Linear Latent World Models in Simple Transformers: A Case Study on Othello-GPT

Dean S. Hazineh* Zechen Zhang* Jeffrey Chiu

Harvard University

Introduction

Do LLMs have world model representations?

We trained a series of transformer models of a variety of complexity to play Othello (Othello-GPT) and probed the neural activations in the residual stream for an interpretable representation. We found that the board states are *linearly* encoded and that linear information is encoded and are used by the Othello-GPT causally in certain layers at certain game-length.

Contributions Overview

- Even small models down to 1L1H can play Othello relatively well, and they possess linearly encoded information about the board state.
- We designed a simple causal intervention technique that directly intervenes at each layer proved its causality.
- Semantic understanding seems to be developed and utilized by the model in *middle* layers.

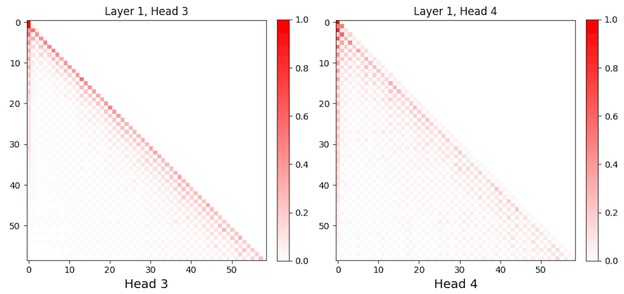
Linear Representation of the Board States

Linear Probe Accuracy

Table 1: Classification Accuracy for Linear Probes Mapping $z \to s$

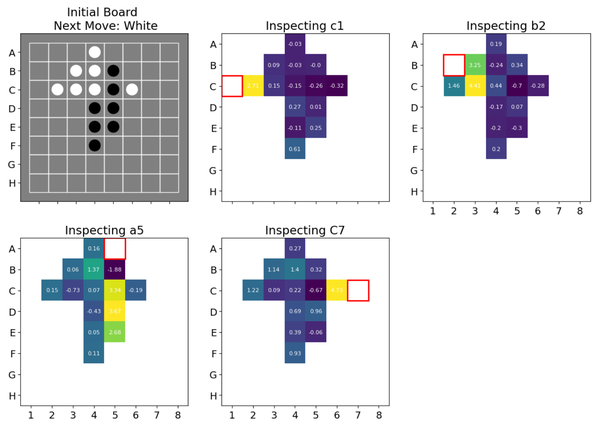
Layer:	1	2	3	4	5	6	7	8
Old (Black,White)	75.7%	75.8%	75.7%	75.7%	75.6%	75.4%	74.9%	74.9%
New (Mine, Yours)	**90.8%**	**94.8%**	**97.1%**	**98.3%**	**99.1%**	**99.5%**	**99.5%**	**99.5%**

"Yours" and "Mine" Attention Heads

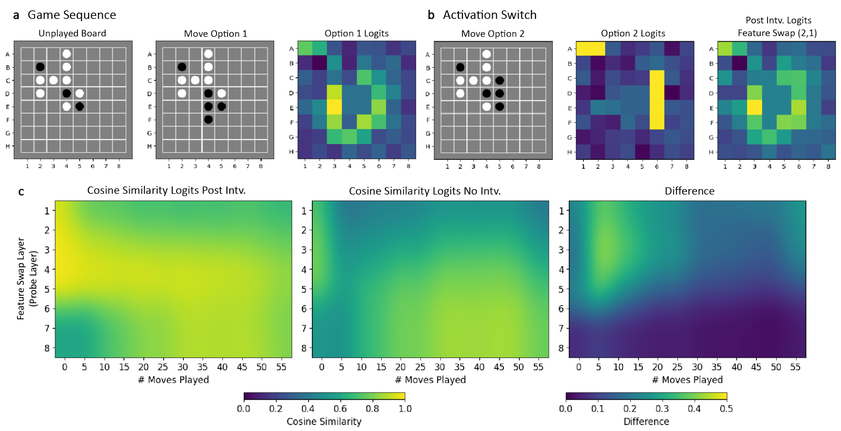


Causal Interventions

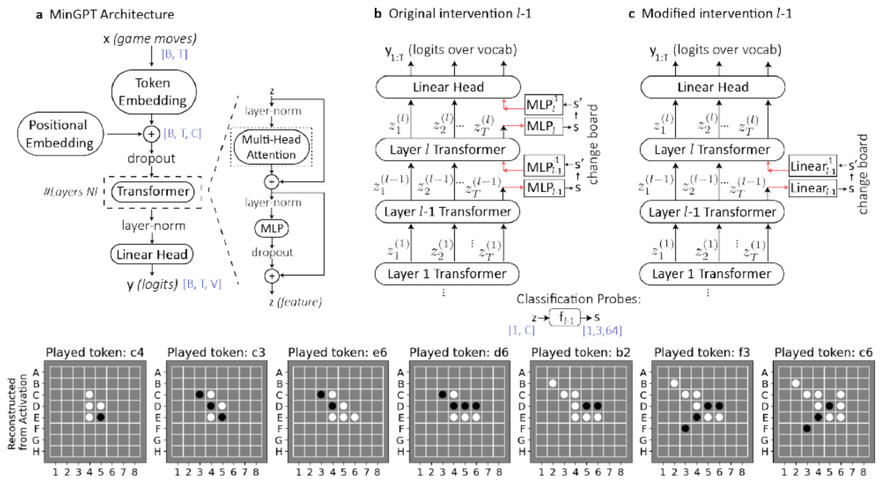
Latent Saliency Map



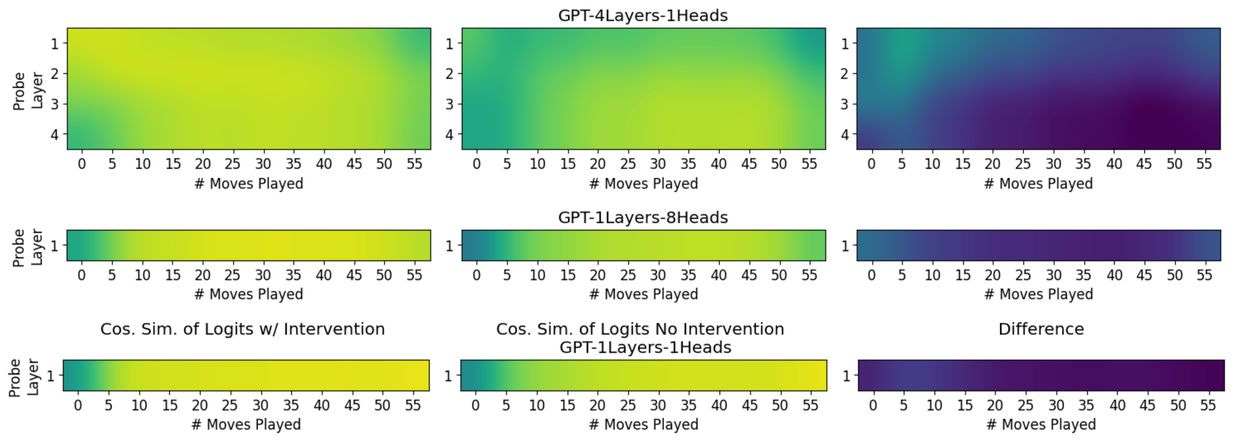
When and Where Casual Interventions are Successful?



Othello-GPT Set-up



Causal Interventions for Shallower Models



Conclusions

- Linear Representation of Board States Encoded in Shallow Transformers
- Deeper Networks Better at Casual Usage of Linearly Encoded Board States
- Board State Information Finalized in Middle Layers

NEURAL INFORMATION PROCESSING SYSTEMS