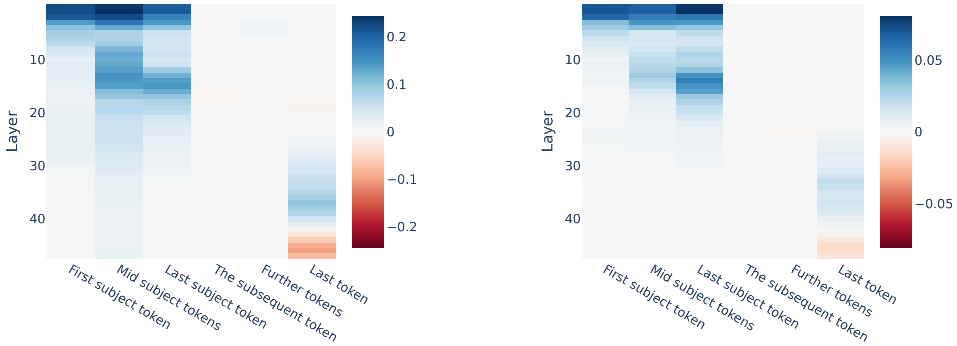
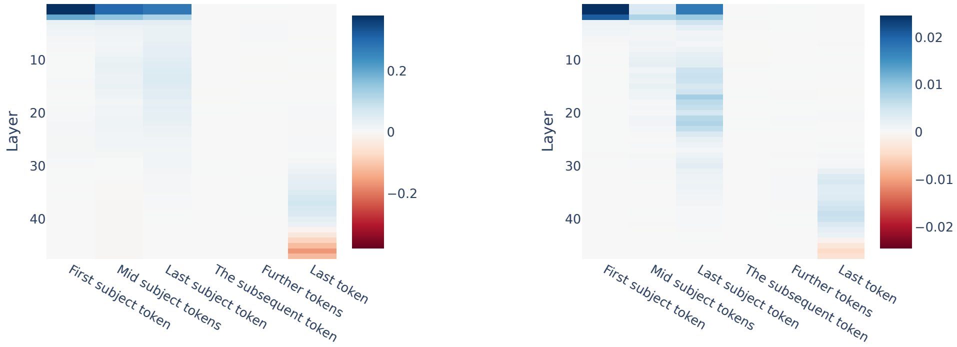


(a) Logit difference (GN)

(b) Probability (GN)

Figure 15: **Activation patching on MLP** across layers and token positions in GPT-2 XL on factual recall prompts. Apply GN corruption and a sliding window of size 3.



(a) Logit difference (STR)

(b) Probability (STR)

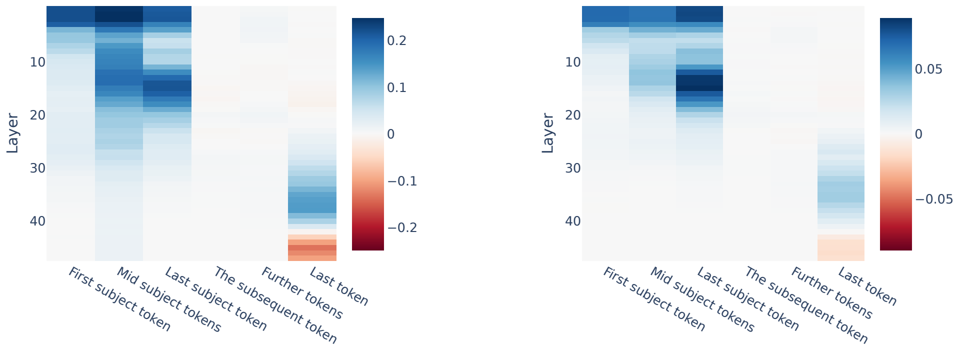
Figure 16: **Activation patching on MLP** across layers and token positions in GPT-2 XL on factual recall prompts. Apply STR corruption and a sliding window of size 3.

I Dataset Samples

Factual data We list a few dataset examples from the PairedFacts dataset used in the factual recall experiments in Figure 31.[6] All the prompts are known true facts.

IOI circuit The detailed templates of constructing the p_{IOI} data distribution can be found in Appendix A of Wang et al. (2023). We perform the same procedure of generating the IOI data by simply reusing their original code.

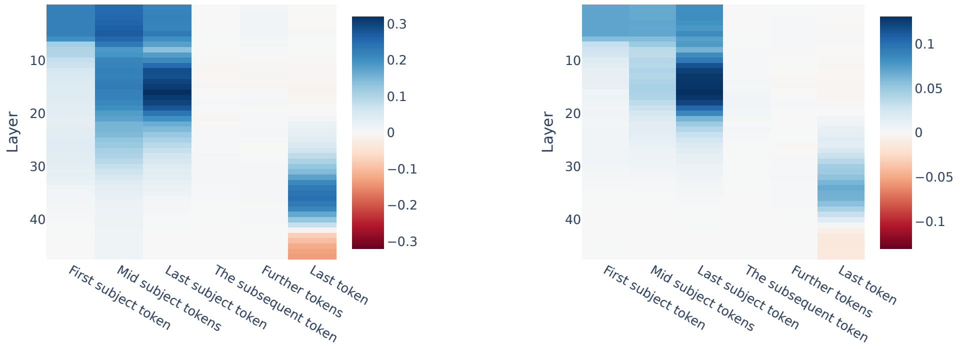
[6] The full dataset is available at https://www.jsonkeeper.com/b/P1GL.



(a) Logit difference (GN)

(b) Probability (GN)

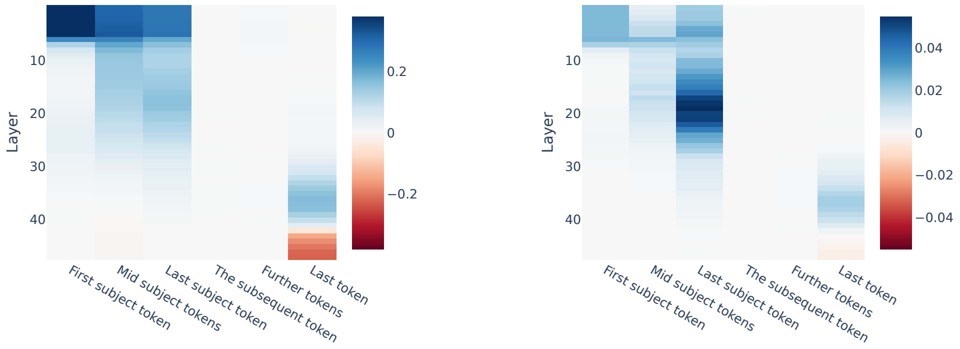
Figure 17: **Activation patching on MLP** across layers and token positions in GPT-2 XL on factual recall prompts. Apply GN corruption and a sliding window of size 5.



(a) Logit difference (GN)

(b) Probability (GN)

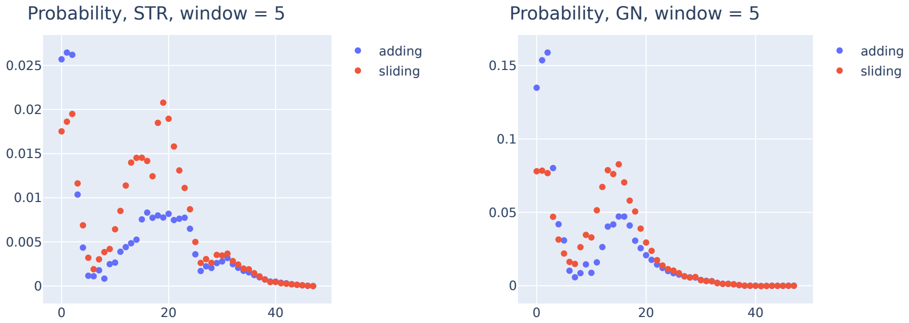
Figure 18: **Activation patching on MLP** across layers and token positions in GPT-2 XL on factual recall prompts. Apply GN corruption and a sliding window of size 10.



(a) Logit difference (STR)

(b) Probability (STR)

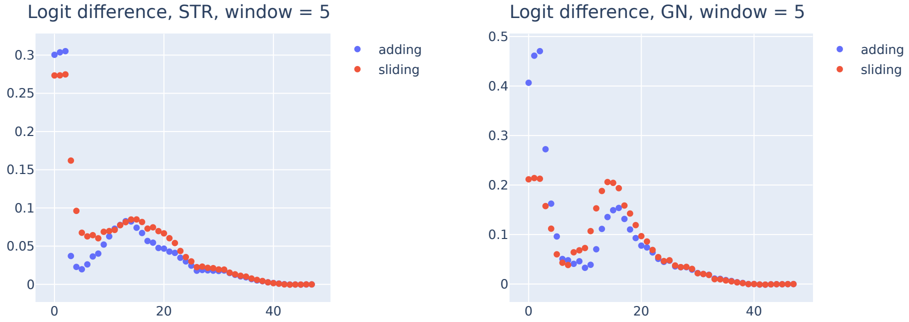
Figure 19: **Activation patching on MLP** across layers and token positions in GPT-2 XL. Apply STR corruption and a sliding window of size 10.



(a) STR corruption

(b) GN corruption

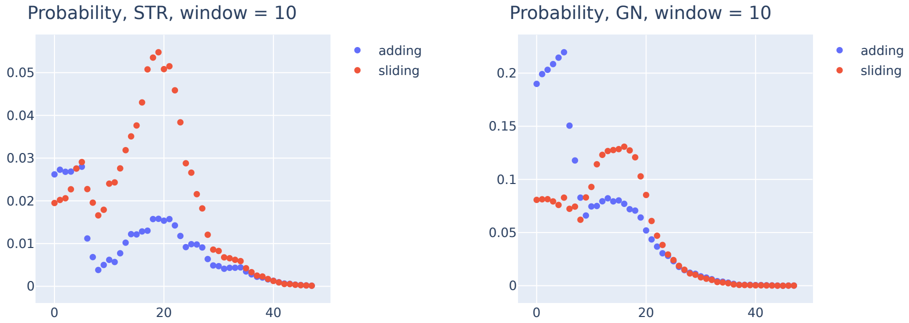
Figure 20: **MLP patching effects, sliding window vs summing up single-layer patching** at last token position in GPT-2 XL on factual recall prompts, with window size of 5. Apply probability as the metric.



(a) STR corruption

(b) GN corruption

Figure 21: **MLP patching effects, sliding window vs summing up single-layer patching** at last token position in GPT-2 XL on factual recall prompts, with window size of 5. Apply logit difference as the metric.



(a) STR corruption

(b) GN corruption

Figure 22: **MLP patching effects, sliding window vs summing up single-layer patching** at last token position in GPT-2 XL on factual recall prompts, with window size of 10. Apply probability as the metric.