

Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals

Francesco Ortu*

University of Trieste

francesco.ortu@studenti.units.it

Zhijing Jin*

MPI & ETH Zürich

jinzhi@ethz.ch

Diego Doimo

AREA Science Park

diego.doimo@areasciencepark.it

Mrinmaya Sachan

ETH Zürich

msachan@ethz.ch

Alberto Cazzaniga†

AREA Science Park

alberto.cazzaniga@areasciencepark.it

Bernhard Schölkopf†

MPI for Intelligent in Systems

bs@tue.mpg.de

Abstract

Interpretability research aims to bridge the gap between empirical success and our scientific understanding of the inner workings of large language models (LLMs). However, most existing research focuses on analyzing a single mechanism, such as how models copy or recall factual knowledge. In this work, we propose a formulation of *competition of mechanisms*, which focuses on the interplay of multiple mechanisms instead of individual mechanisms and traces how one of them becomes dominant in the final prediction. We uncover how and where mechanisms compete within LLMs using two interpretability methods: logit inspection and attention modification. Our findings show traces of the mechanisms and their competition across various model components and reveal attention positions that effectively control the strength of certain mechanisms.¹

1 Introduction

Recent advancements in large language models (LLMs) have brought unprecedented performance improvements to NLP (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023; Anil et al., 2023, *inter alia*). However, the black-box nature of these models obfuscates our scientific understanding of *how these models achieve certain capabilities*, and *how can we trace the problem when they fail at other tasks*. This has brought an increasing focus on interpretability research to help us understand the inner workings of LLMs.

Existing interpretability research has been largely focused on discovering the *existence* of single mechanisms, such as the copy mechanism in induction heads of LLMs (Elhage et al., 2021; Olsson et al., 2022), and factual knowledge recall in the

* Equal contributions.

† Co-supervision.

¹Code: <https://github.com/francescortu/comp-mech>. Data: <https://huggingface.co/datasets/francescortu/comp-mech>.

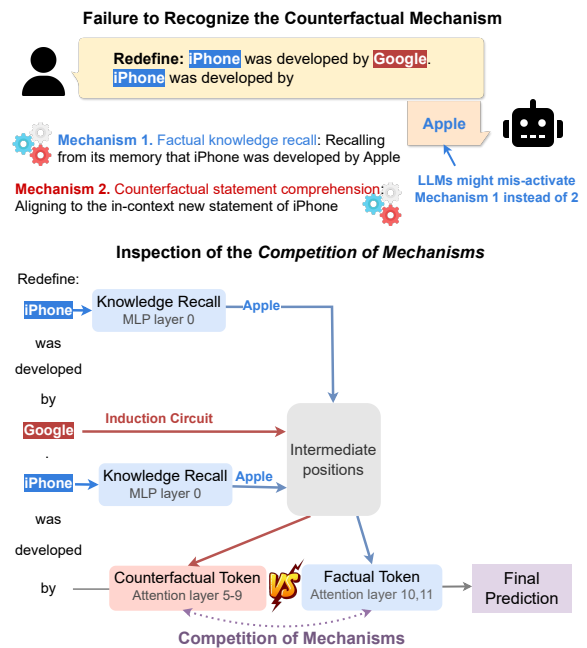


Figure 1: Top: An example showing that LLMs can fail to recognize the correct mechanism when multiple possible mechanisms exist. Bottom: Our mechanistic inspection of where and how the competition of mechanisms takes place within the LLMs.

MLP layers (Geva et al., 2021; Meng et al., 2022; Geva et al., 2023). However, different from discovering *what mechanisms exist in LLMs*, we propose a more fundamental question: *how do different mechanisms interact in the decision-making of LLMs?* We show a motivating example in Figure 1, where the model fails to recognize the correct mechanism when it needs to judge between two possible mechanisms: whether to recall the factual knowledge on who developed the iPhone (i.e., Mechanism 1) or to follow its counterfactual redefinition in the new given context (i.e., Mechanism 2).

We propose a novel formulation of *competition of mechanisms*, which focuses on tracing each mechanism in the model and understanding how one of them becomes dominant in the final prediction

by winning the “competition”. Specifically, we build our work on two single mechanisms that are well-studied separately in literature: (1) the factual knowledge recall mechanism, which can be located in the MLP layers (Geva et al., 2021; Meng et al., 2022; Geva et al., 2023); and (2) the in-context adaptation to a counterfactual statement, which is enabled by the copy mechanism conducted by induction heads of attention layers (Elhage et al., 2021; Olsson et al., 2022). Based on the latest tools to inspect each of these two mechanisms (Nostalgebraist, 2020; Wang et al., 2023; Geva et al., 2023), we then unfold *how and where* the competition of the two mechanisms happen, and how it leads to the overall success or failure of LLMs.

Technically, we deploy two main methods: logit inspection (Nostalgebraist, 2020; Geva et al., 2022) by projecting the outputs of each model component by an unembedding matrix, and attention modification (Geva et al., 2023; Wang et al., 2023). Using these methods, we assess the contributions of various model components, both from a macroscopic view (e.g., each layer) and a microscopic view (e.g., attention heads), and identify critical positions and attention heads involved in the competition of the two mechanisms. Moreover, we locate a few localized positions of some attention head matrices that can significantly control the strength of the factual mechanism. We summarize our main findings as follows:

1. In early layers, the factual attribute is encoded in the subject position, while the counterfactual is in the attribute position (Section 6.1);
2. The attention blocks write most of the factual and counterfactual information to the last position (Section 6.2);
3. All the highly activated heads attend to the attribute position regardless of the specific type of information they promote. The factual information flows by penalizing the counterfactual attribute rather than promoting the factual one (Section 6.3);
4. We find that we can up-weight the value of a few very localized values of the attention head matrix to strengthen factual mechanisms substantially (Section 6.4).

2 Related Work on Interpretability

As deep learning approaches show increasingly impressive performance in NLP, their black-box

nature has hindered the scientific understanding of these models and their effective future improvements. To this end, interpretability research has been a rising research direction to understand the internal workings of these models.

Interpreting the Representations. One major type of work in interpretability has focused on understanding what has been encoded in the representations of deep learning models. This is usually achieved by a *probe*, namely by training a supervised classifier to predict features from its representations (Alain and Bengio, 2016; Conneau et al., 2018; Hupkes et al., 2018; Hewitt and Liang, 2019; Tenney et al., 2019; Jiang et al., 2020; Elazar et al., 2021, *inter alia*), or by geometric methods (Doimo et al., 2020; Valeriani et al., 2024; Park et al., 2024; Cheng et al., 2024). Example features of interest include part of speech (Belinkov et al., 2017), verb tense (Conneau et al., 2018), syntax (Hewitt and Manning, 2019), and factual knowledge (Petroni et al., 2019).

Interpreting the Mechanisms/Functions. Beyond interpreting the representations in the hidden states of the black-box models, another research direction is to interpret the mechanisms or functions that the models have learned, giving rise to the field of mechanistic interpretability (Olah et al., 2020; Elhage et al., 2021; Olsson et al., 2022; Nanda et al., 2023, *inter alia*). Some example mechanisms decoded in recent work include mathematical operations such as modular addition (Nanda et al., 2023) and the greater-than operation (Hanna et al., 2023); natural language-related operations such as the copy mechanism achieved by induction heads in LLMs Olsson et al. (2022) and factual knowledge recall achieved by MLP layers (Geva et al., 2021; Meng et al., 2022; Geva et al., 2023), which we describe below.

The Single Mechanism of Copy: One of the basic actions in LLMs is the copy mechanism, which is found to be operationalized by attending to the copied token in the attention heads and passing it on to the next token prediction (Elhage et al., 2021; Olsson et al., 2022). This foundational mechanism enables further research to decode more complex mechanisms, such as indirect object identification (Wang et al., 2023).

The Single Mechanism of Factual Knowledge Recall: Another major direction is understanding how

LLMs mechanistically recall factual information (Geva et al., 2021; Meng et al., 2022; Geva et al., 2023). For example, Meng et al. (2022) develop the *causal tracing* method to show that the factual information is found in the mid-layer MLP units in GPT-2. A followup work (Geva et al., 2023) shows that MLPs of early layers enrich the subject embeddings with related attributes, and late attention blocks select and write the correct factual information to the sentence’s last position.

Interplay of Multiple Mechanisms: In the final stage of our project in December 2023, we noticed a related study by Yu et al. (2023), which also investigates the role of two different mechanisms in LLMs. Specifically, they inspect a type of prompt whose subjects are the capital cities and whose attributes are the countries, examine the dynamics of the factual recall mechanism and the effect of the in-context counterfactual statement, and find that the subject and attribute frequency in the pre-training set can affect the ability of factual recall. Differently, the methods in our work are applied to a broader set of prompts; moreover, we also establish novel analyses of the underlying mechanistic details of the competition, and precisely localize the path where the information flows at the level of single attention map activations, based on which we discover new findings that are unique to our study.

3 Problem Setup

Following the setup of many existing interpretability studies (Olah et al., 2020; Elhage et al., 2021; Olsson et al., 2022; Nanda et al., 2023, *inter alia*), we look into the next token prediction behavior of autoregressive LLMs in their inference mode, namely

$$P(t_k | t_{<k}), \quad (1)$$

which predicts the k -th token t_k given all the previous tokens in the context.

Next, we design the task to incorporate the competition of mechanisms as in Figure 1. Specifically, for each factual statement $\mathbf{f} := (t_1^f, \dots, t_k^f)$ consisting of k tokens (e.g., “iPhone was developed by Apple.”), we compose a corresponding counterfactual statement $\mathbf{c} := (t_1^c, \dots, t_{k'}^c)$ (e.g., “iPhone was developed by Google.”). Then, we compose a prompt connecting the two statements as “Redefine: \mathbf{c} . $\mathbf{f}_{1:k-1}$.”, such as “Redefine: iPhone was

developed by Google. iPhone was developed by _____”.

The two mechanisms can be traced by inspecting the rise and fall of the factual token t_k^f and the counterfactual token $t_{k'}^c$. For the simplicity of notation, we take the tokens out of the context of their exact position and denote them as t_{fact} and t_{cofa} , respectively, in the rest of the paper.

4 Method and Background

Method 1: Logit Inspection. To inspect the inner workings of the two mechanisms, we trace the *residual stream* (Elhage et al., 2021), or logits of each component in the LLM. Given a text sequence of k tokens, LLMs map it into the residual stream, namely a matrix $\mathbf{x} \in \mathbb{R}^{d \times k}$, where d is the dimension of the internal states of the model. We use the term \mathbf{x}_i^l to specify the residual stream at position i and layer l .

An LLM produces the initial residual stream \mathbf{x}_i^0 by applying an embedding matrix $W_E \in \mathbb{R}^{|V| \times d}$ to each token t_i , where $|V|$ is the size of the vocabulary. Then, it modifies the residual stream by a sequence of L layers, each consisting of an attention block \mathbf{a}^l and MLP \mathbf{m}^l . Finally, after the last layer, it projects the internal state of the residual stream back to the vocabulary space with an unembedding matrix $W_U \in \mathbb{R}^{d \times |V|}$. Formally, the update of the residual stream at the l^{th} layer is:

$$\mathbf{x}^l = \mathbf{x}^{l-1} + \mathbf{a}^l + \mathbf{m}^l, \quad (2)$$

where both the attention and the MLP block take as input the \mathbf{x} after layer normalization norm:

$$\mathbf{a}^l = \mathbf{a}^l(\text{norm}(\mathbf{x}^{l-1})), \quad (3)$$

$$\mathbf{m}^l = \mathbf{m}^l(\text{norm}(\mathbf{x}^{l-1} + \mathbf{a}^l)). \quad (4)$$

To understand which token the residual stream \mathbf{x}^l favors, we follow the common practice in previous work (Halawi et al., 2023; Geva et al., 2023; Dar et al., 2023; Geva et al., 2022) to project it to the vocabulary space using the aforementioned unembedding matrix W_U which maps the latent embeddings to actual tokens in the vocabulary, enabling us to obtain the logits of the factual t_{fact} and counterfactual token t_{cofa} .

Known as the *Logit Lens* (Nostalgebraist, 2020), this method is broadly adopted due to its consistent

success in yielding interpretable results, demonstrating its effectiveness through broad empirical usage. However, it is important to note that it can occasionally fail to reflect the actual importance of vocabulary items, especially in the early layers of the network (Belrose et al., 2023).

Method 2: Attention Modification. Modifying or ablating the activation of a specific model component is also a strategy used to improve the understanding of the information flow within LLMs, including techniques such as causal tracing (Meng et al., 2022) and attention knockout (Wang et al., 2023; Geva et al., 2023).

In our work, we focus on modifying a small number of entries in the attention matrix. Namely, in the attention matrix A^{hl} of the h -th head of the l -th attention layer \mathbf{a}^l , we focus on a certain entry, e.g., at the (i, j) position, where $j < i$, which is the attention value of the token \mathbf{x}_i^l attending to one of its earlier tokens \mathbf{x}_j^l . Following recent work (Yu et al., 2023), the modification is after the softmax layer, so the other attention values of the matrix stay unchanged. For the target entry A_{ij}^{hl} , we scale it up by a multiplier of α :

$$A_{ij}^{hl} \leftarrow \alpha \cdot A_{ij}^{hl}, \quad \text{where } j < i. \quad (5)$$

5 Experimental Setup

Data Creation To compose the factual and counterfactual statements as introduced in Section 3, we adopt COUNTERFACT² (Meng et al., 2022), commonly used dataset to interpret models’ ability of factual knowledge recall. We select 10K data points by considering only examples where the attributes are represented by a single token and where the model completes the sentence in a factually accurate manner.

Each instance of COUNTERFACT expresses a relation r between a subject s and an attribute a : (s, r, a) . For example, in the sentence “iPhone was developed by Apple”, $s = \text{“iPhone”}$, $r = \text{“was developed by”}$, $a = \text{“Apple”}$. Moreover, each (s, r) instance is provided two values of the attribute a , namely a factual token t_{fact} , and a counterfactual token t_{cofa} , representing a false fact.

Using this source data, we compose each instance of our test set in the format of

²<https://rome.baulab.info/data/>

(“Redefine:”, $s, r, t_{\text{cofa}}, s, r, _$), such as “Redefine: iPhone was developed by Google. iPhone was developed by ___”. We preprocess the original dataset by keeping only the data points whose attribute is a single token (for the simplicity of our implementation), and where the model correctly predicts the factual token t_{fact} when completing the sentence $(s, r, _)$. We randomly select 10K test samples into our test set from 219,180 such samples. We open-source our dataset at <https://huggingface.co/datasets/francescortu/comp-mech>.

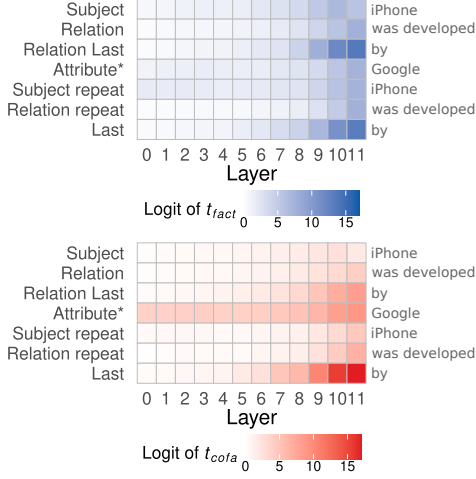
Models For this work, we first choose the GPT-2 small (Radford et al., 2019) model as it is the most commonly used one in previous interpretability studies (e.g., Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023; Hanna et al., 2023). Aligning the same model with those studies can communicate the findings of this work better in the context of existing literature. Then, in addition to GPT-2, we check the generalizability of our work by provide supplemental results of Pythia-6.9B (Biderman et al., 2023) in Appendix A, to show the robustness of our findings across the two LLMs of different architectures and scales. In this way, having similar results across the two very diverse models makes the finding stronger than existing studies, most of which are only on GPT-2.

Implementation Details As for experimental details, GPT-2 small has 117M parameters, consisting of 12 layers with 12 self-attention heads each and a residual stream of 768 dimensions. Pythia-6.9B has 32 layers with 32 self-attention heads each and a model dimension of 4,096, with a 30x increase in the number of parameters. For all our experiments, we deploy the pre-trained models from the Huggingface Hub (Wolf et al., 2020), and inspect the residual streams by the LogitLens tool in the TransformerLens library (Nanda and Bloom, 2022).

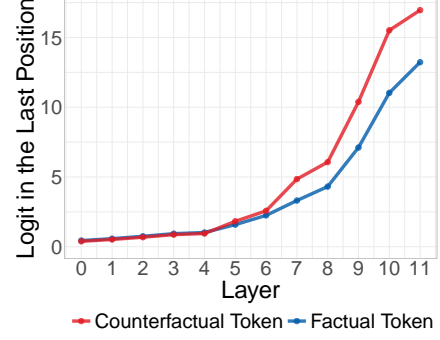
6 Results and Findings

In this section, we trace the competition of the mechanisms within the LLM via the two methods introduced in Section 4, i.e., inspecting the residual stream and intervening on the attention. We provide mechanistic analyses on five research questions in the following subsections:

1. Macroscopic view: Which layers and token positions contribute to the two mechanisms? (Section 6.1)



(a) The logit values for the factual token t_{fact} (blue), and counterfactual token t_{cofa} (red).



(b) The average logits of t_{fact} (blue) and t_{cofa} (red) in the last token position.

Figure 2: Logits of the factual token t_{fact} and counterfactual token t_{cofa} across different positions and layers in GPT-2. The logit of t_{fact} is higher in the subject position in the initial layers and in the last position of the premise and second sentence in the final layers. The logit of t_{cofa} is higher in the attribute position in the first layers and in the last position of the second sentence at the end of the network.

2. Intermediate view: How do we attribute the prediction to attention and MLP blocks? (Section 6.2)
3. Microscopic view: How do individual attention heads contribute to the prediction? (Section 6.3)
4. Intrinsic intervention: Can we edit the model activations to modify the strength of a certain mechanism? (Section 6.4)
5. Behavioral analysis: What word choice varies the strength of the counterfactual mechanism in the given context? (Section 6.5)

6.1 Macroscopic Inspection across Layers and Token Positions

In the main model that we inspect, GPT-2, we find that it can usually identify the counterfactual mechanism in 96% of the 10K test examples. This means that, in the last sequence position, at the output of the network, the counterfactual token, t_{cofa} , gets most of the times a higher probability than t_{fact} . In the following, we will inspect how the “winning” of the counterfactual mechanism happens across the layers of the LLM.

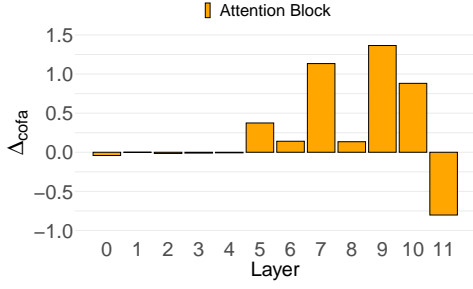
Method. We study how t_{fact} and t_{cofa} are encoded in the residual stream using the logit inspection method described in Section 4. Specifically, for a given token position i and a layer l , we project the embedding \mathbf{x}_i^l , i.e., the residual stream in Eq. (2), to the vocabulary space by

$\tilde{\mathbf{x}}_i^l = W_U \cdot \text{norm}(\mathbf{x}_i^l)$, where W_U is the unembedding matrix and norm is the normalization of the last layer of the model. By varying l , we measure the values of the logits of t_{fact} and t_{cofa} as they evolve in the residual stream after the first attention block.

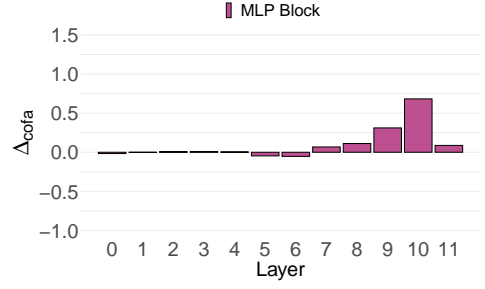
Results. Our results reveal the prevalence of each mechanism by varying the layer l and position i .

Finding 1: Information flows from different tokens for different mechanisms. We analyze the role of previous context at various token positions with respect to different depths of the layers. In Figure 2a, the blue heatmap above shows the logits of the factual token t_{fact} , and the red heatmap below shows those of the counterfactual token t_{cofa} .

Looking at the blue heatmap, we see that the *subject* position is the main contributor to the logits of t_{fact} in early layers, which is consistent with a previous finding (Geva et al., 2023). Specifically, we also locate the factual attribute in the subject positions by the first MLP layer, and find they increase on average the value of t_{fact} from 0.38 to 0.74 in the premise and from 0.9 to 1.93 in the second sentence. Then, in the later layers, the strongest contributor is the last tokens before the attribute, as the last token position is used to predict the attribute. From the red heatmap, we see the evolution of t_{cofa} ’s logits. The observations of later layers are similar across two mechanisms, in that the last



(a) Logit difference Δ_{cofa} of the last token position after the attention block in each layer of GPT-2.



(b) Logit difference Δ_{cofa} of the last token position after the MLP block in each layer of GPT-2.

Figure 3: Contributions of the attention and MLP blocks to the competition of the mechanisms. The attention blocks (left) contribute more to the marginal win of the counterfactual mechanism than the MLP blocks (right).

token contributes the most. However, in early layers, the counterfactual mechanism’s t_{cofa} token is best encoded in the *attribute* position instead of the subject position for the factual mechanism.

Such information flow between different token positions suggests a major role played by the attention mechanism in moving such information to the last position, resonating with observations in Geva et al. (2023).

Finding 2: Both the individual mechanisms and competition take place in late, but not early layers. We trace the competition of the two mechanisms across the layers by plotting in Figure 2b the scale of the logits corresponding to the two mechanisms in the last token position. The first observation is that the strength of each individual mechanism increases monotonically across the layers, from a relatively small logit below 1 in early layers to large values of around 15 in the final layer.

Another observation is that, although both mechanisms increase in strength, stronger signals of the competition (where the counterfactual mechanism prevails the factual one) start after the fifth layer, and this prevalence gradually grows in later layers. The logits of the counterfactual mechanism are, in most of the examples, the highest in the 50K-dimensional vocabulary of GPT-2, making t_{fact} dominant in 96% of the examples.

6.2 Intermediate Inspection of Attention and MLP Blocks

Behind the overall win of the counterfactual mechanism, we want to trace the contributions from the attention and MLP blocks in each layer.

Method. For each attention or MLP block, it processes the input embedding and outputs the logits

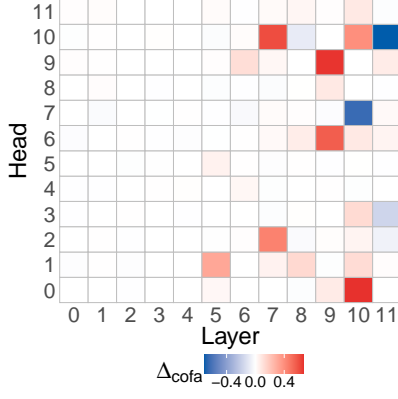
of t_{fact} and t_{cofa} to be added to the residual stream. We can consider the contribution of each block as its added logit values to the residual stream. Intuitively, if the added logit value for t_{cofa} is higher than that of t_{fact} , then this block pushes the overall prediction to lean towards the counterfactual mechanism; otherwise, this block suppresses the counterfactual mechanism.

Hence, we inspect the margin of the added logit of t_{cofa} over that of t_{fact} in each block, represented by $\Delta_{\text{cofa}} := \text{BlockLogit}(t_{\text{cofa}}) - \text{BlockLogit}(t_{\text{fact}})$. To this end, we apply the logit inspection method to analyze the logit distribution at W_{UAN}^l and W_{UMN}^l , where N denotes the last token position in the sequence. The logit contribution of the attention block is the sum over that of all the attention heads. As for the result, a positive value of Δ_{cofa} for a block means that it supports the counterfactual mechanism in the competition, and a negative value indicates suppression.

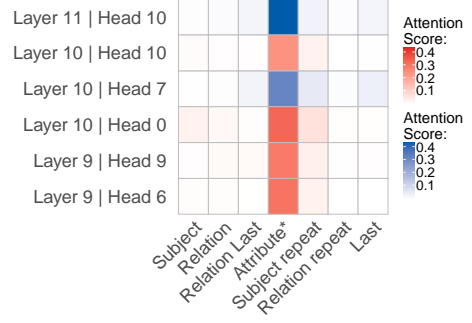
Results. We quantify the contribution of each block in each layer by plotting the Δ_{cofa} values in Figure 3.

Finding 1: The attention blocks play a larger role in the competition of mechanisms than the MLP blocks. Contrasting the Δ_{cofa} margin of the added logits of the attention blocks in Figure 3a and MLP blocks in Figure 3b, we see that the size of Δ_{cofa} is almost always larger in the attention blocks than in MLP blocks. This is consistent with the work of Geva et al. (2023) showing that the attention blocks adds most of the information about t_{cofa} to the residual stream.

Finding 2: Only late but not early layers contribute to the competition of the mechanisms. We find that the early layers have almost no contribution to the



(a) Direct contribution to Δ_{cofa} of all heads in GPT-2. Heads favoring t_{fact} are colored in blue, and those favoring t_{cofa} in red.



(b) Attention scores of relevant heads for the last position. A large attention score in the attribute position is found in all highly activated heads.

Figure 4: Attention pattern for relevant attention heads.

competition of the mechanisms, reflected by the close-to-zero margin Δ_{cofa} in Layer 0-4 for both types of blocks. However, later layers contribute to substantially to the increase of the margin Δ_{cofa} , by a relatively smaller rate for the MLP blocks, and a larger overall rate for the attention blocks, together with a large variance.

Note that we observe a negative Δ_{cofa} around -0.8 in the last attention block, somewhat favoring t_{fact} , which might be since the factual information is moved to the last position in the last layers, as already noted by Geva et al. 2023.

6.3 Microscopic Inspection of Individual Attention Heads

Beyond the overall contributions of the attention block, we further study the contribution of each individual attention head in this section.

Method. We analyze the effect of each individual attention head with the logit inspection method by projecting the outputs of each attention head to the last sequence position N in the vocabulary space. Formally, we consider $\Delta_{\text{cofa}} = \text{HeadLogit}(t_{\text{cofa}}) - \text{HeadLogit}(t_{\text{fact}})$ with the logits from the projection $W_U \mathbf{a}_N^{h,l}$ of each head h . Here \mathbf{a}_N is the output of the attention head h after it has been processed by the output matrix of the Attention Block but before its sum to the residual stream.

Results. We plot the contributions of individual attention heads to Δ_{cofa} in Figure 4, and introduce the main findings as follows.

Finding 1: A few specialized attention heads contribute the most to the competition. As we can see

from the overall contributions of all attention heads across all the layers in Figure 4a, several attention heads (e.g., L9H6, L9H9, L10H0, and L10H10) strongly promote the counterfactual mechanism, i.e., with a positive value of Δ_{cofa} colored in dark red, and two attention heads (L10H7 and L11H10) strongly support the factual mechanism instead, reflected by the large negative Δ_{cofa} in dark blue.

For example, the sum of L7H2 and L7H10 equals 75% of the large positive Δ_{cofa} contribution of Layer 7. The sum of L9H6 and L9H9 explains 65% of the Δ_{cofa} at Layer 9.

On the other hand, the two attention heads, L10H7 and L11H10, explain almost the 70% of the total negative contribution to Δ_{cofa} in the entire network (33% and 37% respectively). This also explains the reason behind the negative Δ_{cofa} in Figure 3a of the previous section. Our study is consistent with McDougall et al. (2023) showing that these two heads are responsible for suppressing the copy mechanisms in GPT-2 small. In our setting, the joint ablation of these two heads decreases the factual recall of GPT-2 small from 4.13% to 0.65%.

Finding 2: All the highly activated heads attend to the same position – the attribute token. Focusing on the heads with large absolute values of Δ_{cofa} , we show the attention scores of the last position N to different tokens in Figure 4b. Expectedly, the major heads supporting the counterfactual mechanism (those in red) attend to the attribute position because they need to copy this token for the prediction, which echoes the findings in Section 6.1.

However, it is surprising to see the other heads sup-

porting the factual mechanism (those in blue) also mainly attend to the counterfactual attribute token. We find that those heads read from the attribute position to give a lower value to the logit of t_{cofa} , which might be an easier operation for it to learn than increasing the logit of the factual token. The evidence is that, in these two heads, the logit of t_{fact} is smaller than the mean of the two layers, but the logit of t_{cofa} (which is -1.13 for L10H7 and -1.05 for L11H10) are the lowest of all the heads in the network.

We include supplementary analyses showing the consistency of Finding 2 on Pythia in Appendix A, and provide the full attention maps with attention scores between every pair of tokens for these heads in Appendix B.2.

6.4 Intrinsic Intervention by Attention Modification

After *understanding* where the two mechanisms take place, we use the insights to *intervene* on the model internal states. Specifically, we perform model editing to alter the factual mechanism, which concentrates on a few strongly activated attention heads (L10H7 and L11H10 in GPT-2, and mostly L17H28, L20H18, and L21H8 in Pythia, see Appendix A.3), and has most of the information flowing from the attribute position (see Figure 4-right and Section 6.2). In the following, we show that enlarging the value of a few well-localized attention values can largely improve the factual recall of the model.

Method. We utilize the attention modification method in Eq. (5) to apply a multiplier of α to the attention weights of the last token to the attribute position in L10H7 and L11H10 for GPT-2, and L17H28, L20H18, and L21H8 in Pythia. To choose the α value, we perform a grid search over [2, 5, 10, 100] to maximize the factual recall rate of the model. We find that $\alpha = 5$ is the best value for both GPT-2 for Pythia.

Results. We highlight the effect of our model editing method on the strength of the factual recall mechanism in Figure 5. Originally, GPT-2 has only 4% of the cases where the factual mechanism prevails the counterfactual one, and Pythia only 30%. However, after modifying the attention weights of the entries mentioned above, the strength of the factual mechanism increases drastically that it wins over the other mechanism in 50% of the cases for

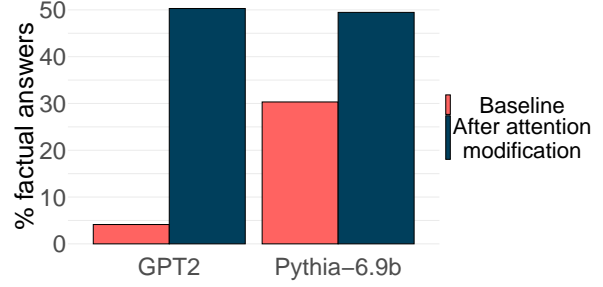


Figure 5: The factual recall mechanism increases substantially across GPT-2 and Pythia after attention modification.

both models. This result is remarkable since we modify only two entries in the attention map out of the 33,264 attention values of GPT-2 (117M parameters) and three entries out of the 270,848 attention values of Pythia (6.9B parameters). This highlights the importance of the interpretability analysis in Sections 6.2 and 6.3, which enables us to find the detailed role played by the individual units of each transformer layer.

6.5 What Word Choices Intensify the Competition?

After the intrinsic intervention to edit the internal states of the model, we explore how the similarity between t_{fact} and t_{cofa} in our dataset affects the mechanism described in the previous sections.

Method. We divided the dataset into 10 equal bins based on the similarity between the vectors for t_{fact} and t_{cofa} , with each bin containing 1000 items. Starting from the lowest, each group represents a 10% segment of the dataset, arranged by increasing similarity scores. For our word similarity metric, we calculate the cosine similarity of the 300-dimensional word embeddings from the pre-trained Word2Vec model (Mikolov et al., 2013) implemented in the Gensim Python package (Řehůřek and Sojka, 2010).

Results. As a result of the varying similarity of the two tokens, we see a drastic change in the dominance of the factual mechanism in Figure 6.

Finding 1: Similar tokens confuse the model more easily. Consistently across all the models, the more similar the two tokens are, the more likely the model is to be confused and mistakenly let the factual mechanism dominate, to predict t_{fact} as its output.

Finding 2: Larger models suffer more from such

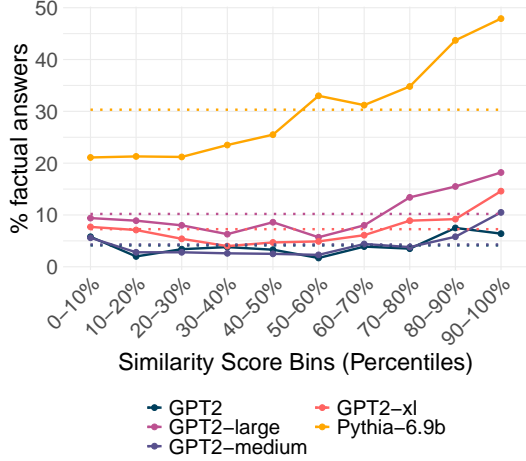


Figure 6: Prediction frequency of factual token by similarity level. We show the percentage of t_{fact} predictions within each bin compared to the entire dataset (represented by a dotted line) across various model sizes. We can notice that more similar t_{fact} and t_{cofa} are, and the factual mechanism is stronger.

confusion. For example, the largest one, Pythia-6.9B, demonstrates a very strong attachment to the factual information, letting the factual mechanism win almost 45% of the cases when the token similarity reaches 90%. Even when the similarity is low, larger models are still more likely to confuse and lean towards the factual mechanism. This finding resonates with the observations from the inverse scaling prize (McKenzie et al., 2023) that larger models have a greater capacity to store and retrieve factual information, thus more influenced by the factual mechanism.

7 Discussion and Future Work

Situating Our Findings in Related Work. Our findings about the late attention blocks are consistent with Geva et al. (2023), showing that late attention blocks write most of the information to the last layer when adding a counterfactual premise. Surprisingly, however, we find that the largest contribution to the factual prediction of the network mostly comes from the suppression of the counterfactual token read from the attribute position rather than the promotion of the factual token from the subject position.

Consistently with McDougall et al. (2023), we find that few highly specialized heads suppress the counterfactual information. Moreover, we make a unique contribution up-weighting only two or three attention entries of these heads to increase

substantially the number of factual responses of the model.

With an approach similar to ours, Yu et al. (2023) find that more heads can promote the factual mechanism, also in early layers, but found it challenging to improve the factual responses by scaling up the weights of the attention maps. This discrepancy can be due to the broader set of topics we include in our prompts, which allowed us to select fewer, more specialized heads, to the different ways the prompts are framed, or also to our more focused modification of the attention maps.

Future Work For future research directions, we aim to analyze more in depth how our findings depend on the prompt structure and whether the promotion of factual responses by suppressing the counterfactuals generalizes to larger models and a more comprehensive variety of datasets.

8 Conclusion

In this work, we have proposed the formulation of the *competition of mechanisms* as a powerful interpretation when LLMs need to handle multiple mechanisms, only one of which leads to the correct answer. We deployed two mechanistic interpretability tools, logit inspection and attention modification, and identified critical positions and model components involved in competing for the mechanisms. Finally, we discovered a few localized positions in the attention map, which largely control the strength of the factual mechanism. Our study sheds light on future work on interpretability research for LLMs.

Limitations

Limited models: Our study aligns with most existing work in mechanistic interpretability to use GPT-2 small. However, we understand that this is a small model with far fewer parameters than current state-of-the-art LLMs. Future work is welcome to extend to larger-sized models, which might generalize our conclusion to a certain extent, and also reveal interesting behavior once the models get beyond a specific size, maybe also seeing a U-shaped curve (Wei et al., 2023) for the dominance of the counterfactual mechanism.

Interpretability method: Furthermore, our experiments and insights are heavily grounded in the interpretability within the embedding space of the

model’s inner components. This approach is reliable and extensively employed in mechanistic interpretability research (Dar et al., 2023; Geva et al., 2022; Halawi et al., 2023). The logit inspection method, although commonly employed in previous work, can occasionally fail to reflect the actual importance of some vocabulary items, especially in the early layers of the network (Belrose et al., 2023).

Simplicity of the prompts: Our prompts have a relatively simple structure for the controllability of the counterfactual information tracing, as it is very challenging to follow the information flow in a more diversified set of prompts. We welcome future work to explore methodological advances to enable analyses over more diverse prompts.

Ethical Considerations

The aim of our study is to enhance comprehension of the interplay among mechanisms within language models that may yield unforeseen and undesirable outcomes. Additionally, our research serves as a conceptual demonstration of methods to guide model behavior under such conditions. We believe that recognizing and dissecting the mechanisms by which LLMs produce unpredictable responses is crucial for mitigating biases and unwanted results. Moreover, understanding the competitive dynamics under investigation is critical for improving the safety of LLMs. Specifically, inputting a prompt with an inaccurate redefinition may lead the model to inadvertently reveal sensitive factual information.

Acknowledgment

We thank Alessandro Stolfo and Yifan Hou for their insightful suggestions, including the pointer to MLP layers for knowledge recall, and many relevant studies. We also thank the audience at the BlackBoxNLP Workshop at EMNLP 2023 for discussions and suggestions on various aspects of this project as well as ethical implications.

This material is based in part upon works supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by the John Templeton Foundation (grant #61156); by a Responsible AI grant by the Haslerstiftung; and an ETH Grant (ETH-19

21-1). Alberto Cazzaniga and Diego Doimo are supported by the project “Supporto alla diagnosi di malattie rare tramite l’intelligenza artificiale” CUP: F53C22001770002. Alberto Cazzaniga received funding by the European Union – NextGenerationEU within the project PNRR ”PRP@CERIC” IR0000028 - Mission 4 Component 2 Investment 3.1 Action 3.1.1. Zhijing Jin is supported by PhD fellowships from the Future of Life Institute and Open Philanthropy.

Author Contributions

The paper originates as the Master’s thesis work of Francesco Ortu hosted jointly at the Max Planck Institute of Intelligence Systems, Tuebingen, Germany, and Area Science Park, Trieste, Italy. Zhijing Jin closely supervised the development of the technical idea and the design of the experiments. In the meantime, Francesco developed the technical skills in mechanistic interpretability and conducted the experiments with lots of resilience.

Professors Alberto Cazzaniga and Bernhard Schölkopf co-supervised this work, and gave insightful research guidance. Diego Doimo closely monitored the execution of the experiments and helped substantially with the design of the word choice experiment and the improvement of the factual recall experiments. Professor Mrinmaya Sachan provided helpful research suggestions throughout the project. All of Francesco, Zhijing, Diego, and Alberto contributed significantly to the writing of this paper.

References

- Guillaume Alain and Yoshua Bengio. 2016. [Understanding intermediate layers using linear classifier probes](#). *ArXiv*, abs/1610.01644. 2
- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. [Gemini:](#)

A family of highly capable multimodal models. *CoRR*, abs/2312.11805. 1

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. [What do neural machine translation models learn about morphology?](#) In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics. 2

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. [Eliciting latent predictions from transformers with the tuned lens](#). *CoRR*, abs/2303.08112. 4, 10

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR. 4

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 1

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. 2024. Emergence of a high-dimensional abstraction phase in language transformers. *arXiv preprint arXiv:2405.15471*. 2

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. [Towards automated circuit discovery for mechanistic interpretability](#). *CoRR*, abs/2304.14997. 4

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single \\$&!#* vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics. 2

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. [Analyzing transformers in embedding space](#). In *Proceedings of the 61st Annual Meeting of the Associ-*

ation for Computational Linguistics (Volume 1: Long Papers), *ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16124–16170. Association for Computational Linguistics. 3, 10

Diego Doimo, Aldo Glielmo, Alessio Ansuini, and Alessandro Laio. 2020. Hierarchical nucleation in deep neural networks. *Advances in Neural Information Processing Systems*, 33:7526–7536. 2

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Trans. Assoc. Comput. Linguistics*, 9:1012–1031. 2

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2021/framework/index.html>. 1, 2, 3

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. [Dissecting recall of factual associations in auto-regressive language models](#). *CoRR*, abs/2304.14767. 1, 2, 3, 4, 5, 6, 7, 9

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. [Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 2, 3, 10

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 1, 2, 3

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2023. [Overthinking the truth: Understanding how language models process false demonstrations](#). *CoRR*, abs/2307.09476. 3, 10

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. [How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model](#). *CoRR*, abs/2305.00586. 2, 4

John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics. 2

- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics. 2
- Dieuwke Hupkes, Sara Veldhoen, and Willem H. Zuidema. 2018. [Visualisation and ‘diagnostic classifiers’ reveal how recurrent and recursive neural networks process hierarchical structure](#). *J. Artif. Intell. Res.*, 61:907–926. 2
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) *Transactions of the Association for Computational Linguistics*, 8:423–438. 2
- Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2023. [Copy suppression: Comprehensively understanding an attention head](#). *CoRR*, abs/2310.04625. 7, 9
- Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgajt, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2023. [Inverse scaling: When bigger isn’t better](#). *CoRR*, abs/2306.09479. 9
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in gpt](#). *arXiv preprint arXiv:2202.05262*. 1, 2, 3, 4
- Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119. 8
- Neel Nanda and Joseph Bloom. 2022. Transformerlens. <https://github.com/neelnanda-io/TransformerLens>. 4
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. [Progress measures for grokking via mechanistic interpretability](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. 2, 3
- Nostalgebraist. 2020. [interpreting gpt: the logit lens](#). Accessed: Nov 2023. 2, 3
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. [Zoom in: An introduction to circuits](#). *Distill*, 5(3):e00024–001. 2, 3
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. [In-context learning and induction heads](#). *CoRR*, abs/2209.11895. 1, 2, 3
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774. 1
- Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2024. [The geometry of categorical and hierarchical concepts in large language models](#). 2
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics. 2
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8). 4
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. <http://is.muni.cz/publication/884893/en>. 8
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. [What do you learn from context? probing for sentence structure in contextualized word representations](#). *ArXiv*, abs/1905.06316. 2
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *CoRR*, abs/2302.13971. 1
- Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2024. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36. 2
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#). In *The Eleventh International*

Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net. 2, 4

Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. 2023. [Inverse scaling can become u-shaped](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15580–15591. Association for Computational Linguistics. 9

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. 4

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. [Characterizing mechanisms for factual recall in language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics. 3, 4, 9

A Experiments for Pythia-6.9b

This section extends the experimental analysis conducted on GPT-2 to Pythia-6.9b. The goal is to replicate the prior methodology and compare the outcomes across the two different models, thus contributing to a broader understanding of model behaviors under similar conditions.

A.1 Macroscopic Inspection across Layers and Token Positions

Figure 7 provides a comparative analysis of the logit values for two specific tokens, labeled as factual and counterfactual, across various positions and layers in Pythia-6.9b.

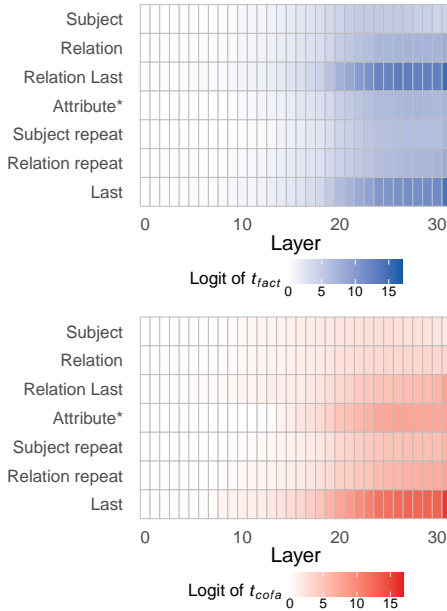


Figure 7: **Layer-wise Position Analysis of Relevant Tokens in GPT-2-small.** The figure presents the logit values for two pertinent tokens across various positions and layers. The left panel illustrates the logit values for the factual token t_{fact} , while the right panel illustrates the logit values for the counterfactual token t_{cofa} .

A.2 Intermediate Inspection of Attention and MLP Blocks

This subsection exposes the contributions of Attention and MLP Blocks to the differences in logit values across layers within Pythia-6.9b. Figure 8 explores how these components influence the computation of logits for two tokens, represented as the difference $\Delta_{\text{cofa}} = \text{Logit}(t_{\text{cofa}}) - \text{Logit}(t_{\text{fact}})$ at the final position of the input. The analysis specifically highlights the distinct effects of these blocks at different stages of the model’s operation.

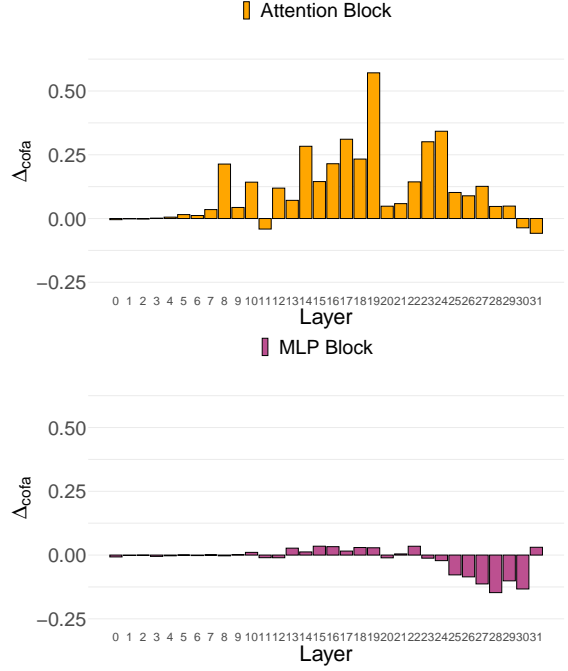


Figure 8: **Attribution of Logit Differences to Attention and MLP Blocks.** delineates the differential impact of Attention and MLP Blocks on logit values at the terminal input position. The attention mechanism is shown to predominantly influence early layer processing in the left panel, while the right panel details the increased contribution of MLP Blocks to the factual token’s logits in the concluding layers, illustrating the dynamic interplay between these fundamental neural network elements.

A.3 Microscopic Inspection of Individual Attention Heads

Figure 9 quantifies the direct contributions of all attention heads to the difference in logit values, labeled as Δ_{cofa} . It specifically identifies heads that preferentially enhance the logits for t_{fact} (shown in blue) versus those favoring t_{cofa} (depicted in red), offering insights into how attention mechanisms differentially prioritize token attributes.

Figure 10 presents the attention patterns of the relevant attention heads at the last token position. It shows the consistent pattern of the relevant heads, with a consistent focus on the attribute position.

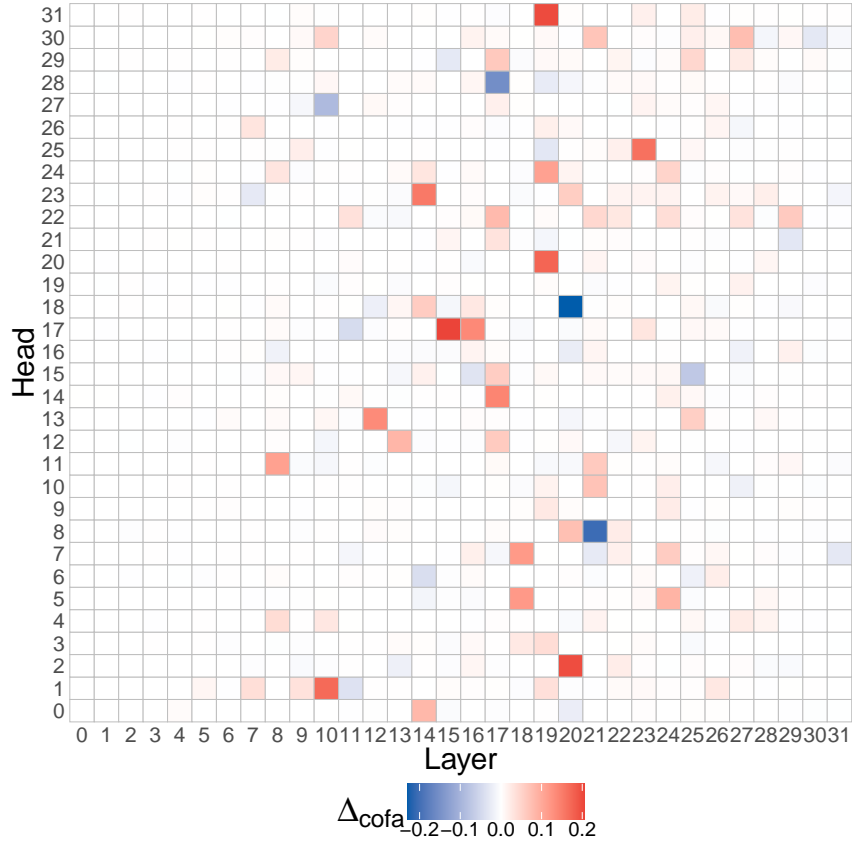


Figure 9: **Direct Contribution of Attention Heads.** The figure displays the direct contribution of all heads in Pythia-6.9b to the logit difference Δ_{cofa} with heads favoring t_{fact} highlighted in blue and those favoring t_{cofa} in red.

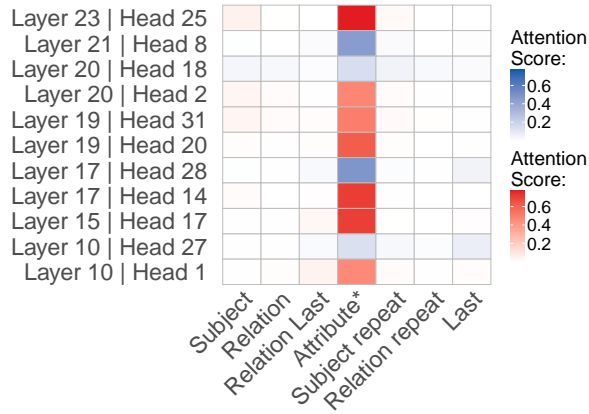


Figure 10: **Attention Pattern for Relevant Attention Heads.** The panel illustrates the attention patterns of relevant heads for the last position, demonstrating consistent attention to the attribute position by both red and blue heads.

B Other Experiment for GPT-2

B.1 Ranks Analysis in the Last Position

We provide additional information in Figure 11 mapping the logits to ranks of the tokens, and find that the rank of t_{cofa} in the projected logit distribu-

tion remains very low: t_{cofa} is among the 20 most likely tokens in the first five layers and between the 20th and the 70th in the last part of the network.

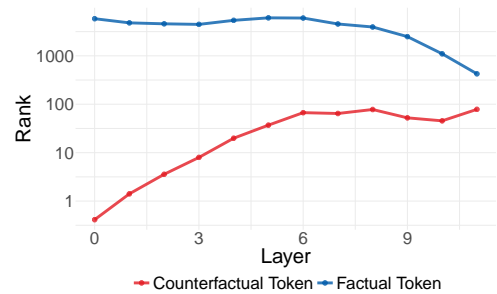


Figure 11: **Rank of Target Tokens for Attribute Position Across Layers in GPT-2.** This figure depicts the trend where the logit rank for the factual token t_{fact} decreases while the rank for the counterfactual token t_{cofa} increases at the attribute position. In the concluding layers, this pattern is evident as t_{fact} typically secures a lower rank, in contrast to t_{cofa} , which shows an upward trajectory in rank. However, it is important to note that t_{cofa} 's rank consistently remains lower than that of t_{fact} .

B.2 Attention Pattern of Relevant Attention Heads

Figure 12 shows the full attention pattern for the relevant attention heads, as identified in Section 6. It is show as the attention pattern is similar between all the relevant attention heads, independently if the heads is favoring t_{fact} or t_{cofa} .

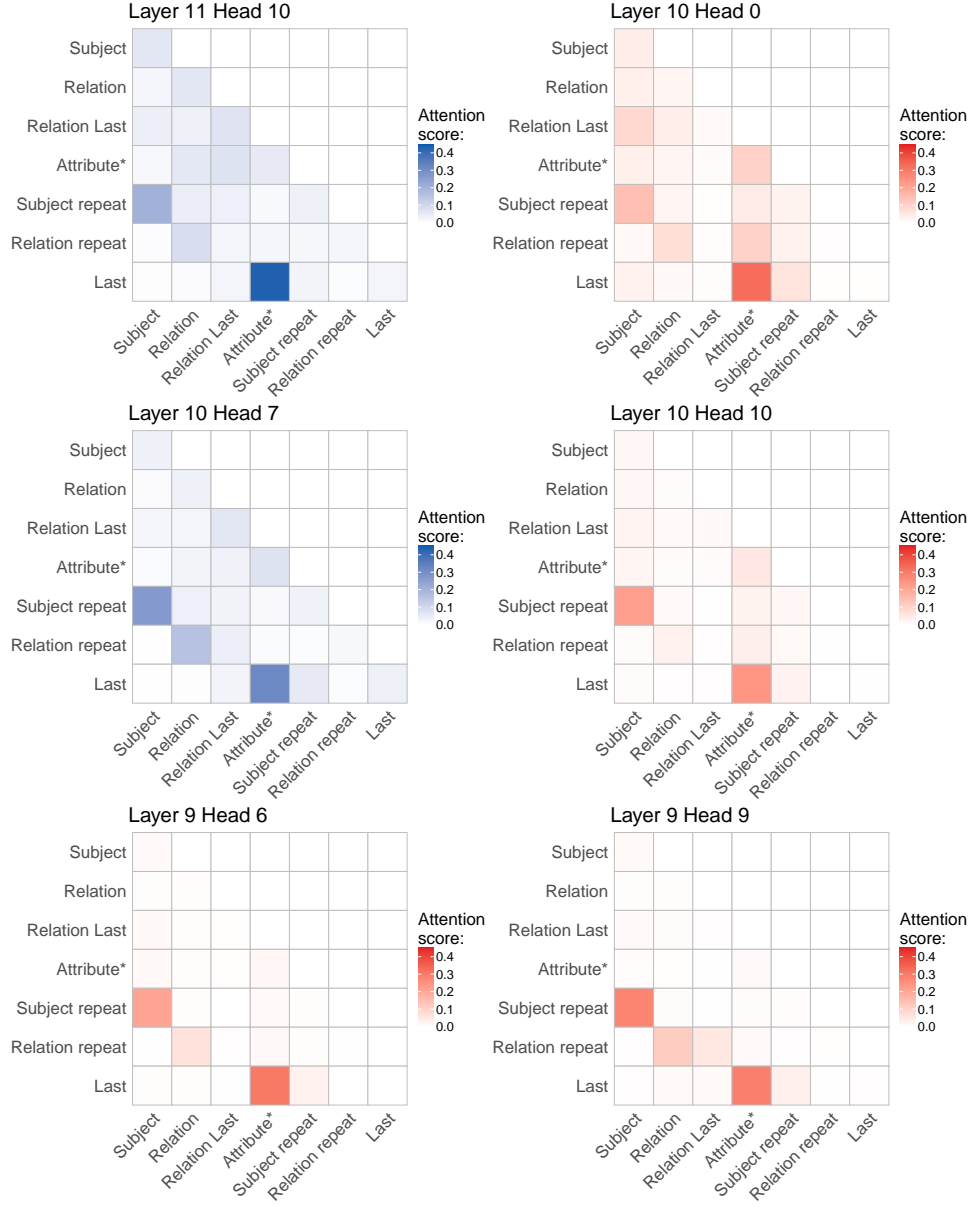


Figure 12: **Attention Pattern of Significant Heads.** This figure illustrates the comprehensive attention pattern of heads substantially influencing $\Delta_{t_{cofa}}$. Notably, a similar pattern emerges for both heads favoring t_{cofa} (depicted in red) and those favoring t_{fact} (illustrated in blue), particularly in the attention edge between the attribute and the final position.