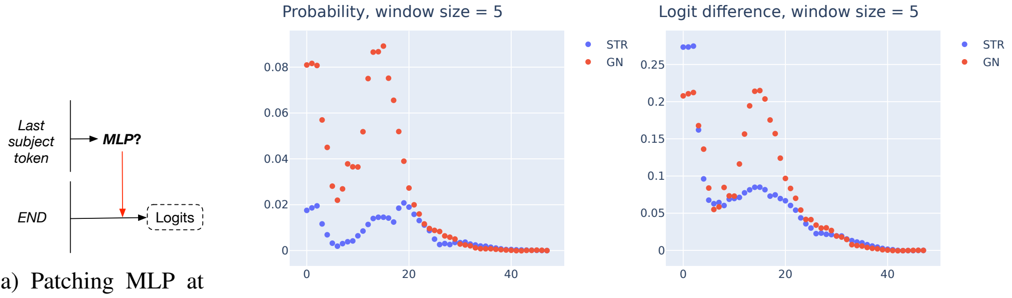


(a) Patching MLP at the last subject token.

(b) Probability as the metric

(c) Logit difference as the metric

Figure 2: **Disparate MLP patching effects for factual recall in GPT-2 XL**. (a) We patch MLP activations at the last subject token. (b)(c) The patching effects using different corruption methods with a window size of 5. STR suggests much a weaker peak, regardless of the evaluation metric.[1]

each layer we restore a set of 5 adjacent MLP layers. (More results on other window sizes can be found in Section G.1. We examine sliding window patching more closely in Section 5.)

For IOI circuit discovery, we follow Wang et al. (2023) and focus on the role of attention heads. Corruption is applied to the S2 token. Then we patch a single attention head's output (at all positions) and iterate over all heads in this way. To avoid relying on visual inspection, we say that a head is *detected* if its patching effect is 2 standard deviation (SD) away from the mean effect.

Dataset and corruption method STR requires pairs of X_{clean} and X_{corrupt} that are semantically similar. To perform STR, we construct PairedFacts of 145 pairs of prompts on factual recall. All the prompts are in-distribution, as they are selected from the original dataset of Meng et al. (2022); see Appendix B for details. GPT-2 XL achieves an average of 49.0% accuracy on this dataset.

For the IOI circuit, we use the p_{IOI} distribution to sample the clean prompts. For STR, we replace S2 by IO to construct X_{corrupt} such that X_{corrupt} is still a valid in-distribution IOI sentence. For GN, we add noise to the S2's token embedding. The experiments are averaged over 500 prompts.

3.1 Results on corruption methods

Difference in MLP localization For patching MLPs in the factual association setting, Meng et al. (2022) show that the effects concentrate at early-middle layers, where they apply GN as the corruption method. Our main finding is that the picture can be largely different by switching the corruption method, regardless of the choice of metric. In Figure 2, we plot the patching effects for both metrics. Notice that the clear peak around layer 16 under GN is not salient at all under STR.

This is a robust phenomenon: across window sizes, we find the peak value of GN to be 2×–5× higher than STR; see Appendix G.1 for further plots on GPT-2 XL in this setting.

These findings illustrate potential discrepancies between the two corruption techniques in drawing interpretability conclusions. We do not, though, claim that results from GN are illusory or overly inflated. In fact, GN does not always yield sharper peaks than STR. For certain basic arithmetic tasks in GPT-J, STR can show stronger concentration in patching MLP activations; see Appendix C.

Difference in circuit discovery We focus on discovering the main classes of attention heads in the IOI circuit, including (Negative) Name Mover (NM), Duplicate Token (DT), S-Inhibition (SI), and Induction Heads. The results are summarized in Table 1 and more details in Appendix H.

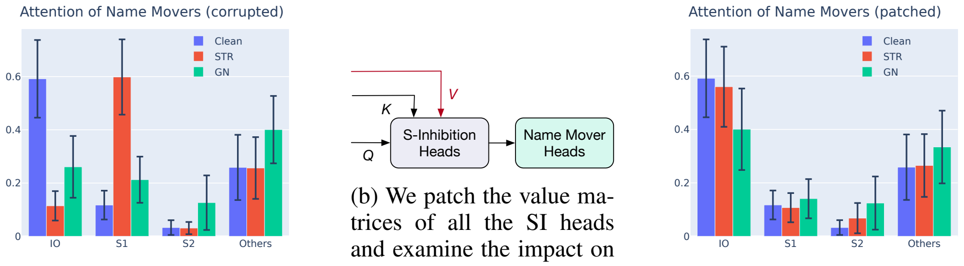
Most importantly, we observe that STR and GN produce inconsistent discovery results. In particular, for any fixed metric, STR and GN detect different sets of heads as important, highlighted in Table 1.

We remark that all the detections are in the IOI circuit as found by Wang et al. (2023). However, the discovery we achieved here appear far from complete, with some critical misses such as NM. This suggests that the extensive manual inspection and the use of path patching, a more surgical patching method, are both necessary to fully discover the IOI circuit.

[1] The effects on the first 3 layers are large simply because MLP0 has significant influence on the model's outputs in GPT-2, regardless of the task (Wang et al., 2023; Hase et al., 2023), so it is not the focus here.

Corruption	**Metric**	NM	DT	SI	Negative NM	Induction
STR	Probability	**1/3**	**0/2**	**3/4**	**1/2**	1/2
GN [†]	Probability	**0/3**	**1/2**	**2/4**	**2/2**	1/2
STR	Logit difference	1/3	**0/2**	3/4	2/2	1/2
GN	Logit difference	1/3	**1/2**	3/4	2/2	1/2
STR	KL divergence	**1/3**	0/2	**3/4**	2/2	1/2
GN [†]	KL divergence	**0/3**	0/2	**2/4**	2/2	1/2

Table 1: **Inconsistency in circuit discovery from activation patching on the IOI task**. We patch the attention heads outputs and list the detections of each class. [†]Also detect 0.10, a fuzzy Duplicate Token Head, as *negatively* influencing model performance. We expect it to be positive (Wang et al., 2023).



(a) Corrupted run

(b) We patch the value matrices of all the SI heads and examine the impact on NMs' attention patterns.

(c) Patched run

Figure 3: **Attention of the Name Movers** from the last token, in corrupted and patched runs.

We also validate our high-level conclusions on the Python docstring (Heimersheim & Janiak, 2023) and the greater-than (Hanna et al., 2023) task. In particular, we find GN can produce highly noisy localization outcomes in these settings; see Appendix D and Appendix E for details.

3.2 Evidence for OOD behavior in Gaussian noise corruption

We suspect that the gaps between the corruption methods can be attributed partly to model's OOD behavior under GN corruption. In particular, the Gaussian noise may break model's internal mechanisms by introducing OOD inputs to the layers. We now give some tentative evidence for this hypothesis. Following the notation of Wang et al. (2023), a head is denoted by "layer.head".

Negative detection of 0.10 under GN Although most localizations we obtain above seem aligned with the findings of Wang et al. (2023), a major anomaly in the GN experiment is the "negative" detection of 0.10. In particular, probability and KL divergence suggest that it contributes negatively to model performance. (Logit difference also assigns a negative effect, though to a lesser degree; see Figure 29b.) This is not observed at all in the experiments with STR corruption.

The detection is in the wrong direction, given the evidence from Wang et al. (2023) that 0.10 *helps* with IOI; on clean prompts, it is active at S2, attends to S1 and signals this duplication. However, by visualizing the attention patterns, we find that this effect largely disappears under GN corruption. We intuit that the Gaussian noise is strongest at influencing early layers, and 0.10's behavior may be broken here, since it directly receives the noised token embeddings from the residual stream.

Attention of Name Movers To exhibit the OOD behavior of the model internals under GN corruptions, we examine the Name Mover (NM) Heads, a class of attention heads that directly affects the model's logits in the IOI circuit (Wang et al., 2023). NMs are active at the last token and copy what they attend to. We plot the attention of NMs in clean and corrupted runs in Figure 3a.

Indeed, on 500 clean IOI prompts, the NMs assign an average of 0.58 attention probability to IO. In the corrupted runs, since STR simply exchanges IO by S1, the attention patterns of NMs are preserved (with the role of IO and S1 switched). On the other hand, with GN corruption, we see that the attention is shared between IO and S1 (0.26 and 0.21). This suggests that GN not only removes the relevant information but also disrupts the internal mechanism of NMs on IOI sentences.

To take a deeper dive, Wang et al. (2023) shows that the output of NMs is determined largely by the values of the S-Inhibition Heads. Indeed, we can fully recover model's logit on IO in STR (logit difference: 1.04) by restoring the values of the S-Inhibition Heads (Figure 3b). The same intervention, however, is fairly unsuccessful under GN (logit difference: 0.49).

Towards explaining this gap, we again examine the attention of NMs. Figure 3c shows that patching nearly restores the NMs' in-distribution attention pattern under STR, but fails under GN corruption. We speculate that GN introduces further corrupted information flowing into the NMs such that restoring the clean activations of S-Inhibition Heads cannot correct their behaviors.

4 Evaluation Metrics

We now study the choice of evaluation metrics in activation patching. We perform two experiments that highlight potential gaps between logit difference and probability. Along the way, we provide a conceptual argument for why probability can overlook negative components in certain settings.

4.1 Localizing factual recall with logit difference

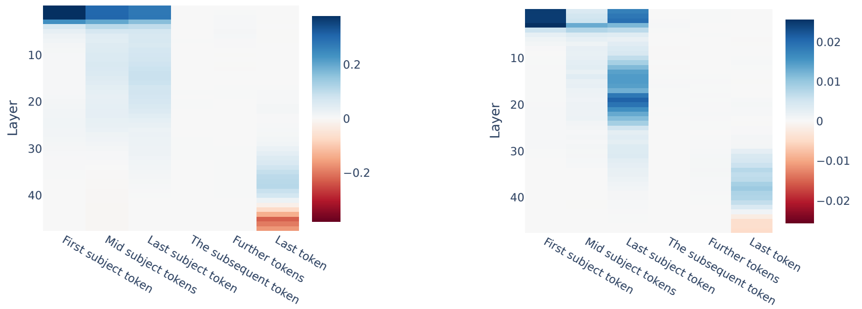
The prior work of Meng et al. (2022) hypothesizes that factual association is processed at the last subject token. Motivated by this claim, we extend our previous experiments to patching the MLP outputs at all token positions and consider the effect of changing evaluation metrics.

Experimental setup We apply the same setting as in Section 3. We extend our MLP patching experiments to all token positions and again use logit difference and probability as the metric.

Experimental results For STR and window size of 5, we plot the patching effects across layers and positions in Figure 4. The visualization shows that probability assigns stronger effects at the last subject token than logit difference. Specifically, we calculate the ratio between the sum of effects (over all layers) on the last subject token and those on the middle subject tokens. In both corruptions, probability assigns more effects to the last subject token than logit difference:

- Using STR corruption, the ratio is 4.33× in probability $>$ 1.22× in logit difference.
- Using GN corruption, the ratio is 1.74× in probability $>$ 0.77× in logit difference.

This observation holds for other window sizes, too, for which we provide details in Appendix G.2. We also validate our findings on GPT-J 6B (Wang & Komatsuzaki, 2021) in Appendix G.5. The results show that the choice of evaluation metrics influences the patching effects across tokens.



(a) Logit difference (STR)

(b) Probability (STR)

Figure 4: **Activation patching on MLP** across layers and token positions in GPT-2 XL, with a sliding window patching of size 5. Note that probability (b) highlights the importance of the last subject token, whereas logit difference (a) displays less effects.

4.2 Circuit discovery with probability

Wang et al. (2023) discovers two Negative Name Mover (NNM) heads, 10.7 and 11.10, that noticeably hurt model performance on IOI. In our previous experiments on STR, both are detected, except when using probability as the metric where 11.10 is overlooked. In fact, the patching effect of 11.10