Othello Mech Int playground [1]

Victor Levoso **Kunvar Thaman**

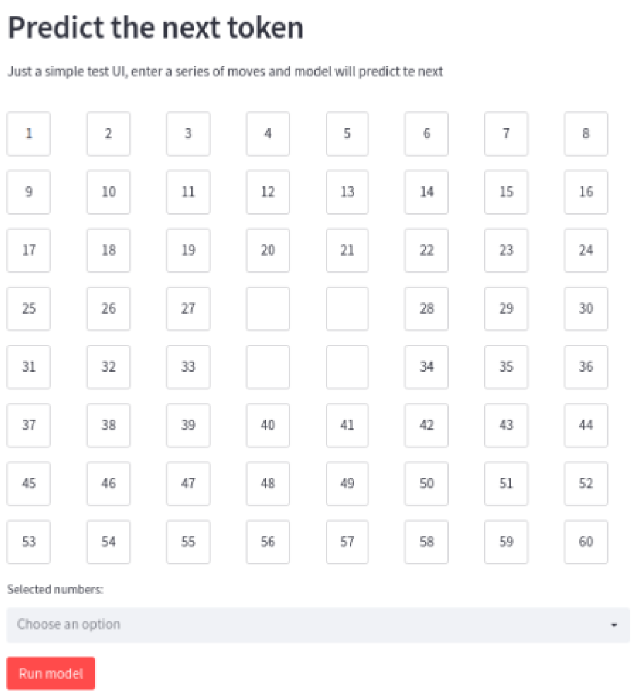
Edoardo Pona **Abhay Sheshadri**

Made an interactive app to run basic interpretability experiments on othello.

Keywords: Mechanistic Interpretability, Othello GPT,streamlit app.

This is a modification of the “Trafo Mech Int playground” project (by Stefan Heimersheim and Jonathan Ng) to work on Othello-GPT instead of LLM.
Maybe available in streamlit but might crash at some point due to memory limitations.
Also available in a github repository to run locally.

Screenshots



Predict the next token

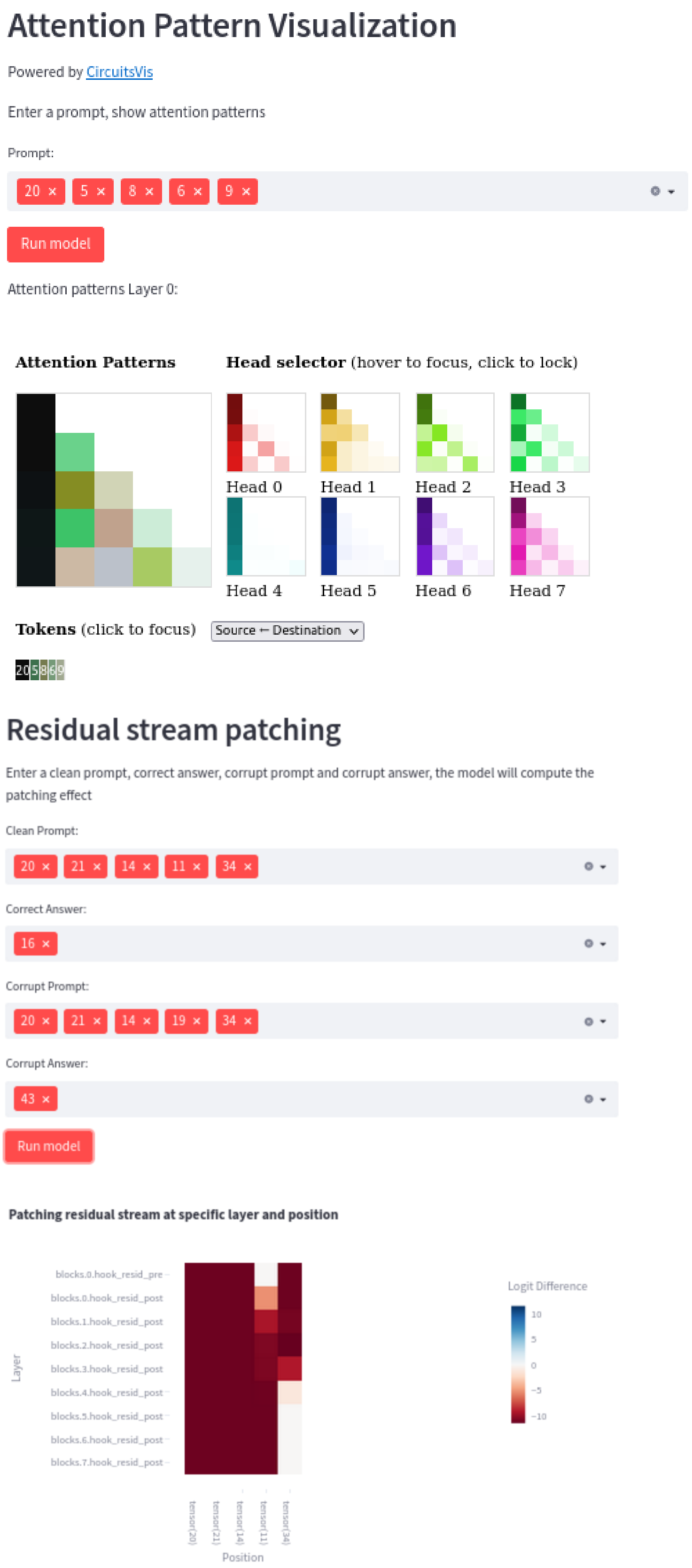
Just a simple test UI, enter a series of moves and model will predict te next

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27			28	29	30
31	32	33			34	35	36
37	38	39	40	41	42	43	44
45	46	47	48	49	50	51	52
53	54	55	56	57	58	59	60

Selected numbers:

Choose an option

[1] Research conducted at the Apart Research Alignment Jam #7 (Interpretability 2.0), 2023 (see https://itch.io/jam/interpretability-hackathon)



Attention Pattern Visualization

Powered by CircuitsVis

Enter a prompt, show attention patterns

Prompt: 20 × 5 × 8 × 6 × 9 ×

Attention patterns Layer 0:

Residual stream patching

Enter a clean prompt, correct answer, corrupt prompt and corrupt answer, the model will compute the patching effect

Clean Prompt: 20 × 21 × 14 × 11 × 34 ×

Correct Answer: 16 ×

Corrupt Prompt: 20 × 21 × 14 × 19 × 34 ×

Corrupt Answer: 43 ×

Patching residual stream at specific layer and position