IS S23: Othello-GPT

Vignesh Pandiarajan
Brown University
vignesh_pandiarajan@brown.edu

Abstract

Some believe that language models are simply word prediction machines, that they have no understanding of what the content that they are trained on and the content they generate. Recently, there has been a push for research supporting the contrary, models can reason and apply knowledge to new concepts. In addition, models are being examined even more closely than before, in this field called Mechanistic Interpretability. Mechanistic Interpretability seeks to reverse engineer neural networks, but due to the scale and complex computations of neural networks, it is often difficult to interpret the results. In an effort to improve transparency, this paper finds that a GPT model playing Othello in fact has an emergent linear representation of the board. This paper goes on to visualize this board and find when it is computed. We also perform intervention experiments and analysis of individual neurons. This work adds to the growing body of work on interpretability.

1 Introduction

Large Language Models (LLMs) are able to complete a wide variety of tasks, such as generating dialogue, correctly identifying world relations, and summarization [6]. Their understanding of how concepts relate to each other is especially intriguing. This advanced multi-task capability is surprising, considering that these models are trained on much simpler tasks.

We know that these models have no access to the real world, they cannot interact and learn from real world concepts. Because of this, some propose that the explanation for their competency is a memorization of ‘surface statistics’ [1]. This memorization of statistics leads to not only opportunity for bias, but possibly confidently incorrect predictions for unfamiliar concepts.

However, even if these models cannot interact with concepts, some work indicates that it may be possible for them to build understanding of them. Recent experiments seek to investigate if we can see where in models there is a representation for direction or color[5].

There has been some significant success in using board games as a playground for understanding the internals of language models. Specifically, there has been some work in both chess and othello [3]. Since these games are constrained, it is easier to see if the models can reason about situations. Li et al. heavily dissects a model predicting Othello moves, motivating this paper.

Our main contributions are:

- We develop a simpler framework for training a linear probe that is much less complex than previous attempts.
- We verify the work of previous researchers and alter and optimize their workflows and create new visualizations to better illustrate concepts.
- We create several probes, both linear and nonlinear, at various points throughout the model comparing their accuracies as well as using them as tools for understanding where state is computed. We perform qualitative analysis of interventions with the aforementioned probes. We then compare these probes directly with cosine similarity and examine what parts of the model contribute the most to the probes.

- We identify a number of meaningful neurons, corresponding to the state of specific spaces in the Othello board.

2 Related Works

We build on recent research in internal representations and mechanistic interpretability.

Patel et al. demonstrate that LLMs such as GPT-3 learn some sort of conceptual space that can in some sense be “grounded” using some training examples. Patel’s findings suggest that (1) this is not memorization of ‘surface statistics’ and (2) there is some understanding of the previously learned concepts that can be applied to the new spaces not included in training. However, they do acknowledge that with purely textual inputs, investigation is difficult [5]. This indicates that there is some representation of concepts being computed when the model is sufficiently large.

Tenney et al. seek to explain syntactic and semantic phenomena within contextualized word embeddings. They introduce an edge probing framework in order to investigate contextual representation of words. They find that the representation of words within models actually contain syntactical information and also encode distant linguistic information [7]. This work indicates that contextualized word embeddings are able to encode not immediately obvious relationships and would lead us to believe that models that use these contextualized word embeddings would have some understanding of these relationships.

Elhage et al. proposes a linear representation hypothesis, hypothesizing that network representations can be described as independent features and features are represented by individual directions. They demonstrate on a toy model, showing that it is possible to interpret individual neurons and their outputs as corresponding to specific features[2]. This helps us understand why a MLP is not needed to probe the model and a linear probe is sufficient.

Li et al. serves as the key motivation for this paper, as they provide pretrained models and a dataset for this task. They also provide a strong framework for creating non-linear probes to understand the internal representations within the network. They intervene to show that these probes are effective and the accuracy of the probe is not coming from the probe itself [3]. This work indicates that an internal representation exists and it is possible to manipulate it.

Nanda refines and expands upon the ideas from Li et al. performing experiments with linear probes to show that the internal representation is actually linear. This serves as further evidence for the linear representation hypothesis. Nanda recreates much of Li’s work with this linear probe, also doing surface level comparisons on the probes [4]. This work provides many ideas for further experimentation in this paper and beyond, relating to neuron level exploration and more refined probe training.

3 Othello-GPT

Othello is a board game, less complicated than Chess and Go, but more complicated than simple games like Mancala. The game is played on an $8x8$ board, with 64 colored discs. The board starts with 2 black discs and 2 white discs at the center, with each color on one diagonal. Black always starts the game. Moves are made by placing a disc of the player’s color in a way that a disc of the opposite players color is sandwiched between two discs of a player’s color. Horizontal, vertical, and diagonal ‘sandwiches’ are all valid. If there are no valid moves for a player, they must pass their turn. The game continues until it is no longer possible for either player to make a move. The winner is the player with the most discs of their color on the board. There are enough possible games such that memorization is unlikely, so Othello is a good candidate to see if LLMs actually learn an internal representation.

3.1 Dataset

There are two datasets provided from Li et al., a championship dataset and synthetic dataset. The championship dataset is collected from online sources, they are real games that have been played competitively. The synthetic dataset is composed of randomly generated valid moves, with no strategy. The championship dataset has around $140,000$ games while the synthetic dataset has around 20

million games. Due to resource constraints and compression issues only $120,000$ of the championship games and 12 million of the synthetic games were usable.

3.2 Model

The Othello-GPT model is trained with only moves. Othello-GPT is not taught any of the rules of the game explicitly, but only provided game transcripts. Each tile index is a word in Othello-GPT's vocabulary and each game is a sentence. This model is an 8-layer GPT model, with an 8-head attention mechanism with a hidden space of 512. The model begins from random initialization. Model checkpoints are from Li et al., where the model has been trained on both datasets. The model trained on championship data often makes strategically good moves while the model trained on the synthetic dataset picks each legal move with roughly equal probability. This makes sense, as the synthetic data is valid random data while the championship data is from players playing with intent to win the game.

4 Methods

4.1 Training Non-linear Probes

We attempt to train non-linear probes on multiple layers in order to see if there is a representation of the board within the model. If the non-linear probe is able to extract an accurate board state at a layer, it would indicate that not only does the model compute an interpretable representation of the board but also that the board state is computed by that specific layer. However, non-linear probes are subject to some scrutiny, because with sufficient complexity, the probe might actually be performing computation instead of simply extracting the results of the computation within the model. If the non-linear probe is not able to extract a board state, it would indicate that this is not deep enough in the model or there is no board state to be extracted. Non-linear probes were trained on a randomly initialized model, the championship model, and the synthetic model.

4.2 Training Linear Probes

We also attempt to train linear probes on multiple layers in an effort to confirm the linear representation hypothesis. If only nonlinear probes were accurate then it would indicate that features were being represented in a nonlinear manner. An alternative explanation is that if only nonlinear probes seem to be accurate then we need to examine the model at a lower level, that these supposed non-linear features are actually composed of simpler linear features. Developing accurate linear probes would also confirm that only the results of the computation within the model are being extracted, since linear probes would not be powerful enough to produce a board state. Li et al. found that linear probes trained on the whole dataset returned extremely poor accuracies and thus proposed that the board representation was nonlinear. However, Nanda found that linear probes trained on every other move ended up being successful. They propose the reason for this is because the model only determines if the discs on the board are of the color of the player whose turn it is or not. That is to say, if it is black's turn, the black discs are interpreted as 'my' discs and the white discs are interpreted as 'opponent' discs. If it is white's turn, the white discs are interpreted as 'my' discs and so on. This means the representation essentially 'flips' every move.

4.3 Interventions & Latent Saliency Maps

After creating these probes, we aim to alter the internal board state to validate the probes. The interventions are essentially just changing activations such that the probe outputs a different prediction for the square in question. Ideally the model's subsequent move predictions will be changed because of this change in board state. If the move predictions are in fact changed, we know that the probe is valid and also that this internal board state is actually being used when computing the move predictions. If the move predictions are unchanged, it could mean that the probe is performing too much computation or that the information changed by the intervention is not being used when computing the prediction for he next move. We explore intervening on multiple layers at the same time as well intervening multiple times on a single layer. We also create latent saliency maps, which we can use to visualize how much the next move prediction is influenced by the pieces currently on the board. We do this by intervening on the pieces currently on the board then examining if the next move prediction has changed and if it has, by how much. This gives us the ability to identify patterns in the model's 'thinking' when it is making a move.

4.4 Neurons

We perform neuron level analysis, trying to figure out what features each neuron corresponds to. The first step is to pass multiple games into the model, then examining when the neuron is excited. The neuron may be excited after a certain combination of moves or a specific board state. This task is difficult because neurons can also be polysemantic [2]. We specifically examine board state and try to understand neurons in relation to if a cell is empty, black, or white.

5 Results

5.1 Linear Probe Accuracy

	Synthetic Dataset			Championship Dataset		
	Layer 4	Layer 5	Layer 6	Layer 4	Layer 5	Layer 6
Full Train	76.2	76.1	76.1	79.3	78.8	78.3
Skip Train	96.5	97.5	98.0	90.0	89.7	89.0

Table 1: Li Recreation Linear Probe Accuracies

Multiple different procedures were used to train the probes. First, we used Li's framework, where he trains a single probe to solve 64 classification problems at the same time. He trains this probe on every move in the dataset. We see that Li's approach yields relatively poor accuracies across both datasets and most layers. However, this probe was then trained on every other move, leading to a huge jump in accuracies. This is inspired by Nanda, but somewhat extended, as Nanda ignores the first few and last few moves in every game. The accuracies also appear to be significantly higher for the synthetic dataset than for the championship dataset. The reason for this is not exactly known, but we hypothesize that the championship model also prioritizes the recognition of specific patterns as well as the board state, possibly making it too difficult to extract the internal representation.

	Synthetic Dataset				Championship Dataset			
	Layer 3	Layer 4	Layer 5	Layer 6	Layer 3	Layer 4	Layer 5	Layer 6
Even Probe	90.8	93.3	97.1	99.3	62.9	64.5	65.3	65.3
Odd Probe	91.7	93.7	97.4	99.4	61.0	63.0	64.0	63.8
All Probe	69.0	70.3	72.3	75.4	56.5	57.3	57.9	58.2

Table 2: Nanda Recreation Linear Probe Accuracies

Additional probes were also trained using Nanda's framework. Nanda trains three different probes at the same time. The first probe is only trained on even moves, the second probe is only trained on odd moves, and the third probe is trained on all moves. The intention of this is to have probes only focus on one color at a time. This approach works extremely well. We see that the even and odd probes far outperform the baseline probe for all layers and both datasets. This provides strong evidence that the model computes board state in terms of the color that is playing and the color that is not playing. We see that accuracy drops even more sharply when probing the championship model. Again, this may reflect more complicated computation going on in the model. It might even be superposition, the concept that multiple features being encoded in one neuron interfere with each other. We must also acknowledge that accuracies seem to increase as we go deeper in the model. This could mean two things, that the internal representation is becoming obvious, or the internal representation has just been computed. It is likely that it is the former, as we will see when we examine nonlinear probe accuracy.

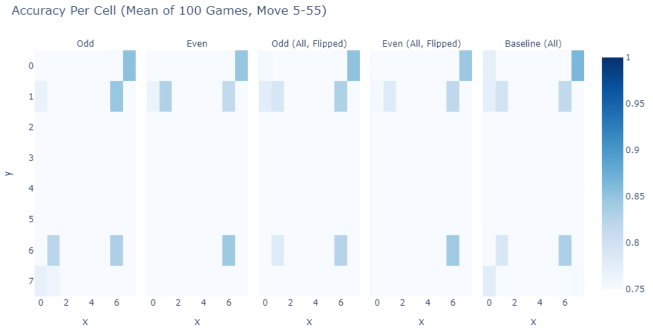


Figure 1: Nanda Linear Probe for Layer 6, Championship Model

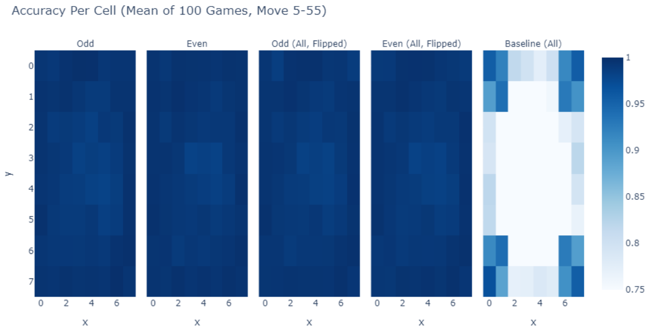


Figure 2: Nanda Linear Probe for Layer 6, Synthetic Model

5.1.1 Nonlinear Probe Accuracies

	Synthetic Dataset			Championship Dataset		
	Layer 4	Layer 5	Layer 6	Layer 4	Layer 5	Layer 6
Full Train	93.7	95.0	96.0	88.4	87.5	86.4
Skip Train	96.3	97.3	97.7	89.9	89.5	88.8

Table 3: Li Recreation Nonlinear Probe Accuracies

The nonlinear probes are trained following Li's framework, with a simple two layer MLP. This MLP easily outperforms the linear models trained on the whole dataset, on both the championship model and synthetic model. This makes sense, as the MLP has the computational ability to simulate the two linear models at the same time. The MLP also gets even better when training on every other move, as it's task becomes simpler. However, we see that the accuracies in the skip category are actually slightly inferior to the linear model. The differences are so slight that there is likely no conclusion to be drawn here. However, we see some interesting results on the championship dataset. The probe performs relatively well on the full dataset, indicating that there is some representation to be extracted and the MLP is better at extracting it and ignoring the possible interference. Still the accuracies are lower than on the probe for the synthetic model, indicating that there is still some interference that is uninterpretable.

5.2 Linear Probe Similarities

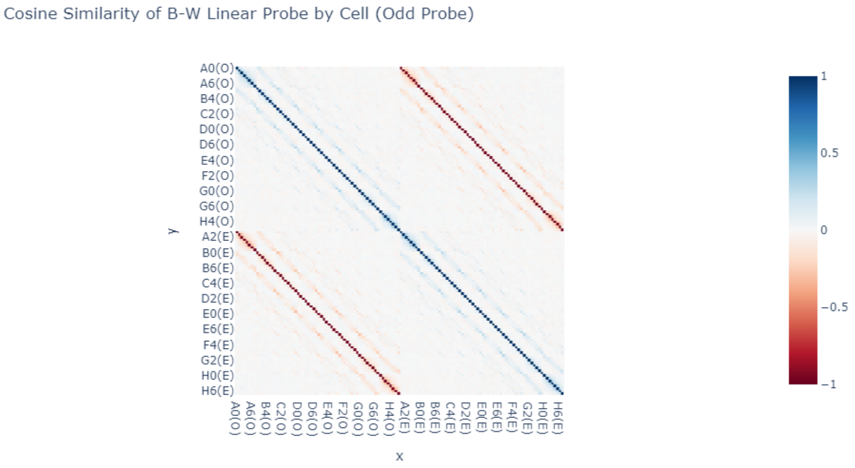


Figure 3: Cosine Similarity of Nanda Probe

We would expect Nanda's even and odd probes to be relative similar, or rather mirror images of each other. This ends up being true, as the even probe flipped performs very well on the odd moves and vice versa. To further investigate this, we plotted the cosine similarity of the black direction minus the white direction for the odd probe and even probe. We see multiple correlations, indicating that these probes do have a relationship and are inverses of each other.

We also perform some similarity experiments on the probes from Li's framework. We see that the cosine similarity between properly trained linear probes and the first layer of an MLP probe is much higher than random correlation. This indicates that similar computations are happening here, which is what we expect. We also compare a randomly initialized linear probe to a championship linear probe and find that the cosine similarity is much lower, indicating that the relationship between the properly trained probes is not spurious. The related figure is located in the appendix.

5.3 Model to Probe Correspondence

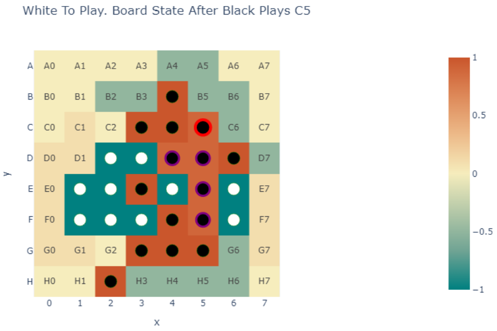


Figure 4: Board for extracting model contributions & intervention experiment

The key figures for this section are figures 4, 5 and 6. These figures are visualizations of what parts of the model contributes to the probe. These visualizations are simply sanity checks. These images are for the 'white to play' probe trained on layer 6, so obviously it makes sense that the major contributions are from layer 6. It is interesting that there is also some strong negative contribution from earlier layers when looking at the white squares. This could be happening because there are specific attention heads devoted to monitoring which white pieces have been flipped or recently unflipped.

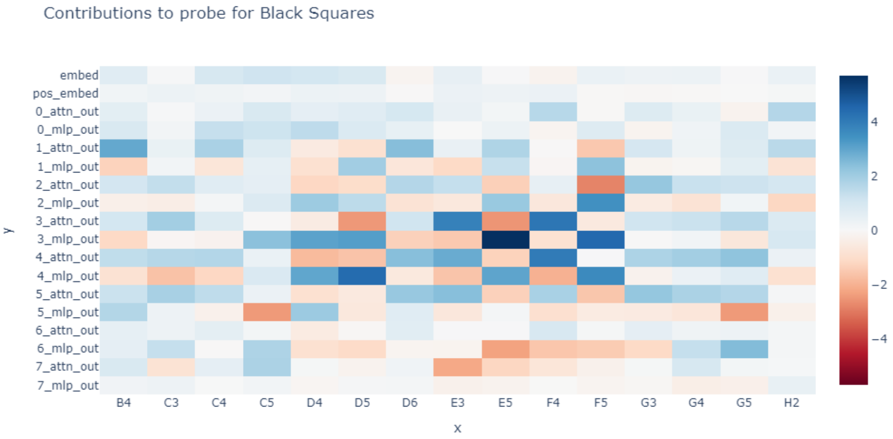


Figure 5: Visualization of model contributions to Nanda black probe

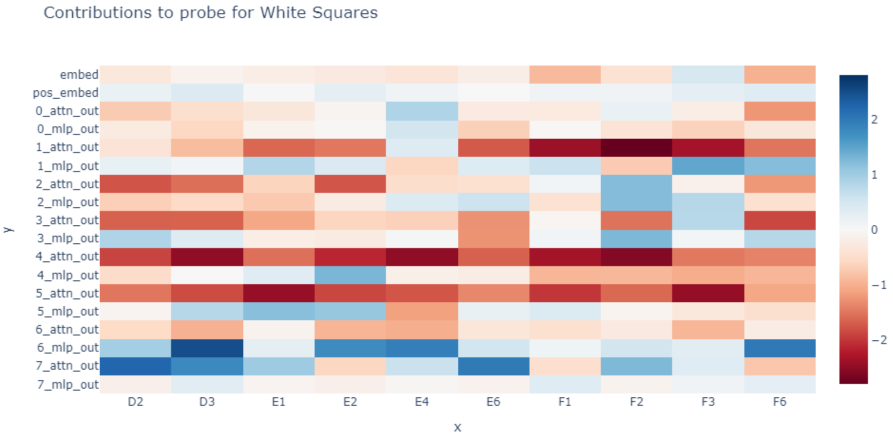


Figure 6: Visualization of model contributions to Nanda white probe

5.4 Interventions

Quality of interventions corresponds closely to probe accuracy. Interventions often fail when using a faulty probe, because if a probe cannot correctly extract the internal representation there is no opportunity to properly modify it. We also observe that interventions at all layers are not required. Intervening at one or two layers seems sufficient if the probe is accurate enough. Probes also seem to generalize well from layer to layer, a probe with good accuracy on layer 6 can also be used to modify layer 7, even up to layer 9 with no obvious issues. Li observed that interventions with linear probes was not possible, but with the new skip procedure, the interventions work well. Interventions with MLPs or linear probes do not seem to have much difference and both seem equally effectively in yielding new legal moves or removing newly illegal moves. Interventions using the probes trained on the championship model exhibit slightly strange behavior, the model often still focus on the same few moves as before. This might be because of the strategy component, it takes multiple interventions to change the optimal move or pattern that has been recognized.

A latent salient map has also been generated, exhibiting reasonable behavior. The black box means this is the predicted square and the more red a square is, the more it contributes to this prediction. This map was generated with a probe trained on the synthetic dataset and so does not seem to be

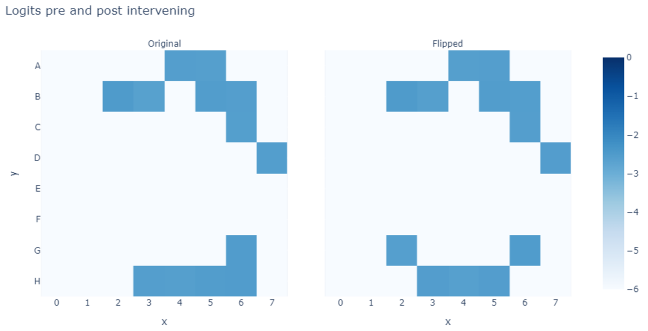


Figure 7: Logits after intervening to turn G5 white

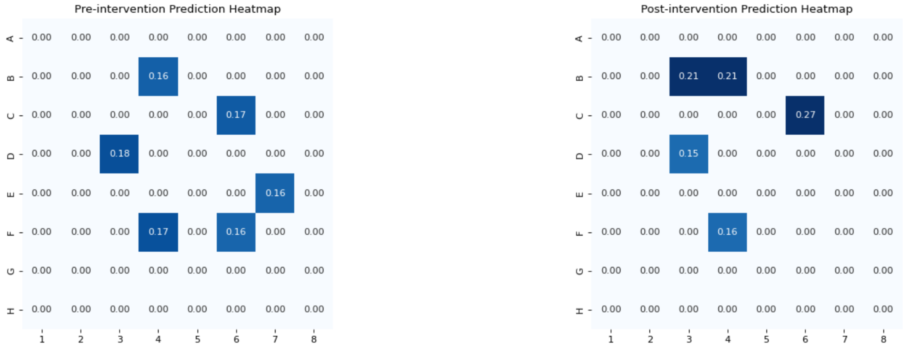


Figure 8: Predictions after intervening to turn G5 white

paying as much attention to strategy as trying to take the obvious move. It pays attention to the immediately relevant blocks, making sure there exists a disc to sandwich and a disc to serve as the other end of the sandwich. This image is also located in the appendix.

5.5 Examples of interesting Neurons

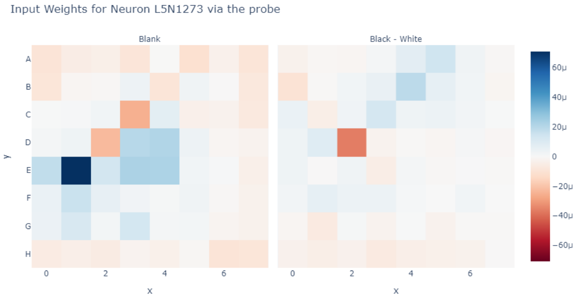


Figure 9: Visualization of Neuron L5N1273

From some exploration, it seems that neurons can be activated upon specific compound conditions. For example, Nanda finds that neuron 1393 in layer 5 seems to be activated on the following condition: (C0==blank) AND (D1==theirs) AND (E2==mine). We contribute to this library by finding Neuron 1273 and 876, also in layer 5. Neuron 1273 seems to be activated when (D2==theirs) AND (E1==blank). Neuron 876 seems to be activated when (D2==mine) OR (D2==blank). This is interesting because Neuron 1393 clearly corresponds to a specific move. However, Neurons 1273 and 876 do not seem to be activated in preparation for any specific reason. Even examining the graphs we can see that the difference in activation across the entire board is not very significant. This is strange, considering that these neurons are among the most highly activated throughout multiple games. This may indicate that these neurons are always active or there is a yet to be discovered pattern.

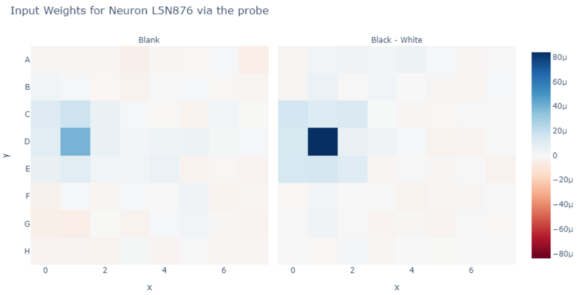


Figure 10: Visualization of Neuron L5N876

6 Discussion

In this research, we aim to recreate and improve on experiments by past researchers. The key ideas is investigating these techniques on this smaller scale so that we can extend these ideas to real LLMs. Of course, we must acknowledge that not all findings from this model can be extrapolated in larger models as Othello-GPT is a far simpler model than an actual GPT model. However, the evidence provided in this paper strongly supports the idea that models have some internal reasoning and there is some representation that can be extracted.

The research in this paper raises many questions. One of the most interesting questions is 'how does the championship model differ from the synthetic model?' With such different probing behavior there must be something going on that we are yet unaware of. Also, how does a model respond when one cell is changing colors multiple times. Does this effect accuracy? We would also like to explore exactly how the model computes if a cell is blank. Is there a single attention head somewhere devoted to this idea? Of course, there also must be some way to train a better and more robust probe.

The prior questions are all specific to this project, but looking at the bigger picture, what other situations can we extract internal representations from? The obvious follow up is different board games like Go, but what if we investigated games that do not have perfect information, say for example Risk, and the fog of war. How would the internal representation be computed in that scenario? We hope some strategies from this paper can be of use in the future experiments.

7 Acknowledgements

I would like to acknowledge Ellie Pavlick for her direction over the course of this project and as a mentor during the majority of my research experience. I would also like to thank the other members of the Lunar Lab for helping me shape this project and always being willing to help when I needed it, specifically Jack Merullo and Tian Yun. This project was conducted for an Independent Study course at Brown University.

References

[1] Emily M. Bender and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463.

[2] Nelson Elhage et al. *Toy Models of Superposition*. 2022. arXiv: 2209.10652 [cs.LG].

[3] Kenneth Li et al. *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task*. 2023. arXiv: 2210.13382 [cs.LG].

[4] Neel Nanda. *Actually, Othello-GPT Has A Linear Emergent World Model*. Mar. 2023. URL: https://neelnanda.io/mechanistic-interpretability/othello.

[5] Roma Patel and Ellie Pavlick. "Mapping Language Models to Grounded Conceptual Spaces". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=gJcEM8sxHK.

[6] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2018). URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[7] Ian Tenney et al. "What do you learn from context? Probing for sentence structure in contextualized word representations". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=SJzSgnRcKX.

A Timeline

I started off the semester in January, looking for a new project for the semester. I met with Mikey and Jack, talking to them about if they had the bandwidth to take on another undergraduate student. I also applied to work with Sam, but unfortunately it seemed like all of them were otherwise busy.

By this time, it was mid February and I decided to at least start by working on the BLOOM project. I ran some experiments attempting to recreate Jack's results. I was somewhat successful and also doing some reading on prompt engineering. I tried out a few approaches to try and improve accuracy, but I was running into problems with the quality of the dataset. I had a brief conversation with Jack and he suggested I try another dataset. I made some efforts to do so, but I started to become far more interested in some of the ideas we were talking about in my deep learning class.

This was around March, when I attempted to change the project I was working on for the semester. I was excited by the idea of circuits and mechanistic interpretability and how accessible it seemed. I had been feeling a bit lost for some time, so this seemed like a great place to jump in. I read some papers by the people at Anthropic and some of Neel Nanda's open problems. I finally was able to decide on a project and even joined the mechanistic interpretability reading group. I was able to have a conversation Jack just about ideation for the project and got a bit of feedback.

Then, it was April, where I presented twice in lab and started puting serious time into the project. I was told that Tian would be helpful to talk to and he might be interested in Othello-GPT. I talked to him a little about the intentions of my project and I was able to get a better sense of what I wanted to learn and explore. Since then, work has been coming along and the report was taking shape.

B Takeaways

I really feel like I learned a lot from this research project. I rewrote tons of code, trying not to rely on the code from the previous authors. I ended up using some of their visualization functions just because they were so good looking. I debugged my own issues and dealt with anything ranging from long training times that suddenly failed and stubborn preprocessing. I even ran into an issue where someone else was running a job on a gpu that I was interacting with. Regardless, I got a lot of really good exposure to little torch tips and transformerlens and am excited for my next project.

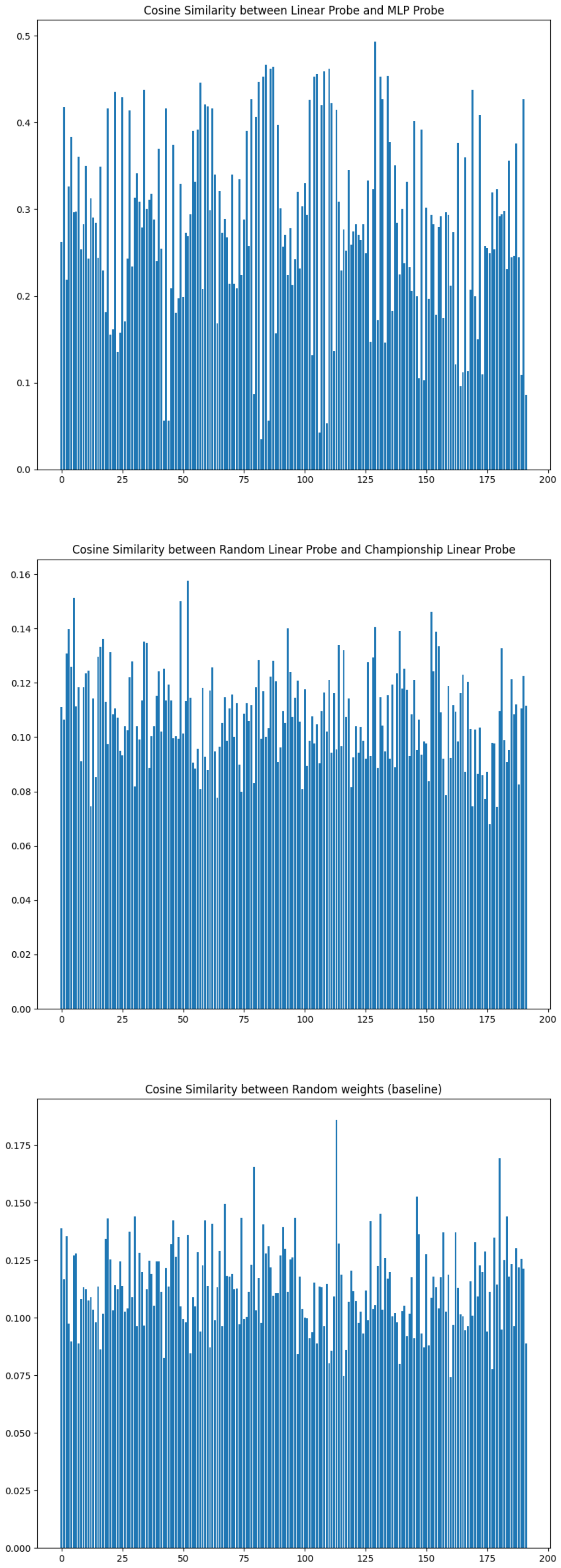


Figure 11: Cosine similarities between various Li Probes

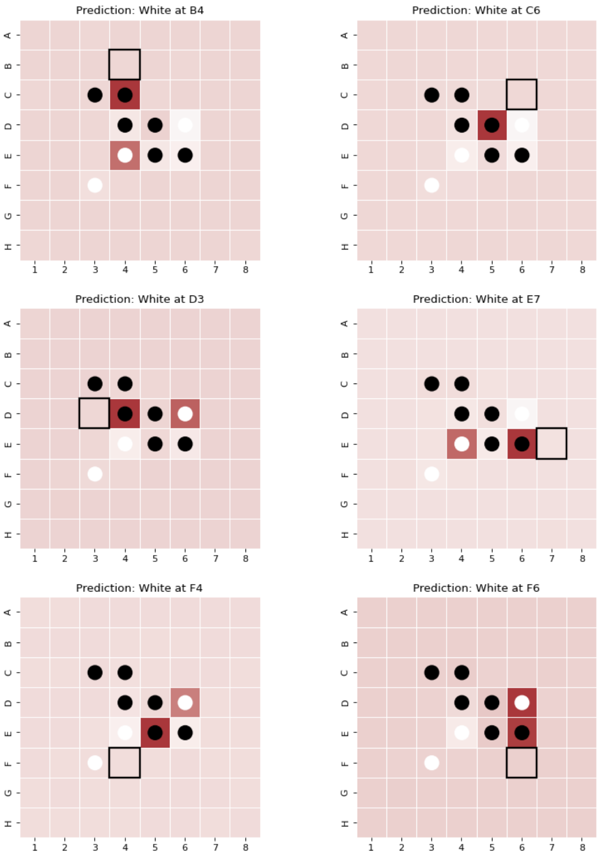


Figure 12: Latent saliency map for Othello-GPT trained on synthetic dataset

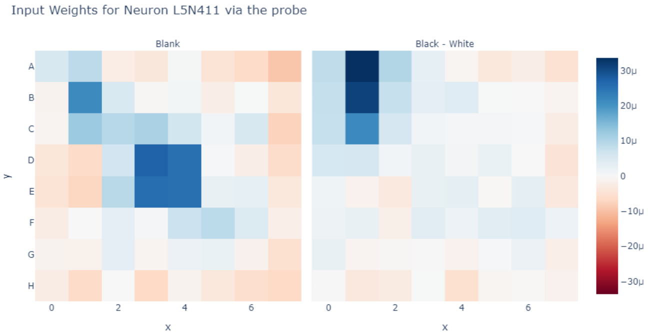


Figure 13: Neuron activating on blank middle, my color top left

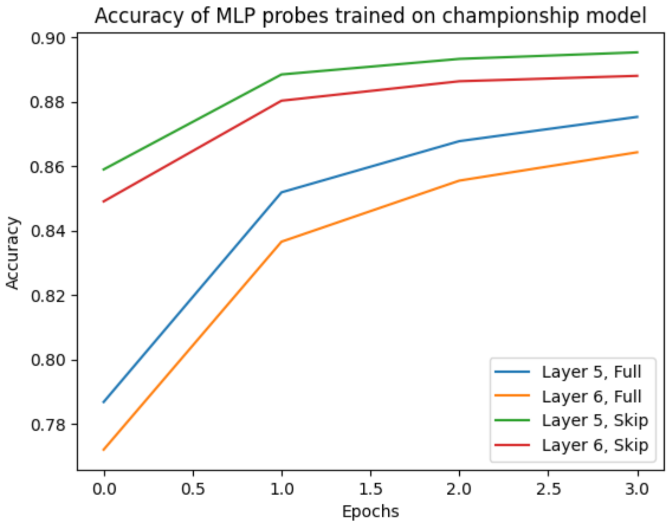


Figure 14: Test accuracies of MLP probes over epochs for championship model

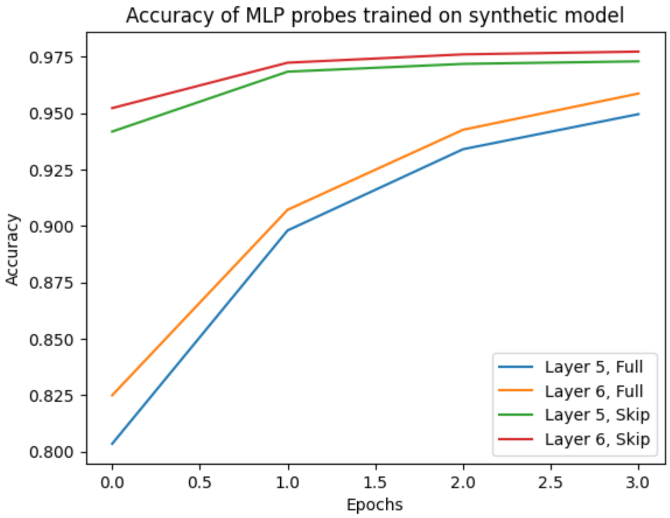


Figure 15: Test accuracies of MLP probes over epochs for synthetic model

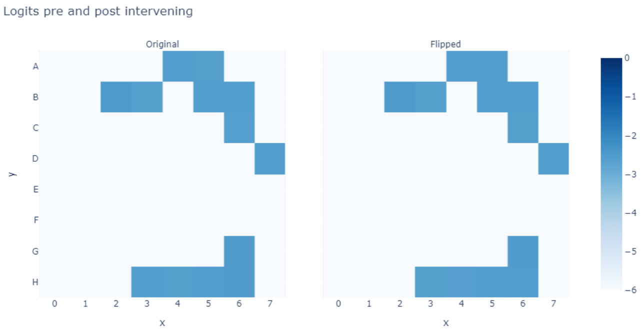


Figure 16: Failed intervention using inaccurate probe

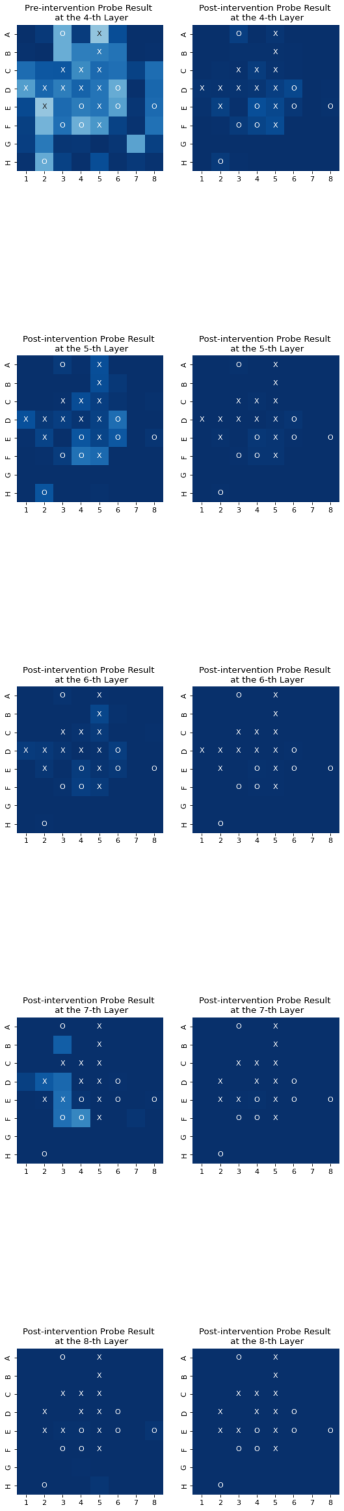
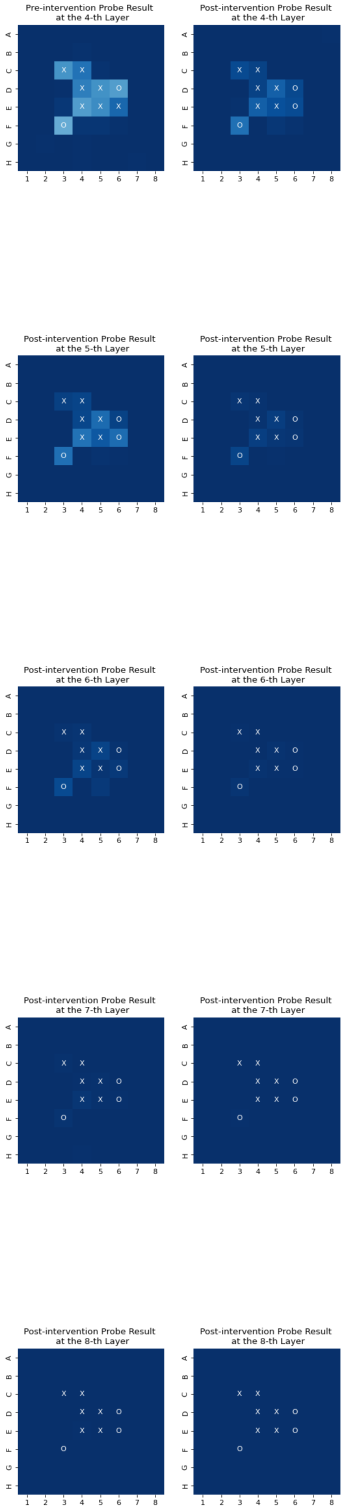


Figure 17: Failed multilayer intervention with MLP probe


Pre-intervention Probe Result at the 4-th Layer

Post-intervention Probe Result at the 4-th Layer

Post-intervention Probe Result at the 5-th Layer

Post-intervention Probe Result at the 5-th Layer

Post-intervention Probe Result at the 6-th Layer

Post-intervention Probe Result at the 6-th Layer

Post-intervention Probe Result at the 7-th Layer

Post-intervention Probe Result at the 7-th Layer

Post-intervention Probe Result at the 8-th Layer

Post-intervention Probe Result at the 8-th Layer

Figure 18: Successful multilayer intervention with Linear probe