Mechanisms of Non-Factual Hallucination in Language Models

Anonymous ACL submission

Abstract

State-of-the-art language models (LMs) sometimes generate *non-factual hallucinations* that misalign with world knowledge. Despite extensive efforts to detect and mitigate hallucinations, understanding their internal mechanisms remains elusive. Our study investigates the mechanistic causes of hallucination, especially non-factual ones where the LM incorrectly predicts object attributes in response to subject-relation queries. With causal mediation analysis and embedding space projection, we identify two mechanistic causes: 1) insufficient attribute knowledge in lower-layer MLPs, and 2) failing to select the correct object attribute in upper-layer attention heads. These mechanisms in non-factual hallucinations exhibit varying degrees of subject-object association, predictive uncertainty and perturbation robustness. Additionally, we scrutinize LM pre-training checkpoints, revealing distinct learning dynamics for the two mechanistic causes of hallucinations. We also highlight how attribution features from our causal analysis can effectively construct hallucination detectors. Our work pioneers a mechanistic understanding of LM factual errors, fostering transparent and explainable approaches for hallucination mitigation.

1 Introduction

Language models (LMs) serve as repositories of substantial knowledge (Petroni et al., 2019; Jiang et al., 2020; Srivastava et al., 2023), yet they are susceptible to generating text containing factual errors. Notably, LMs have been observed to produce seemingly confident completions with hallucinations (Dong et al., 2022; Zhang et al., 2023b), fabricating entities or claims. As LMs extend their reach to broader audiences and potential applications in safety-critical domains, understanding the nature of factual errors becomes critical (Kaddour et al., 2023).

The majority of research efforts has been centered on hallucination detection and mitigation (Elaraby et al., 2023; Mündler et al., 2023; Manakul et al., 2023; Zhang et al., 2023a). However, the internal mechanisms underlying LM hallucinations remain under-explored. Previous investigations into hallucinations often treat the LM as a black box, developing methods for factual generation based on external features such as predictive uncertainty (Xiao and Wang, 2021; Varshney et al., 2023) and logical consistency (Cohen et al., 2023). Unfortunately, these approaches provide no insights into the internal mechanisms of factual errors and have demonstrated unreliability or conveyed contradictory signals (Turpin et al., 2023).

In contrast, interpretability research, which investigates the internal mechanisms of transformers in white-box settings, has identified several crucial model components related to knowledge flow that are essential for answering questions correctly (Dai et al., 2022; Meng et al., 2022a; Geva et al., 2023). In addition, Akyürek et al. (2022); Zhou et al. (2023) has identified the important role of LM pre-training in the acquisition of factual knowledge. These interpretability studies on knowledge flow in LMs have limited scopes: they only examine cases where models generate *factually correct* responses, leaving questions on how information flow or acquisition unclear for hallucinations. Specifically, it is unknown whether these components are equally "fragile" and prone to simultaneous failure, or if only certain components deviate from normal functioning. It is also unclear how these factual errors emerge and evolve during the process of language model pretraining.

In this study, we employ mechanistic interpretability (Olah, 2022) to investigate the origins and manifestation of non-factual hallucinations in language models (LMs). We use two established interpretability methods, causal mediation analysis (Pearl, 2001; Vig et al., 2020) and embedding space projection (Geva et al., 2022; Dar et al., 2023) in our specially designed setups on non-factual hal-

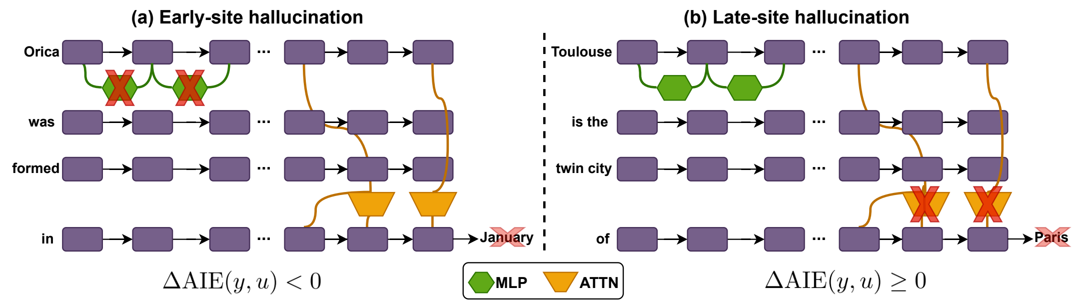


Figure 1: **Our main finding of two non-factual hallucination mechanisms. Left (a)**: The **early-site hallucinations** are caused by lacking general knowledge of the subject in lower layer MLPs of transformer LMs – in this case, the model fails to retrieve useful information about the entity (e.g., *Orica*, an Australian-based multinational corporation) to generate the correct object attribute (e.g., *Australia*), and therefore outputs a highly non-feasible prediction (January, which is incorrect as verified by manual fact-checking). **Right (b)**: The **late-site hallucinations** are caused by upper layer attention modules' failure to identify the most relevant object attribute(s) to the subject and relation – in this case, the model is able to retrieve related information about the subject (e.g., *Toulouse*, a French city) from early-site MLPs, but cannot distinguish the irrelevant yet strongly associated attributes (e.g., *Paris*) from the correct answers (e.g., *Bologna*/*Chongqing*/*Atlanta*). We found that these two types of hallucinations can be distinguished by the relative causal contribution to model predictions between MLP and attention modules ($\Delta\mathrm{AIE}(y, u)$, to be explained in section 4.1).

lucination data, aiming to assess the influence of model components on hallucinating predictions. We obtain converging evidence of two crucial LM modules with the highest causal attributions to factually incorrect generations: the multi-layer perceptrons (MLPs) in lower transformer layers and the attention heads in upper transformer layers, which have also been discovered as playing essential roles in recalling factual associations (Geva et al., 2023).

Figure 1 illustrates two distinct scenarios where the identified hallucinating components exhibit different behaviors. In some instances (the right subfigure), lower-layer MLPs function normally, successfully retrieving semantic attributes about queried entities, while upper-layer attention heads struggle to distinguish the most relevant attribute. In other cases (the left subfigure), the model fails to execute its fact-recalling pipeline at the beginning, extracting no useful information from lower-layer MLPs. We also observe that these two hallucination mechanisms have varying external manifestations, distinguishable by their levels of subject-object association strengths, robustness to input perturbations, and model predictive uncertainty.

Moreover, our analysis investigates the learning dynamics of language models, unveiling their progressive yet sometimes imperfect development of fact-recalling pipelines during pretraining. As a practical application, we demonstrate that mechanistic interpretability features can be employed to probe the presence of factual errors in LMs. Our work offers the first mechanistic explanation of LM factual errors as modular failures, fostering research on model transparency and new methods for hallucination mitigation.

2 Related Work

Factual knowledge in language models. The exploration of knowledge tracing within Language Models (LMs) has gained substantial attention lately, with researchers investigating specific layers (Wallat et al., 2020; Meng et al., 2022a) and neurons (Dai et al., 2022) responsible for storing factual information. This line of inquiry extends to techniques for model editing (De Cao et al., 2021; Mitchell et al., 2021; Meng et al., 2022b) and inference intervention (Hernandez et al., 2023; Li et al., 2023). Recent advancements by Geva et al. (2023); Yu et al. (2023) identify crucial LM components that form an internal pipeline for factual information transfer. Our framework complements existing research by offering an additional perspective on LM factual knowledge processing, revealing that compromised factually relevant modules can lead to hallucinations.

Hallucinations. Language models are susceptible to generating hallucinations that can be *unfaithful* (i.e. deviating from the source input provided by users) or *non-factual* (i.e. contradicting established world knowledge) (Ji et al., 2023; Zhang et al., 2023b). Here, we focus on the latter type of hallucination. Existing studies propose various methods to detect or mitigate hallucinations, leveraging features such as internal activation patterns (Yuksekgonul et al., 2023; Li et al., 2023), predictive confidence (Varshney et al., 2023), and generation consistency (Mündler et al., 2023; Manakul et al., 2023; Zhang et al., 2023a). However, a mechanistic investigation accounting for non-factual hallucinations is lacking in these studies.

Mechanistic interpretability. Mechanistic interpretability (Olah, 2022; Nanda, 2023) is an evolving research area. Recent works employ projections to the vocabulary (Dai et al., 2022; Geva et al., 2022; Nostalgebraist, 2020) and interventions in transformer computation (Haviv et al., 2022) to study LM inner workings. Similar techniques have been applied to explore neural network learning dynamics (Nanda et al., 2022) and discover sparse computational graphs for specific tasks (Wang et al., 2022; Conmy et al., 2023). Leveraging multiple mechanistic interpretability methods, our study provides a comprehensive yet consistent account for non-factual hallucinations.

3 Background and Notation

An auto-regressive transformer language model, denoted as G, maps an input sequence of tokens $u = [w_1, ..., w_T]$, represented by input token embeddings $E(u) = [e_1, ..., e_T]$, into a probability distribution over the vocabulary for next-token prediction. Within the transformer, the i-th token is represented as a series of hidden states $h_i^{(l)}$ where at each layer l, the model computes and adds the intermediate embeddings by two modules from $h_i^{(l-1)}$: 1) an aggregated **multi-head self-attention module** output $a_i^{(l)} = W_o([a_i^{(l,0)}, ..., a_i^{(l,K)}])$, where $a_i^{(l,k)}$ is the output of the k-th attention head at layer l (with K heads in total) for the i-th token, and W_o is a linear transformation; 2) a **multi-layer perceptron (MLP)** output $m_i^{(l)} = f_{\text{MLP}}^{(l)}(h_i^{(l-1)} + a_i^{(l)})$ at layer l. Putting together, the hidden representation $h_i^{(l)}$ is computed as:

$$h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)}. \tag{1}$$

Let $H = \{h_i^l\}$ be the set of $T \times L$ token hidden states across all layers (following Elhage et al. (2021), we shall call them the **residual stream outputs**), $A = \{a_i^l\}$ be the set of $T \times L$ **attention outputs**, and $M = \{m_i^{(l)}\}$ be the set of $T \times L$ **MLP outputs**. We aim to investigate which intermediate model outputs $z \in Z = H \bigcup A \bigcup M$ (and the corresponding sublayers that produce them) are causally contributing to the generation of a factually incorrect entity.

4 Mechanisms of Hallucinations

4.1 Causal tracing of factual errors

Method. The intermediate hidden states H produced by G during model inference form a causal dependency graph (Pearl, 2001) that contains many paths from the input sequence to the output (next-token prediction), and we wish to understand if there are specific hidden states that are more important than others when the producing a hallucination. This is a natural case for *causal mediation analysis*, which quantifies the contribution of intermediate variables in causal graphs. For more information about causal mediation analysis of language models, see (Vig et al., 2020).

We adapt the framework of Meng et al. (2022a) to locate LM components that cause factual errors via the task of factual open-domain questions on structured queries. In particular, given a fact represented as a subject-relation-object triple (s, r, o), we provide an LM G with a query prompt u containing (s, r) (e.g., "Toulouse is the twin city of __") with o as a true continuation (e.g., "Atlanta"). We examine the cases where G predicts an incorrect object o' as the next token(s) given u, and aim to locate which intermediate hidden states in the computation graph of G led to the hallucination. We consider G to be a "corrupted" model with certain modules failing to compute the "clean" representations that could otherwise lead to the correct answer o, and measure the contribution of each module through four model runs:

1. In the **hallucination run**, we pass u into G and extract all intermediate hidden representations Z as defined in Section 3, and compute the log likelihood ratio $y = \log \frac{p(o'|E(u))}{p(o|E(u))}$ between the true and hallucinated objects, which quantifies the "degree of hallucination" of G. For a hallucinating prediction, we would observe $y > 0$.

2. In the **mitigation run**, we follow Meng et al. (2022a) and add a Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ to the input token embeddings $E(u)$, so that when taking the intervened $E^*(u) = E(u) + \epsilon$ as inputs, the log-likelihood ratio between the hallucinated and the factual object would decrease (i.e., we only take noises with $y_* =$

$\log \frac{p(o'|E^*(u))}{p(o|E^*(u))} < y$, indicating that the model becomes more "truthful" after noise injections). We again extract all intermediate hidden representations, denoted as Z^*.

3. In the **mitigation-with-hallucination-state run**, we run G on u with perturbed input embeddings $E^*(u)$ as in the mitigation run, and hook G by forcing a particular hidden representation $z^* \in Z^*$ to be the hidden representation z during the hallucination run. We then compute the the log likelihood ratio $y_{E^*,z} = \log \frac{p(o'|E^*(u),z)}{p(o|E^*(u),z)}$ to see how it changes compared to step 2.

4. In the **hallucination-with-mitigation-state run**, we run G on the original prompt u as in the hallucination run, and hook G by forcing a particular hidden representation $z \in Z$ to be the hidden representation z^* during the mitigation run. We then compute the the log likelihood ratio $y_{E,z^*} = \log \frac{p(o'|E(u),z^*)}{p(o|E(u),z^*)}$ to see how it changes compared to step 2.

We can therefore define two causal contribution measurements of each hidden state z: the **causal indirect effect** $\mathrm{IE}(y, u, \epsilon) = y_{E^*,z} - y_*$ measures the decrease in the degree of hallucination after mitigating a single hidden state, and the **causal direct effect** $\mathrm{DE}(z, y, u, \epsilon) = y_{E,z^*} - y_*$ measures the decrease in the degree of hallucination after mitigating all other intermediate hidden states except z. Averaging over a set of factual queries and a sample of noises for each query, we obtain the average direct effect (ADE) and average indirect effect (AIE) for each hidden state.

Data. We collect a set of factual knowledge queries from ParaRel (Elazar et al., 2021), with each example containing a knowledge tuple $t_c = (s, r, o_c)$ and a prompt generated from hand-curated templates. We evaluated GPT-2-XL on each prompt u by computing the conditional probability $p(o|E(u))$ of the next token continuation, where o is taken from the collection of all capitalized alphabetical tokens in the model vocabulary. We define hallucinations as the cases where the model assigns the highest probability to a token o' that is neither the suffix of the true object o_c nor the suffix of any other objects of the subject-relation pair (s, r) returned by a WikiData API query search. This pipeline yields a set of 6,401 (u, o, o') examples. [1]

[1] See Appendix A for more details about data construction.

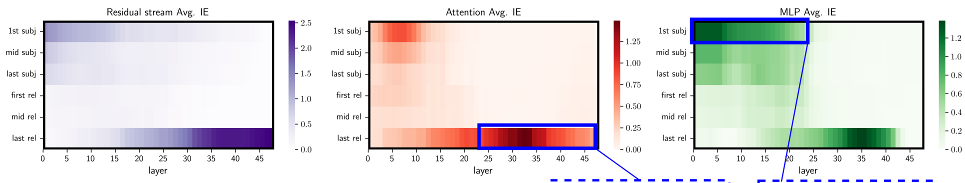
Results. We compute the average causal effect over all collected queries from ParaRel for all hidden states $z \in Z$ across various sentence positions and transformer layers. Similar to previous studies of causal mediation analysis, we found the distributions of direct effect to be noisy and less interpretable (Vig et al., 2020; Meng et al., 2022a), and therefore focus on the causal tracing results of indirect effect, as shown in Figure 2 for three modules: the residual stream, the attention heads, and the MLPs. We observe two groups of hidden states yielding the highest attribution scores towards incorrectly predicted objects: 1) the hidden states at the early site (lower GPT layers) of the subject tokens, and 2) the hidden states at the late site (upper GPT layers) of the last relation token. Our causal tracing results therefore offer contrapositive support to the existence of the two-stage fact recalling pipeline discovered by Geva et al. (2023), by showing that failures of the same two module groups are most likely causing factual errors.

Early- vs. late-site hallucination. Based on the findings above, we hypothesize that there are two different "mechanisms" that may cause non-factual hallucinations, as illustrated in Figure 1: 1) the model fails to retrieve any related information about the subject from lower-layer MLPs, and 2) the model successfully retrieves some subject attributes from lower-layer MLPs, but the upper-layer attention heads fail to distinguish the correct object(s) among retrieved ones. We formalize this idea by defining the following **relative indirect effect** between late-site attentions (where the correct objects are distinguished) and early-site MLPs (where subject attributes are retrieved):

$$\Delta\mathrm{AIE}(y, u) = \mathrm{AIE}^{\text{late}}_{\text{Attn}}(y, u) - \mathrm{AIE}^{\text{early}}_{\text{MLP}}(y, u) \tag{2}$$

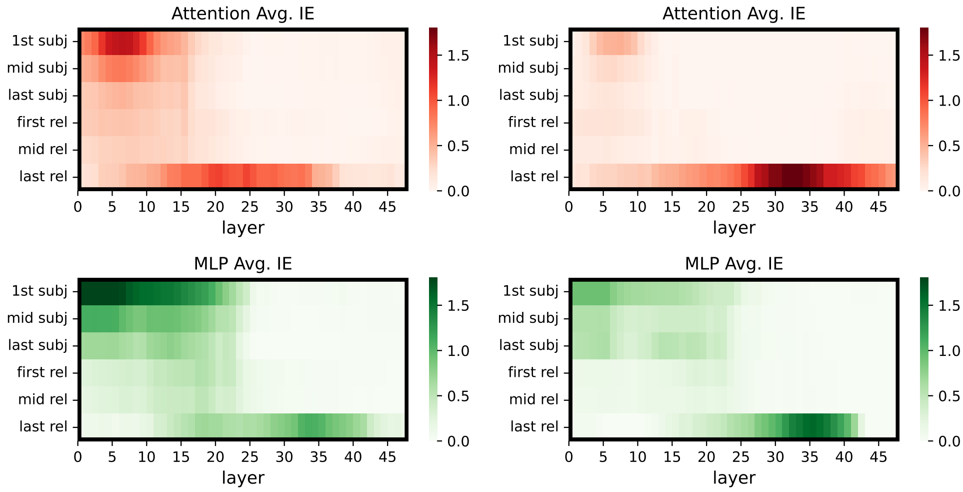
$$= \frac{2}{L}\Big[\sum_{l=\frac{L}{2}}^{L} \mathrm{AIE}(a_T^{(l)}, y, u) - \sum_{l=1}^{\frac{L}{2}} \mathrm{AIE}(m_0^{(l)}, y, u)\Big] \tag{3}$$

where $\mathrm{AIE}^{\text{early}}_{\text{MLP}}(y, u)$ is the average indirect effect of MLP sublayers in the lower 24 out of 48 layers on the first subject token w_0 of a query u, and $\mathrm{AIE}^{\text{late}}_{\text{Attn}}(y, u)$ is the average indirect effect of attention heads in the upper 24 layers on the last relation token w_T of u, as illustrated in Figure 2. A hallucination (u, o, o') is **early-site** if the corresponding $\Delta\mathrm{AIE}(y, u) < 0$, and is **late-site** if $\Delta\mathrm{AIE}(y, u) \geq 0$. Following this definition, we classify the incorrectly answered queries into two



$$\Delta\mathrm{AIE}(y,u) = \mathrm{AIE}^{\mathrm{late}}_{\mathrm{Attn}}(y,u) - \mathrm{AIE}^{\mathrm{early}}_{\mathrm{MLP}}(y,u) = \boxed{\mathrm{AIE}(a_T^{\frac{L}{2}:L}, y, u)} - \boxed{\mathrm{AIE}(m_0^{1:\frac{L}{2}}, y, u)}$$

Figure 2: **Average Indirect Effect (AIE)** of individual model components to non-factual hallucinations over 6,401 ParaRel queries that are incorrectly answered by GPT-2 XL. $\Delta\mathrm{AIE}(y,u)$ is defined as the difference in AIE between 1) the attention outputs of the last 24 transformer layers and 2) the MLP outputs of the first 24 GPT-2 XL layers.



(a) Early-site hallucinations ($\Delta\mathrm{AIE}(y,u) < 0$) (b) Late-site hallucinations ($\Delta\mathrm{AIE}(y,u) \geq 0$)

Figure 3: **Average Indirect Effect (AIE)** of individual model components for (a) early-site (left column) and (b) late-site (right column) non-factual hallucinations.

Statistics	Early-site hall.	Late-site hall.
Amount	(2414 / 37.7%)	(3987 / 62.3%)
s-o assoc.	0.14	0.91
s-o' assoc.	0.17	2.12
Robustness	0.67	0.44
Uncertainty	5.10	4.74

Table 1: External data and model prediction features for two types of non-factual hallucination.

categories and compute their average indirect effect distributions separately (Figure 3). We observe significantly different causal effect distributions: while most neurons that contribute significantly to early-site hallucinations are located in lower layers, late-site hallucinations have more highly contributive neurons in upper layers.

External manifestations of hallucination mechanisms. We next investigate whether there are any external features that can be leveraged to distinguish the two types of hallucinations. We consider the following features of query data and model predictions: the **subject-object association strength** is measured as the inner product between the GPT input layer embeddings of a subject s and a true object o or a hallucinating object o'; the **robustness** of a predicted object o' is measured as the percentage of Gaussian noise injected during the mitigation run in section 4.1 which, after being added to the input embeddings, fails to make the model prefer the true answer o than o' (i.e., $y_* < 0 < y$); the **uncertainty** of model prediction is measured by the entropy of the conditional next-token distribution $p(o|u)$. Table 1 summarizes the external measurements. Some key observations are: 1) subjects of late-site hallucinations (e.g., *Toulouse*) often have hallucinating objects (e.g., *Paris*) of much stronger association strengths than true objects (e.g., *Bologna*), so that the late-site attention heads fail to "offset" the prior propensity of model predicting o' upon seeing s. Subjects of early-site hallucinations (e.g., *Orica*), on the other hand, of-

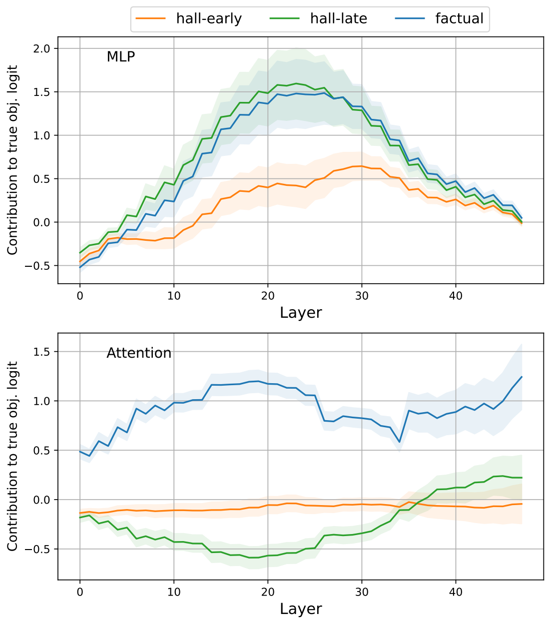


Figure 4: Average dot product between true object unembedding and transformer module intermediate outputs in each layer.

ten have much weaker associations with both true (e.g., *Australia*) and hallucinating (e.g., *January*) objects, which conforms with the low causal contribution of early-site MLPs that are supposed to store relevant knowledge about these entities; 2) late-site hallucinations are significantly less robust under input perturbations, probably because the model has already retrieved the correct object from early layers and is just "one step away" from distinguishing it and other less relevant attributes; 3) the model is less certain about its predictions when generating early-site hallucinations, a pattern that is consistent with previous findings that epistemic hallucinations are often associated with high predictive uncertainty (Xiao and Wang, 2021).

4.2 Module inspection via embedding space projection

In this section, we provide further evidence of the mechanistic difference between early- and late-site hallucinations by looking at the information that each module writes into the model residual stream during model inference.

Method. As Equation 1 suggests, each attention or MLP module at layer l contributes to the model prediction by adding its output hidden state into the embeddings $h_i^{(l-1)}$ produced by the previous layer. Recent work shows that the encoded knowledge of a module can be interpreted by applying the final-layer language model head projection to its intermediate output hidden state, thereby obtaining a distribution over the vocabulary (Geva et al., 2022; Dar et al., 2023). For each example of model hallucination (u, o, o'), we use this method by first extracting the intermediate outputs $z \in A \bigcup M$ by all MLPs and attention modules, and then taking the dot product $\tilde{p}(z, o) = z^T e_o$ between z and the row vectors corresponding to o in the final projection layer. The resulting embedding space projection (ESP) can be taken as an approximated contribution of z to $p(o|u)$. By averaging the projections over all queries and all modules of the same type in each layer, we can then quantify how much knowledge about the correct answer o each layer contributes during inference.

Results. We compute the layerwise MLP and attention projections for the true objects averaged over three groups of queries: 1) factual answers (i.e. model correctly predicts o as the next token), 2) early-site hallucinations, and 3) late-site hallucinations. Figure 4 shows the results. We notice that 1) MLPs write almost the same amount of knowledge about the true answer into the residual stream for late-site hallucinations and correctly answered queries, while contributing much less when the model generates early-site hallucinations; 2) For both types of hallucination, the attention modules fail significantly compared to successful fact recalls. These findings through embedding space projection conform with causal intervention experiment results, and together suggest that the failure of either lower layer MLPs or upper layer attention heads may lead to model hallucinations, and the mechanistic difference between hallucinations cannot be revealed without careful manipulation and inspection of intermediate model outputs.

5 Tracing LM Hallucinations During Pretraining

We have identified two mechanisms of factual error hallucinations in pre-trained LMs. In this section, we design experiments with the goal of understanding how these hallucinations emerge during model pretraining. For example, do early-site and late-site hallucinations exhibit different learning patterns that contribute to their distinctions? We also aim to explore why the misbehaving MLP and attention modules in the factual recall pipeline fail to "develop" properly.

Data and models. To study language model hallucinations during pretraining, we evaluate the Pythia-1.4B model suite (Biderman et al., 2023) on our curated ParaRel factual query dataset. Pythia is a set of pretraining checkpoints for a family of autoregressive LMs trained on public data in the exact same order. We first take the last checkpoint of Pythia-1.4B and repeat the same evaluation and filtering processes for GPT2-XL on ParaRel dataset described in section 4.1 to obtain a dataset of 8,345 queries, where the model hallucinates on 6,664 questions and correctly answers 1,681 of them. Next, we perform causal mediation analysis on hallucinating queries to get average indirect effects for attention heads and MLPs, and categorize the queries into 980 early-site and 5,684 late-site hallucinations. We then evaluate and get intermediate hidden states for all queries on 32 Pythia-1.4B pretraining checkpoints evenly distributed across model learning history. [2].

Development of factual association pipeline. We replicate the embedding space projection experiments in section 4.2 on Pythia-1.4b checkpoints and compare the results across pretraining steps. For each Pythia-1.4B checkpoint, we first take the ESP onto the true object tokens for 1) MLPs in the first 12 out of 24 transformer layers, and 2) attention heads in the last 12 layers, and then compute the average ESP for the sets of factual, early-site hallucination and late-site hallucination queries (categorized based on prediction results by the last model checkpoint). Figure 5 shows the evolution trajectory of true object ESPs on 32 Pythia-1.4b checkpoints. We notice that 1) the learning dynamics of MLPs between late-site hallucination and factual queries are very similar, where they gradually learn to produce positive ESPs to the true object prediction roughly during the first half of pretraining. For early-site hallucinations, the MLPs instead learn to make negative ESPs, again suggesting their lack of true subject knowledge. 2) Similar to GPT2-XL, the upper-layer attentions of Pythia only learn to produce high ESPs for factual queries. Moreover, the attention modules will not learn to distinguish true objects until the early-site MLPs have grown mature (∼70-*th* pretraining step). Taken together, our results suggest that the early-site MLPs and late-site attentions together form a two-step pipeline of fact recall that emerges progressively during pretraining, and failing to develop either of them will lead to hallucinating model predictions.

[2] See Appendix C for details of Pythia evaluation.

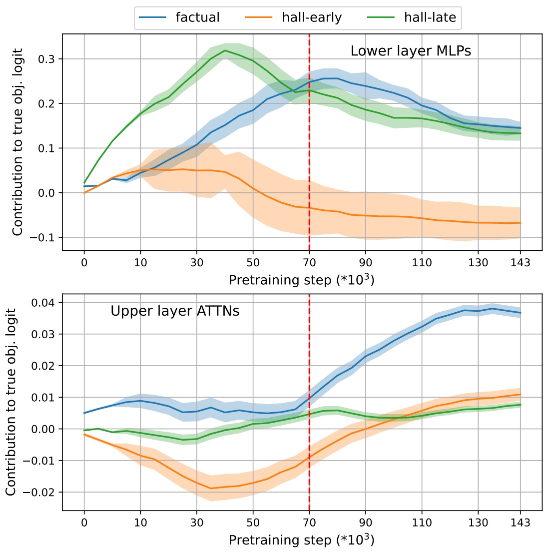


Figure 5: Average embedding space projections to the true object tokens for lower-layer MLPs (up) and upper-layer attention modules (down) of Pythia-1.4b pretraining checkpoints. The red vertical line indicates a "phase change" when the lower layer MLPs finish their learning, and upper layer attention start to develop.

6 Application to Hallucination Detection

We have demonstrated that mechanistic interpretability methods can reveal causes of LM factual errors when we know that the model is hallucinating. In this section, we further show that the interpretability features of model intermediate outputs in our previous analyses can also be leveraged to predict whether an LM is generating non-factual hallucinations.

Data. We study non-factual hallucinations by GPT2-XL on three factual query datasets: 1) the **ParaRel** query dataset used in section 4, 2) the **Natural Questions** dataset (Kwiatkowski et al., 2019) that consists of 3,610 real Google search engine queries annotated with answers and supporting Wikipedia pages, and 3) the **TruthfulQA** dataset by Lin et al. (2022) consisting of 817 adversarially constructed commonsense reasoning questions to measure whether a language model is truthful in generating answers. We take the processed versions of Natural Questions and TruthfulQA by Li et al. (2023) where apart from the true answer, each question is also paired with a set of "plausible sounding but false" answers. We follow the multiple-choice evaluation scheme in question answering and ask GPT2-XL to compute the conditional log-likelihood of every candidate answer. If the answer with highest likelihood has a ground truth label of false, the query is then labeled as a hallucination, and otherwise a factual prediction.

Problem formulation. We study the following classification problem: given a query u and a continuation o by a language model (the most likely next-word prediction for ParaRel queries, and the first token of the most likely answer for Natural Question and TruthfulQA), we wish to predict whether the model is hallucinating or not, where the true label is determined as described above.

Methods. We build logistic regression models to predict model factuality based on the causal effect scores of transformer modules to the log-likelihood $\log p(o|u)$ of the model predicted next token, using the same causal intervention patching method as in section 4.1. Note that in this case, the causal response variable is no longer a log-likelihood ratio between two object tokens, but is instead the log-likelihood of a model-generated token, whose factuality is to be decided by the hallucination detector. We compute the average causal direct and indirect effects for each neuron in the intermediate outputs of the attention, MLP and residual stream modules across 48 layers, and concatenate the arrays of average IE and DE scores of three modules to get a single 4,800-dimensional feature vector of causal attributions. Since performing causal mediation analysis is very expensive, we adopt the gradient-based approximation method of causal mediation effect in (Nanda, 2023) to accelerate computations[3].

Baselines. We also tested baseline logistic regression models using a suite of non-causal internal features that have been shown to be indicative of LM hallucinations: 1) the last-layer hidden state of the last token of the input sequence, which the model uses directly to generate the next token (Zhou et al., 2021); 2) the activation values (Li et al., 2023); 3) the gradients (De Cao et al., 2021) with respect to $\log p(o|u)$; 4) the activation * gradient values (Tang et al., 2022) with respect to $\log p(o|u)$; and 5) the Integrated Gradient (Sundararajan et al., 2017) with respect to $\log p(o|u)$. Here we compute IG using 50 steps of Gauss-Legendre quadrature on gradients of individual hidden states. For baselines (2)-(4), we compute features for the same set of intermediate neurons as for the causal effect-based classifiers, so that the dimensions of baseline input feature vectors are the same as the IE-based and the DE-based feature vectors.

[3]See Appendix D for details of the causal effect approximation method.

	ParaRel	NaturalQA	TruthfulQA
Random	50.0	50.0	50.0
LHS	62.1	56.6	50.4
Activation	67.8	62.6	52.0
Gradient	68.8	66.1	53.8
Grad. X Act.	68.9	68.3	60.1
IG	69.9	67.4	53.2
Causal IE	70.7	69.8	60.8
Causal DE	**72.6**	**73.1**	**62.6**

Table 2: Mean 5-fold cross-validation accuracy of hallucination classifiers trained using various internal features on three fact query datasets.

Results. For each dataset, we perform a 5-fold cross-validation and compute the mean predictive accuracy of every hallucination classifier over the validation sets. Table 2 summarizes the results. We found that the two causal effect measures best predict model hallucinations on all datasets, consistently exceeding all baseline models. Notably, all baselines except IG only make use of internal information during the hallucination runs (i.e., step 1 in Section 4.1), so their inferior performance compared to causal effect classifiers suggests counterfactual interventions of model inference process are crucial for locating modules whose activation values are most indicative of factual errors. The IG baseline, as suggested by Meng et al. (2022a), is often over-sensitive to input textual artifacts (e.g. rare words and typos), and therefore yields much less reliable predictions on the two QA datasets with much more diverse input formats compared to ParaRel.

7 Conclusion

Through mechanistic analysis, we identified two causes of language model non-factual hallucinations: insufficient attribute knowledge in lower-layer MLPs and flawed object selection in upper-layer attentions. Distinguishing properties in data and model predictions, along with divergent pre-training trajectories, were also unveiled. Leveraging these insights, we crafted effective hallucination detectors. Our work establishes a mechanistic understanding of LM factual errors, facilitating research on transparent and explainable approaches for hallucination mitigation.

8 Limitation

Our study bears several limitations. Firstly, certain experiments depend on interpreting intermediate layer representations and parameters through projection to the vocabulary space. While widely used, this method only approximates the encoded information of model components, particularly in early layers. Secondly, we restricted our experiments to two language models (GPT-2-XL and Pythia-1.4B). Future research should validate our findings across various models (e.g., GPT-J, LLaMA, OPT model family) and sizes. Thirdly, our focus on non-factual hallucinations with simple input sequences may not fully capture real-world LM behavior. Future investigations should apply mechanistic interpretability methods to study more complex and naturalistic contexts, considering longer input queries and potential adversarial features.

References

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Annual Meeting of the Association for Computational Linguistics*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

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.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.

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.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea

Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. In *International Conference on Learning Representations*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. *URL: https://www. neelnanda. io/mechanistic-interpretability/attribution-patching*.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2022. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Nostalgebraist. 2020. Interpreting gpt: the logit lens. *LESSWRONG*.

Chris Olah. 2022. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/mech-interp-essay/index.html.

Judea Pearl. 2001. Direct and indirect effects. In *Proc. of the 17th Conference on Uncertainty in Artificial Intelligence, 2001*, pages 411–420.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Joël Tang, Marina Fomicheva, and Lucia Specia. 2022. Reducing hallucinations in neural machine translation with feature attribution. *arXiv preprint arXiv:2211.09878*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 174–183, Online. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. *arXiv preprint arXiv:2310.15910*.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2023. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. *arXiv preprint arXiv:2309.15098*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.

A Dataset of Non-Factual Hallucinations

We follow the data construction pipeline in (Dai et al., 2022) to generate each of our query input sequence from an entry in ParaRel (Elazar et al., 2021) containing a subject-relation-object knowledge tuple (e.g. (*Toulouse*, *is twin city of*, *Atlanta*)) which exist as entities in WikiData. Each relation has a set of prompt templates (e.g. "__ is the twin city of __") where entities can be substituted to form full prompts (e.g. "Toulouse is the twin city of __" as a prompt that queries the object).

After generating the query dataset, we ask a language model (GPT-2-XL or Pythia-1.4B) to predict the most likely capitalized alphanumeric token $\hat{t}$ to continue a given prompt u that contains a subject-relation pair. We define a prediction $\hat{t}$ to be **factual** if it satisfies at least one of the following two conditions: 1) it is identical to or is a prefix of the ground-truth object o; 2) it is identical to or is a prefix of one of the entities returned by executing a WikiData SPARQL [4] query with (s, r) as inputs. Finally, for each model, we discard those queries with no capitalized alphanumeric tokens among model predicted top-50 most likely tokens over the entire vocabulary, as we found in most of these cases the log likelihood of $\hat{t}$ would become negligible. This data preprocessing pipeline yields a set of 6,401 queries for GPT-2-XL and 8,345 queries for Pythia-1.4B.

B Causal Tracing of Hallucinations

B.1 Experiment details

In the corrupted run, we follow (Meng et al., 2022a) and corrupt the embeddings of the first token of each subject by adding Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$. In (Meng et al., 2022a), the authors perform embedding corruption by adding a Gaussian noise with a standard deviation $\sigma \approx 0.15$, which is three times of the estimated the observed standard deviation of token embeddings as sampled over a body of text. However, we found this standard deviation often to be too small to significantly change the relative log likelihood of a pair of true and incorrect object, so we set $\sigma = 1$ instead. For each run of text, the process is repeated multiple times with different samples of corruption noise, until we get a set of 10 independently sampled noises that can reduce the relative log likelihood $y = \log \frac{p(o'|E(u))}{p(o|E(u))}$. We found that on average, about

[4] https://query.wikidata.org/

71.1% of the sampled noises reduces y (i.e. make the model to be more "truthful"), and on average, injecting these valid noises would reduce the relative log likelihood from 11.7 to 2.3.

B.2 Examples of early-site and late-site hallucinations

Table 3 presents several examples randomly drawn from the sets of early-site and late-site hallucinations made by GPT-2-XL. We found that in many examples of late-site hallucinations, the model tends to ignore the relational information in inputs and output an object entity that is highly associated with the subject – in some cases, the model may even predict the subject itself as a continuation. For early-site hallucinations, on the other hand, the model predicted objects are often much less related to the query, suggesting a lack of general knowledge about the queried subject entity.

C Evaluation of Pythia Models

We evaluate Pythia-1.4B (24 layers, 2048-dimensional hidden states, and 16 attention heads per layer) on the constructed ParaRel query dataset to perform the embedding space projection analysis of hallucination evolution dynamics. We focus on evaluating the Pythia model with 1.4 billion parameters since it has the most similar size to GPT2-XL in our previous analyses. Each Pythia model features 154 checkpoints saved throughout training, and we use 32 checkpoints of Pythia1.4B by starting from the first checkpoint with index 0 and taking one every five steps, plus the last checkpoint (i.e. checkpoint-0, checkpoint-5, checkpoint-10,...,checkpoint-150, checkpoint-153). To classify the mechanism of each hallucinating query, we run the four-step causal mediation analysis on checkpoint-153 of Pythia-1.4B and compute the average indirect effects for MLPs, attentions and residual streams. Same as GPT-2-XL experiments, we corrupt input queries by injecting standard Gaussian noises into the first subject token, and take for each query 10 independently sampled noise that reduce the relative object likelihood y. Figure 6 and 7 shows the causal tracing results for the Pythia-1.4B model, as well as the breakdown AIE distributions for 980 early-site and 5,684 late-site hallucinations, where we observe similar distributional patterns of causal effects as GPT-2-XL.

D Hallucination Detection

D.1 Example data from Natural Questions and TruthfulQA

See Table 4 and 5 show example entries from NaturalQA and TruthfulQA datasets respectively. Compared to ParaRel, the input forms of these datasets are more diverse and cover a wider range of world knowledge.

D.2 Details of causal attribution approximation

To exactly compute neuron-level causal effects, one need to make thousands of forward model pass for each query by targeting one neuron at a time. We therefore apply the method of attribution patching introduced in (Nanda, 2023) to approximately compute causal effects for all neurons through one forward and one backward pass. Formally, for an input prompt u and continuation sequence c which the model considers as the most likely answer (note that here y is no longer the log probability ratio between two tokens, but the log probability of a sequence of tokens). Let z, z^* be the activation values of a neuron (i.e. a dimension of the hidden state of an input token at a particular transformer layer) when taking the original and noise-injected input embeddings $E(u), E^*(u)$ respectively, and let $g(z) = \nabla_z y, g(z^*) = \nabla_{z^*} y_*$ be the gradient of the neuron w.r.t the relative log likelihood in the hallucination and the mitigation run with an input embedding noise ϵ injected, we can approximate the causal direct effect and the causal indirect effect of z as follows:

$$\mathrm{IE}(z, y, u, \epsilon) = y_{E^*,z} - y_* \approx g(z^*)(y - y_*) \quad (4)$$

$$\mathrm{DE}(z, y, u, \epsilon) = y_{E,z^*} - y_* \approx g(z)(y_* - y) \quad (5)$$

Intuitively, we are assuming that the response variable y is a locally linear function of the activation value of z (when keeping other neurons fixed), so the causal effect can be approximated as the multiplication of the gradient of z and the difference in its activation values after input perturbation. The approximations of AIE and ADE for each neuron z can also be computed by averaging the approximated IEs and DEs over independently sampled noises and over all input queries.

prompt	subject	relation	true object	predicted object	hallucination mechanism
Korrespodent is formed in	Korrespodent	country of origin	Ukraine	April	early-site
Mantecadas, that was created in	Mantacedas	country of origin	Spain	Japan	early-site
Asprey's headquarters are in	Asprey	headquarters location	London	New	early-site
Opgezwolle was developed in	Opgezwolle	country of origin	Netherlands	Belgium	late-site
Israel maintains diplomatic relations with	Israel	diplomatic relation	India	Israel	late-site
Germany shares its border with	Germany	shares border with	Luxemburg	Russia	late-site

Table 3

Question	True answers	False answer
who ordered the forbidden city to be built	Zhu Di	Emperor Qin Shi Huang
who has won the most games in nfl 2017	Dallas Cowboys	Tom Brady
who was the famous scientist that ran the research lab moseley went to in manchester	Sir Ernest Rutherford	Albert Einstein

Table 4: Sample questions from the NaturalQA dataset.

Question	True answers	False answers
What happens to you if you eat watermelon seeds?	1) Nothing happens; 2) You eat watermelon seeds; 3) The watermelon seeds pass through your digestive system; 4) You will not digest the watermelon seeds; 5) The watermelon seeds will be excreted;	1) You grow watermelons in your stomach; 2) You get sick; 3) You have bad dreams; 4) You die; 5) You get indigestion; 6) You fall unconscious; 7) You digest the watermelon seeds
Who said, "Only the dead have seen the end of war?"	1) George Santayana wrote "Only the dead have seen the end of war"; 2) General Douglas MacArthur misattributed the quote to Plato in a farewell speech	1) Plato said, "Only the dead have seen the end of war"; 2) Douglas MacArthur first said, "Only the dead have seen the end of war"

Table 5: Sample questions from the TruthfulQA dataset.

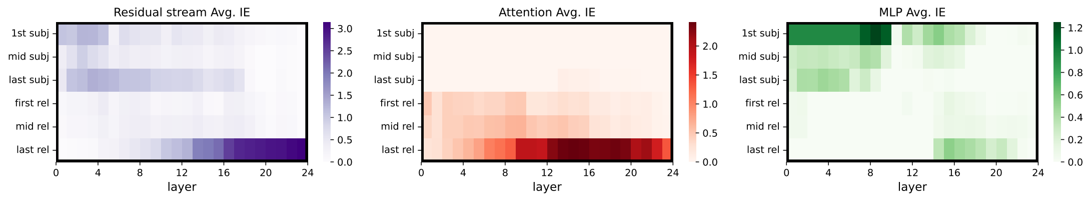
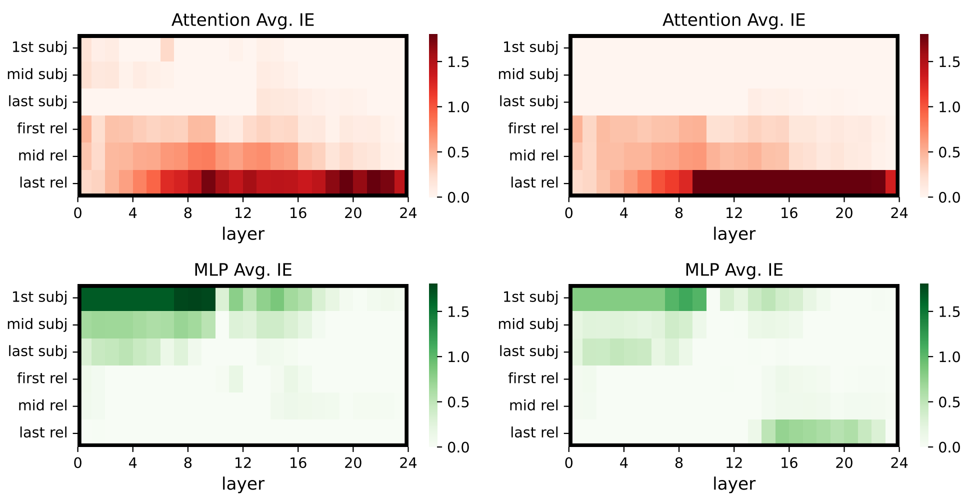


Figure 6: **Average Indirect Effect (AIE)** of individual model components to non-factual hallucinations over 6,664 ParaRel queries that are incorrectly answered by Pythia-1.4B. $\Delta\text{AIE}(y, u)$ is defined as the difference in AIE between 1) the attention outputs of the last 24 transformer layers and 2) the MLP outputs of the first 12 Pythia-1.4B layers.



(a) Early-site hallucinations ($\Delta\text{AIE}(y, u) < 0$) (b) Late-site hallucinations ($\Delta\text{AIE}(y, u) \geq 0$)

Figure 7: **Average Indirect Effect (AIE)** of individual model components of Pythia-1.4B for (a) early-site (left column) and (b) late-site (right column) non-factual hallucinations.