Tracing Facts or just Copies? A critical investigation of the Competitions of Mechanisms in Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

This paper presents a reproducibility study examining how Large Language Models (LLMs) manage competing factual and counterfactual information, focusing on the role of attention heads in this process. We attempt to reproduce and reconcile findings from three recent studies by Ortu et al. [16], Yu, Merullo, and Pavlick [21] and McDougall et al. [8] that investigate the competition between model-learned facts and contradictory context information through Mechanistic Interpretability tools. Our study specifically examines the relationship between attention head strength and factual output ratios, evaluates competing hypotheses about attention heads' suppression mechanisms, and investigates the domain specificity of these attention patterns. Through this analysis, we aim to provide a clearer understanding of how different model components contribute to managing conflicting information in LLMs.

1 Introduction

The rise of Large Language Models (LLMs) has transformed our interaction with information, both through relying on the parametric knowledge that a language model learned through training [19] and through using approaches like retrieval augmented generation (RAG) [6] to retrieve facts from a given database. While these models achieve impressive results across many tasks, they often operate as "black boxes," making it challenging to understand how they arrive at their outputs. Built on transformer architectures and trained on vast datasets, LLMs have become increasingly sophisticated, yet their decision-making processes remain largely opaque. Yet, understanding whether a generated answer comes from model memory or from given context is essential for guaranteeing the trustworthiness of information access systems [20].

The two competing mechanisms, the use of parametric memory (factual recall) and the adaptation of in-context information, come into direct conflict when counterfactual statements opposing the model's parametric memory are provided as input. Hence, for understanding the origin of the provided answer we need to understand the interactions between these mechanisms.

Recent years have seen significant efforts to interpret these complexe language models. While early work focused on investigating input-output relationships through perturbation analysis [7], newer research suggests that understanding LLMs requires examining their internal operations. Mechanistic Interpretability[2] investigates how different transformer modules contribute to model outputs, focusing on specific components such as neurons, attention heads, and circuits (subnetworks of interconnected components) that drive particular model behaviors.

Prior work has investigated the competition between factual and counterfactual information in language models, particularly focusing on how different attention heads influence this competition. Yu, Merullo, and Pavlick [21], used a capital cities dataset to study how attention heads in Pythia-1.4B support either factual or counterfactual tokens, while Ortu et al. [16] investigate the same mechanism for GPT-2 and Pythia-6.9B using the CounterFact dataset. McDougall et al. [8] specifically examined the copy suppression role of one specific attention head in GPT-2. These studies collectively suggest that certain attention heads play crucial roles in mediating between factual and counterfactual information, and that manipulating the strength of these heads can influence the model's tendency to generate factual versus counterfactual responses. Yet they disagree on the exact role that different model compents play in this competition.

Our reproducibility study aims to validate and extend the findings of these three papers in three key directions. First, we examine the generalizability of the relationship between attention head strength and factual output ratios. Second, we investigate competing hypotheses about the mechanism through which these attention heads operate, specifically whether they function through targeted suppression of counterfactual tokens or through general copy suppression. Finally, we explore the domain specificity of these attention heads, building on observations from Yu, Merullo, and Pavlick [21] and Millidge and Black [10] about the potential specialization of attention heads to particular domains. Through these investigations, we aim to provide a more comprehensive understanding of how language models manage competing factual and counterfactual information.

2 Scope of reproducibility

In the scope of reproducibility, We focus on the works from Ortu et al. [16], Yu, Merullo, and Pavlick [21], McDougall et al. [8] and Millidge and Black [10]. We hereby extract their claims in different scopes and introduce some of their techniques, which we replicated or used. We also come up with some hypothesis to test based on the claims. The details of the techniques are introduced in section 3.1.

2.1 Existence of competition between factual and counterfactual tokens

When the prompt contains a counterfactual statement, the model has to make a choice between repeating this counterfactual statement or recalling the fact during generation. We focus on studying the situation when the factual or counterfactual statement is a single token, hence we denote the factual token by t_{fact} and the counterfactual token by t_{cofa}. The role of different attention heads in making the choice has been studied by Yu, Merullo, and Pavlick [21], Ortu et al. [16] and McDougall et al. [8].

Yu, Merullo, and Pavlick [21] used a dataset about capital cities of countries to investigate how each attention head in Pythia-1.4B[3] supports t_{fact} or t_{cofa} with the logit attribution techniques [11][13]. They also investigated the effect of multiplying the corresponding attention heads by a scalar, and found that decreasing the strength of an attention head supporting t_{fact} will significantly decrease the ratio of factual output compared with counterfactual output.

Moreover, to investigate the competition in GPT-2 [18] and Pythia-6.9B [3], Ortu et al. [16] applied similar techniques using the CounterFact dataset[9], a collection of counterfactual assertions paired with generation prompts. They found that increasing the strength at specific positions of attention heads supporting t_{fact} would increase the ratio of factual output to approximately 50% on both models.

McDougall et al. [8] examined the copy suppression role of the seventh head in the tenth layer (L10H7) of GPT-2, which is one of the heads analyzed by Ortu et al. [16].They suggest that this head plays a critical role in the generation of factual responses when there is a related counterfactual statement in the context.

ased on the combined findings from Yu, Merullo, and Pavlick [21], Ortu et al. [16], and McDougall et al. [8], we propose the following hypothesis:

Hypothesis 2.1. In a transformer model, a small number of attention heads contribute most significantly to the competition between t_{fact} and t_{cofa}, with each head supporting either t_{fact} or t_{cofa}. When a counterfactual statement exists in the prompt, increasing the strength of heads supporting t_{fact} will increase the proportion of factual outputs, while decreasing their strength will reduce the proportion of factual outputs.

We try to further test how this claim generalizes using methods described in section 3.4.1, and the results can be found in section 4.2.1.

2.2 The mechanism of the competition

While prior work generally agrees on the existence of such attention heads, there is disagreement on the exact underlying mechanisms that support them. Based on the findings of these papers, we formulate the following two hypotheses:

Suppression of counterfactual tokens. The work from Ortu et al. [16] called the attention heads that support t_{fact} 'factual recall heads', and suggests that those heads increase the ratio of factual responses by suppressing the copying of counterfactual tokens. Hence, we come up with the following hypothesis about the mechanism:

Hypothesis 2.2.1. The attention heads supporting t_{fact} increase the ratio of factual responses by suppressing the copying of counterfactual tokens, which means that it is possible to enhance the factual recall by modifying specific attention heads.

The anti-induciton effect. Olsson et al. [15] introduced the term 'induction' head, which "completes the pattern" by copying and completing sequences that have occurred before, and discovered some 'anti-induction' heads, which have negative scores on the induction effect. McDougall et al. [8] further conducted experiments on various datasets and showed that L10H7 in GPT-2 has a general anti-induction effect. These findings may indicate that L10H7 has a more general role in copy suppression beyond just suppressing counterfactual statements. We have also noticed that, when Yu, Merullo, and Pavlick [21] increased the overall strength of t_{fact} supporting attention heads for Pythia-1.4B, the ratio of factual responses never exceeded 20%, suggesting that the heads may only have a limited capability for selective suppression of counterfactual statements. Hence, an alternative hypothesis is:

Hypothesis 2.2.2. The attention heads supporting t_{fact} increase the ratio of factual responses by suppressing the induction effect (i.e., copying from the context by matching content and structure). Consequently, when a factual statement appears in the prompt, these attention heads will also suppress the copying of that statement.

We compare these two hypotheses using methods described in section 3.4.2, and the results can be found in section 4.2.2.

2.3 Domain specialization of heads engaged in the mechanism

Ortu et al. [16] did not specifically study how different attention heads work on different domains, and McDougall et al. [8] primarily focused on a single attention head. However, Yu, Merullo, and Pavlick [21] found that their identified attention head could not generalize well to datasets with more domains, and using the SVD technique from Millidge and Black [10], they discovered that the heads they identified were highly related to geography. The work from Millidge and Black [10] itself also suggests that different attention heads correspond to different domains. Based on these findings, we propose the following two hypotheses:

Hypothesis 2.3.1 (Existence of domain specialization). Attention heads do not uniformly suppress the induction effect or the copying of general counterfactual statements across different domains. Instead, their effects vary depending on the type of information required to answer the question.

Hypothesis 2.3.2 (Domain specialization and model size). As model size increases, attention heads become more selective and specialized in their suppression patterns.

We attempt to verify or refute these two claims using the methods described in section 3.4.3, and the results are presented in section 4.2.3

3 Methodology

Ortu et al. have provided their code on GitHub which we adapt and use for our experiments. They have also shared the dataset they are using. They have not shared their running results. As such, we first try to replicate their results by doing the same experiments.

Our main goal is to test the validity and generalisability of hypothesis 2.1, to compare hypothesis 2.2.2 and hypothesis 2.2.1, and to justify or unjustify hypothesis 2.3.1 and 2.3.2.

3.1 Analysis techniques

3.1.1 Logit attribution techniques

We, like Ortu et al., use the transformer_lens package to extract the outputs of the individual attention heads in each layer. The resulting output is grouped by position to obtain the outputs per layer per head at a certain position. The grouped results are unembedded and then we look for the logits of t_{cofa} and t_{fact}. A head at a certain layer promotes t_{fact} if in the unembedded output the logit for t_{fact} is increased after the head when compared to before. In many experiments the measure used to compare the logits of the tokens is the difference of t_{cofa} and t_{fact}, defined as $\Delta_{\text{cofa}} := \text{BlockLogit}(t_{\text{cofa}}) - \text{BlockLogit}(t_{\text{fact}})$.

3.1.2 Attention modification

Yu, Merullo, and Pavlick [21] and Ortu et al. [16] used attention modification to study the role of attention heads. More specifically, Ortu et al. [16] used the following setup: Let A^{hl} be the attention matrix of the h-th attention head in layer l, we then modify a position (i, j), where i is a higher value than j. This corresponds to the i-th token in layer l attending to the j-th token in layer l. The full equation for attention modification now becomes:

$$A'^{hl}_{ij} = \alpha \cdot A^{hl}_{ij}$$

where α is the amount of modification.

It is also possible to conduct this operation to all different j values, increasing the over strength of the attention head among all positions. Yu, Merullo, and Pavlick [21] did not specify whether they modified the attention values for all positions or just for the position of t_{cofa} explicitly, but considering the difference between their results and the results from Ortu et al. [16] who only modified the attention values for the position of t_{cofa}, we have reason to believe that they used a different designation than **ortu**, modifying the attention values for all positions.

In our study, we stick with the method used by Ortu et al. [16], modifying only the attention score from the position of t_{cofa} to the position of t_{fact}.

3.1.3 SVD analysis on OV matrices of attention heads

The product of the Value and Output weight matrices in a head is known as the OV matrix [5]. This matrix is interpretable by decomposing it with the Singular-Value-Decomposition (SVD) [10]. Namely, the singular vectors that are of the embedding dimension can be unembedded and detokenized to show what words are most encoded by the head. We utilize this technique in experiment 3.4.3 to investigate the roles individual heads of interest.

3.2 Model descriptions

For our experiments we also used GPT-2 and Pythia-6.9B. GPT-2 has 12 heads and 12 layers, with a total of 117 million parameters [18]. Pythia-6.9B has 32 heads and 32 layers with a total of 6.9 billion parameters [3]. Both models are pretrained. These are the same models that were used in the original paper.

More specifically, according to the results from Ortu et al. [16], Yu, Merullo, and Pavlick [21], McDougall et al. [8] and others, we are the most interested in the following attention heads supporting t_{fact}, where LxHy denotes the y-th head in the x-th layer:

- For GPT-2, we are the most interested in L10H7 and L11H10.
- For Pythia-6.9B, we are the most interested in L17H28, L20H18, and L21H8.

3.3 Original Datasets

In their paper, Ortu et al. construct their dataset using the pattern “Redefine: {base prompt} {falsehood}. {base prompt}”. An example of an entry in the dataset is “Redefine: The official language of Australia is Indonesian. The official language of Australia is". They source the falsehoods, associated correct facts and sentence patterns from the CounterFact dataset [9]. The CounterFact dataset is a large set of these base prompts, with associated counterfactual answers and factual answers. The model is queried for a single token which is then classified as either the copied falsehood, or the correct fact. Only prompts which the LLM could answer correctly in a single token without the preceding counterfactual statement were considered, and of those they chose to randomly sample 10 thousand prompts. They did this for both the models they used, so they ended up with two dataset of 10 thousand entries. One for GPT-2 and one for Pythia-6.9B.

We used the same datasets as the authors used for replicating their results. Both the GPT-2 and the Pythia-6.9B datasets contain $10,000$ entries. Each entry is comprised of 5 fields:

- "base_prompt" contains the original prompt from the CounterFact dataset.
- "template" contains the original prompt, but now duplicated with a field to be filled in before the original prompt, where redefine is always placed, and a field to be filled in before the duplicated sentence, where the counterfactual token is always placed.
- "target_true" contains the factual token.
- "target_new" contains the counterfactual token that is placed in the sentence.
- "prompt" contains the prompt to be given to the model, which is the template but with the values filled in.

There is also a variant of each dataset that contains a subject field with the subject of the sentence. This is used to determine the position and length of the subject in each prompt, allowing us to group the subject tokens together when showing results.

3.4 Experimental setup and code

To better understand the roles of the attention heads of interest mentioned in the work of Ortu et al. [16], Yu, Merullo, and Pavlick [21], and McDougall et al. [8], we conducted several experiments across different scopes. Specifically, we examine:

1. Whether these heads are sensitive to specific wordings and patterns (Section 3.4.1).
2. The general role and mechanism of these heads. That is, whether these heads are suppressing the copying of counterfactual tokens, or suppressing the induction behavior (Section 3.4.2).
3. Whether these heads perform identical roles or specialize, notably in domain knowledge (Section 3.4.3).

3.4.1 Experiments on the sensitivity to specific wordings and patterns

Different premises. We change the words before the semicolon in each prompt from redefine to one of the following: "Repeat", "Define", "Update", "New fact", "Updated fact", "This is true", "The following is true", "Pretend the following is true", and "Accept the following statement to be true".

We then measure for each head at each layer in the last position how much it promotes either the factual or the counterfactual token. We also test attention modification on the heads L11H10 and L10H7 for GPT-2.

Different sentence structures. Most of the sentences from the original dataset share a fixed pattern, namely "{subject} {relationship} {object}.", which means that we cannot eliminate the possibility that the attention heads of interest are focusing on specific grammatical patterns instead of recalling facts. Hence, we modified the sentence structure to test the generalization ability of the model on different sentence structures.

For each combination of two patterns selected from a predefined set of three, a dataset is generated by respectively applying both patterns to the two sentences in each entry of the data. The three patterns are as follows:

- Normal: "{subject} {relationship} {object}."
- Consider: "Consider {subject}. It {relationship} {object}."
- PeopleKnow: "Many people know {subject}. It {relationship} {object}."

We test model performance on the new patterns by measuring factual recall and counterfactual copying on the new structures. We also perform attention modification on the same heads and with the same alpha values as the original paper.

3.4.2 Experiments on the general role and mechanism

Different presence of hints. Many of the sentences from the original dataset (62.3%) contain the correct answer in the subject explictly, which means that we cannot eliminate the possibility that the attention heads of interest are in charge of copying from a specific part, instead of suppressing counterfacts or the induction mechanism. Hence, we split the dataset into two parts, one contains all sentences with the true target as a substring of the subject, called the hinted dataset, while another contains the remaining sentences, called the no-hint dataset.

Effect of changing the counterfacts to facts in the prompt. We generated datasets from the original datasets by replacing the counterfactual targets in the prompt with factual targets, which means that in these constructed datasets, the prompt will first explicitly state the fact, then ask the model to repeat the factual sentence.

If the attention heads of interest are heads suppressing the copying of counterfactual tokens as hypothesis 2.2.2 stated, increasing their strength should not significantly affect the performance of the model, while if those heads are heads with anti-induction effects as hypothesis 2.2.1 stated, increasing their strength should negatively affect the performance because the factual token is suppressed.

3.4.3 Experiments on domain specialization of heads

Effect of subject and answer categories on strongly contributing heads. To test hypothesis 2.3.1 We wish to see if heads have a bias towards certain topics or domains of knowledge, or if they are general induction or anti-induction attention heads playing identical roles. We do this by splitting our dataset into possible domains of knowledge. To categorize the answers, we first extracted the types of text patterns in the dataset, and matched them with the original patterns in the ParaRel dataset [4], a dataset from which the CounterFact dataset was derived. We then used GPT-4o to classify the categories of subjects and answers of each type of text pattern in the dataset and manually checked the classification results. Then, we add the category information to each entry in the dataset and filter the dataset by categories. We then perform logit inspection of the attention block for each head of each layer in the final token position of filtered subsets of the dataset. Furthermore, to test hypothesis 2.3.2 we observe how the behavior extends to larger models.

SVD analysis of relevant attention heads. Using the SVD technique [10] described in section 3.1.3, we inspect the attention heads of interest found in the previous experiment, to verify that any variation in results between categories is corroborated by the knowledge encoded in those respective heads.

3.5 Computational requirements

We ran our experiments involving GPT-2 on Ubuntu 22.04.2 on a computer with an RTX 4070 12Gb and an Intel(R) Core(TM) i5-9600K CPU @ 3.70GHz. Experiments involving Pythia-6.9B were conducted on a single NVIDIA A100 GPU from the Snellius cluster. In total we used about 10 hours of GPU time on the RTX 4070, and about 14 hours on the A100.

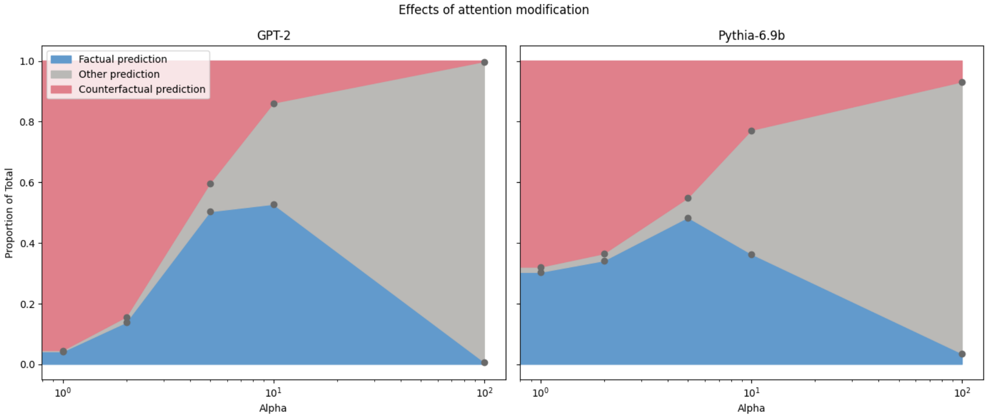


Figure 1: Results of attention modification on GPT-2 and Pythia-6.9B.

The RTX 4070 has a total graphics power of 200 watts [14], hence 10.0 hours of usage will cost 2.00 kWh. Also, according to the statistics provided by Snellius, we used roughly 10.2 kWh of power on the Snellius cluster. So, in total the experiments cost 12.2 kWh. The power intensity in the Netherlands in January 2025, when most of the experiments were conducted, was 0.284 kgCO$_2$/kWh [12]. So the carbon emission for the experiments is 3.47 kgCO$_2$. This is equivalent to 28.4 km of driving a new passenger car, assuming an emission of 0.122 kgCO$_2$/km [17]. Assuming that a mature tree can absorb approximately 21.8 kg (48 pounds) CO$_2$ per year [1], this amount of CO$_2$ requires approximately 1.91 months for a mature tree to absorb.

4 Results

4.1 Results reproducing original papers

Competition of mechanisms. We have successfully reproduced most of the work from Ortu et al. [16] by re-running their original experiments. Most of the results of re-running the original experiments align with the the original results, showing a decent reproducibility on the results of the original paper.

Additionally, we generated figure 1 as a supplementary result, which shows the proportions of factual and counterfactual predictions under different α values. The blue area represents the percentage of model generating the factual token, while the pink area is for the given counterfactual token. The gray area represents the case where the model generates neither factual nor the given counterfactual token.

From the figure, we can see that as the value of α increases, the factual prediction ratio first increases from the drop of counterfactual prediction, then decreases due to the model being unstable and generating other predictions. Hence, a moderate α value gives us the best result. However, there is a slight mismatch in the detail:

Mismatch on best α value for GPT-2. Figure 1 shows the proportions of factual and counterfactual predictions under different α values. For GPT-2, the maximum factual prediction rate is obtained at $\alpha = 10$, being contradictive to the statement in the original paper that the best α value for GPT-2 is 5. However, $\alpha = 5$ also yields a similar factual prediction rate, so this difference is not critical.

Singular value decompositions of transformer. We have also reproduced the singular value decomposition used by Yu, Merullo, and Pavlick [21], using the methods and steps by Millidge and Black [10]. Although the details do not exactly match, the domains and meanings of the obtained tokens are similar. Considering the fact that SVD is numerical and the choice of sign of a singular vector is arbitrary, the reproducibility of their works is decent.

4.2 Results beyond original papers

4.2.1 Experiments on the sensitivity to specific wordings and patterns

Our results show that hypothesis 2.1 holds when performing the same experiments. However, when changing the sentence structure the results become less pronounced than in the original setting.

Different premises. The supported answer of the significant heads in GPT-2 does not change when modifying the premise. This supports hypothesis 2.1. The strength of the heads however does appear to vary between premises in a limited degree.

Different sentence structures. As shown in table 1, when the sentence structure is changed, boosting the attention of the factual token promoting heads by $\alpha = 5$ or $\alpha = 10$ can still provide an increase to factual recall, being consistent with the original result. This indicates that hypothesis 2.1 can be generalized with respect to sentence structure changes.

The effect of changing the attention heads is lower if the two sentences have the same structure, and is higher if the two sentences have different structures. We hypothesize that when the first sentence is of a different structure than the second sentence, it is not as clear for the model that we expect it to copy t_{cofa}, leading to a less strong bias towards it.

Also when the first sentence and the second sentence are of equal structure, but different from the original structure, the total number of times t_{fact} is predicted is significantly lower than when compared to the original structure. This could be due to the alternate sentence templates being of a longer length. The longer length causes the sentences to show an example of copying for more tokens, making the model more likely to also want to copy t_{cofa}.

First Sentence	**Second Sentence**	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$	$\alpha = 5$	$\alpha = 10$	$\alpha = 100$
NORMAL	NORMAL	65	412	1377	5032	**5274**	70
	CONSIDER	87	603	2209	**6175**	3924	13
	PEOPLEKNOW	98	749	2640	**6066**	3474	5
CONSIDER	NORMAL	506	2186	4895	**7651**	5427	9
	CONSIDER	6	45	267	2896	**4481**	61
	PEOPLEKNOW	19	259	1279	**5636**	4794	12
PEOPLEKNOW	NORMAL	451	1928	4465	**6988**	3817	2
	CONSIDER	79	761	2452	**6126**	3264	3
	PEOPLEKNOW	2	10	82	2990	**3297**	15

Table 1: Table with grouped sentence structures and alpha values. The values indicate amount of times the factual token was predicted among the $10,000$ entries.

4.2.2 Experiments on the general role and mechanism

Our results favor hypothesis 2.2.1 more than 2.2.2, showing that the attention heads supporting t_{fact} are more likely to be related with the overall suppression of the induction effect, rather than the selective suppression of counterfactual statements.

Different presence of hints. The results are shown in Figure 2 and 3. When using the originally optimal value of $\alpha = 10$, GPT-2 predicted t_{fact} in 71% of the time when t_{fact} was a substring of the prompt (hinted), whereas it did so only in 23% of the time when it was not (no-hint).

Moreover, in the dataset where t_{fact} was not a substring, the model predicted neither t_{cofa} nor t_{fact} in 45% of cases. These results suggest that these attention heads are not particularly important for factual recall on their own, which was also supported by the original authors, and McDougall et al. [8].

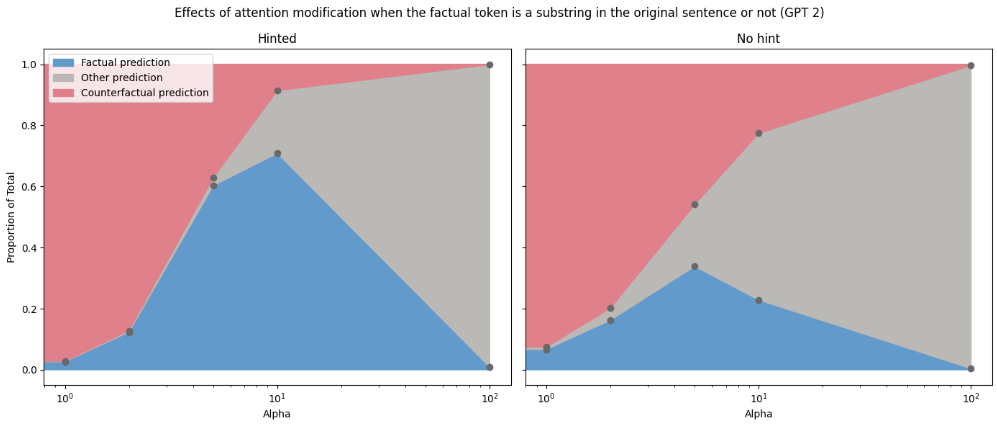


Figure 2: Results of GPT-2 attention modification when splitting the dataset based on whether or not t_{fact} was part of the subject or not. The hinted dataset has 6229 entries and the no-hint dataset has 3771 entries.

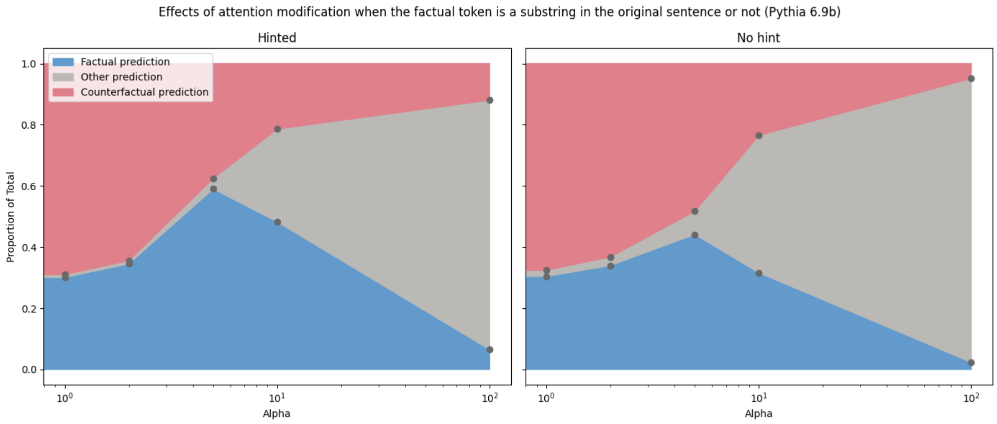


Figure 3: Results of Pythia-6.9B attention modification when splitting the dataset based on whether or not t_{fact} was part of the subject or not. The hinted dataset has 2855 entries and the no-hint dataset has 7145 entries.

Comparing this to the Pythia-6.9B results, we observe a similar pattern: the model predicts t_{fact} more often when it can copy from earlier in the sentence. However, the difference between the hinted and no-hint datasets is much smaller than in GPT-2, suggesting that Pythia-6.9B exhibits more factual recall. The attention modification in the no-hint cases increases factual predictions in Pythia-6.9B from about 30% to about 40%.

Another key difference between Pythia-6.9B and GPT-2 is the dataset composition. In Pythia-6.9B, the hinted dataset contains 2,855 entries, while the non-hinted dataset has 7,145 entries. We theorize that because the authors selected entries from CounterFact that the models could answer correctly, the GPT-2 dataset included more prompts where factual recall was unnecessary. In contrast, the Pythia-6.9B dataset contained more prompts requiring factual recall. This is likely due to the model's larger size allowing it to recall/record more difficult facts than GPT 2.

Effect of changing the counterfacts to facts in the prompt. The attention modification experiment result on the constructed dataset that replaces the counterfacts with facts is shown in Figure 4. The proportion of factual responses steadily decreases when the α value increases, even if $\alpha \leq 5$.

For GPT-2, when $\alpha = 5$ or $\alpha = 10$, the model answers the factual prediction for less than 90% or 70% of the time respectively. For Pythia-6.9B, the model answers the factual prediction for less than 70% of the time when $\alpha = 10$, mostly returning "the" or "also" as the response token when it does not answers the factual prediction.

This may indicate that the heads of interest are not heads suppressing solely counterfacts, but are heads with general anti-induction effects, and editing the model by increasing the strength of the specific attention heads cannot generally guarantee the increase of factual recalls as hypothesis 2.2.2 suggests.

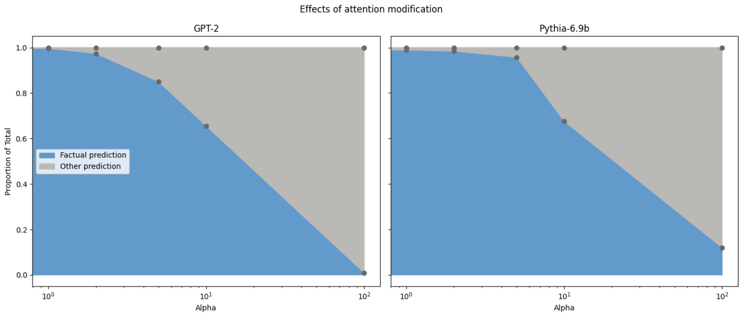


Figure 4: Results of attention modification after changing the counterfactual targets to factual targets in prompts.

4.2.3 Experiments on domain specialization of heads

Effect of subject and answer categories on strongly contributing heads. The heads in GPT-2 that were found by Ortu et al. [16] to most promote the counterfactual recall were L9H6, L9H9, L10H0, and L10H10, and the heads that most promote the factual recall were L10H7 and L11H10. Figure 5 illustrates the contribution of each of those heads to Δ_{cofa} in each slice of the categorized dataset. Evidently, subject and answer categories have a strong effect on the heads responsible for competition. To test if the effect is based on subject or answer category, we compare datasets filtered on both categories, differing in only one category of either subject or answer from a baseline. In our case we picked the dataset slices with most remaining results after filtering. The filtered dataset slices are truncated to be of equal size, 124 entries. The latter two heatmaps in Figure 5 demonstrate that the change in subject category is far less significant than the change of answer category.

These findings show that answer categories highly determine the contribution of heads to the competition, confirming hypothesis 2.3.1.

To address how this effect changes with model size, we perform this on Pythia-6.9B. The heads we compare were chosen after inspecting the contribution of heads in different answer categories. For that analysis, see appendix A.1. We choose the heads that are most influential in the competition, and heads that have interesting variation between categories that we would like to investigate. The chosen heads are compared

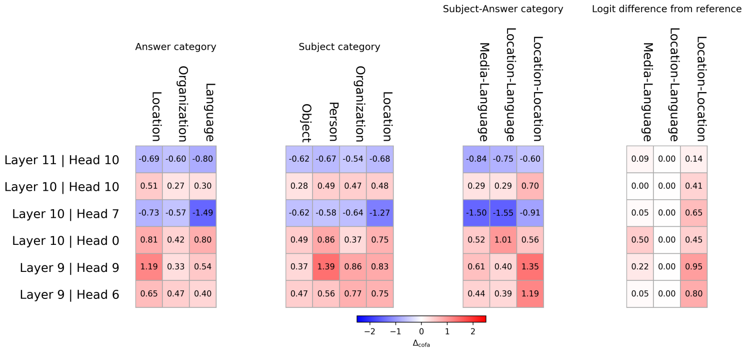
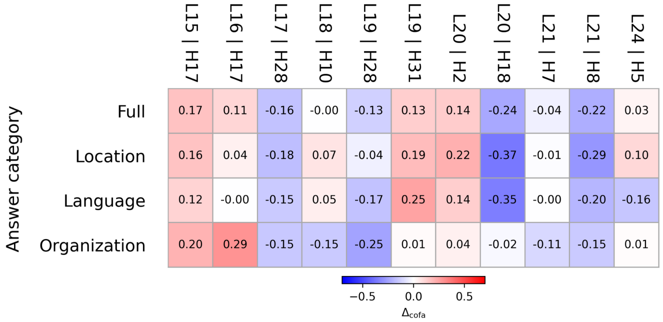


Figure 5: Attention scores of relevant heads for the most common categories that GPT-2 can predict.



Answer category	L15 | H17	L16 | H17	L17 | H28	L18 | H10	L19 | H28	L19 | H31	L20 | H2	L20 | H18	L21 | H7	L21 | H8	L24 | H5
Full	0.17	0.11	-0.16	-0.00	-0.13	0.13	0.14	-0.24	-0.04	-0.22	0.03
Location	0.16	0.04	-0.18	0.07	-0.04	0.19	0.22	-0.37	-0.01	-0.29	0.10
Language	0.12	-0.00	-0.15	0.05	-0.17	0.25	0.14	-0.35	-0.00	-0.20	-0.16
Organization	0.20	0.29	-0.15	-0.15	-0.25	0.01	0.04	-0.02	-0.11	-0.15	0.01

Figure 6: Heatmap of the srenght of the logit difference for different heads and categories.

in figure 6. The heads most supporting t_{cofa} and t_{fact}, are L15H17 and L20H18 respectively. Our results show that some influential heads do not vary much, but others vary extremely, much more than in GPT-2. The influence of heads appears more sparse in their effect on categories, with many categories being affected negligibly by certain heads that are highly influential on other categories (L16H17, L19H31, L20H18, L21H7, L24H5). Some heads (L18H10, L24H5) even significantly support t_{cofa} in certain categories and t_{fact} in others. Regarding these two heads, because their mean Δ_{cofa} over the total dataset is close to zero (0.00 and 0.03 respectively), they may have been disregarded as not relevant in previous literature when investigating only the full dataset, despite on average having a large influence on individual samples (standard deviation of L24H5 on full dataset: 0.78). Our observed patterns in this experiment are in contrast to the results on GPT-2, where the most influential heads vary in strength of contribution, but regardless of category contribute a significant amount to either t_{fact} or t_{cofa}. These results confirm hypothesis 2.3.2.

SVD analysis of relevant attention heads. To verify that the patterns in the previous experiment can be attributed to categories, we perform the SVD techniques described by Millidge and Black [10] on various heads of interest, illustrated in Figure 7, and Figure 10 which can be found in the appendix.

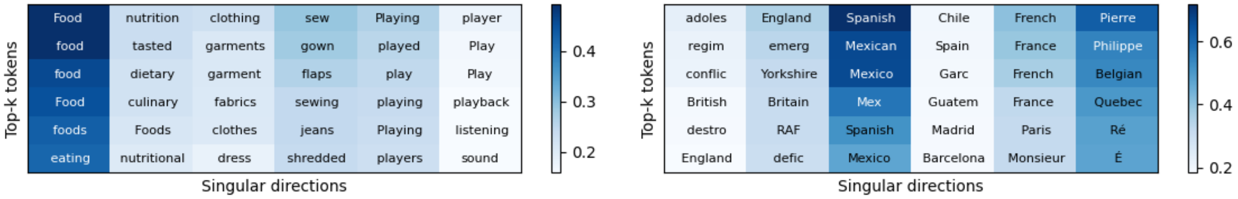
Starting with L15H17 (fig 10b), a head that globally promotes t_{cofa}. We do not see many tokens associated with meaning beyond low level grammatical structure of language. This is consistent with the finding of Yu, Merullo, and Pavlick [21] that their head contributing most to t_{cofa} was not related to any specific domain knowledge. Millidge and Black [10] did however find that earlier layers more commonly contain such heads encoding punctuation or less interpretable tokens.

Continuing to look at t_{cofa} promoting heads, L16H17 (fig 10b) and L19H31 (fig 10a) seem to behave as each other's complement in categories. Namely only being effective in the Organization category or only in the non-Organization categories respectively. Do note that Location and Language are related (e.g. England, English), so this is not too surprising. We see that in this case, there is a minor association with the domain for L19H31, containing some singular directions with geography information, particularly of the united states, but also some with organic and non-organic materials, such as 'aluminum', 'algae' and 'almond'. These do not have a trivial association with the topic of Organizations either however. L16H17 is less interpretable, but notably contains no geographic information.

Looking at t_{fact} promoting heads, we have L21H7 (fig 7a), L20H18 (fig 7b). Again, with L21H7 and L20H18 we see the complement in categories of Organization and non-Organization respectively. In this case, the words encoded by the layers are far more interpretable and coherent than for the t_{cofa} promoting heads. The words encoded by L20H18 very clearly relate to the topics Location and Language as expected. The words encoded by L21H7 do not directly relate to organizations, rather to the topics of food, clothing and hobby. It should be noted however that L21H7 is much less contributing to the competition than L20H18.

Our last head, L24H5 (fig 10d), is very interesting, as it significantly supports t_{cofa} for the Location category but t_{fact} for the Language category, and has little bias for the Organization category. We see that the words are very interpretable and geographically related, but curiously, almost all words are specifically related to English speaking commonwealth nations.

These results remain consistent with our findings in the previous experiment, and further support that the heads do not always contribute to the competition between t_{cofa} and t_{fact} in general, but are dependent on the domain of information required to correctly answer the prompt.



Food	nutrition	clothing	sew	Playing	player
food	tasted	garments	gown	played	Play
food	dietary	garment	flaps	play	Play
Food	culinary	fabrics	sewing	playing	playback
foods	Foods	clothes	jeans	Playing	listening
eating	nutritional	dress	shredded	players	sound

(a) Layer 21 Head 7

adoles	England	Spanish	Chile	French	Pierre
regim	emerg	Mexican	Spain	France	Philippe
conflic	Yorkshire	Mexico	Garc	French	Belgian
British	Britain	Mex	Guatem	France	Quebec
destro	RAF	Spanish	Madrid	Paris	Ré
England	defic	Mexico	Barcelona	Monsieur	É

(b) Layer 20 Head 18

Figure 7: Heatmap of the top-k tokens of the singular vectors sorted by their singular values.

5 Discussion

Generally, our experiment results support the hypothesis 2.1 that when the prompt contains a counterfactual token t_{cofa}, a small number of attention heads contribute most significantly to the competition between t_{fact} and t_{cofa}, and moderately manipulating the strengths of the attention heads will affect the proportion of factual and counterfactual outputs accordingly, as the results supporting this hypothesis are reproducible, and the hypothesis can generalize onto different premises and sentence structures.

However, this does not necessarily mean that moderately increasing the strengths of these heads will guarantee the generation of more factual responses under all circumstances. When we replace the the counterfactual token with the factual token in the original dataset, these heads now suppress the factual token. This suggests that these heads are more likely involved in copy suppression rather than specifically counterfactual suppression, supporting hypothesis 2.2.2 and challenging hypothesis 2.2.1. This also suggests that manipulating the strength of specific attention heads to increase the factual recall rate may not be generally applicable, or need more improvements and modifications.

Furthermore, we have gained insights into the effect of domain knowledge on how heads contribute to the competition. We have confirmed our hypotheses 2.3.1 and 2.3.2 and we have investigated further into the information encoded by relevant heads. We find that some heads, most commonly t_{fact} promoting heads, encode information corresponding to the domain of t_{fact} and t_{cofa}. Further work could be done by using a counterfact dataset where the domain of the factual and counterfactual token differ.

While our research covered multiple scopes and topics, a limitation of our experiments is the fineness of our experiments. More specifically, we did not run the experiments on more models, and we also did not study the attention modification of different heads on different domains. The future work can be in the direction of applying the methods onto different models, and study how each attention head interact with different domains. Considering that the dataset we used is mostly about locasions, languages, organizations and people, constructing new datasets on other domains like STEM and applying similar techniques on them is also a possible direction for the future work.

Our code is available here: `https://anonymous.4open.science/r/bzxwqrcn-CB70/`.

References

[1] U.S. Department of Agriculture. *The Power of One Tree: The Very Air We Breathe.* CO absorbed by a mature tree, Accessed: 2025-01-31. 2025. URL: `https://www.usda.gov/about-usda/news/blog/power-one-tree-very-air-we-breathe`.

[2] Leonard Bereska and Efstratios Gavves. “Mechanistic Interpretability for AI Safety–A Review”. In: *arXiv preprint arXiv:2404.14082* (2024).

[3] Stella Biderman et al. “Pythia: A suite for analyzing large language models across training and scaling”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 2397–2430.

[4] Yanai Elazar et al. “Measuring and Improving Consistency in Pretrained Language Models”. In: *Transactions of the Association for Computational Linguistics* 9 (2021). Ed. by Brian Roark and Ani Nenkova, pp. 1012–1031. DOI: `10.1162/tacl_a_00410`. URL: `https://aclanthology.org/2021.tacl-1.60/`.

[5] Nelson Elhage et al. “A mathematical framework for transformer circuits”. In: *Transformer Circuits Thread* 1.1 (2021), p. 12.

[6] Patrick Lewis et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474.

[7] Jiwei Li, Will Monroe, and Dan Jurafsky. “Understanding neural networks through representation erasure”. In: *arXiv preprint arXiv:1612.08220* (2016).

[8] Callum Stuart McDougall et al. “Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads”. In: *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Yonatan Belinkov et al. Miami, Florida, US: Association for Computational Linguistics, Nov. 2024, pp. 337–363. DOI: `10.18653/v1/2024.blackboxnlp-1.22`. URL: `https://aclanthology.org/2024.blackboxnlp-1.22/`.

[9] Kevin Meng et al. “Locating and Editing Factual Associations in GPT”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 17359–17372. URL: `https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf`.

[10] Beren Millidge and Sid Black. “The singular value decompositions of transformer weight matrices are highly interpretable”. In: *LessWrong* (2022). URL: `https://www.lesswrong.com/posts/mkbGjzxD8d8XqKHzA/the-singular-value-decompositions-of-transformer-weight`.

[11] Neel Nanda and Joseph Bloom. *TransformerLens*. `https://github.com/TransformerLensOrg/TransformerLens`. 2022.

[12] NED. *Electriciteitsmix afgelopen 12 maanden*. Accessed: 2025-01-31. 2025. URL: `https://ned.nl/nl/dataportaal/energie-productie/elektriciteit/totale-elektriciteitsproductie#datasets`.

[13] nostalgebraist. “interpreting GPT: the logit lens”. In: *LessWrong* (2020). URL: `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

[14] NVIDIA. *GeForce RTX 4070 Family Graphics Cards*. Total Graphics Power: 200W, Accessed: 2025-01-31. 2025. URL: `https://www.nvidia.com/nl-nl/geforce/graphics-cards/40-series/rtx-4070-family/`.

[15] Catherine Olsson et al. “In-context Learning and Induction Heads”. In: *Transformer Circuits Thread* (2022). https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[16] Francesco Ortu et al. “Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals”. In: *arXiv preprint arXiv:2402.11655* (2024).

[17] European Parliament. *CO emissions from cars: Facts and figures (infographics)*. CO emissions per km, Accessed: 2025-01-31. 2025. URL: `https://www.europarl.europa.eu/topics/en/article/20190313STO31218/co2-emissions-from-cars-facts-and-figures-infographics`.

[18] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: (2019).

[19] Adam Roberts, Colin Raffel, and Noam Shazeer. “How Much Knowledge Can You Pack Into the Parameters of a Language Model?” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 5418–5426.

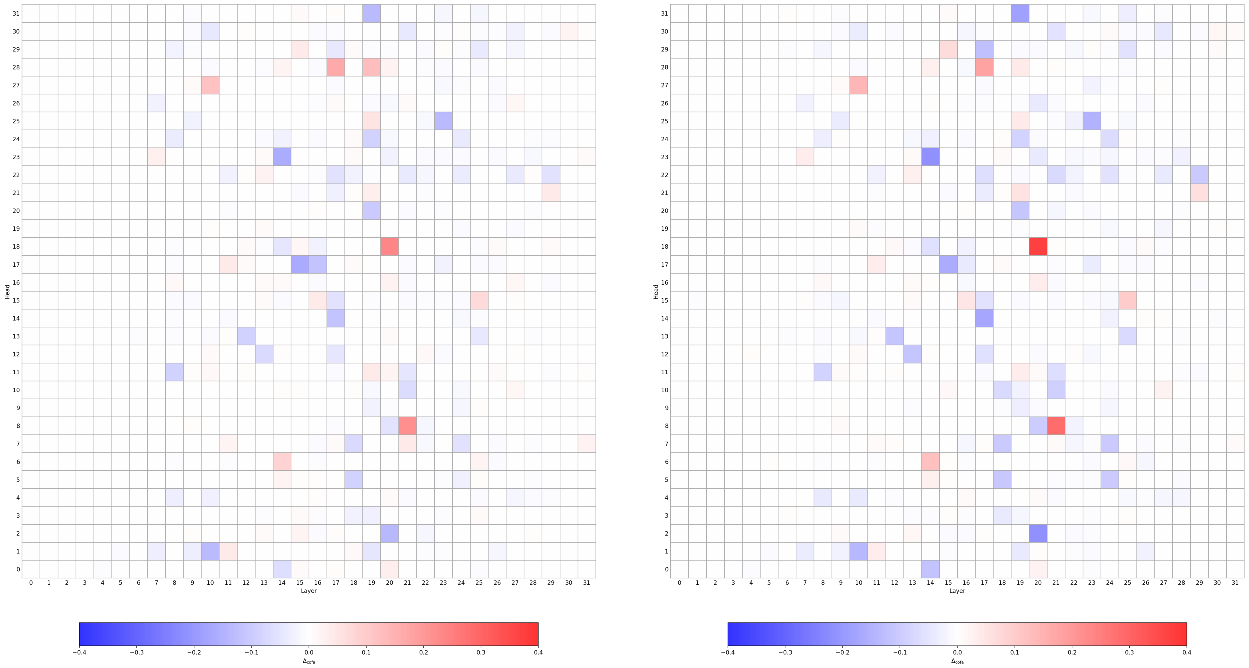
[20] Jonas Wallat et al. “Correctness is not Faithfulness in RAG Attributions”. In: *arXiv preprint arXiv:2412.18004* (Dec. 2024).

[21] Qinan Yu, Jack Merullo, and Ellie Pavlick. “Characterizing mechanisms for factual recall in language models”. In: *arXiv preprint arXiv:2310.15910* (2023).

A Appendix

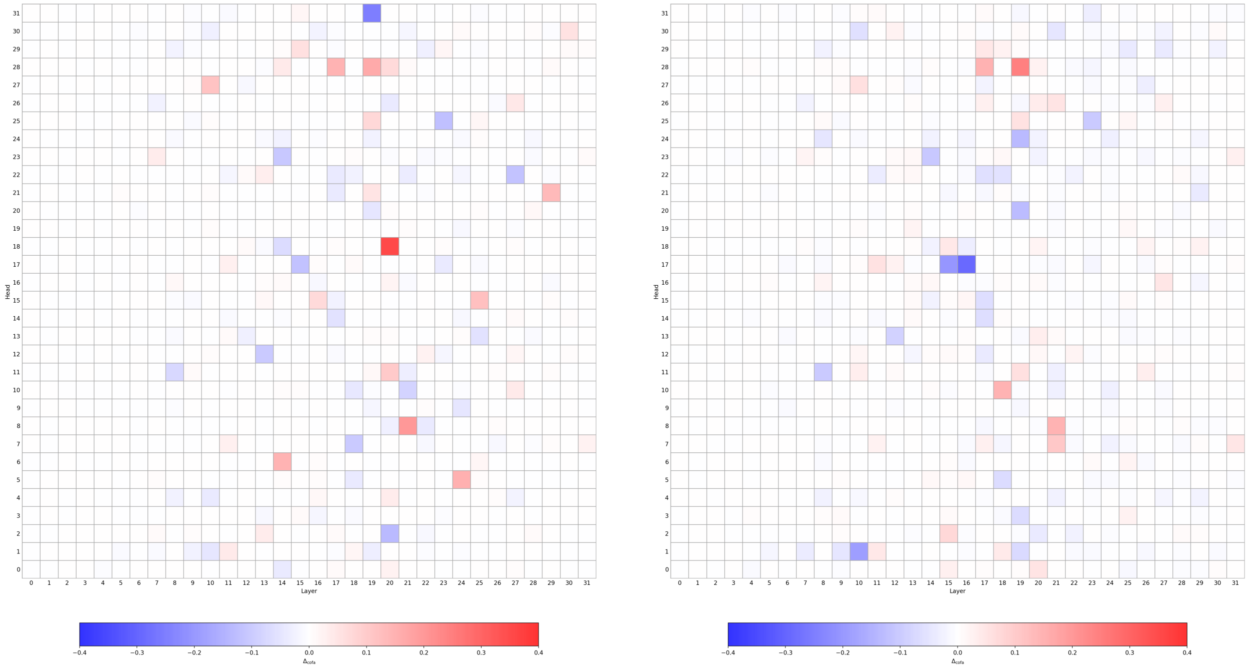
A.1 Head selection for comparison in experiment 3.4.3

Figure 8 and 9 were used to identify important heads within and between categories. The standard



(a) Full dataset

(b) Location category



(c) Language category

(d) Organization category

Figure 8: Mean Δ_{cofa} values of each head in Pythia-6.9B for the analysed categories

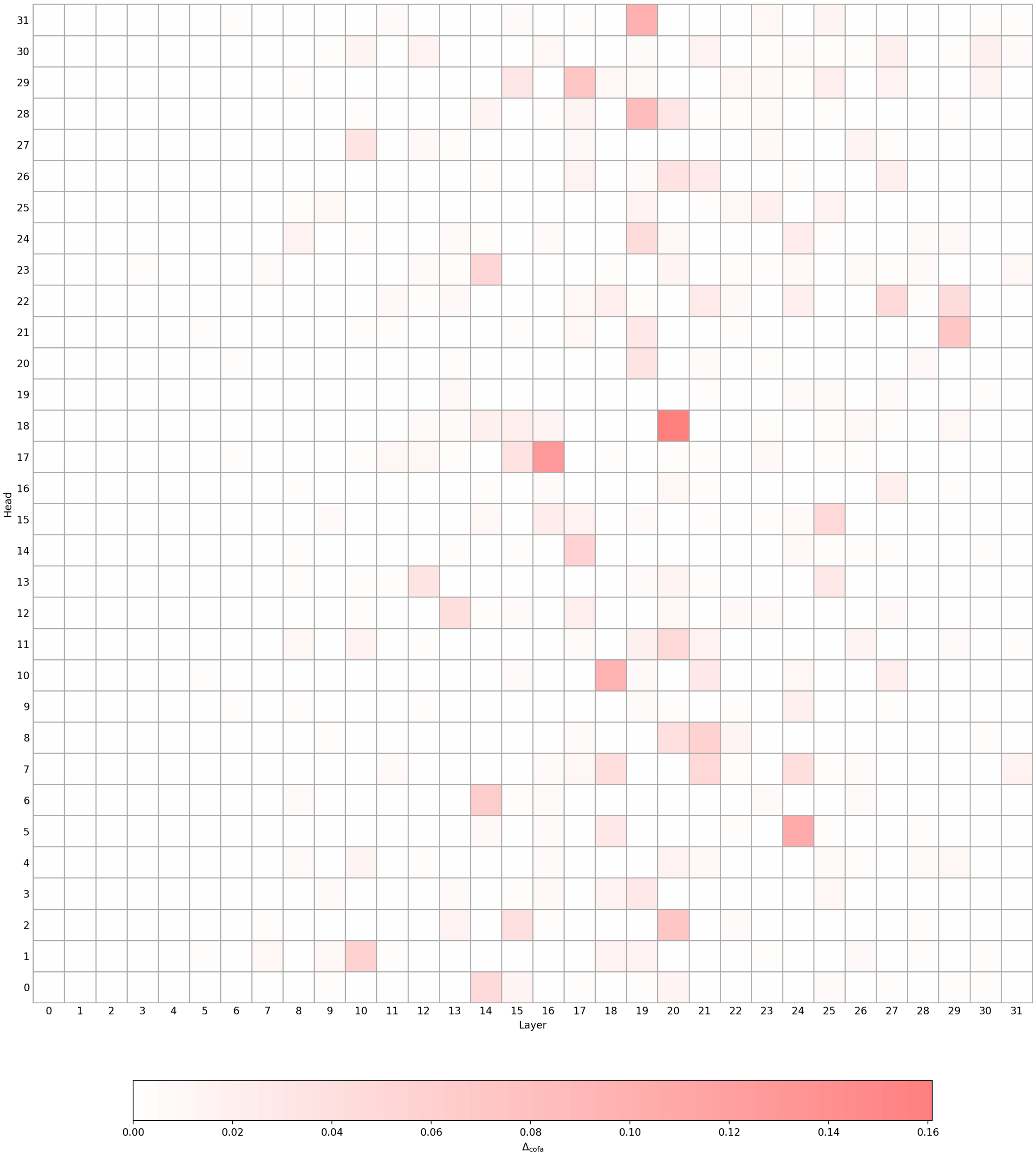
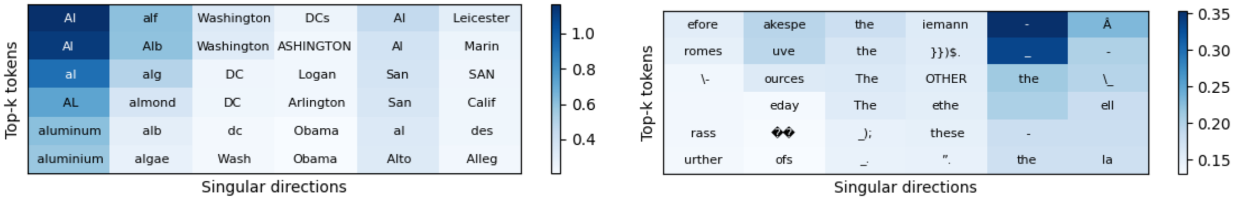
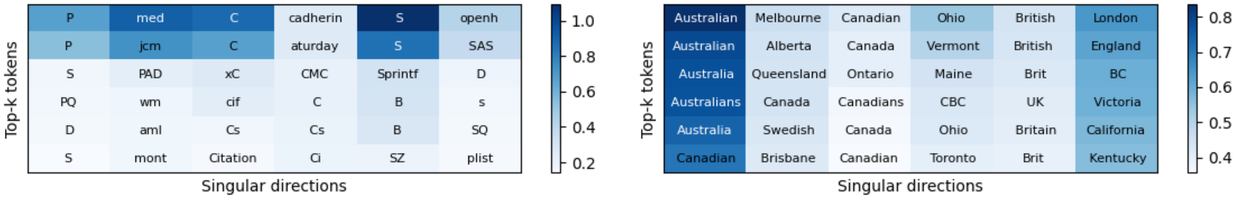


Figure 9: Standard deviation between the mean Δ_{cofa} value of the three analyzed categories of the dataset



(a) Layer 19 Head 31

(b) Layer 15 Head 17



(c) Layer 16 Head 17

(d) Layer 24 Head 5

Figure 10: Heatmap of the top-k tokens of the singular vectors sorted by their singular values.