Othello GPT

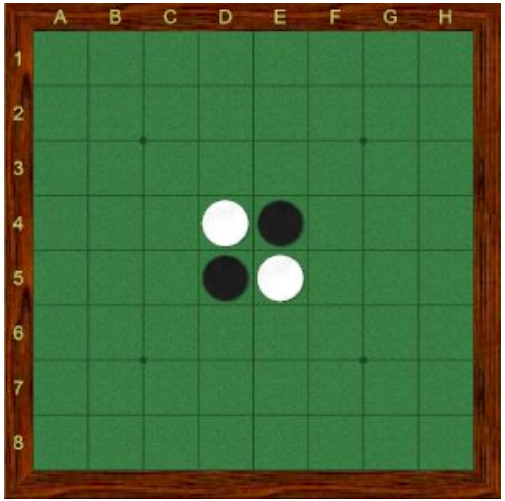
EMERGENT WORLD REPRESENTATIONS: EXPLORING A SEQUENCE MODEL TRAINED ON A SYNTHETIC TASK

Motivation and What's Ahead

- **Language Models** are capable, but **source** of competence unknown!
  - Memorization of Surface Statistics or Reliance on Internal Representations of Generation Process
- Experiment in a **synthetic** setting.
  - Othello GPT: Predicting next legal move in **Othello**
- Evidence of emergent non-linear internal representation of board state.
  - Interventional Experiments
  - Discovery of a possible bijection between **Board State** and **World Model**
  - Bijection allows for interesting representations: **Latent Saliency Maps**

The What and Why of Othello

- Rules:
  - 8x8 board
  - Black begins
  - Initial Board State on the right
  - Every move is ‘outflanking’ and ‘flipping’ opponent disc color
  - Goal: To have more discs of your own color
  - Stop: When no more legal moves left
- Why Othello?
  - Game is very simple: Ideal for synthetic setting
  - Sufficiently large game tree: Avoids memorization

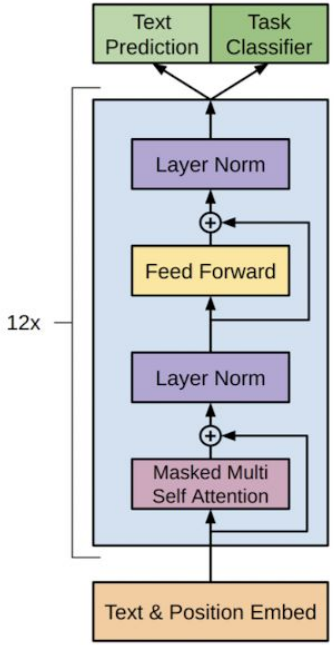


World Models

- A thought experiment
- World Models
  - An understandable model of the process generating the data
  - Example: Chess/Othello Engines

Othello GPT

- Model Architecture:
  - GPT Architecture with 8 Decoder Blocks
  - Each block with 8 Attention Heads
  - Embeddings are 512-D vectors
- Dataset:
  - Championship: Othello championship games from online
    - ~20,000 games
  - Synthetic: Uniform sampling from Othello game tree
    - ~ 23,700,000 games
- Training:
  - Tokens/Vocabulary: 60 words for 60 tiles
  - Pre-Training: Next Move Prediction
- No Memorization:
  - Good performance on skewed dataset



Probes

- Classifiers with constrained capacity:
  - More informative inputs => Better accuracy
  - Activations as input => Predicting properties of sequences as output
    - POS Tag, Parse Tree Depth etc
- 64 multi-class classifiers
  - 1 classifier for every tile
  - 3 possible outputs: White, Black, Empty
- 2 types of Probes:
  - 1-Layer MLP: Linear Probe
  - 2-Layer MLP: Non-Linear Probe

Performance

	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

Table 1: Error rates (%) of linear probes on randomized Othello-GPT and Othello-GPTs trained on different datasets across different layers (x^i represents internal representations after the i-th layer).

	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

Table 2: Error rates (%) of nonlinear probes on randomized Othello-GPT and Othello-GPTs trained on different datasets across different layers. Standard deviations are reported in Appendix H.

Probe Geometry

- Using concept vectors from probes:
  - **Draped Cloth on a Ball** Geometry, resembling Othello Board

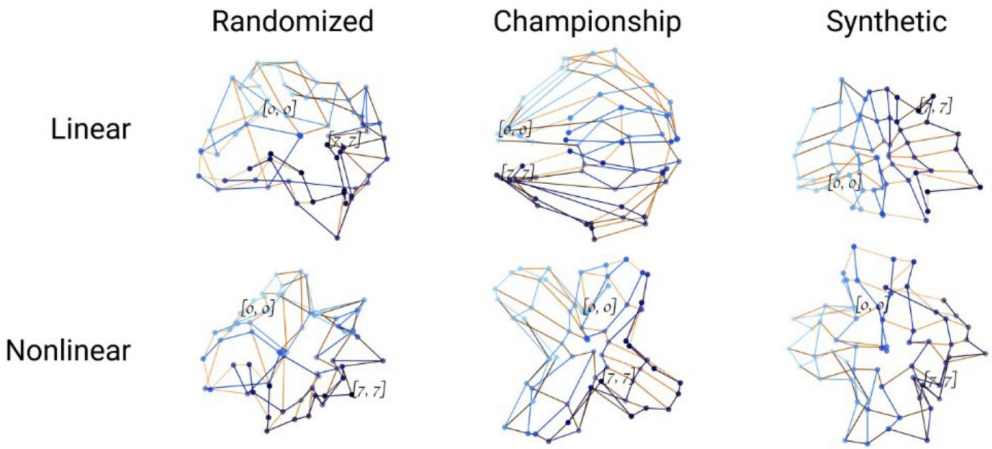
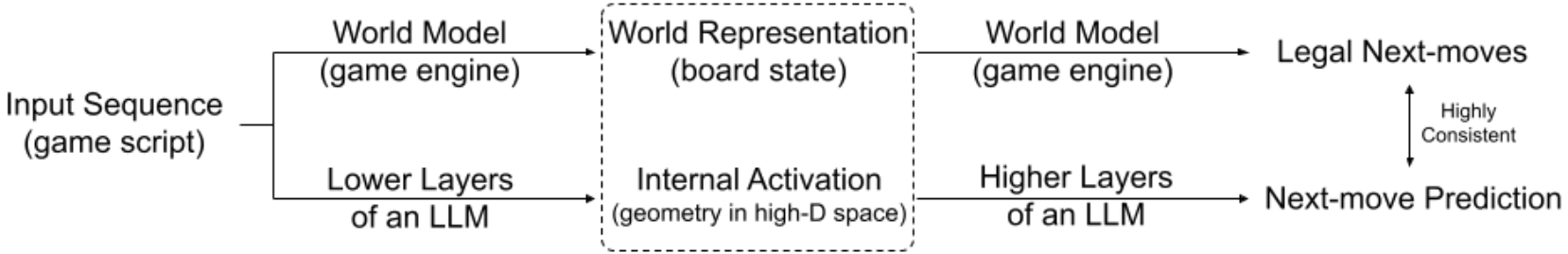


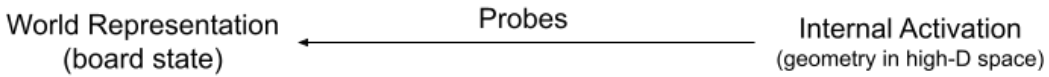
Fig 3: Left: probe geometry of a randomly-initialized Othello-GPT; right: probe geometry of a trained Othello-GPT.

The Big Picture

- 2 unrelated processes:
  - Human Understandable World Model: Othello Engine
  - Black-box Neural Network with Internal Activations: Othello GPT
  - Both almost always agree with each other. Why?

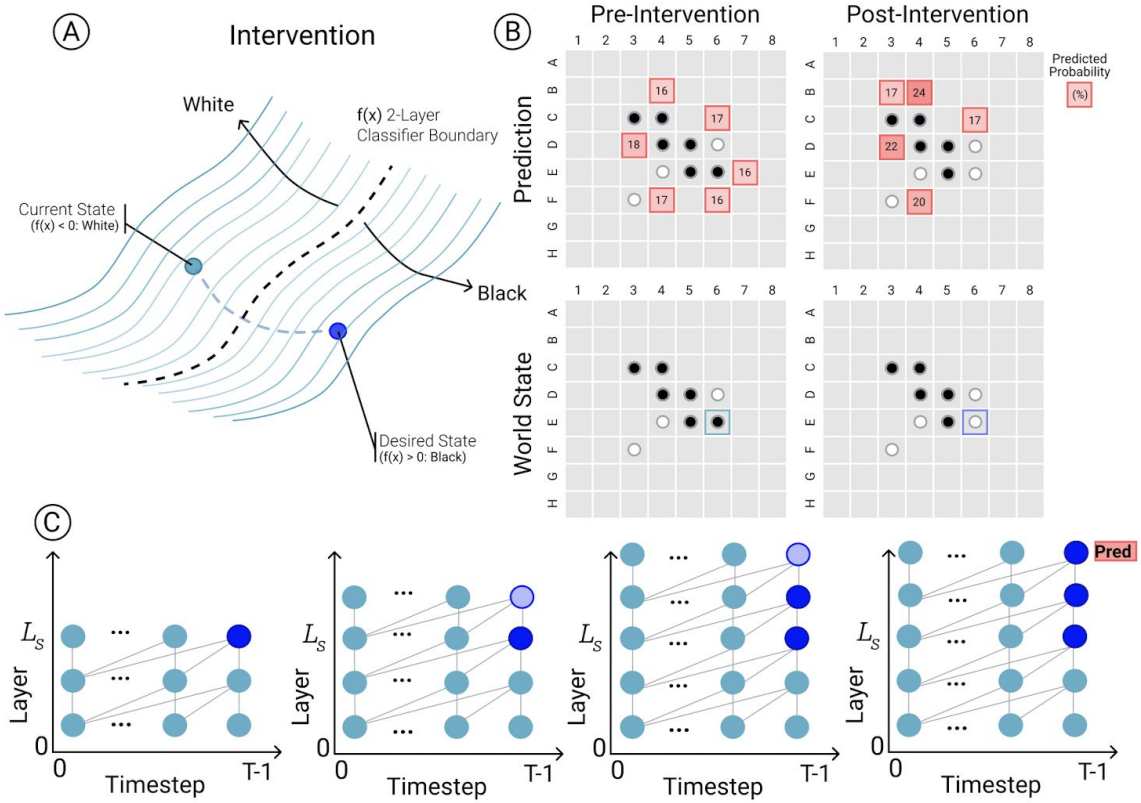


Probe Interventions



- Experiment:
  - Modified Board State => Modified Activation
  - Use Modified Activations to produce next legal move
  - Check if this matches with expectation w.r.t. modified board state
- How to modify the activations?
- Which layers to modify?
- Evidence for Causal Role: Natural vs Unnatural Datasets

Example


A
Intervention
White
f(x) 2-Layer
Classifier Boundary
Current State
(f(x) < 0: White)
Black
Desired State
(f(x) > 0: Black)
B
Pre-Intervention
Post-Intervention
Predicted
Probability
(%)
Prediction
World State
C
Layer
Timestep
0
T-1
Pred

Intervention Results

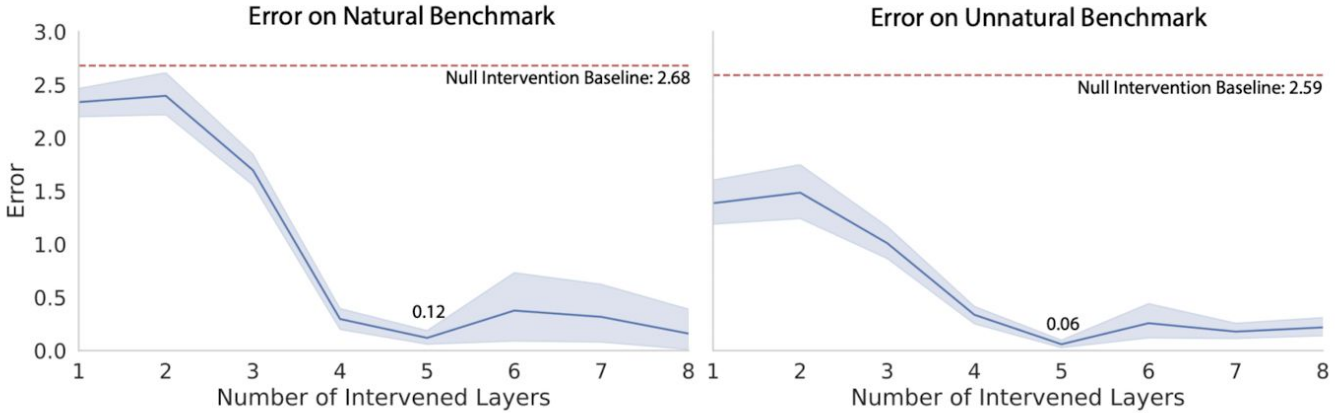
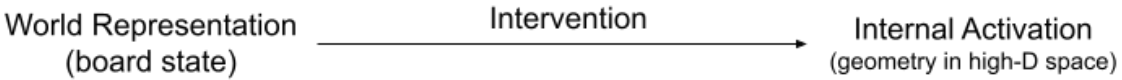


Figure 3: Intervention experiment results. Red dashed lines represents average number of errors by testing pre-intervention predictions on post-intervention ground-truths, representing a null intervention method for contrast. The shaded area represents the 95% confidence interval.

Attribution via Interventions



- Experiment:
  - Attribution Square: Next Move tile
  - Attribution Scores
    - Scores associated with all the remaining occupied tiles
    - Measure of how current state of tile is affecting the next move
  - Latent Saliency Maps: Heatmap of Attribution Scores
  - Importance given to 2 types of tiles:
    - Tiles of same color making the sandwich
    - Tiles of opponent occupied between the sandwich

Latent Saliency Maps

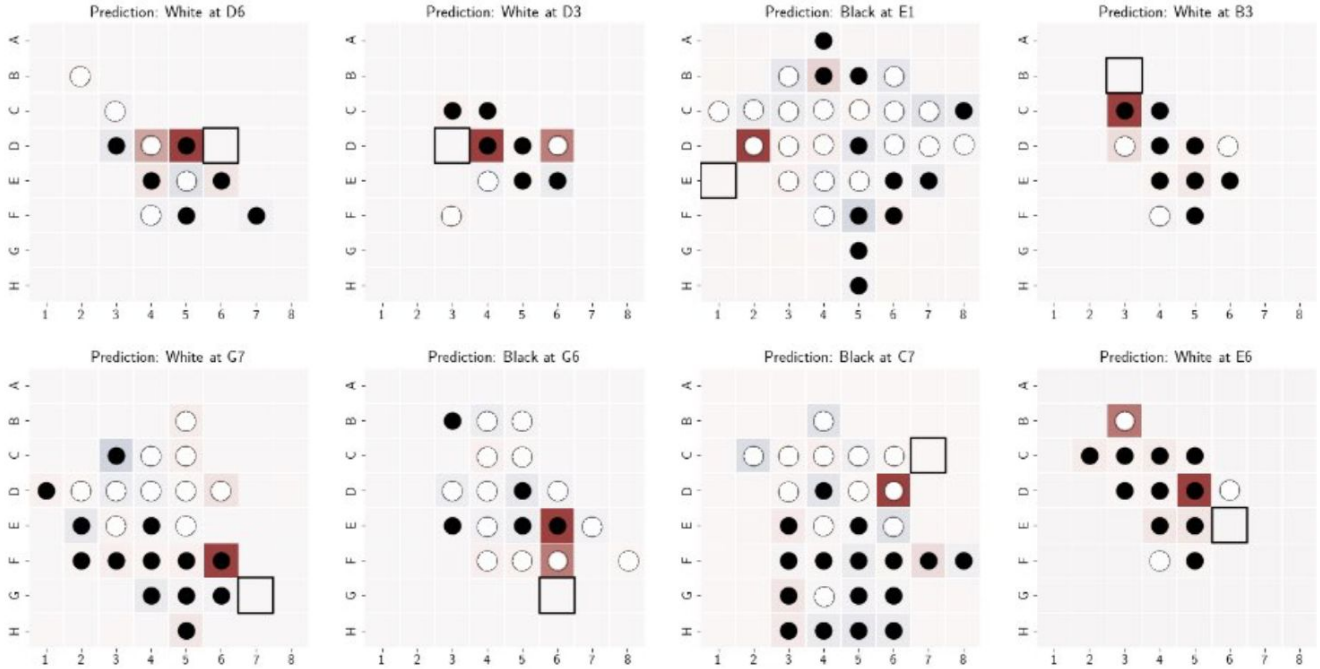
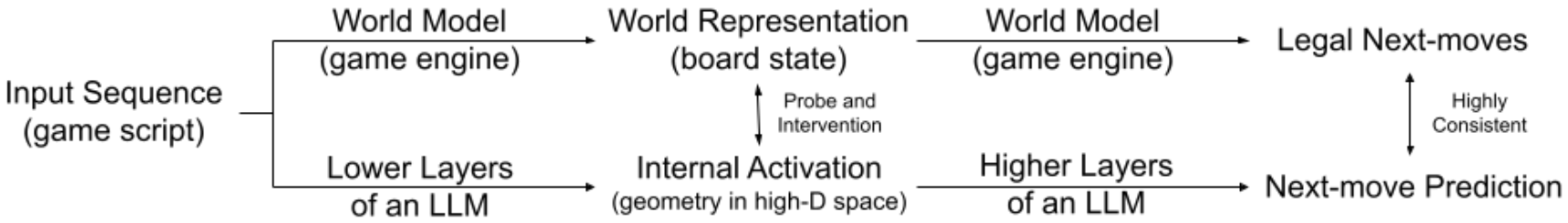


Fig 5: for each of the 8 plots, the text above is the next-move it is attributing (also enclosed). For other tiles on the board, the darker red, the more important it is for the attributed move. For example, in the upper left plot, D5 contributes the most to the prediction of D6.

Result


Input Sequence
(game script)
World Model
(game engine)
Lower Layers
of an LLM
World Representation
(board state)
Probe and
Intervention
Internal Activation
(geometry in high-D space)
World Model
(game engine)
Higher Layers
of an LLM
Legal Next-moves
Highly
Consistent
Next-move Prediction

Further Exploration

- Existence of Linear Probes
- Embedding Space
  - Smaller embedding spaces
  - Simpler network
  - More ablation experiments
- Generalization to Natural Languages:
  - Discovery of further structures from activations
- Further Interventional Experiments:
  - Exploring activations further

Thank You