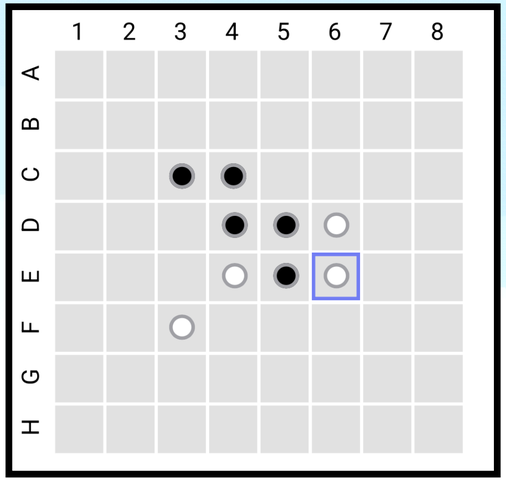
Othello-GPT

Manipulating Activations

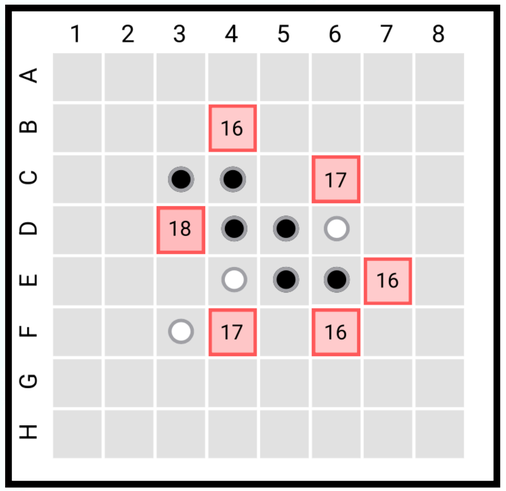
Current board state

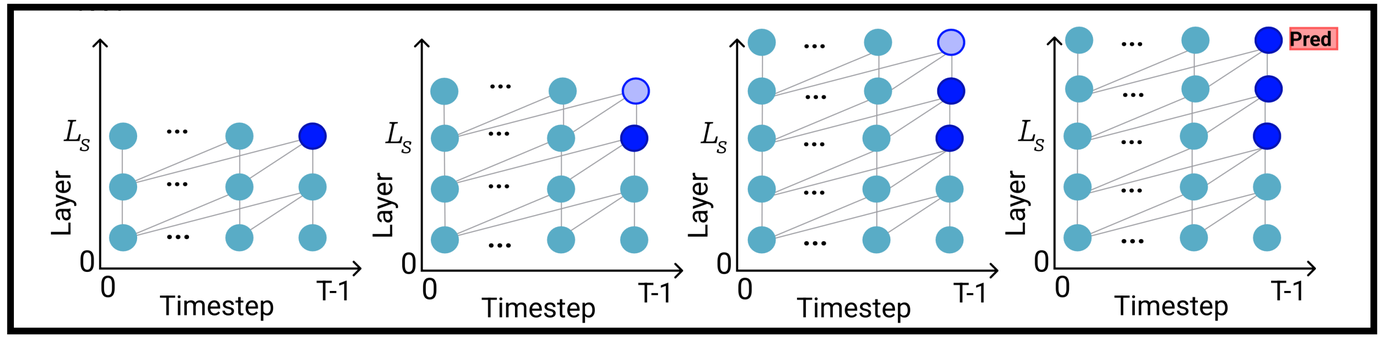


New board state



Current predictions

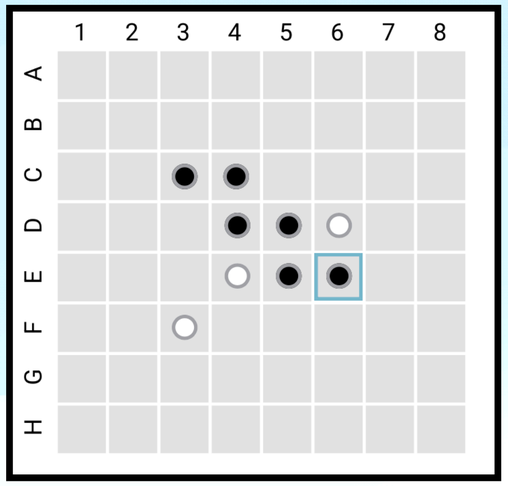




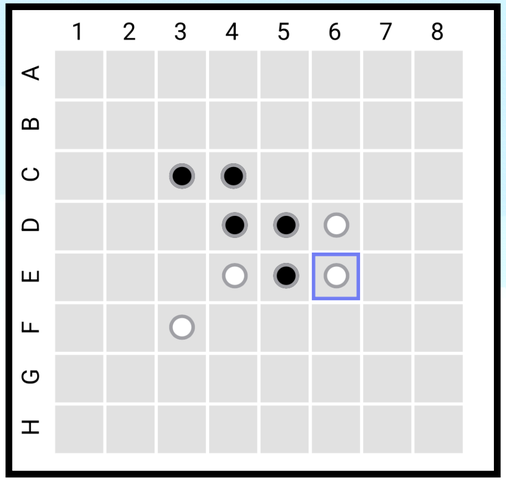
Othello-GPT

Manipulating Activations

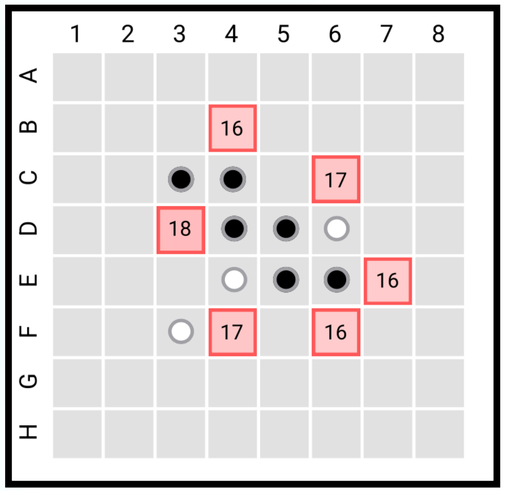
Current board state



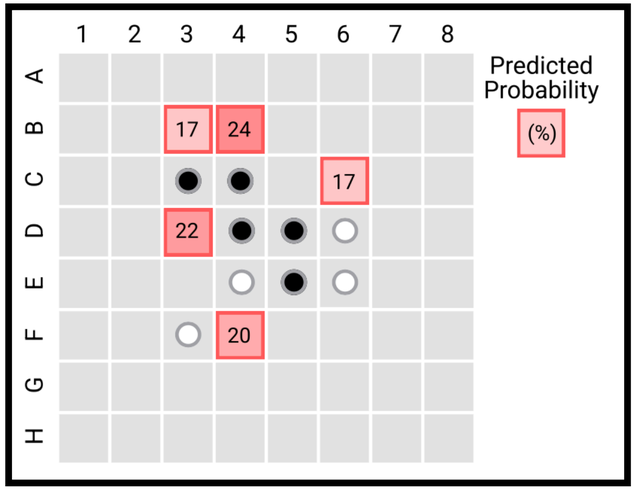
New board state



Current predictions



New predictions

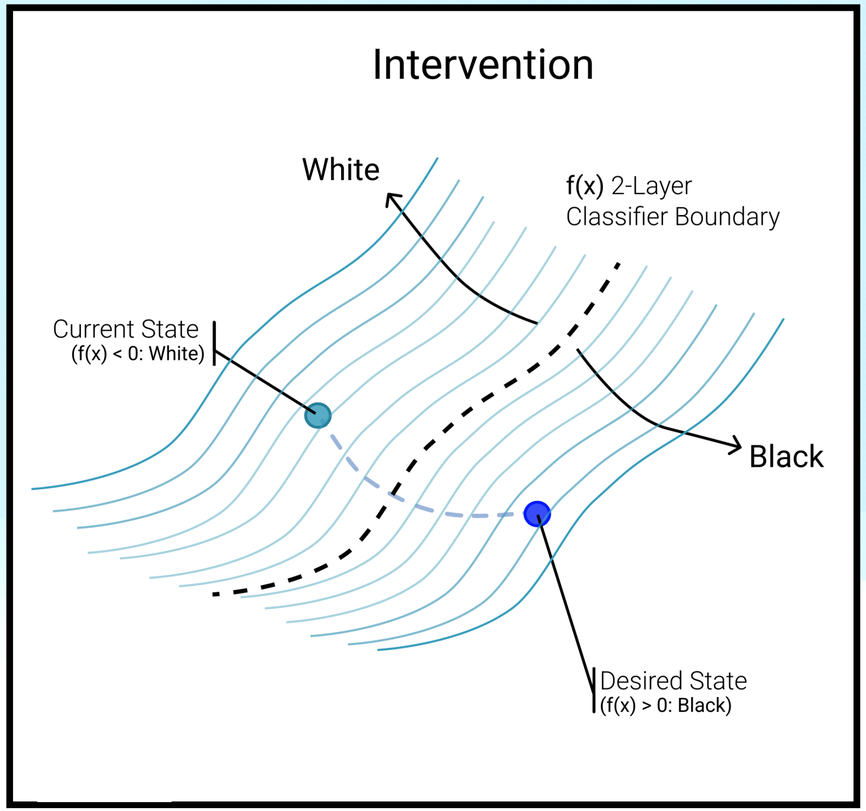


Othello-GPT

Manipulating Activations

- Use GD:

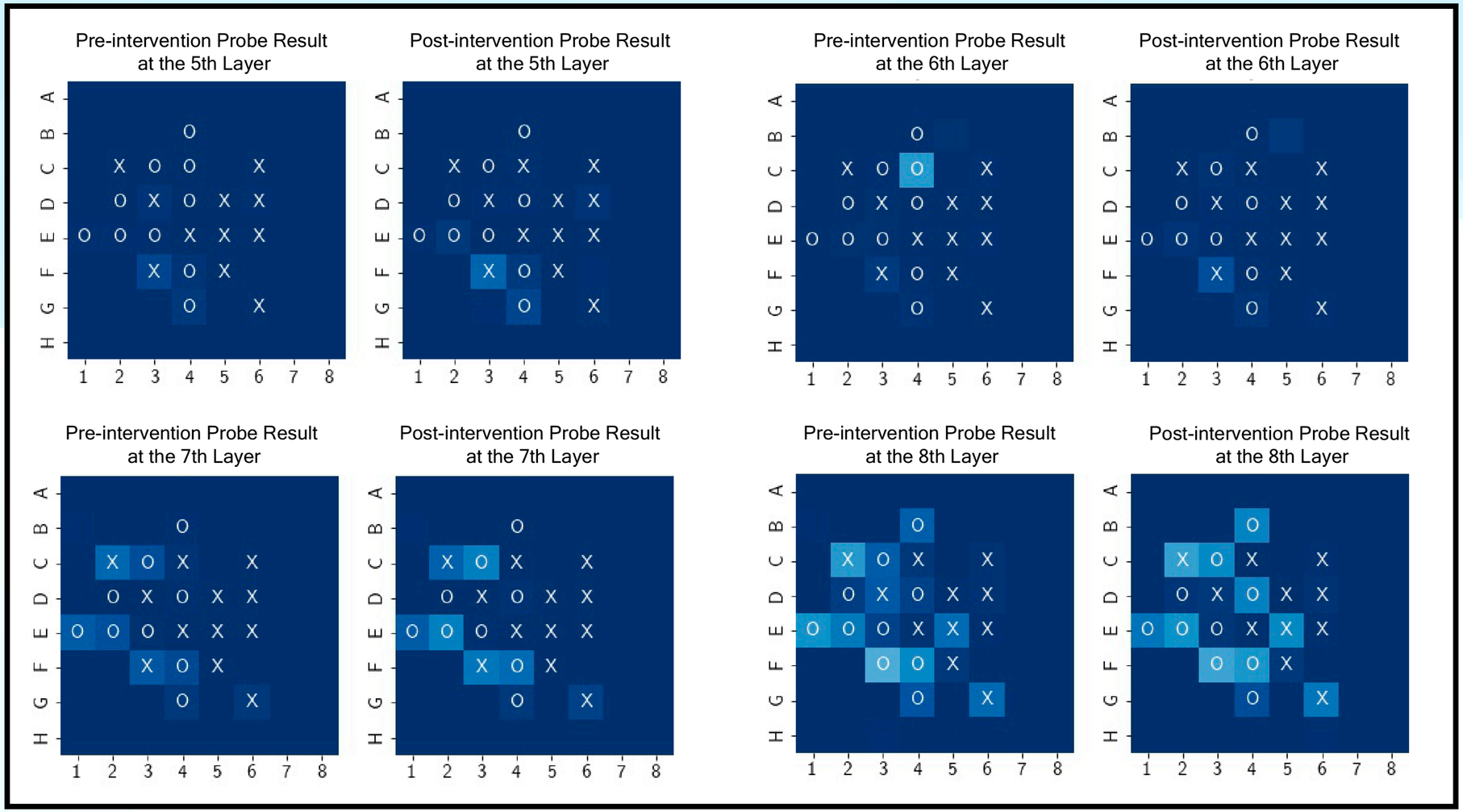
$$x' \leftarrow x - \alpha \frac{\partial \mathscr{L}_{CE}(p_\theta(x), B')}{\partial x}$$



- Don’t update the weights of the probe! Update the internal activations of Othello-GPT!

Othello-GPT

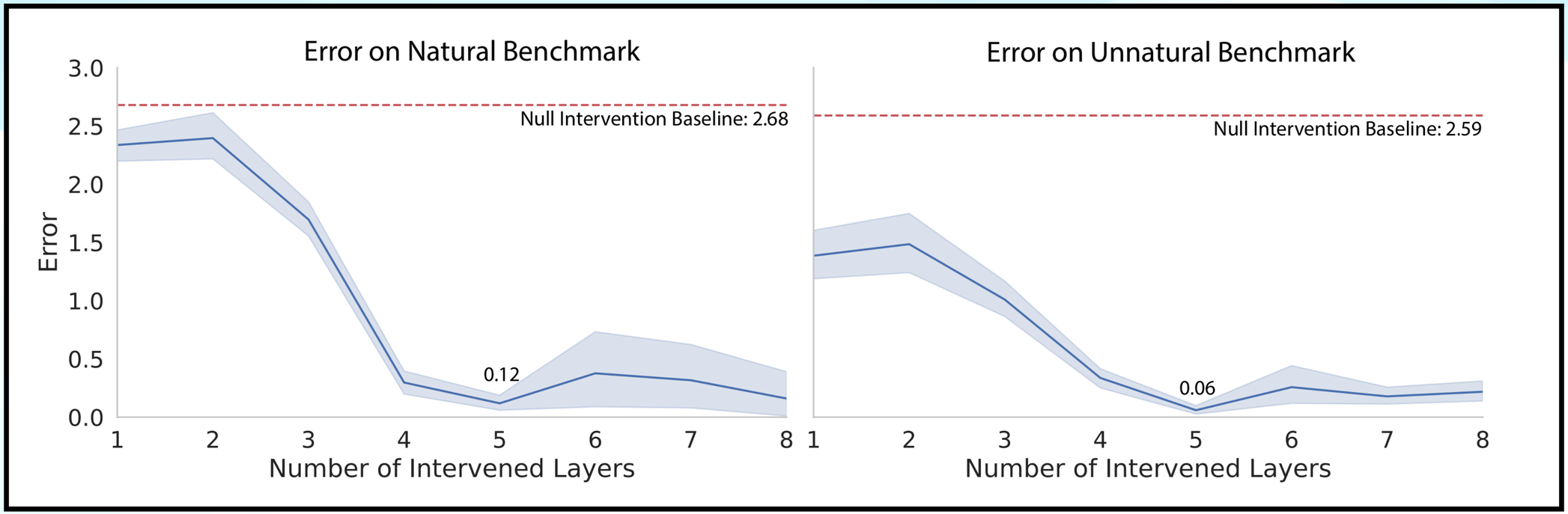
Manipulating Activations



Othello-GPT

Intervention Evaluation

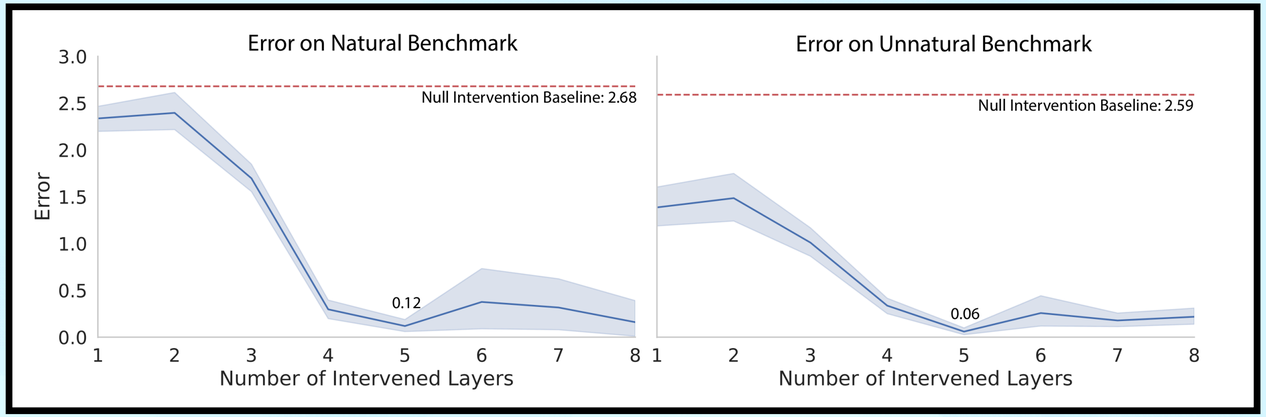
- Middle Layers strike again

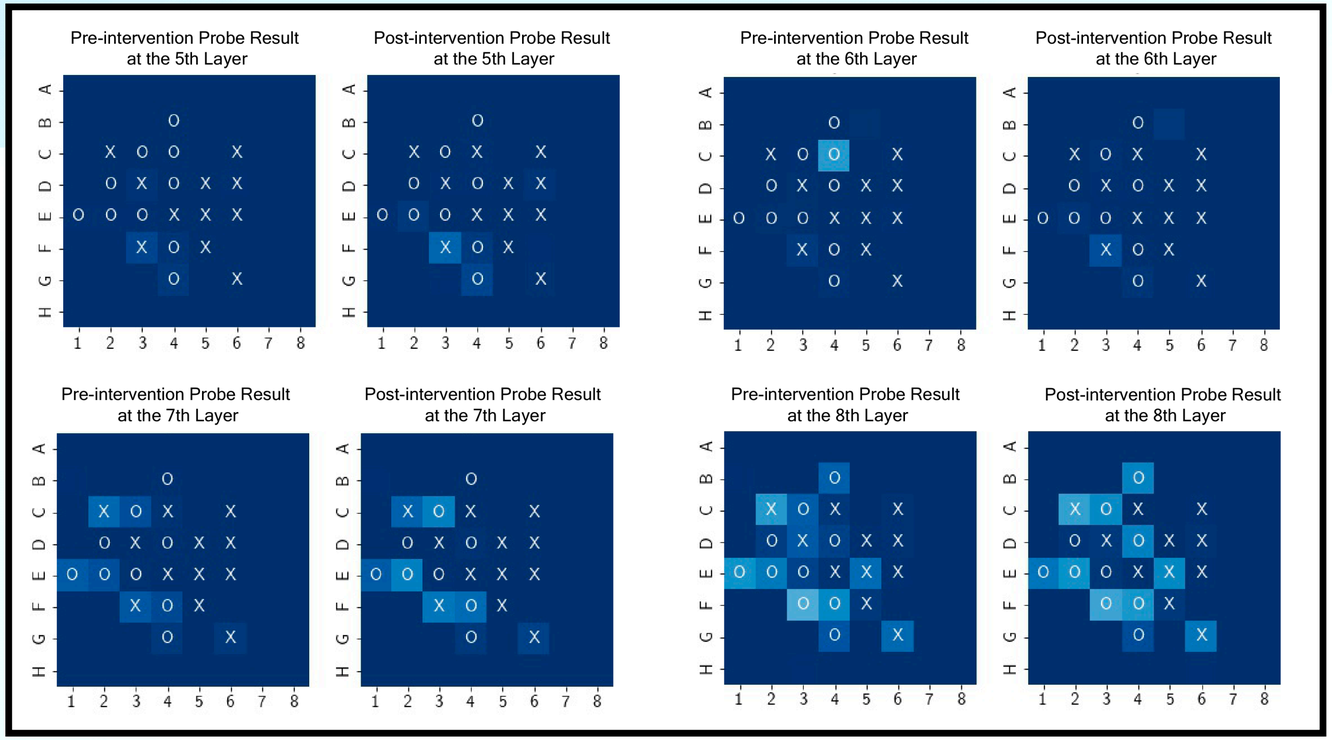


Othello-GPT

Intervention Evaluation

- Middle Layers strike again

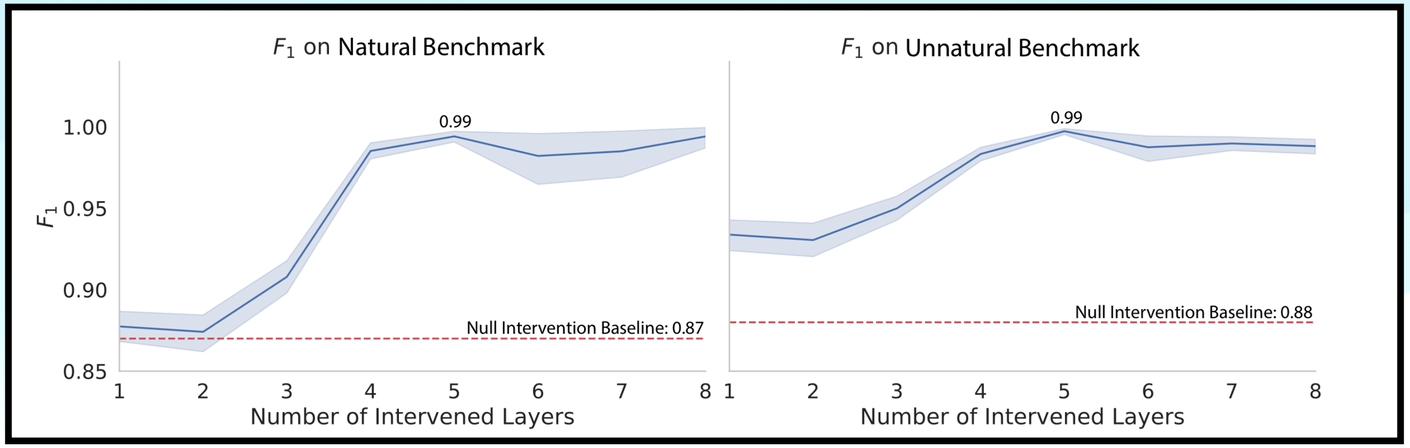


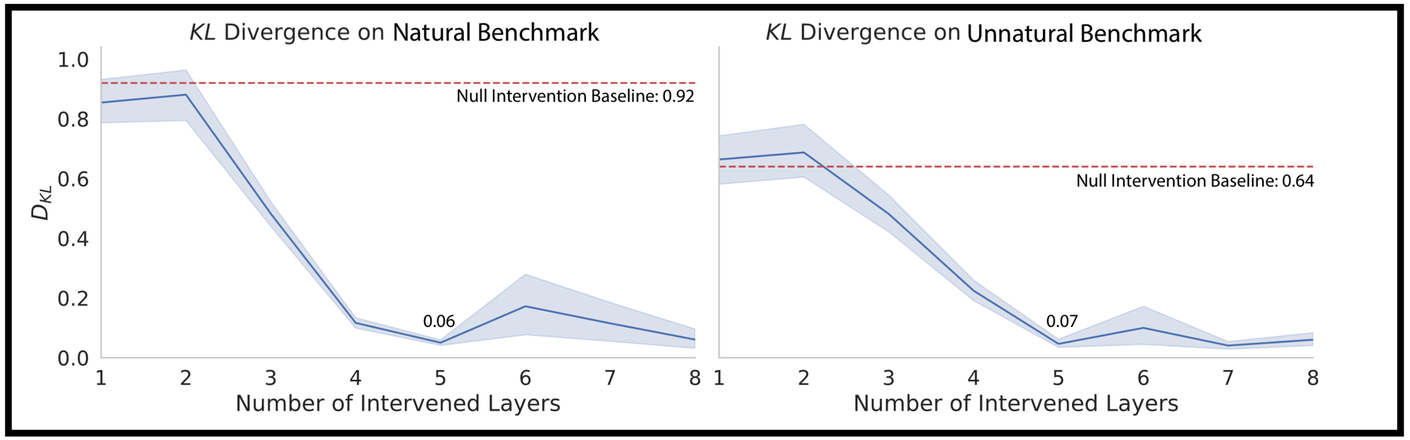


Othello-GPT

Intervention Evaluation

- Middle Layers strike again

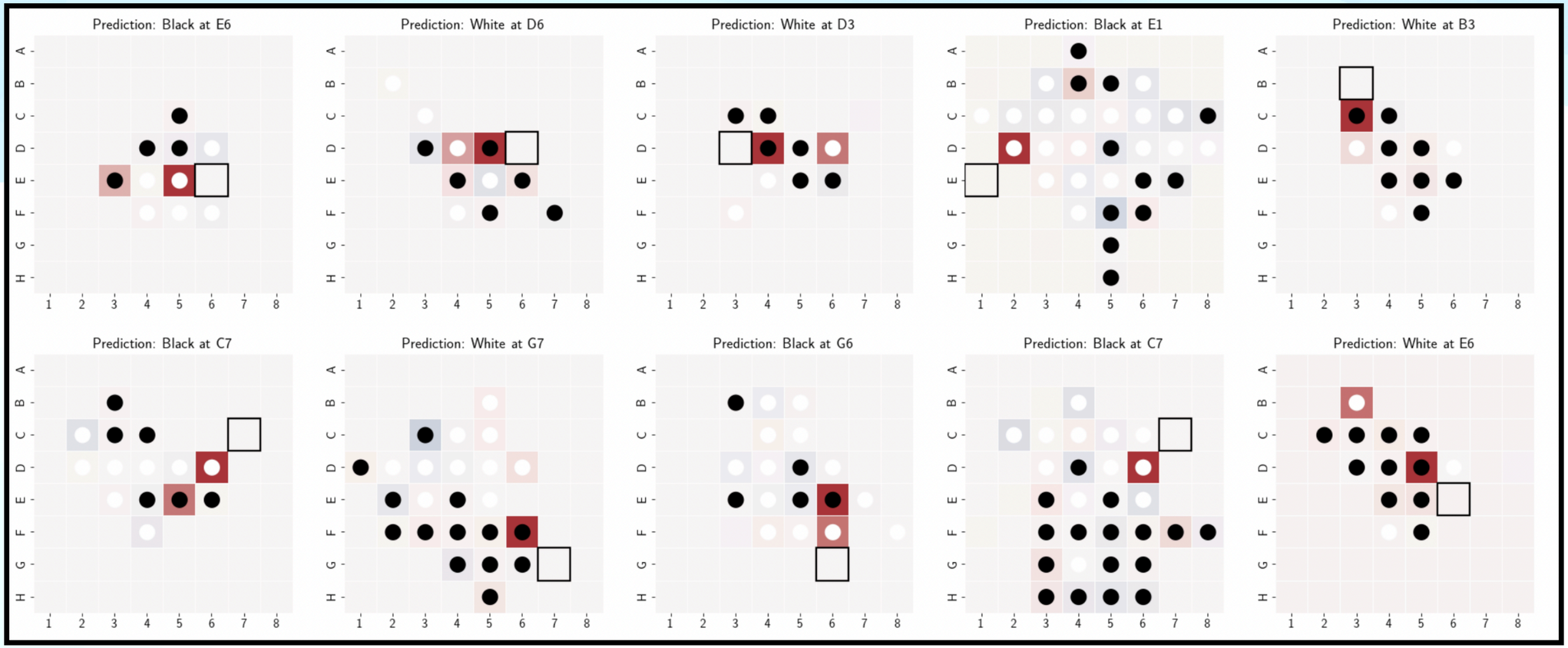




Othello-GPT

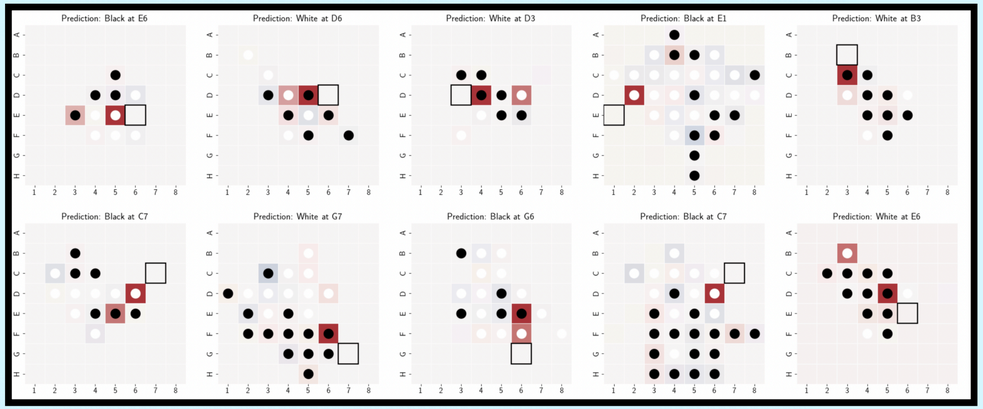
Intervention Evaluation

Latent Saliency Maps (Synthetic):

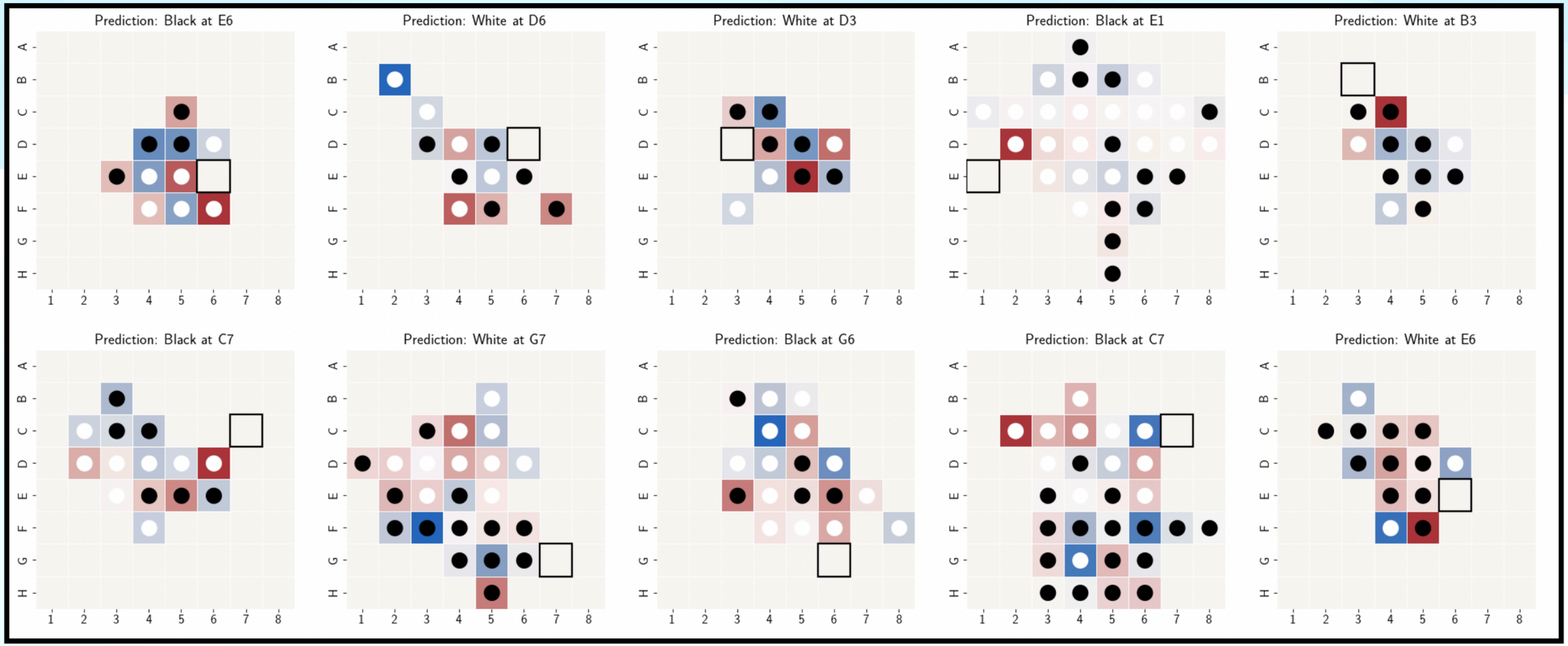


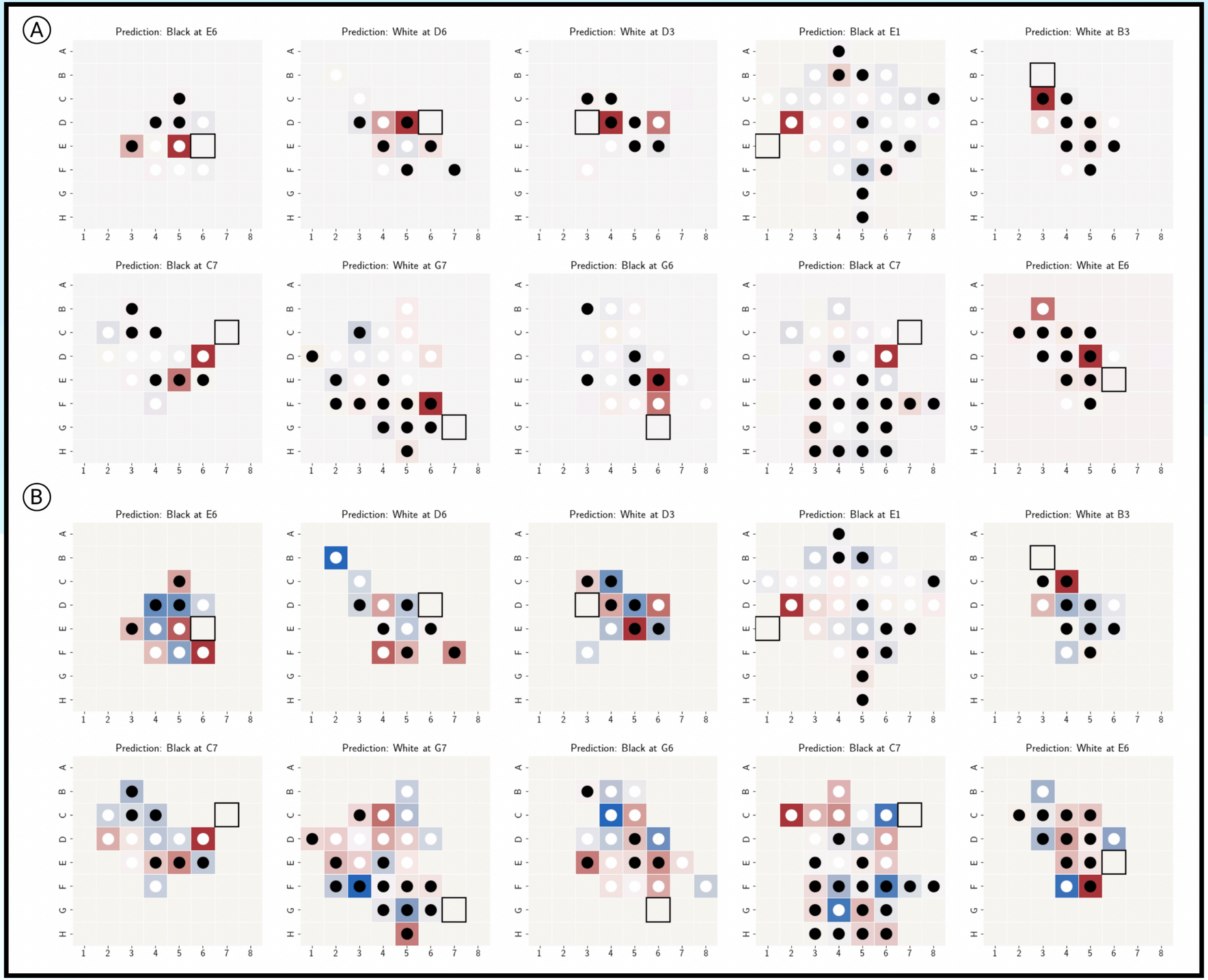
Othello-GPT

Intervention Evaluation



Latent Saliency Maps (Championship):




A
Prediction: Black at E6
Prediction: White at D6
Prediction: White at D3
Prediction: Black at E1
Prediction: White at B3
Prediction: Black at C7
Prediction: White at G7
Prediction: Black at G6
Prediction: Black at C7
Prediction: White at E6
B
Prediction: Black at E6
Prediction: White at D6
Prediction: White at D3
Prediction: Black at E1
Prediction: White at B3
Prediction: Black at C7
Prediction: White at G7
Prediction: Black at G6
Prediction: Black at C7
Prediction: White at E6

Othello-GPT

Representation

There is one flaw …

Othello-GPT

Representation

- The **number** of discs, and their **colours** are inherently linked!
  - Black, White, Black, White, Black, White, …
  - Turn: 1, 2, 3, 4, 5, 6, …
  - Odd, Even, Odd, Even, Odd, Even, …

Othello-GPT

Change of Representation

- Instead of encoding:
  - **{*Black*, *White*, *Empty*},**
- Use instead:
  - {*Me*, *You*, *Empty*} or
  - {*My_Discs*, *Opponents_Discs*, *Empty*} or
  - **{*Mine*, *Yours*, *Empty*}**

Othello-GPT

Linear Representation

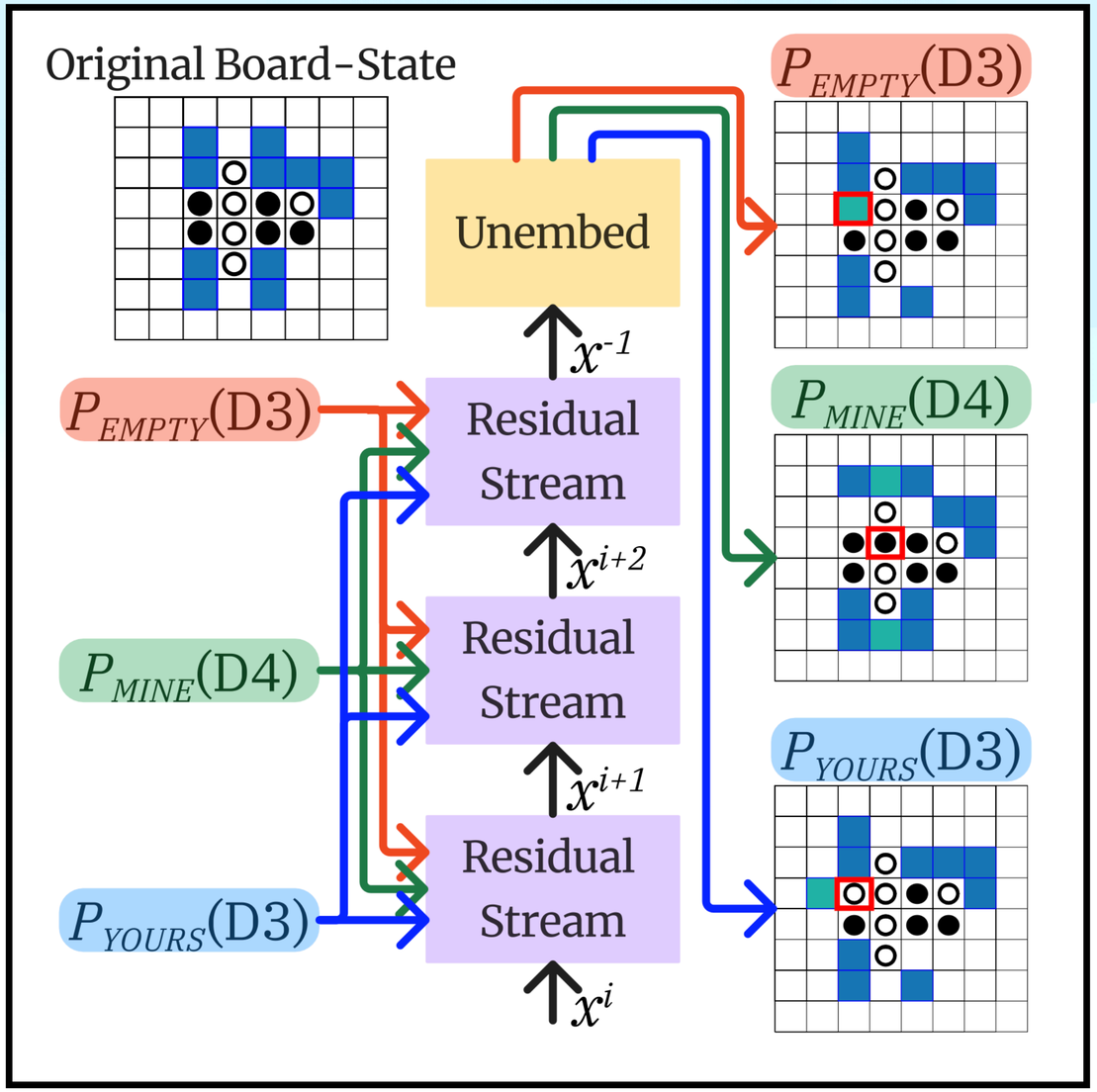
- **{*Mine*, *Yours*, *Empty*}** allows linear encoding!
- Probe ERs:

	x^0	x^1	x^2	x^3	x^4	x^5	x^6	x^7
Randomized	37	35.1	33.9	35.5	34.8	34.7	34.4	34.5
Probabilistic				61.8				
Linear {BLACK, WHITE, EMPTY}	62.2	74.8	74.9	75.0	75.0	74.9	74.8	74.4
Non-Linear {BLACK, WHITE, EMPTY}	63.4	88.6	93.3	96.3	97.5	98.3	98.7	98.3
Linear {MINE, YOURS, EMPTY}	**90.9**	**94.8**	**97.2**	**98.3**	**99**	**99.4**	**99.6**	**99.5**

Othello-GPT

Linear Intervention

- Same Intervention idea as before, but
- also use *Empty* state, and
- **Linearly** intervene!
- Vector addition:
  - $x' \leftarrow x + \alpha \; p_d^{\lambda}(x)$



Othello-GPT

Linear Intervention

- Prediction Error Rates (Top N):

Flipping colours	Avg. # Errors
Null Intervention Baseline	2.723
Non-Linear Intervention	0.12
Linear Probe Addition	**0.10**
Erasing	**Avg. # Errors**
Null Intervention Baseline	2.73
Non-Linear Intervention	0.11
Linear Probe Addition	**0.02**

Othello-GPT

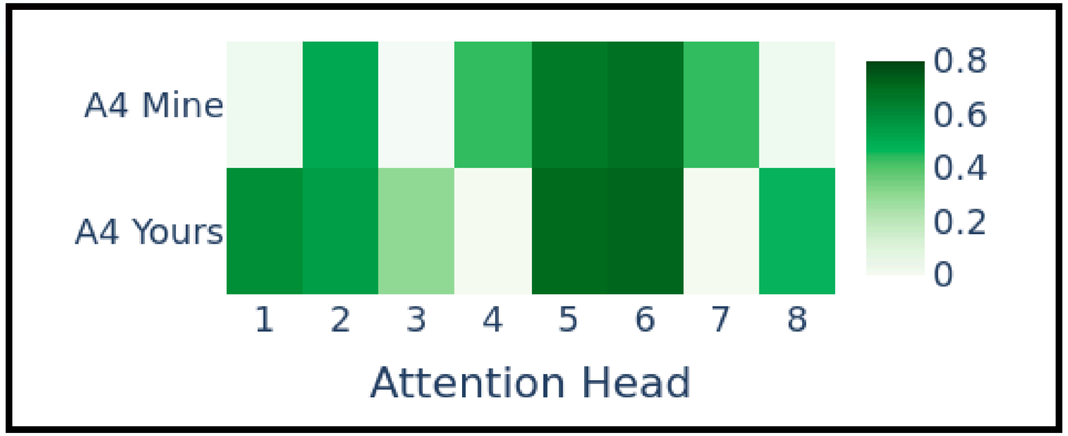
Linear Representation

- Remember, linearity in our models would be awesome ✨
- Unlocks:
  - Modularity (i.e. circuits)
  - Attribution,
  - …

Othello-GPT

Interpreting the model

- The **Empty Circuit**:
  - Circuit: *“[A] sub part of a model that does some understandable computation to produce some interpretable features”* (Nanda, 2022)
  - Othello-GPT learns which tiles are empty in the attention-part of the **first layer**!



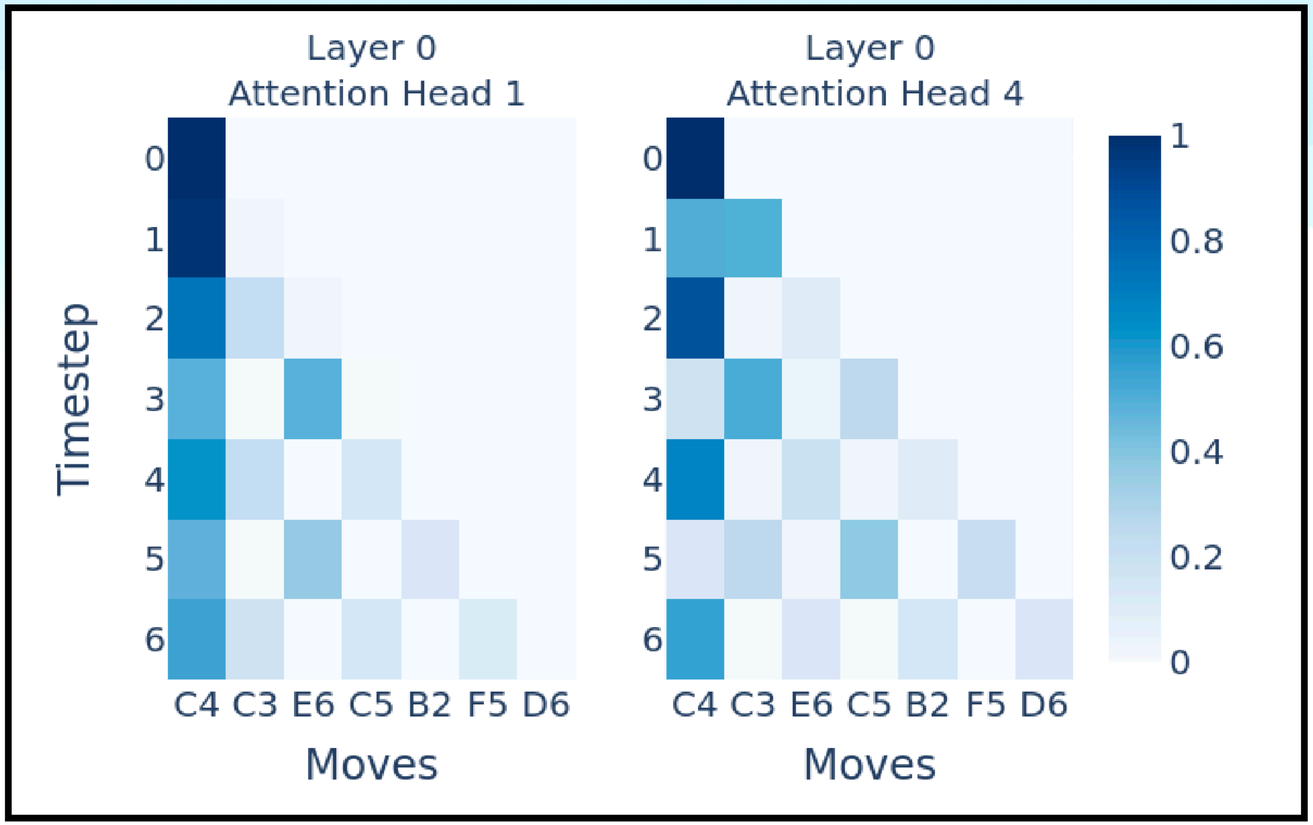
Othello-GPT

Interpreting the model

- *My* moves

  VS

  Your moves:



Othello-GPT

Interpreting the model

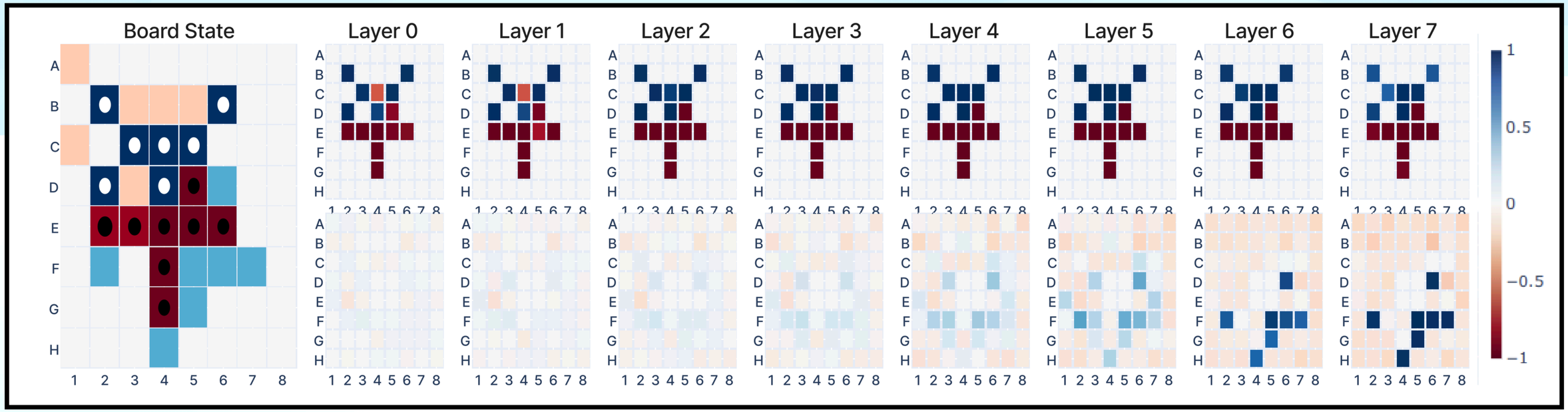
- Which tiles are flipped after a turn?
  - *Flipped* vs *Not-Flipped* Intervention:

	x^0	x^1	x^2	x^3	x^4	x^5	x^6	x^7
Linear {FLIPPED, NOT-FLIPPED}	74.76	85.75	91.62	94.82	96.44	97.13	96.82	96.3

Othello-GPT

Interpreting the model

- Iterative refinements:



Othello-GPT

Interpreting the model

However

...

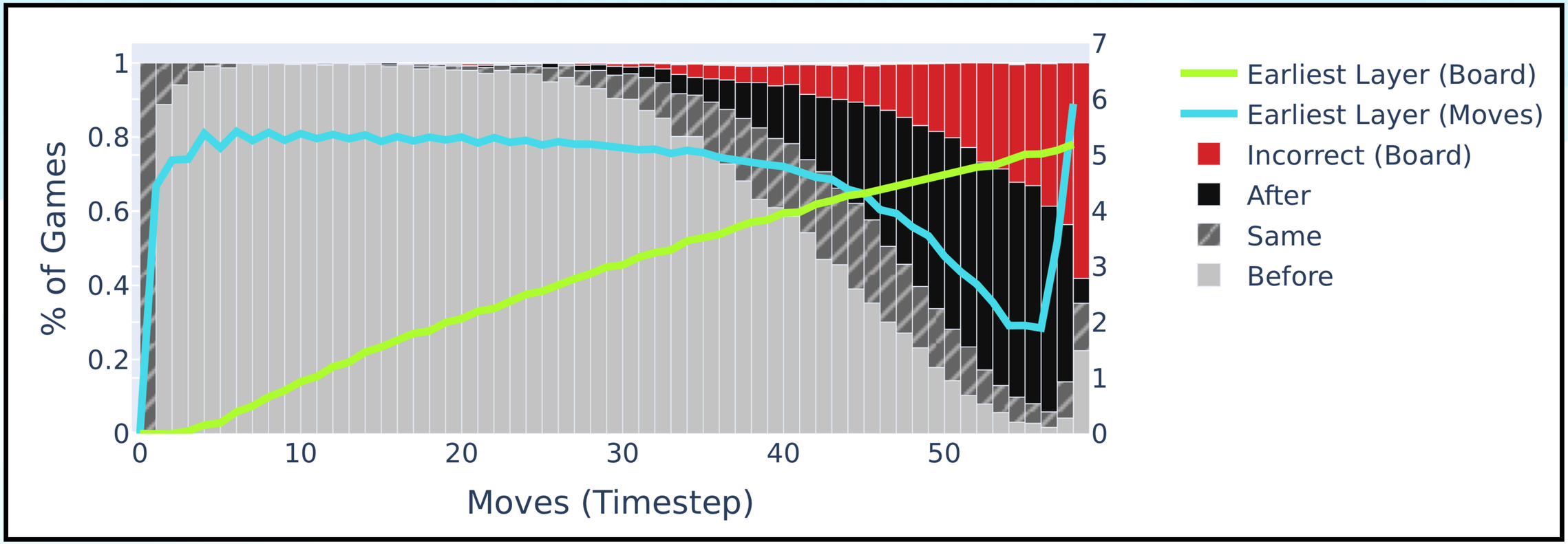
Remember!

Correlation ≠ Causation!

...

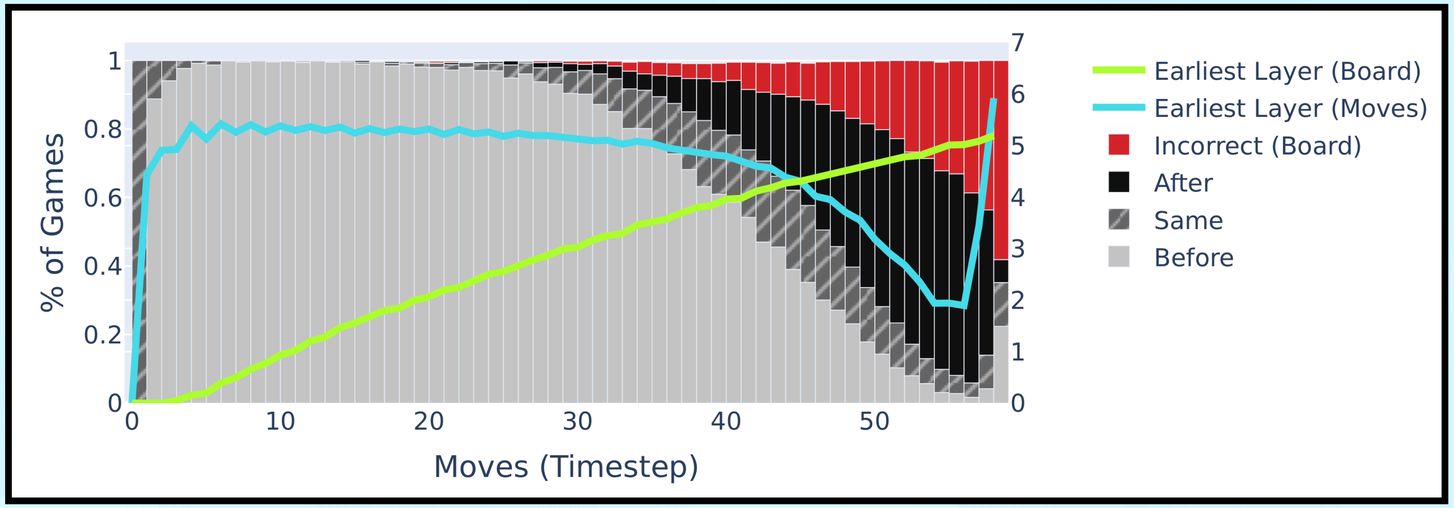
Othello-GPT

Truly causal?



Othello-GPT

Truly causal?



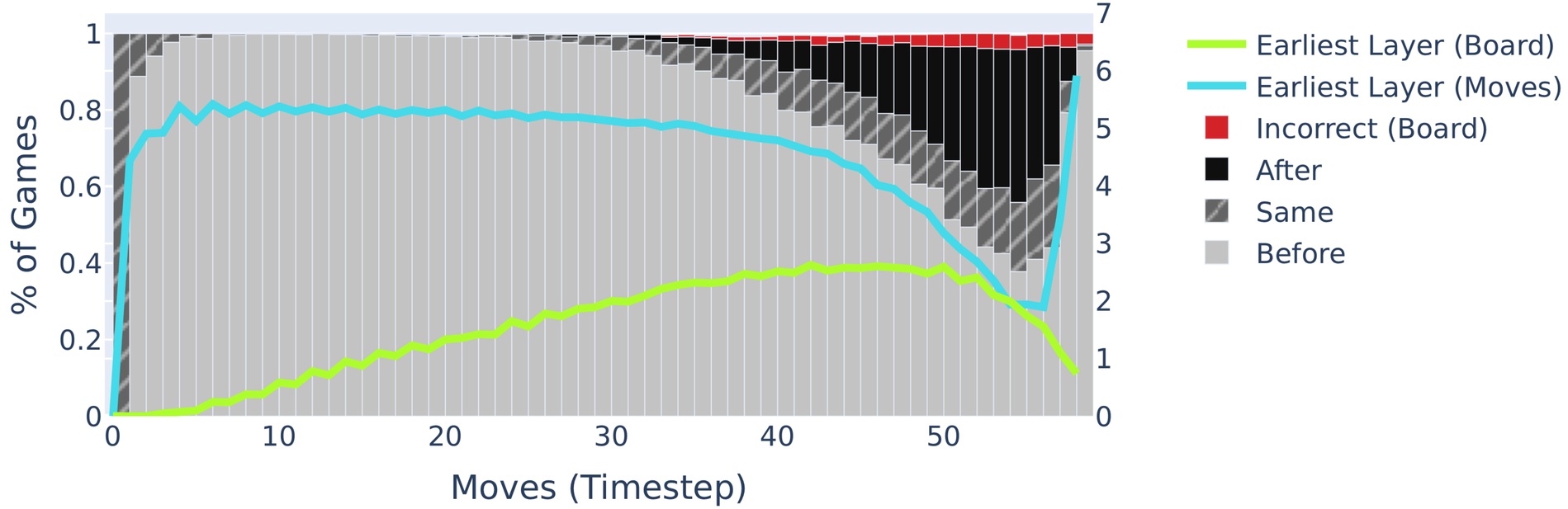


Figure 9: Percentage of times the **"minimum set"** of necessary board state is computed before/after move predictions are made.

Some thoughts

- Impact of model sizes (Othello & Probes)?
- Does the representation of the 4 center tiles differ?
- Only using Error Rates is a bit weak:
  - What about Set Hamming Distance or Jaccard Similarity?
- I’m still a bit sceptic …
- Statistical Tests would be nice
- But very cool results!
- I liked it very much!
- Neel Nanda has interesting opinions … (meant in a mostly positive way)
- Truly open source!
- Othelloscope looks cool

Sources

- Nanda, N.. (2022). A Comprehensive Mechanistic Interpretability Explainer & Glossary.
- Neel Nanda, Andrew Lee, & Martin Wattenberg. (2023). Emergent Linear Representations in World Models of Self-Supervised Sequence Models.
- Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viegas, Hanspeter Pfister, & Martin Wattenberg (2023). Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *The Eleventh International Conference on Learning Representations*.

Discussion & FAQs

Selected questions

1. How do we ensure that the probe dataset accurately reflects the target features and avoids introducing bias? How to deal with category imbalance or feature sparsity when constructing a dataset?

2. When performing ablation experiments, how should you select the neurons to be ablated? Is it based on their activation patterns, strength of association with a particular feature, or other criteria?

3. Might the relatively high saliency for **illegal** tiles of championship Othello point to the model merely mimicking "favorable" edge and corner placement moves from peaks in the multimodal data distribution? Does this not also adhere to spurious correlations learned?

4. How would the representations learned in this synthetic setting translate to real-world tasks?

Discussion & FAQs

Selected questions

5. How might different types of probing methods (such as contrastive or hierarchical probing) provide additional insights into Othello-GPT’s world representation?

6. How does the intervention technique used to alter internal activations provide evidence for the causal role of emergent world representations in Othello-GPT’s predictions?

7. Would you mind giving another example for an interventional experiment, possibly not for a game?