Transformers can **learn to implement look-ahead** in a single forward pass.

Evidence of Learned Look-Ahead in a Chess-Playing neural network

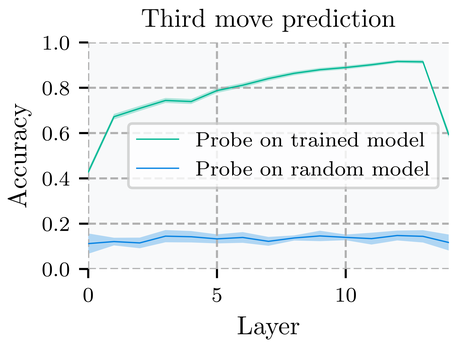
Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, Scott Emmons, Stuart Russell

Center for Human-Compatible Artificial Intelligence

Berkeley UNIVERSITY OF CALIFORNIA

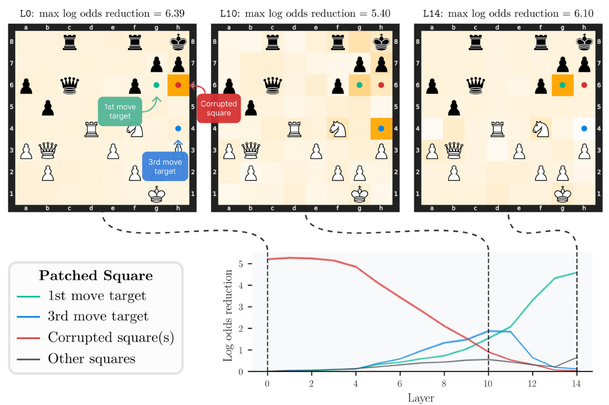
Three lines of evidence

1



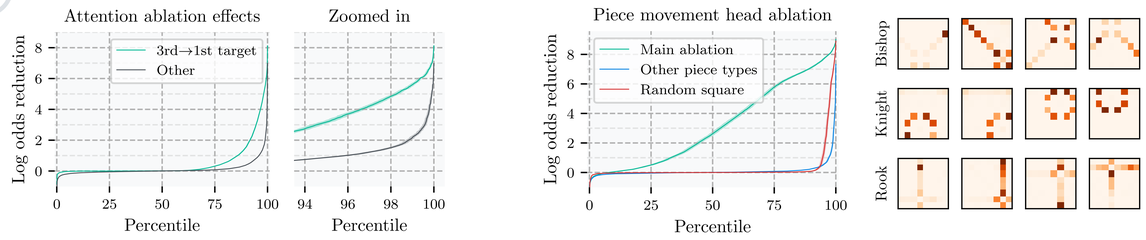
Simple **probes** can predict the **optimal move 2 turns into the future** with **92%** accuracy.

2



Activations on **squares of future moves** store **important information**.

3



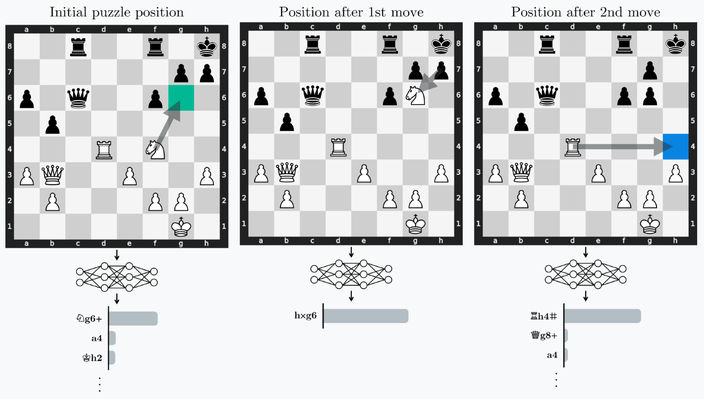
Specific **attention heads** seem **involved in look-ahead**.
Carefully targeted **ablations** in these heads have **outsized effects**.

(a) L12H12 moves information **"backward in time"** from the target square of a future optimal move to the target square of the current optimal move.

(b) "Piece movement heads" help compute **consequences of future moves**, looking **"forward in time."**

Setup details

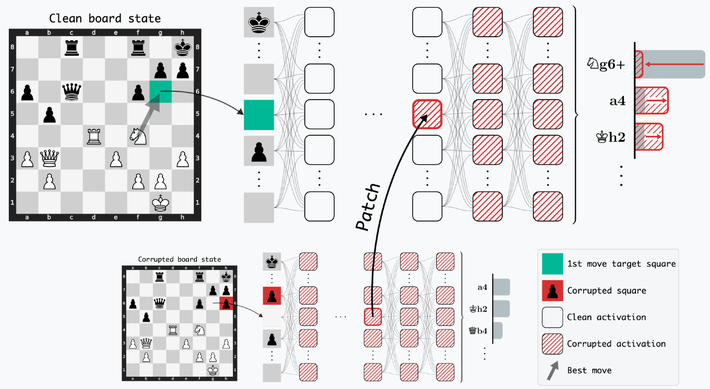
Model and dataset



Dataset: board states from puzzles, automatically filtered to be **complex** but still **solvable** for the network. Annotated with a unique *principal variation* of at least three moves.

Network: policy net of Leela Chess Zero (strongest MCTS engine). **Transformer** that maps a board state to a move distribution. It **treats every square like a token in a language model**.

Activation patching



Mechanistic interpretability method that lets us determine **how important** a given **model component** is.
Patch an activation from the forward pass on a "corrupted" state into the forward pass on the original state.
Then **measure how much this intervention affects the output**.
We use **automatically generated corruptions**.