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

Dean S. Hazineh*
Department of Applied Physics
Harvard University
dhazineh@g.harvard.edu

Zechen Zhang*
Department of Physics
Harvard University
zechen_zhang@g.harvard.edu

Jeffrey Chiu
Department of Computer Science
Harvard University
jeffrey_chiu@college.harvard.edu

Abstract

Foundation models exhibit significant capabilities in decision-making and logical deductions. Nonetheless, a continuing discourse persists regarding their genuine understanding of the world as opposed to mere stochastic mimicry. This paper meticulously examines a simple transformer trained for Othello, extending prior research to enhance comprehension of the emergent world model of Othello-GPT. The investigation reveals that Othello-GPT encapsulates a linear representation of opposing pieces, a factor that causally steers its decision-making process. This paper further elucidates the interplay between the linear world representation and causal decision-making, and their dependence on layer depth and model complexity. We have made the code public.

1 Introduction

Auto-regressive language models, esteemed for their predictive capabilities, have showcased unprecedented potential in logical reasoning and complex task resolution (Srivastava et al. (2022) Creswell et al. (2022)). This leads to a discourse (Bender et al. (2021)) regarding whether their performance is merely a reflection of memorization of "surface statistics," namely the correlations between tokens unrelated to the original sequence's causal process. Contrarily, recent research endeavors examine the inherent development of interpretable world models within such language models. World models, as referred throughout, signify the encoding of semantic information about the true causal process within the network's activations. Instances of world representations include the location and type of game pieces for a language model trained in chess (Toshniwal et al. (2021)) and object part classification in a visual transformer (Amir et al. (2021)). Exploring the unsupervised development of internal world representations in these models is paramount for comprehending the emergence of intelligence in foundation models.

In this research, the emergent world principle is scrutinized in a series of simple transformer models, trained to play Othello and termed as Othello-GPT (Appendix A). These trained models (Figure 1a) exhibit proficiency in legal move execution. Utilizing linear probes to decode neuron activations across transformer layers, coupled with causal interventions, this paper underscores the enhanced world model of Othello-GPT and its causal utility for decision-making.

* Equal contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Contributions: In this ongoing work,

1. We show that even very small models down to 1 layer and 1 attention head can play Othello, and they possess **linearly** encoded information about the board state. We find that world representations may be linearly encoded in the model's activations to an extent that generally increases with layer depth for deeper models.
2. We design a simple novel causal intervention technique that directly intervenes at each layer with the linear representation and demonstrate that at certain layers the linear world model is **causal** in influencing the model's next move prediction. In particular, it appears that semantic understanding in the model is developed and utilized about halfway through the model.

2 Related Work

Othello-GPT Li et al. (Li et al. (2023)) inaugurated the research on an 8-layer transformer model for Othello, demonstrating the non-linear encoding of board state information within each layer's activations and introducing a sophisticated causal intervention scheme. A subsequent blog post (Nanda (2023)) highlighted the potential for linear encoding of the board state, albeit without an interpretable representation.

Probing Classifiers The employment of classifiers for probing neural networks has gained prominence as an instrumental tool for understanding internal representations within neural networks. This methodology has facilitated the understanding of sentence embeddings in transformers (Clark et al. (2019)), hierarchical structures in recurrent networks (Hupkes et al. (2018)), and visual concepts within vision models (Alain and Bengio (2018)). Despite these advancements, debates continue (Belinkov (2021)) regarding the actual causal utility of the encoded information by the network. This research contributes a meticulously designed causal intervention study to ascertain the network's utilization of the encoded representation.

Mechanistic Interpretability Complementary research (Olah et al. (2017) Olah et al. (2020) Olsson et al. (2022) Nanda (2023)) endeavors to decipher the internal mechanism of networks by identifying "circuits", network segments responsible for computing certain interpretable features. This paper lays a robust foundation for further mechanistic interpretability work on Othello-GPT.

3 Linear Representation of the Board State

3.1 Linear Probe Uncovers a Natural Board State Representation

In this section, we first report our findings that there is an intuitive world representation that is linearly encoded within the activation vectors of the trained sequential model.

The original work of Li et al. (2023) postulated a world representation corresponding to the placement of white or black discs on a game board. While this is how most humans would visualize the board when playing the game of Othello, it is important to note that the sequential model differs slightly from human play in that the model alternates between acting as the white vs. black player depending on the length of the input sequence. As the rules dictating a legal move require only an understanding of which piece is yours and the opponents', a more natural choice for a world representation in this case is instead the classification of a tile as containing "my piece" vs. "your piece" (or empty). We term the "Black, White, Empty" representation as the old representation, and "Yours, Mine, Empty" as the new representation.

To quantify this, we first train a multi-headed attention model (diagrammed in Figure 1a) with 8 transformer blocks, referred to as layers, and 8 attention heads on the synthetic Othello dataset that consists of random legal moves. Once trained to predict legal next moves, the model's weights are frozen. We then train a separate linear transformation L (referred to throughout as a linear probe) to map the activation vector of a particular layer to the world representation, i.e., the mapping $f : \mathbb{R}^{C=512} \rightarrow \mathbb{R}^{64x3}$ where C is the token embedding dimension and the output corresponds to the three-state classification logits for each of the 64 tiles. The accuracy for the linear probe at each layer for the old vs. new world representations is shown in Table. 1. We observe that the new

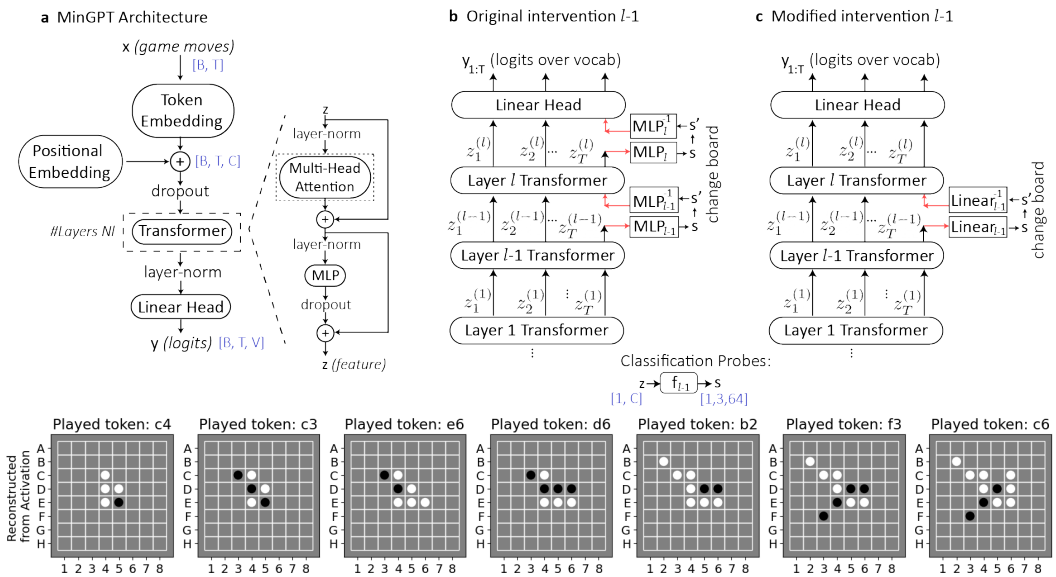

Figure 1: Overview of the principles in section 3. (a) The neural architecture utilized in this paper, where the number of layers refers to the number of transformer blocks. B=Batch Size, T=Length of Tokens, C=Hidden Layer Dimension, V= Vocabulary Size. (b) The original intervention scheme of Li et al. (2023) is replaced by an alternate version shown in (c), whereby intervention is applied to a single layer–see text for details. (Bottom Panel) Example of an extracted world representation where the game board state is obtained from the activation vectors.

representation is linearly encoded in the activation vectors, particularly within the deeper layers, and can be recovered from the model's activations by a simple linear transformation with near unity accuracy on a withheld set. Surprisingly, the linear probe is successful even for a 1-layer Othello-GPT, as will be discussed in Appendix B.

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**%

3.2 "Yours" and "Mine" Attention Heads

In addition to the linearly encoded board states that represent "yours" and "mine" pieces, we also find regularities in the attention heads that support such representations. For example, Figure 2 shows that there exist attention heads that only keep track of "your" first-placed pieces and "my" first-placed pieces, resulting in alternating patterns. In fact, the "yours" and "mine" attention heads occupy a large percentage of the attention heads in the 8L8H Othello-GPT, and these attention heads emerge even in a simple 1L transformer (Appendix C).

4 Causal Interventions

To see if the linear world model is used by the Othello-GPT in making next-move predictions, we intervene at each layer of the model at certain game length, to trick the model into believing that it has another board state, as shown in Figure 1c. In general, we take the activation z that corresponds to the board state s per the linear probe, then change it to some other board state s' and map it back to an activation z' that is associated with it, ie.

$$L(z') = s',$$

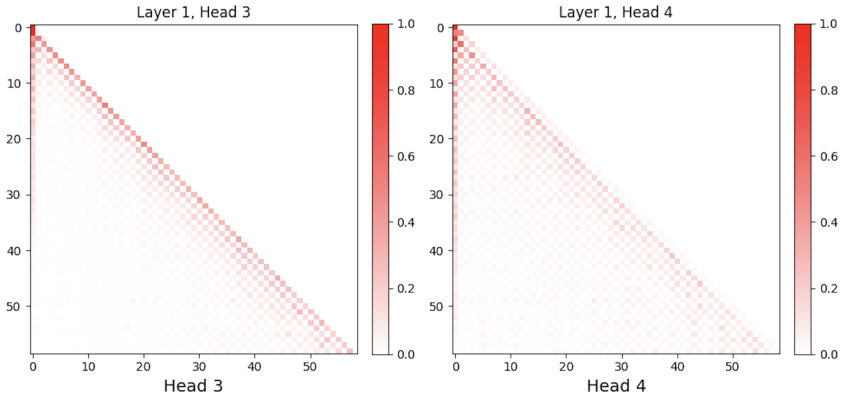

Figure 2: Example attention heads of the 8L8H model. Note that both heads show alternating patterns, which we interpret as processing information from pieces first placed by the same player. For example, Head 3 is keeping track of all my historical moves and Head 4 is keeping track of all the opponent's historical moves.

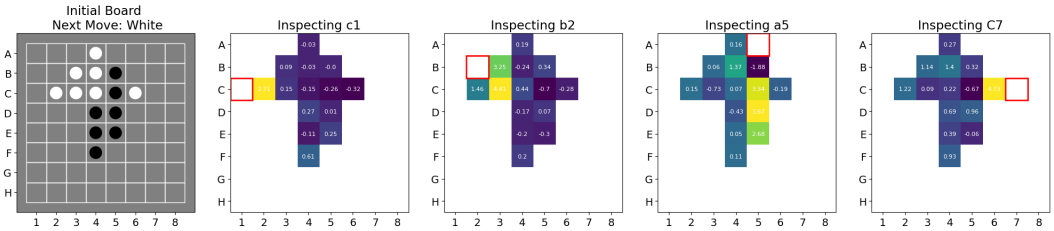

Figure 3: Latent saliency map for a particular game, interventions at layer 6. m^* is highlighted with a red box which indicates the move that we calculate logits for and the numbers indicate the logit changes given interventions at that tile. If a possible next move is currently illegal, we observe that interventions that would make the move legal produce a positive change (bright yellow); alternatively, interventions that make a currently legal move now illegal have a negative change (dark blue).

where L is the linear transformation induced by the linear probe. In the interventions, we train another linear map with L^{-1} cross-entropy loss on the same board state – activation pairs as the pseudoinverse of L and find z' by $z' = L^{-1}(s')$.

4.1 Latent Saliency Map

First, we demonstrate that the interventions from changing the activation z to z' is indeed causal in certain layers by plotting the latent saliency map of interventions. The map is obtained by flipping each piece at a certain game length and recording the change of logits of a particular move m^*. If the intervention is indeed causal, we would expect that some particular interventions would encourage or discourage the move m^*, and this is indeed what we observe in Figure 3. It is worth reporting that the interventions are successful only at certain layers, which we will discuss more in the next section.

4.2 When and Where Causal Intervention is Successful

The latent saliency map is limited as it is only a qualitative demonstration and is also restricted to single-flipping-piece interventions. To see when (game length) and where (layer) the linear board state is used by the model causally, we look at the logit distribution of a model tricked into believing it has another board state. Specifically, at game length t (termed unplayed board), we introduce two move options s_1, s_2 and intervene with the second move option at each layer to change activations z_2 to z_1 that corresponds to the board state s_1. If the intervention is indeed causal, then we would see that the post-intervention logit distribution of move option 2 is very similar to that of move option 1, as shown in Figure 4a,b.

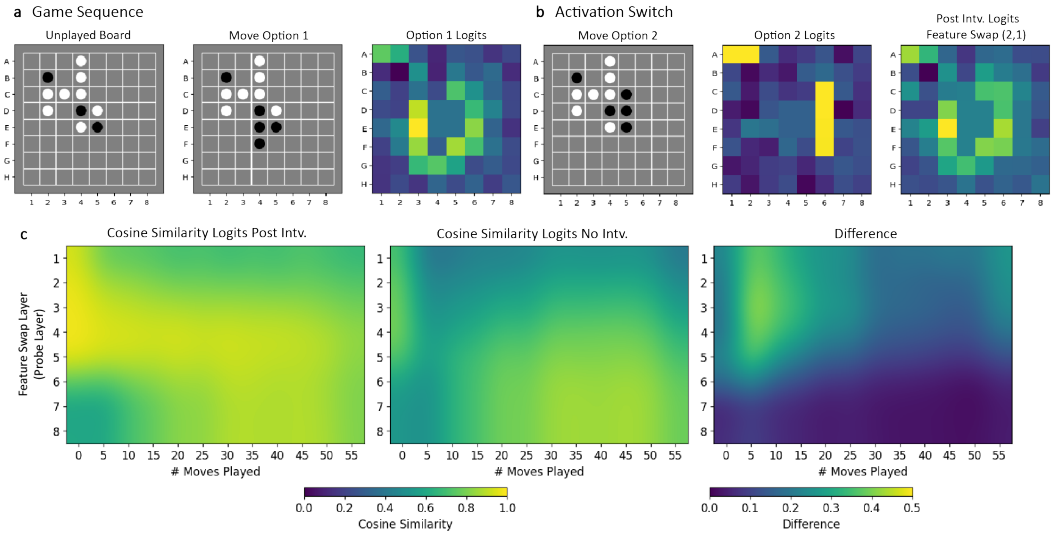

Figure 4: Comparison of logit distributions with causal interventions. (a) Logit distribution for move option 1 at a particular game length (unplayed board). (b) Logit distribution of move option 2 pre- and post-interventions. The logit distribution post intervention resembles that of move option 1. (c) Cosine similarities between the two options pre and post interventions, averaged over 50 sample games. The difference between the the two cosine similarities shows when and where intervention is most influential.

We plot the cosine similarity between the logit distribution of move option 1 and that of move option 2 post-intervention in 4c, with cosine similarity of the two move options without interventions as a reference. We find that the intervention is mostly successful in middle layers and in early games. In particular, it seems that the intervention at the last two layers is not causally affecting the next-move decision at all, which is surprising since the linear probe has near peak accuracy at last few layers according to Table 1. In addition, in appendix B we show that shallow networks are worse at using the board state representation causally.

Taking account of the attention head distribution (Appendix C), we hypothesize that the linear world model is fully developed in the middle layers and the residual stream is mostly using the information in the middle layers to make next-move decisions. In late games, however, Othello-GPT might only rely on "surface statistics" as the game tree is vastly shrinking.

5 Conclusion

In this paper, we showed that the Othello-GPT neurons do encode information of the board state linearly, even for a shallow 1-layer network. Furthermore, the board state representation is used to make predictions with deep networks but not with shallow networks. Lastly, we demonstrated that the linear board state information is already finalized in the middle layers of the Othello-GPT and is used casually for decision-making, indicating that the transformer recalls information in the middle layers to make predictions.

In general, our work provides a firm playground for future explorations on understanding when, where, and how foundation models utilize the internal world model to make predictions. Future work involves using mechanistic interpretability tools to understand the circuits within the Othello-GPT and breaking down its decision-making process.

6 Acknowledgement

The authors would like to thank Neel Nanda and Kenneth Li for open-sourcing the Othello-GPT models. ZZ is supported by the Center for Brain, Minds, and Machines at MIT.

A The Rules of Othello and Board Representations

Here, we provide a brief summary of the rules of the game. Othello, also known as Reversi, is a classic strategy board game for two players, typically played on an 8x8 grid with two-sided pieces called "discs" that are black on one side and white on the other. The objective is to have more discs of your color on the board than your opponent at the end of the game. All games start with an initial setup of four discs placed at the center of the board, with two white discs forming a diagonal and two black discs forming the opposite diagonal. One player will play black discs while the other will play white. Players then take turns placing one disc of their color on an empty square on the board, with the white player always going first.

- Flipping discs: A move is valid only if it "sandwiches" (or "flanks") one or more of your opponent's discs between the disc you are placing and another of your discs already on the board. This must occur in a straight line (horizontal, vertical, or diagonal). After placing the disc, flip all the sandwiched opponent's discs to your color. If a player cannot make a valid move, they must pass their turn.
- Win Condition: The game ends if neither player can make a valid move or if the board is full. The player with the most discs of their color on the board wins the game. If both players have the same number of discs, the game is a draw.

B Emergence of Linear Representations in Smaller Othello-GPT Models

In this section, we report our findings on linear world representations for smaller models than the 8-layer-8-head Othello-GPT.

To this end, we trained a series of shallow models of the same architecture (Figure 1a) but with a number of attention blocks (layers) between 1 and 8 and with either 1 or 8 attention heads*. The models are trained on a dataset of over 20 million simulated games (where legal moves are played but without strategy) and the training losses are shown in Figure 5 (top row). During training, we periodically benchmark the models performance by computing the percent of predicted next moves that are legal for a validation set (500 game sequences/29,500 move predictions). Notably, we find that even the smallest model tested, GPT 1-Layer 1-Head, learns to play legal moves approximately 95% of the time, while the larger models predict legal next-moves over 99% of the time.

Next, we trained linear probes that map the activation vector to the new world representation for each layer and each model (see the discussion in section 3.1), with the training results shown in Figure 5 (bottom rows). The findings are similar to before in that the mapping is most accurate for deep layers rather than shallow (see for example the layer 1 probe vs the final layer probe). Interestingly, it appears that all Othello-GPT models tested present a similar ability to encode the world representation for a given layer; for example, the first layer probe accuracy and losses are similar for the GPT models with 1, 4, and 8 layers.

As in section 4.2, we conduct causal intervention studies for these models and the results are shown in Figure 6. We see that though there is a linear representation of the board states in shallow networks (1-layer), the next-move decision-making is not using the information causally. However, the 4L1H model shows similar results as the 8L8H model, suggesting that the minimum model for causal linear world representations can be really small.

C Visualization of Attention Heads

In this section, we provide more visualizations of the attention heads in different models. As discussed in section 3.2, there exist "yours" and "mine" attention heads. We see how they keep track of the moves in Figure 7 and 8 respectively, as the game progresses. We see clearly that heads 3 and 4 are each tracking one-half of the game pieces, which corresponds to my and the opponent's historical moves.

*Our naming convention is slightly misleading. We refer to GPT models in the sense that our trained neural architectures are similar to the GPT-2 architecture. We are not utilizing "pre-trained transformers" since all models are randomly initialized and trained from scratch.

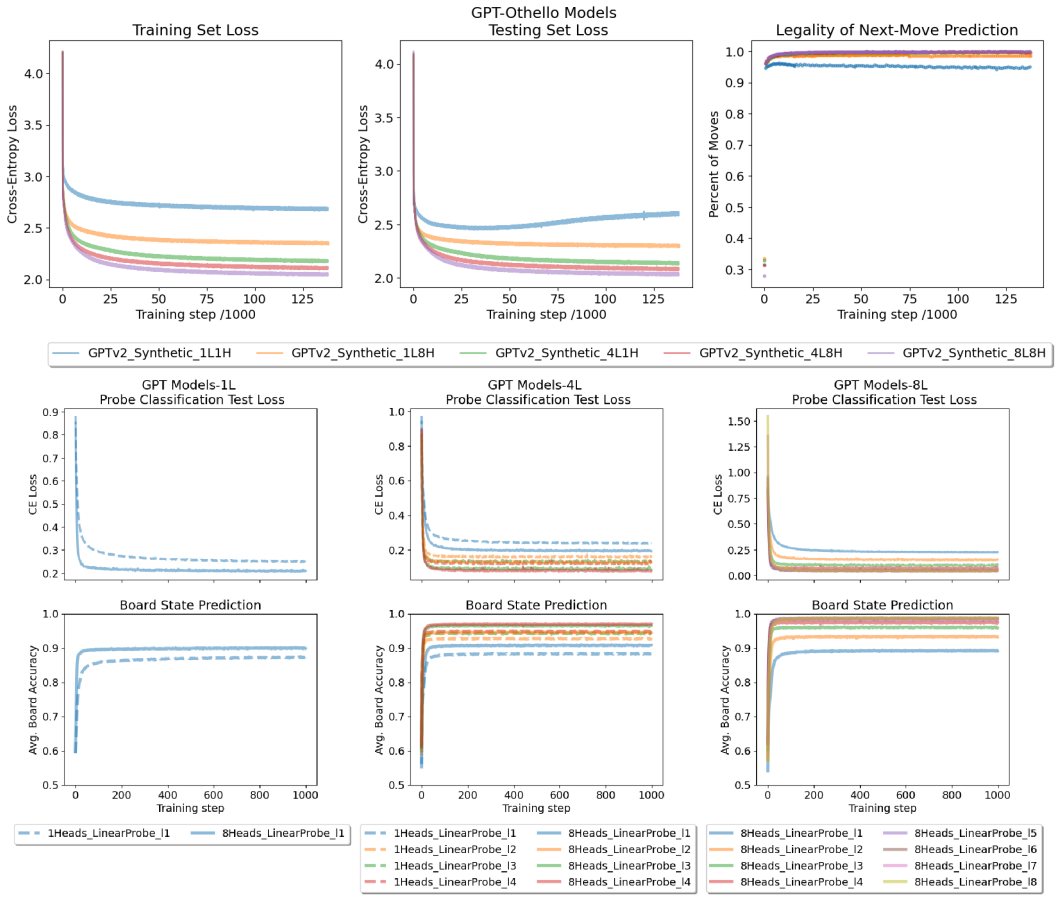

Figure 5: (First Row) Losses for various sequential models trained to play legal moves in Othello. The average percent of legal next-moves played on a validation set are [94.9%, 98.6%, 99.7%, 99.6%, 99.9%] for the models in the legend. (Bottom rows) Linear probe accuracy for different models at each layer.

Furthermore, we present in Figure 9 all of the attention heads of the 8L8H Othello-GPT model on a sample game. They can be roughly categorized as in Figure 10.

The attention heads roughly fit into four categories:

Table 2: Distribution of Attention Head Types across Layers

Layer	"Last" Head	"First" Head	"Mine" Head	"Theirs" Head	Other
Layer 1	0.00%	0.00%	37.50%	25.00%	37.50%
Layer 2	0.00%	0.00%	24.99%	49.98%	25.03%
Layer 3	0.00%	0.00%	49.98%	49.99%	0.03%
Layer 4	0.00%	0.00%	86.95%	12.50%	0.55%
Layer 5	12.50%	0.00%	62.27%	25.00%	0.23%
Layer 6	0.00%	0.01%	49.90%	37.50%	12.61%
Layer 7	0.00%	28.80%	10.19%	11.54%	49.47%
Layer 8	23.36%	25.79%	0.16%	26.49%	24.20%

- **First Token Heads**: These attention heads focus only on the first token of the sequence; examples include Layer 7 Head 1, Layer 6 Head 1, and Layer 8 Head 2 (Figure 9).

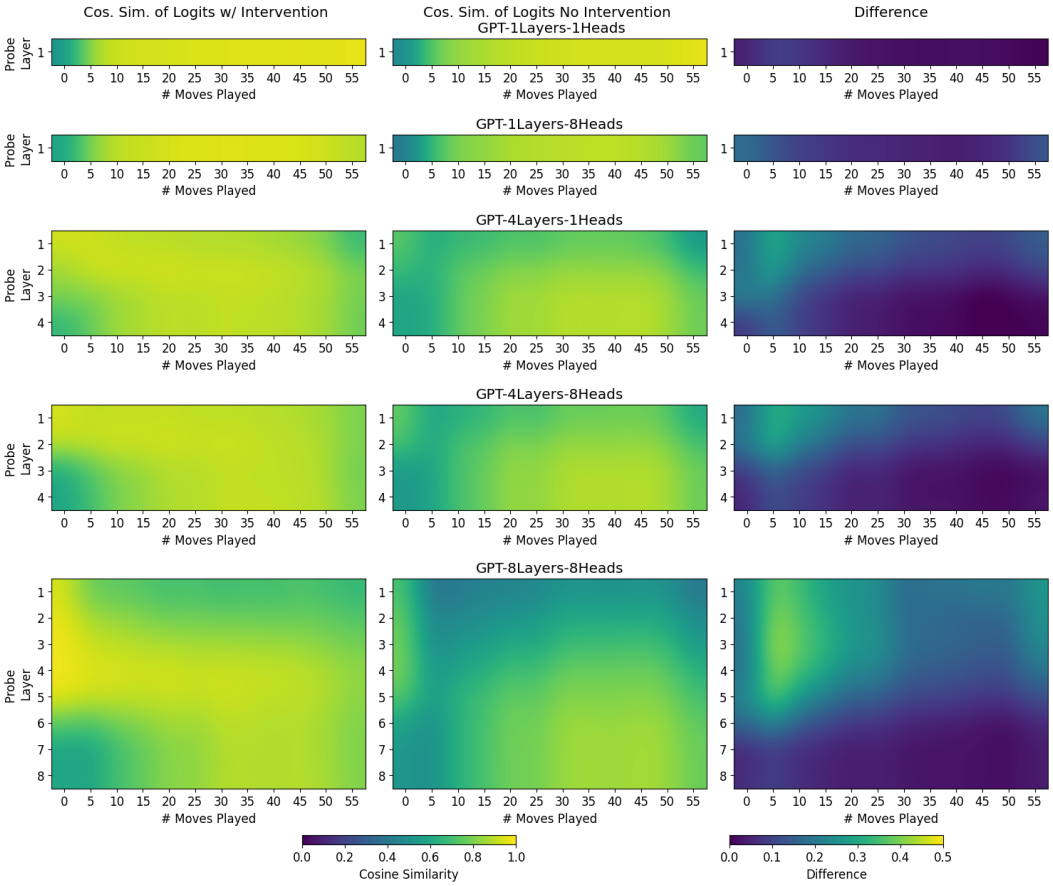

Figure 6: Cosine similarities for different Othello-GPT models with causal interventions. As in Figure 4 the data displayed is the cosine similarity of two logit vectors (ground truth vs. post-intervention logits or ground truth vs. pre-intervention logits), averaged over 50 games. Interventions are conducted for different probe layers and for game sequences of different lengths. The rightmost column displays the difference between the first and second columns and reveals where/when the interventions are most impactful to the model's next-move predictions.

- **Last Token Heads**: These attention heads focus on the second-to-last (position of last move you played) or last token (position of last move your opponent played); examples include Layer 5 Head 6, Layer 8 Head 3, and Layer 8 Head 1 (Figure 9).
- **My Token Heads**: These attention heads focus on positions corresponding to moves made by the color to play (for example, if white to play, it will pay attention to moves in the sequence that were made by white). Examples include Layer 1 Head 1, Layer 2 Head 3, and Layer 3 Head 2 (Figure 9).
- **Their Token Head**: These attention heads focus on positions corresponding to moves made by the color that just played (for example, if white to play, it will pay attention to moves in the sequence that were made by black). Examples include Layer 1 Head 4, Layer 2 Head 1, and Layer 2 Head 5 (Figure 9).

Table 2 shows the attention head types across layers. We see that the heads in Layers 1 through Layer 4 only consist of my token heads or their token heads. Additionally, we notice that in Figure 9, the "context size" of the attention heads for my token heads and their token heads seem to decrease over time despite still following the behavior we enumerated above; we see that the attention heads in earlier layers tend to pay attention to moves across the entire context, whereas attention heads in later layers only pay attention to the moves closest to them.

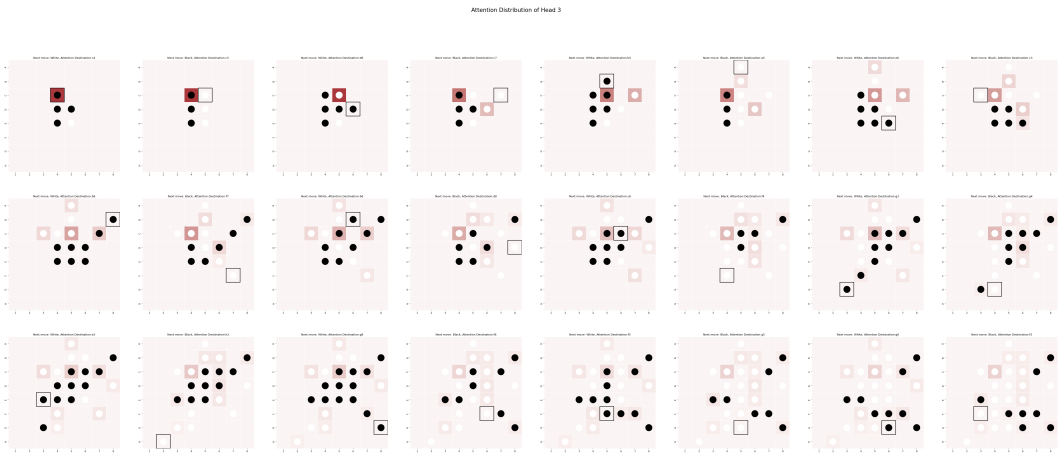

Figure 7: Attention distribution of head 3 on Othello board for the first 24 moves. For the same game as in Figure 2, we show the board state and attention distribution for the last move on the board, which is labeled by the black square. Dark color corresponds to large weights and light color corresponds to small attention weights.

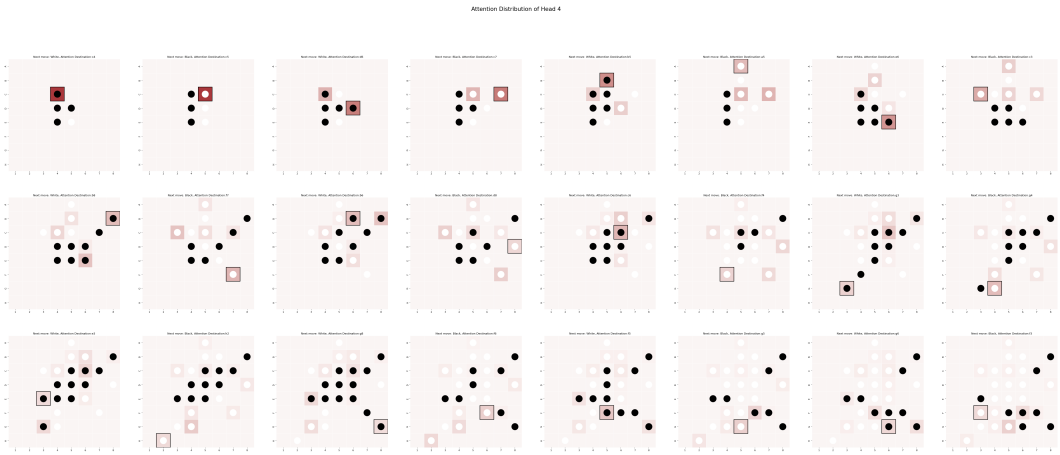

Figure 8: Attention distribution on Othello board for head 4. Head 4 is tracking a different set of pieces than head 3.

Another interesting finding is that the Last token heads only appear in Layers 5 and 8 and First Token Heads only in Layers 7 and Layers 8. Taking all this into account, we hypothesize that the necessary representations for the board state (my piece vs. your piece) are largely already finalized by the later layers (Layer 5 and onwards) and thus the causal interventions fail in deeper layers.

Lastly, we show that "yours" and "mine" attention heads emerge even for the 1-layer transformer, shown in Figure 11. This is also consistent with our findings that 1L8H network has a linear representation of the board states.

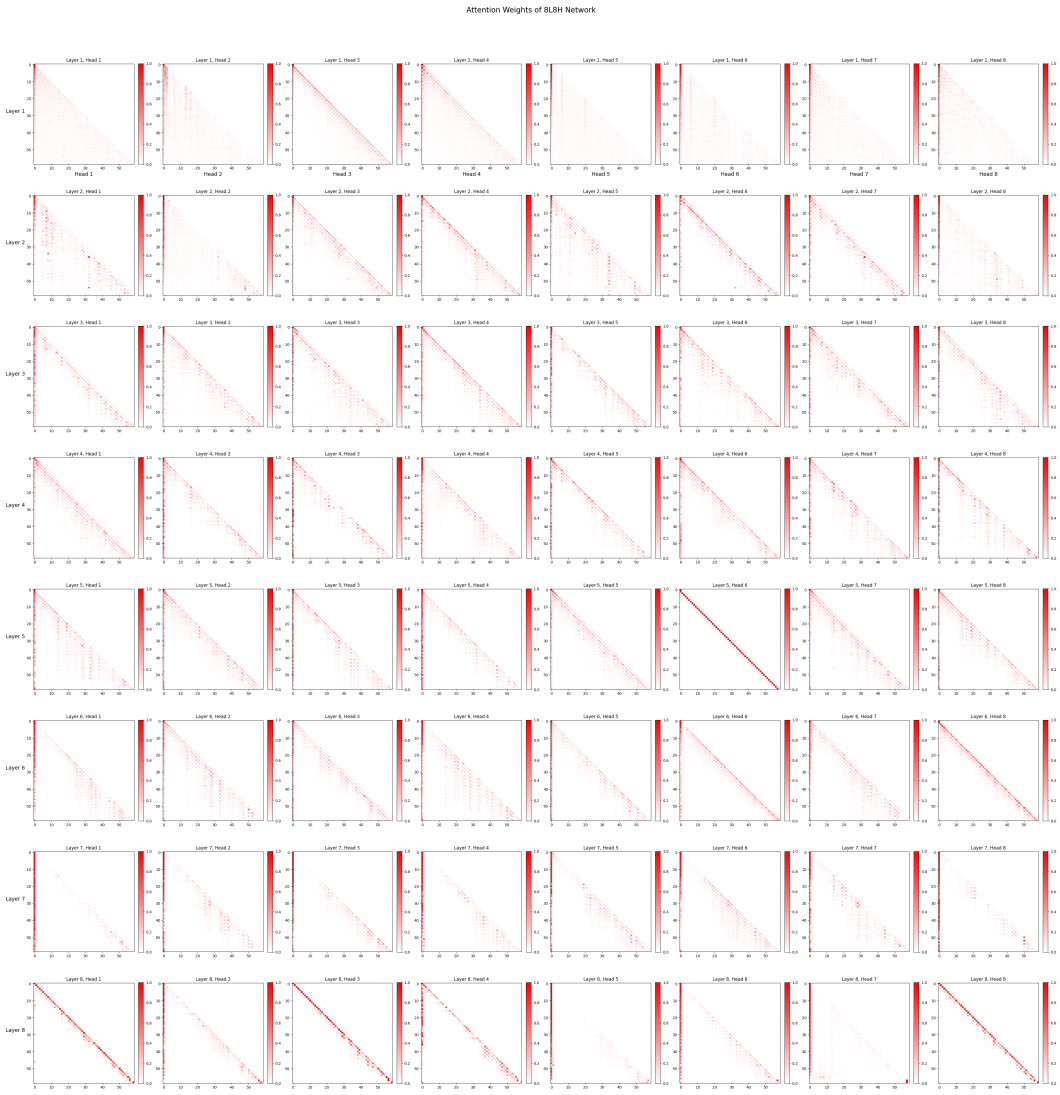

Figure 9: Attention heads for the 8L8H Othello-GPT on a sample game.

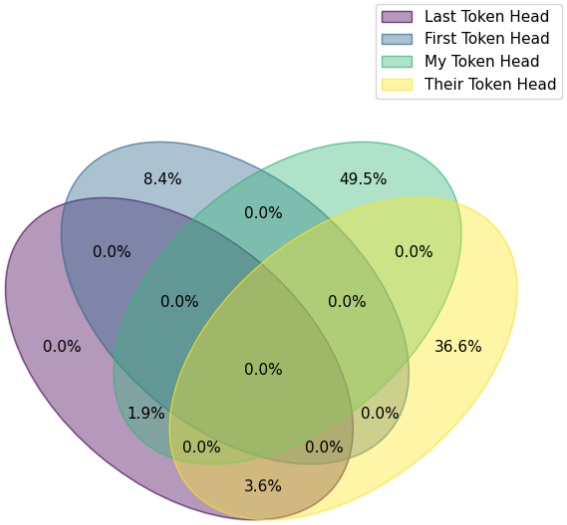

Figure 10: Venn diagram of attention head types. There are four type of circuits we find in our Othello model, last token heads (which looks at the last played move), first token heads (which look at the first played move), my token heads (which look at moves I made), and their token heads (which look at moves my opponent made).

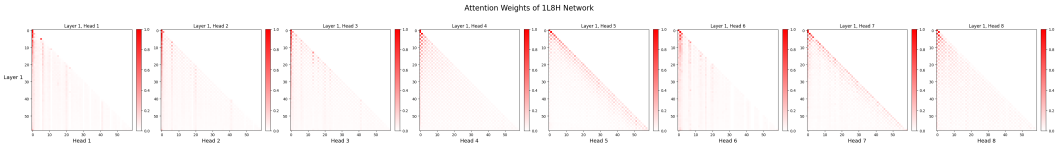

Figure 11: Attention heads for the 1L8H Othello-GPT on a sample game.

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644 [stat.ML]

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. 2021. Deep ViT Features as Dense Visual Descriptors. *CoRR* abs/2112.05814 (2021). arXiv:2112.05814

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. arXiv:2102.12452 [cs.CL]

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy, 276–286. `https://doi.org/10.18653/v1/W19-4828`

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712* (2022).

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research* 61 (2018), 907–926.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *The Eleventh International Conference on Learning Representations*. `https://openreview.net/forum?id=DeG07_TcZvT`

Neel Nanda. 2023. Actually, Othello-GPT Has A Linear Emergent World Model. `<https://neelnanda.io/mechanistic-interpretability/othello>`

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill* 5, 3 (2020), e00024–001.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* (2017). `https://doi.org/10.23915/distill.00007` https://distill.pub/2017/feature-visualization.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context Learning and Induction Heads. arXiv:2209.11895 [cs.LG]

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. Learning Chess Blindfolded: Evaluating Language Models on State Tracking. *CoRR* abs/2102.13249 (2021). arXiv:2102.13249