# Locating Domain-specific Facts in GPT

Yuyang Jiang
Student ID: 12407830

May 24, 2024

## 1 Introduction

My proposal is inspired by the paper I presented, 'Locating and Editing Factual Associations in GPT' [3]. This proposal builds on two main questions: (1) How can we refine the definition of 'factual' associations, particularly by breaking them down into domain-specific facts, and will the storage locations within GPT change based on these distinctions? (2) Typically, fine-tuning is used to develop domain-specific models; how does this approach compare to domain-specific model editing in terms of outcomes?

Previous studies have noted intriguing findings with spatial and temporal facts [2], suggesting that integrating spatial representations can enhance language model capabilities in next-word prediction and geospatial tasks by developing an internal model of space [1, 2]. Additionally, 'time neurons' have been identified that can encode temporal information [2]. There is also evidence that the initial logit distributions in large vision-language models (LVLMs) contain essential hidden knowledge that can improve content generation by identifying problematic questions and mitigating attacks [4]. This finding could potentially guide the storage of knowledge in medical imaging, though its application is challenging.

For this project, I will focus on exploring location of different domain-specific facts in GPT, including Spatial Facts, Temporal Facts, Medical Facts and Mathematical Reasoning.

Key findings are as follows:

- The observation that Medical Fact exhibits a similar pattern to the findings reported in the original study, where the early site of MLP modules exhibits the most pronounced causality, reinforces the consistency of our results.
- For both Spatial and Temporal Facts, we identified strong causal relationships at the early site of both MLP and Attention modules at the initial token. Furthermore, when transitioning to a one-shot prompt, the predominant causality shifts towards the one-shot information, diverging from the causality observed with the original prompts.
- In the domain of Mathematical Reasoning, our analysis did not reveal any dominant causal states for the recall of facts, suggesting the potential failure of ROME in mathematical fact editing.

## 2 Methodology

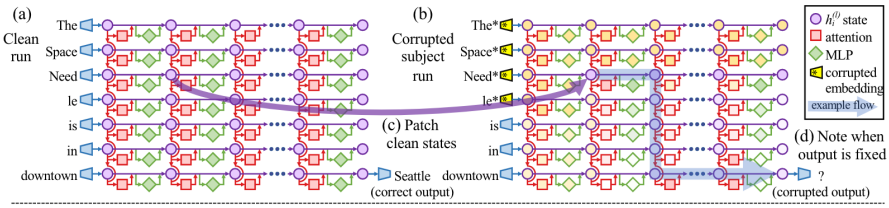


Figure 1: Causal Tracing [3]

In the pursuit of understanding how large pretrained autoregressive transformers pinpoint factual information, our methodology commences by examining specific hidden states within the model that exert a significant causal impact on the predictions of particular facts, following the approach outlined by Meng et al. (2022). As illustrated in Figure 1, numerous pathways span from the input nodes on the left to the output node responsible for the next-word prediction on the right. Our goal is to determine whether certain hidden states are particularly crucial in the recall of facts.

To explore this, we employ a technique known as causal tracing, which involves two simultaneous interventions on the transformer model: firstly, we introduce corruption to a subset of the input by adding random noise to the subject tokens. This corruption impairs the model's ability to accurately complete prompts about the subject. Secondly, we aim to restore specific hidden states by scanning across all layers and tokens to identify those states essential for the model to regain its fact-completing capability. Restoration involves copying clean, non-noised states from a pristine batch instance to corrupted instances within the batch.

This experiment employs a standard setup where the first element of the batch remains uncorrupted, serving as a control, while the remaining elements undergo corruption. These corrupted elements may receive Gaussian noise in their embeddings, or simply different input tokens, depending on the specified strategy in the input batch. To robustly represent corrupted behavior, we generally conduct multiple (ten) corrupted runs simultaneously in the batch.

During the execution, we restore certain hidden states to their original state from the uncorrupted run, as specified in the 'states_to_patch' list, which includes [(token_index, layer_name), ...] pairs. This list can detail a single token and layer pair to observe the impact of restoring one specific state, or it can include multiple pairs to study the effect of restoring several states, thereby tracing their combined influence on the transformer's output.

Furthermore, we apply this causal tracing strategy to specify domain-specific fact storation in GPT-2-xl.

# 3 Experiments

## 3.1 Spatial Facts

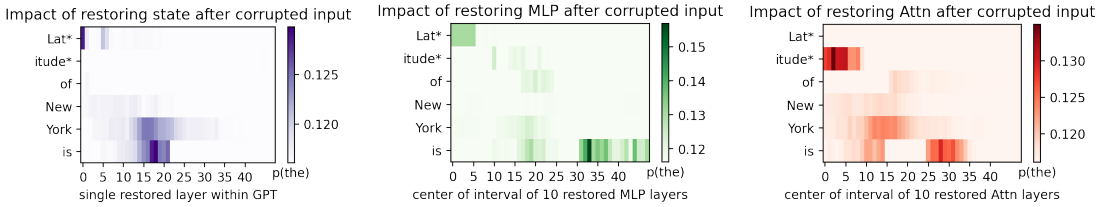


Figure 2: Causal Effect of "Spatial" Neuron Activations on Zero-shot Prompt: `Latitude of New York is ___.`

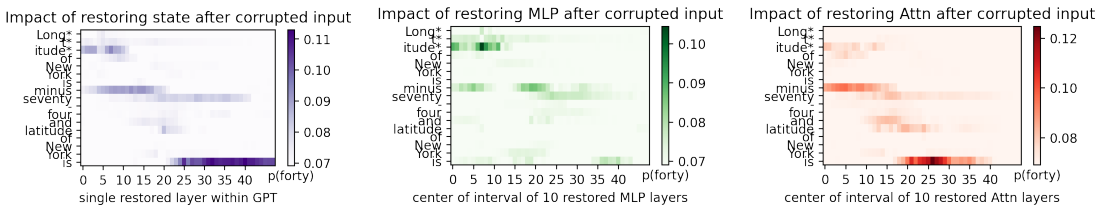


Figure 3: Causal Effect of "Spatial" Neuron Activations on Context Enhancement Prompt: `Longitude of New York is minus seventy-four and latitude of New York is ___."`

**Zero-shot Prompt** We first try zero-shot prompt to implement next-word prediction. However, GPT-2-xl fails to recall facts in its predicted token. Besides last token, strong causality at the token

`Latitude` at an "early site" in the initial layers of both MLP and Attention is a bit different from the orignal paper.

**One-shot Prompt** Considering that gpt-2-xl can hardly return an accurate number with the limit of one-word prediction, we further try providing longitude as supplementary information and format instruction. This time we can find out that provided longitude seems to have stronger causal states, especially at the early site at the token of `Longitude` and `minus seventy-four`. This actually differs from the paper's view that the last subject matters most.

## 3.2 Temporal Facts

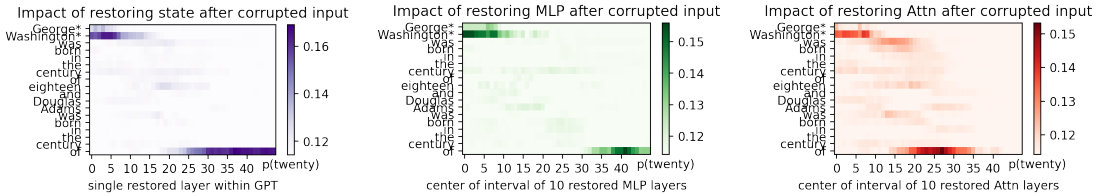


Figure 4: Causal Effect of "Temporal" Neuron Activations on One-shot Prompt: `George Washington was born in the century of eighteen and Douglas Adams was born in the century of ___`.

Considering the limitation of next-word prediction, we simplify the task to return `century` rather than accurate time. Since gpt-2-xl still fails to recall facts with one-word prediction, we further try one-shot prompting as Figure 4. Similar to spatial facts, heatmaps show a strong causality in the "early site" at the first token `George Washington` in both MLP and Attentions. This result is a bit counter-intuitive and doubtful for the robustness of future model editing, considering that it is doubtful whether editing will still work if we reconstruct the one-shot.

## 3.3 Medical Facts

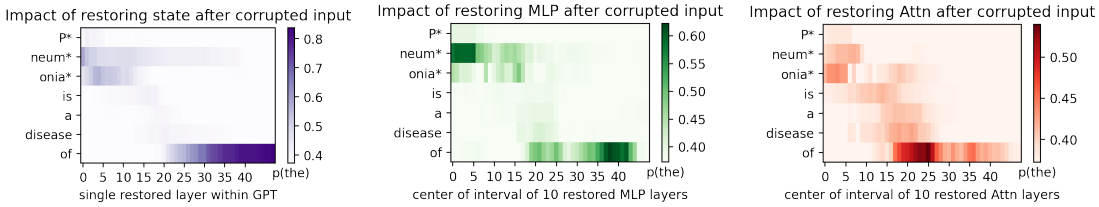


Figure 5: Causal Effect of "Medical" Neuron Activations on Zero-shot Prompt: `Pneumonia is a disease of ___`.

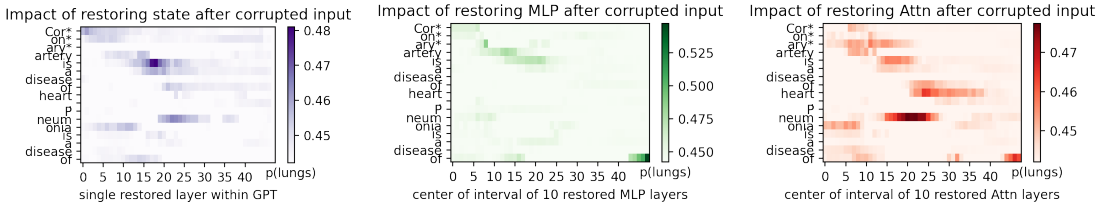


Figure 6: Causal Effect of "Medical" Neuron Activations on One-shot Prompt: `Coronary artery is a disease of heart. Pneumonia is a disease of ___`.

**Zero-shot Prompt** Although gpt-2-xl still fails to recall facts in one-word prediction, its causal states are more similar to the original paper's discovery. At the token `Pneumonia`, MLP shows a dominant causal state at the "early site".

**One-shot Prompt** After changing to one-shot prompt, the dominant contribution of causal states changes to Attentions. `Pneumonia` at the middle site seems to have strongest causality while one-shot information also plays an important role in deciding the next-word prediction.

### 3.4 Mathematical Reasoning

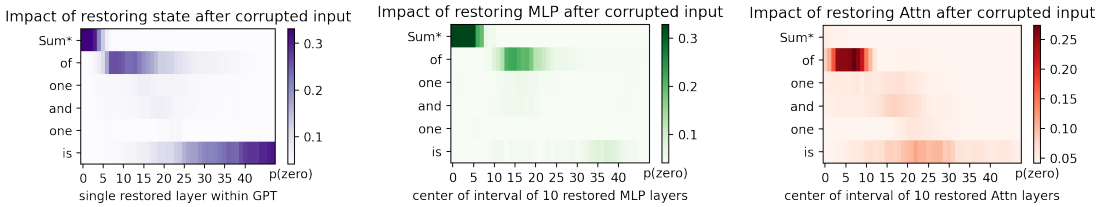


Figure 7: Causal Effect of "Math" Neuron Activations on Zero-shot Prompt: `Sum of one and one is ___.`

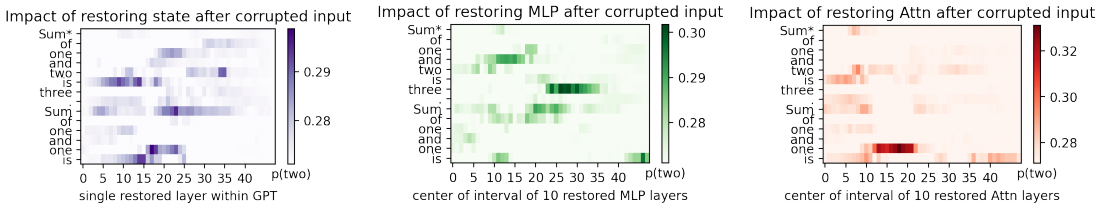


Figure 8: Causal Effect of "Math" Neuron Activations on One-shot Prompt: `Sum of one and two is three. Sum of one and one is ___.`

According to Figure 8, we can see that gpt-2-xl fails to return correct answer for zero-shot mathematical reasoning prompt. This can be explained by the observation that the strongest causal states lie in the token `sum` and `of`, which cannot provide any effective information for the final calculation. When changing to one-shot prompt, similarly, we can find out the important role of one-shot example. However, dominant states become more ambiguous to specify. This is reasonable because mathematical reasoning needs almost all information in the prompt to derive the final answer. This might suggest the difficulty of editing mathematical facts in the model.

## 4 Conclusions

Our study confirms the consistency of earlier findings, particularly noting that Medical Fact exhibits a similar causality pattern where early MLP modules display pronounced effects, as reported in the initial study. Additionally, we observed that both Spatial and Temporal Facts demonstrate strong causal relationships at the initial stages of the MLP and Attention modules. Intriguingly, the causality focus shifts towards one-shot information when prompts transition to a one-shot format, diverging from the original prompt-based causality. In contrast, our analysis within the domain of Mathematical Reasoning reveals no dominant causal states for recalling facts, suggesting the limitations of ROME in mathematical fact editing.

Addressing the challenges in constructing effective prompts for next-word predictions, where models often fail to recall facts, and managing the counter-intuitive storage of information in models, is critical. These findings pose important considerations for future model editing and highlight potential underlying issues in model architecture that could impact performance and outcomes in unexpected ways.

# References


[1] Yida Chen, Yixian Gan, Sijia Li, Li Yao, and Xiaohan Zhao. More than correlation: Do large language models learn causal representations of space? *arXiv preprint arXiv:2312.16257*, 2023.

[2] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[3] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[4] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? *arXiv preprint arXiv:2403.09037*, 2024.