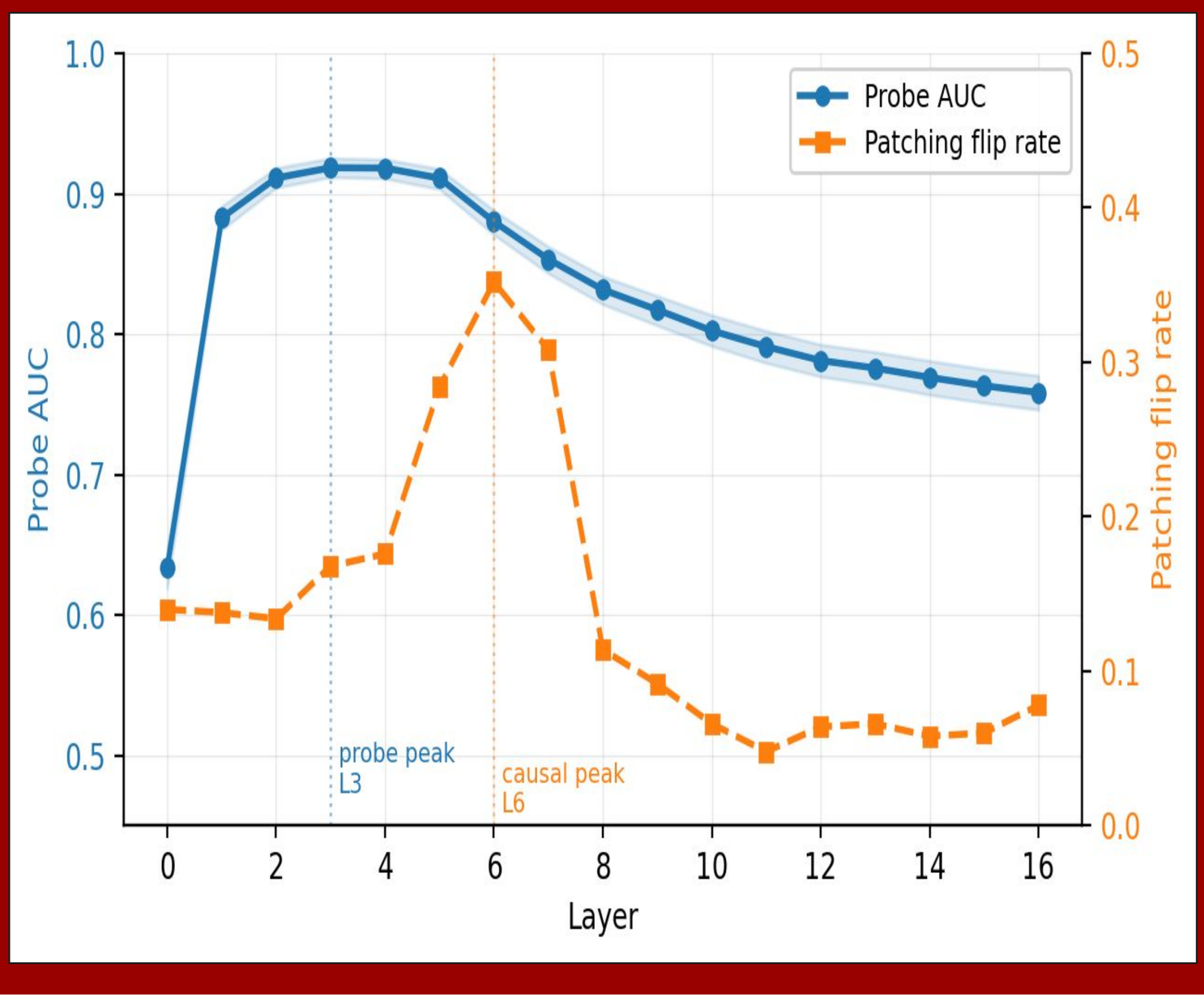


Standard Interpretability Evidence Can Overstate Behavioral Causation

Linear Pin Representations and the Limits of Patching Evidence in a Searchless Chess Transformer

Lucas Fugate

1 Abstract

MAIN 3 TAKEAWAYS

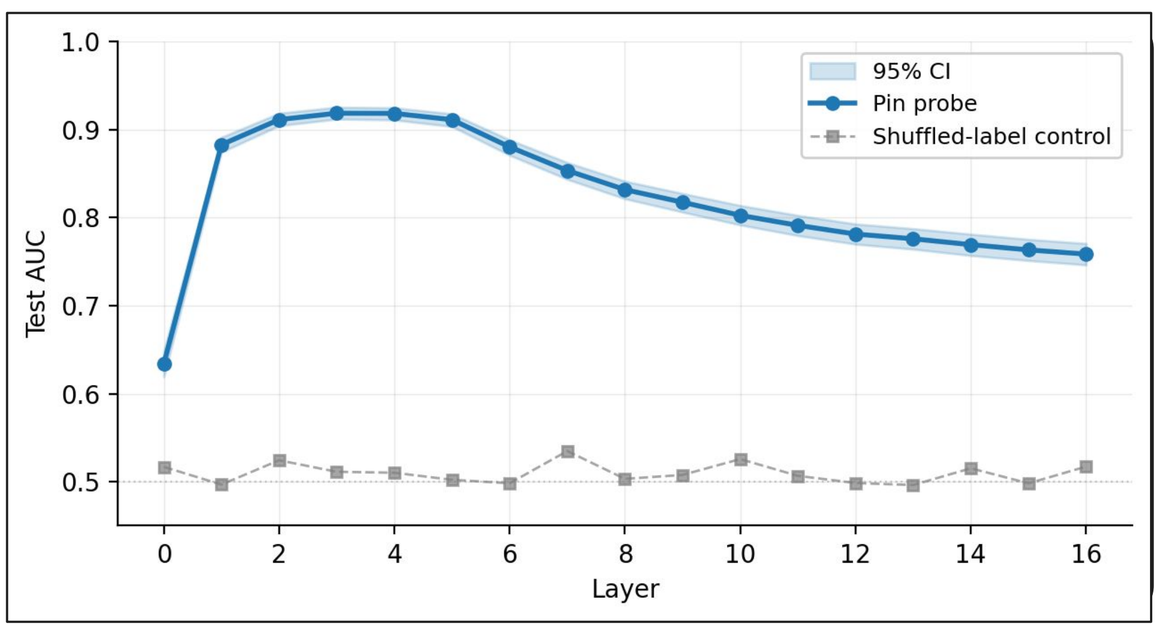
- The searchless chess transformer linearly represents absolute pins beyond board geometry.
- Probe and patching peaks occur at different layers in the 270M and 9M engine variants.
- Probe accuracy does not localize causal computation and directional specificity does not establish concept-specific encoding.

2 Background / Motivating Question

- By the linear representation hypothesis, advanced semantic concept often correspond to directions in activaton space.
- We extend prior research on engines like AlphaZero and OthelloGPT.
- Tactical motifs have been largely unexplored in such models, with some previous studies on positional ideas instead.

3 Probing for Pin Representation

- The 270M probe peaks at layer three with AUC 0.919, indicating that information is linearly decodable in early layers. Using a bitboard baseline, we confirm this is not simply from board geometry.
- Most of the probe gain happens after a single transformer block, indicating that relational structure requires additional computation beyond the embedding layer.



4 Activation Patching

- Intervention substantially reshapes the value distribution, but actually rarely changes which move the model plays.
- The pin direction encodes pin-related information distinct from line-attack information, but may not be the dominant encoding direction.

Layer	Mean KL	Flip rate	95% CI	Logit diff
0	0.146	0.140	[0.112, 0.173]	0.014
1	0.196	0.138	[0.111, 0.171]	0.016
2	0.155	0.134	[0.107, 0.167]	0.017
3	0.196	0.168	[0.138, 0.203]	0.015
4	0.029	0.176	[0.145, 0.212]	0.003
5	0.029	0.284	[0.246, 0.325]	0.000
6	0.042	**0.368**	[0.327, 0.411]	0.002
7	0.040	0.308	[0.269, 0.350]	0.002
8	0.005	0.114	[0.089, 0.145]	0.006
9	0.003	0.092	[0.070, 0.121]	0.009
10	0.002	0.066	[0.047, 0.091]	0.011
11	0.001	0.048	[0.032, 0.070]	0.014
12	0.002	0.064	[0.046, 0.089]	0.017
13	0.002	0.066	[0.047, 0.091]	0.017
14	0.002	0.058	[0.041, 0.082]	0.016
15	0.002	0.060	[0.042, 0.084]	0.015
16	0.002	0.078	[0.058, 0.105]	0.015

Direction	Norm	Patch layer	Flip rate
Δh_3	144	3 (own)	0.156
Δh_6	1094	6 (own)	**0.368**
Δh_3	144	6 (cross)	0.014
Δh_6	1094	3 (cross)	0.702

- The presence of a probe-patching gap seems to indincate distinct layer-specific encodings of the same concept.
  Probe accuracy can be a misleading proxy for the casual role of a feature, with reliable perturbation being further downstream then earlier indicated.

5 Limitations and Discussion

- Our placebo tests indicate that the pin direction uniquely encodes pins, as other meaningful directions, such as line-attack, exist.
- We solely study absolute pins, as relative pins require symbolic reasoning beyond python-chess.
- Internal representation of the Ruoss Model is particularly interesting, as its training, with Stockfish action-values, no search, and no self-play, is different from previous engines.

PAPER

CODE

lucasfugate@gmail.com