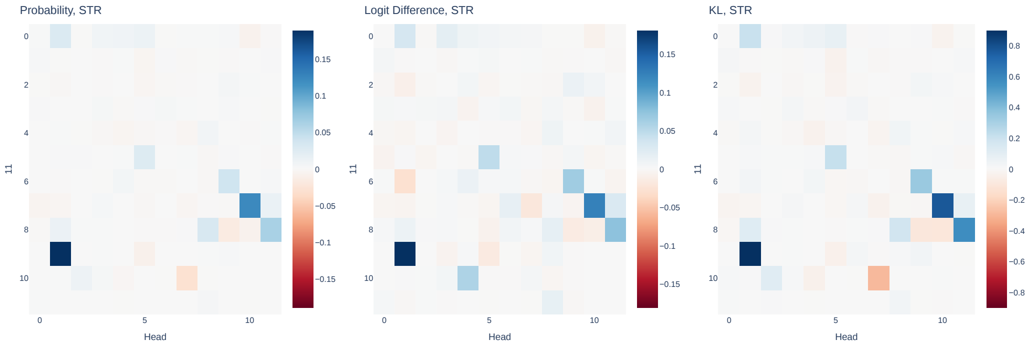


(a) Probability as the metric (b) Logit difference as the metric (c) KL divergence as the metric

Figure 10: **The effects of patching attention heads** in GPT-2 small on the greater-than task, using STR corruption. This gives clearly localized results.

Corruption	Metric	Positive	Negative
STR	Logit difference	9.6, 9.9, 10.0	10.7, 11.10
STR	Probability	9.9	
STR	KL divergence	9.6, 9.9, 10.0	10.7, 11.10
GN	Logit difference	9.6, 9.9, 10.0	10.7, 11.10
GN	Probability	9.6, 9.9, 10.0	
GN	KL divergence	9.6, 9.9, 10.0	10.7, 11.10

Table 4: **Detections from activation patching by corrupting S1 and IO** in IOI. The Name Mover Heads are 9.6, 9.9, 10.0 and the Negative Name Mover Heads are 10.7 and 11.10, based on Wang et al. (2023). No other heads, including the S-Inhibition Heads, are noticed with this approach.

Overall, the experiment suggests that exactly which token is corrupted affects the localization outcomes. Intuitively, varying the corrupted token(s) allows activation patching to trace different information within the model's computation paths; see Section 6 for a discussion.

G Further details on factual association

The plots of subsection Appendix G.1 to G.3 are produced on GPT-2 XL and with the PariedFacts as dataset. Following that, we also experiment with the GPT-2 large (Radford et al., 2019) and GPT-J (Wang & Komatsuzaki, 2021) model in Appendix G.4 and G.5.

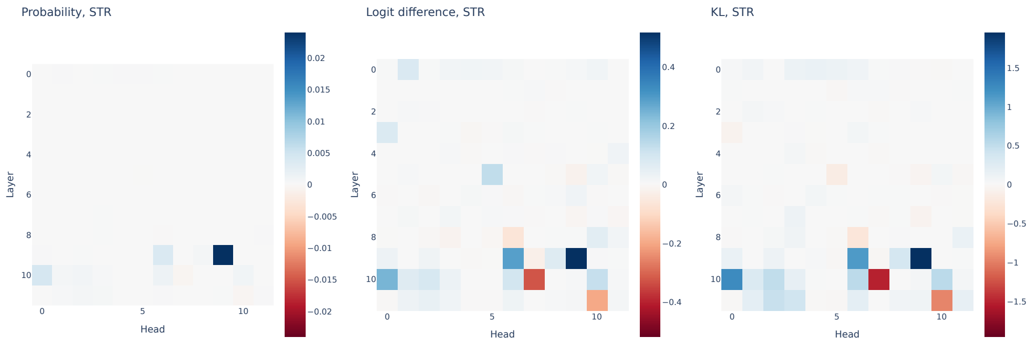
G.1 Plots on MLP patching at the last subject token in GPT-2 XL

First, we perform single-layer patching of MLP activation at the last subject token and examine the effects in Figure 12. We observe that the experiment suggests weak or no peak at middle MLP layers, across metrics and corruption methods.

Also, see Figure 13 and Figure 14 for plots with sliding window size of 3 and 10. Again, activation patching is applied to the MLP activations at the last subject token. We find again that GN yields significantly more pronounced peak.

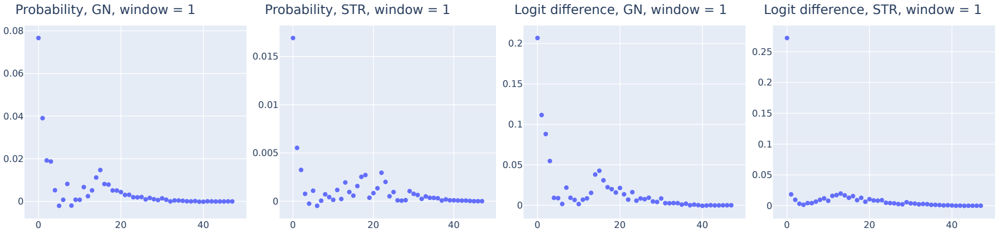
G.2 Plots on MLP patching at all token positions in GPT-2 XL

See Figure 15–Figure 19 to plots on MLP patching at all token positions in GPT-2 XL, across window sizes of 3, 5, 10. We observe that the right-side plots, using probability as the metric, highlights the last subject token as important. In contrast, the left-side figure using logit different does it to lesser degree.



(a) Probability as the metric (b) Logit difference as the metric (c) KL divergence as the metric

Figure 11: **The effects of patching attention heads** in GPT-2 small on IOI sentences, using STR corruption on S1 and IO.



(a) Probability (GN) (b) Probability (STR) (c) Logit difference (GN) (d) Logit difference (STR)

Figure 12: **Patching single MLP layers at the last subject token** in GPT-2 XL on factual recall prompts. None of them suggest a strong peak at the middle MLP layers.

G.3 Plots on sliding window patching in GPT2-XL

We provide further plots from our experiment that compares the sliding window patching with individual patching aggregated via summation over windows. See Figure 20–Figure 23.

G.4 Plots on activation patching of MLP layers on GPT-2 large

We perform activation patching on MLP layers of GPT-2 large in the factual association setting. Following our experiments in Section 3.1, we focus the effects at patching the MLP activation of the last subject token. We validate the high-level finding of Section 3.1, where we observe the disparity of GN and STR applied to MLP activation in the factual prediction setting. In particular, GN gives more pronounced concentration at early-middle MLP layers. We apply sliding window patching of size 3 and 5; see Figure 24 and Figure 25 for the resulting plots.

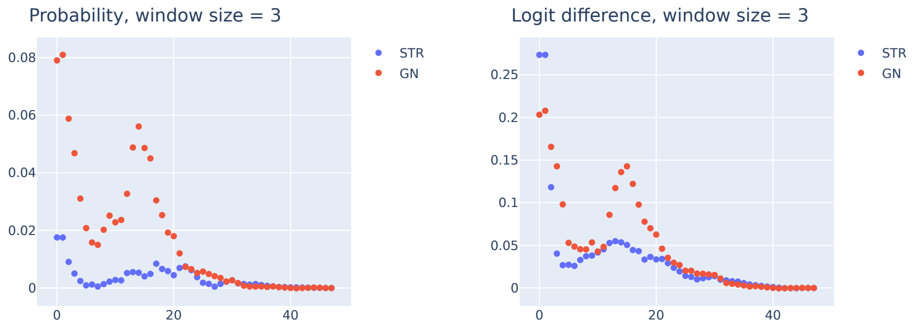
G.5 Plots on activation patching of MLP layers in GPT-J

We perform activation patching on MLP layers of GPT-J (Wang & Komatsuzaki, 2021) in the factual association setting. We patch the MLP activations across all token positions and verify that probability tends to highlight the importance of the last subject token than logit difference. We focus on a sliding window patching of size 5 and the plots are given in Figure 27 (GN) and Figure 26 (STR). This complements our results in Section 4.

H Further details on IOI circuit discovery

H.1 Detailed plots on activation patching

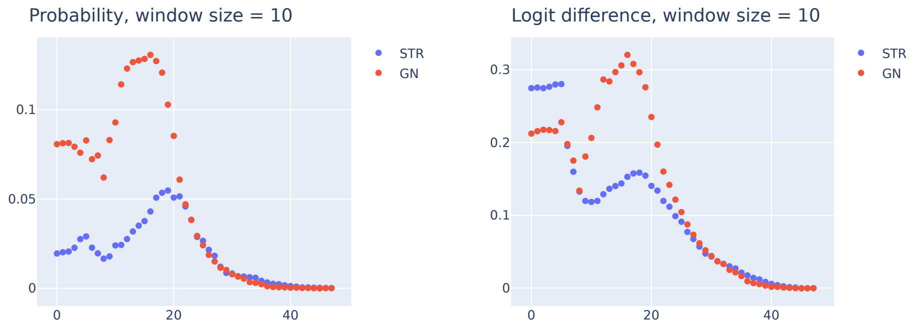
We now provide the detailed plots from the activation patching experiments on the IOI circuit discovery task (Wang et al., 2023); see Figure 28 and Figure 29.



(a) Probability as the metric

(b) Logit difference as the metric

Figure 13: **MLP patching effects at the last subject token position** in GPT-2 XL on factual recall prompts, with window size of 3.



(a) Probability as the metric

(b) Logit difference as the metric

Figure 14: **MLP patching effects for factual recall** at the last subject token position in GPT-2 XL on factual recall prompts, with window size of 10.

H.2 Details on detections

We provide a detailed list of detection from attention heads patching in the IOI circuit setting (Section 3); see Table 5.

Corruption	**Metric**	Negative heads	Positive heads
STR	Logit difference	10.7, 11.10	5.5, 7.9, 8.6, 8.10, 9.9
STR	Probability	10.7	5.5, 7.9, 8.6, 8.10, 9.9
STR	KL divergence	10.7, 11.10	5.5, 7.9, 8.6, 8.10, 9.9
GN	Logit difference	10.7, 11.10	3.0, 5.5, 7.9, 8.6, 8.10, 9.9
GN	Probability	0.10, 10.7, 11.10	3.0, 5.5, 7.9, 8.6
GN	KL divergence	0.10, 10.7, 11.10	5.5, 7.9, 8.6

Table 5: **Detailed results from attention heads patching** in GPT-2 small on IOI sentences. A head is detected if the patching effect is two standard deviation from the mean effect. Negative heads are heads with negative patching effects, suggesting they hurt model performance.

H.3 Detailed plots on fully random corruption

We provide the plots on fully random corruption, termed p_{ABC} in Wang et al. (2023). We perform activation patching on all attention heads, using both probability and logit difference as the metric in order to draw contrasts between them. See Figure 30. In particular, we notice that there is no negative head in the plot. This is natural and totally expected, as we explained in Section 4.