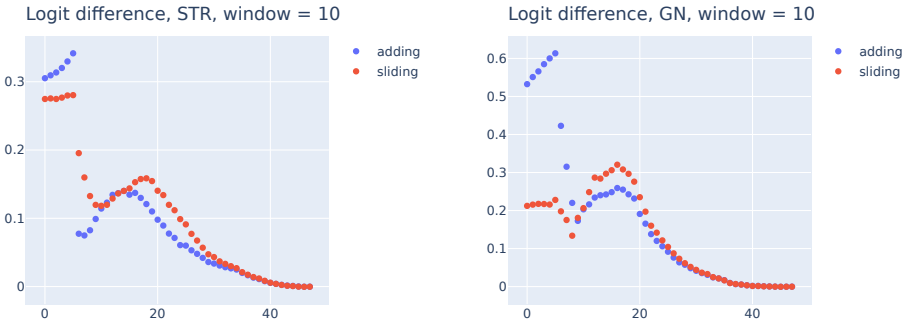
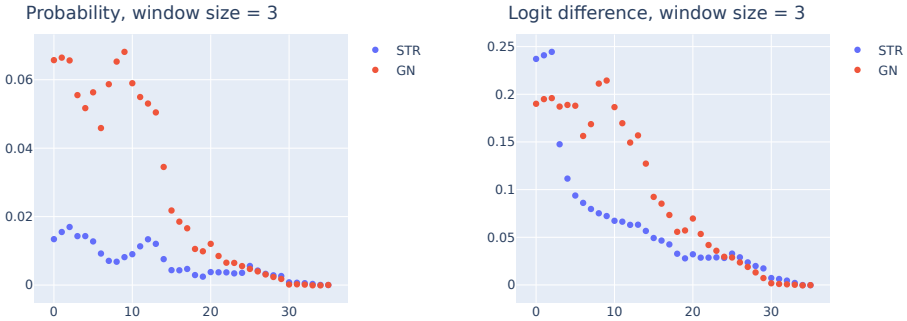


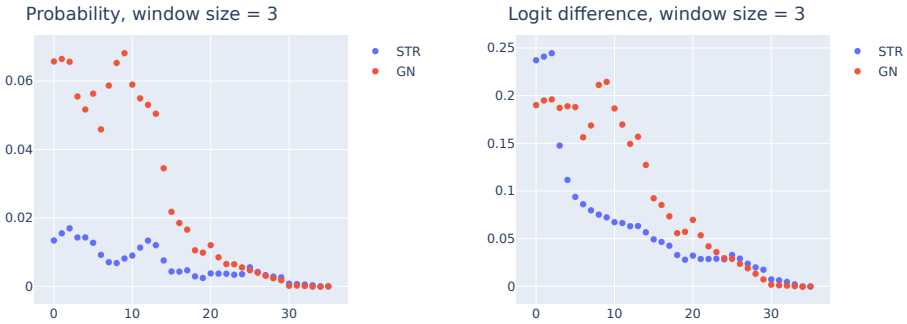
(a) STR corruption (b) GN corruption

Figure 23: **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 logit difference as the metric.



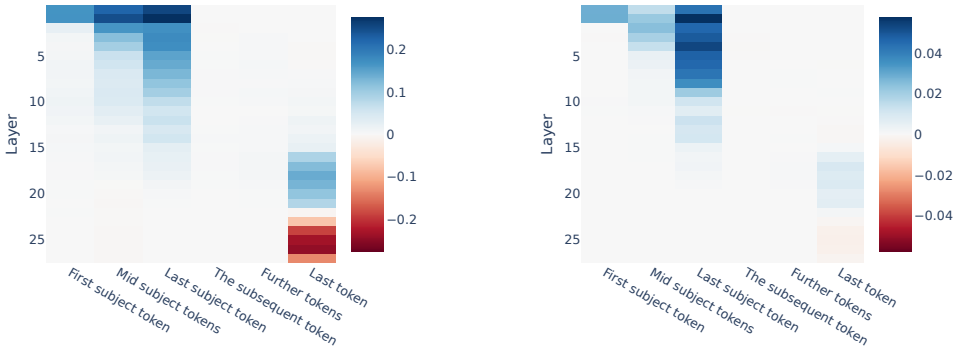
(a) Probability as the metric (b) Logit difference as the metric

Figure 24: **MLP patching effects at the last subject token position** in GPT-2 large on factual recall prompts, with window size of 3.



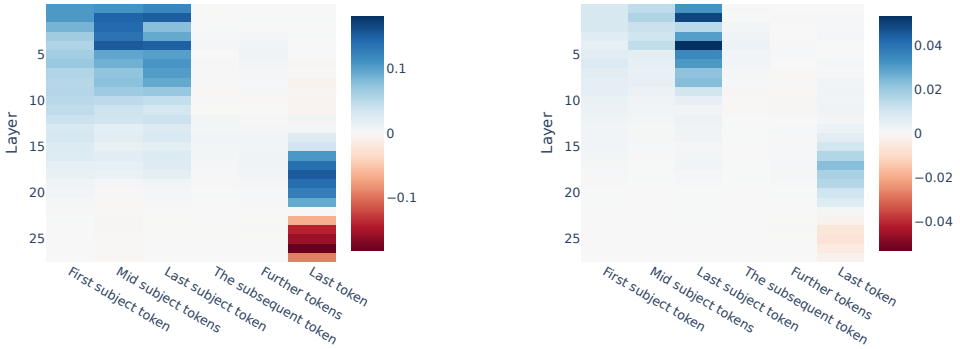
(a) Probability as the metric (b) Logit difference as the metric

Figure 25: **MLP patching effects at the last subject token position** in GPT-2 large on factual recall prompts, with window size of 5.



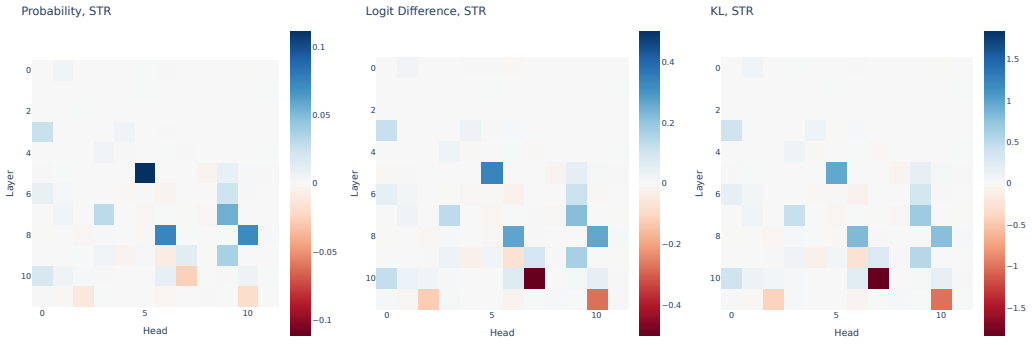
(a) Logit difference (GN) (b) Probability (GN)

Figure 26: **Activation patching on MLP across layers and token positions in GPT-J** on factual recall prompts. Apply STR corruption and a sliding window of size 5.



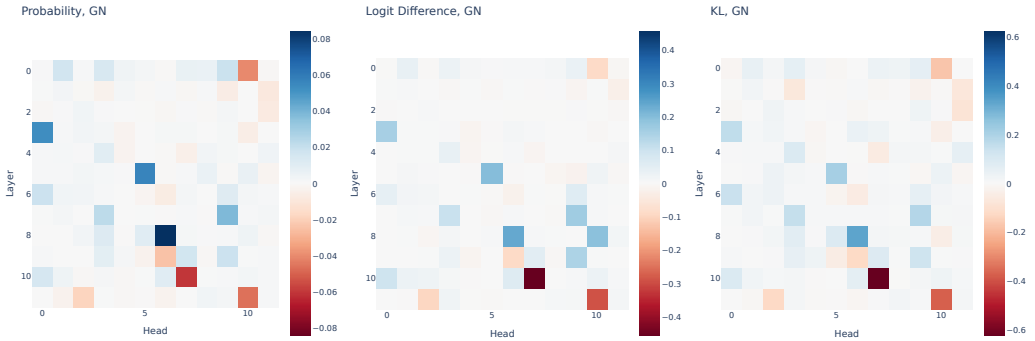
(a) Logit difference (GN) (b) Probability (GN)

Figure 27: **Activation patching on MLP across layers and token positions in GPT-J** on factual recall prompts. Apply GN corruption and a sliding window of size 5.



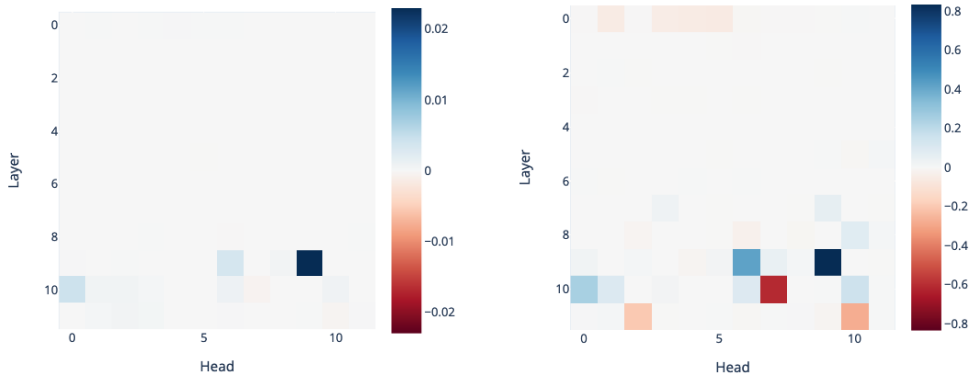
(a) Probability as the metric (b) Logit difference as the metric (c) KL divergence as the metric

Figure 28: **The effects of patching attention heads** in GPT-2 small using STR corruption on IOI sentences.



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

Figure 29: **The effects of patching attention heads** in GPT-2 small using GN corruption on IOI sentences.



(a) Probability as the metric (b) Logit difference as the metric

Figure 30: **The effects of patching attention heads** in GPT-2 small using fully random corruption on IOI sentences, with S1, S2 and IO replaced by three random names (denoted by p_{ABC} in Wang et al. (2023)).


```
{
  "pair": [
    "Honus Wagner professionally plays the sport of",
    "Don Shula professionally plays the sport of"
  ],
  "answer": [
    " baseball",
    " football"
  ],
  "length": 9,
  "category": "athletes"
}
```

```
{
  "pair": [
    "Wii MotionPlus is developed by",
    "Chromebook Pixel is developed by"
  ],
  "answer": [
    " Nintendo",
    " Google"
  ],
  "length": 8,
  "category": "developers"
}
```

```
{
  "pair": [
    "Schreckhorn belongs to the continent of",
    "Afghanistan belongs to the continent of"
  ],
  "answer": [
    " Europe",
    " Asia"
  ],
  "length": 9,
  "category": "continents"
}
```

```
{
  "pair": [
    "The Eiffel Tower is in the city of",
    "Kinkakuji Temple is in the city of"
  ],
  "answer": [
    " Paris",
    " Kyoto"
  ],
  "category": "city_landmarks",
  "length": 11
}
```


Figure 31: **Sample text prompts** from the PAIREDFACTS dataset. The `length` field refers to the sequence length of the prompt under GPT-2 tokenizer.